CAIDA participation in PREDICT

- Provider role: what data are we collecting, how are we curating and serving?
- Host role: what data are we hosting, who are we trying to recruit?
- Researcher role: anything security-related and/or useful being done with the data?
- Fires burning brightest: what are priorities for this year?

what data do we collect?

- OC192 backbone: 6 TB
  - monthly one hour anonymized packet header traces since March 2008
- UCSD telescope: 3.6 TB (30 day window, castrated subsets shared via PREDICT)
- OC48 traces: 1.7TB on SAN 149GB on web
  - 3 traces 2002 (3 hour) and 2003 (2x1 hour)

what data do we collect? (cont)

- topology: 10.9 TB
  - 1998 - present
- routed ipv4: 1.6TB (in PREDICT)
  - 4.4 billion traces by Ark since Sept 2007
- routed ipv6: 123MB
  - since Dec 2008 (need stats)

Total: 22TB (as of 31 August 2009)

how do we curate the data?

- OC192 backbone: capture, strip payload, transfer, anonymize, archive (aggregated links)
- OC48 traces: strip payload/L1/L2, anonymized w (prefix-preserve) cryptopan
- UCSD telescope: filter legitimate traffic at the router, 30 days on disk, curate backscatter and worm data separately.
- topology: see cybersecurity project

(http://www.caida.org/home/legal)

how many requests for the data?

- Passive traces: 685 requests, 502 approved, 419 accessed data
- UCSD telescope: 462 requests, 243 approved, 218 accessed data
- topology: 554 requests, 331 approved, 267 accessed data

Total: 1701 requests cumulative since 2003

why so many rejected?

- Rejected requests are mostly commercial cases asked to resubmit with an academic email address if they appear to be from academic users.
- A small fraction get rejected because of export restrictions or association with foreign military.

how do we serve the data?

- OC192 backbone: report generator

http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA (also traces to academics who sign AUP)

- OC48 traces: to academics who sign

http://www.caida.org/data/passive/anon_internet_traces_request.xml

- UCSD telescope: to academics who sign

http://www.caida.org/data/passive/network_telescope.xml#access

- topology: to academics who sign

http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml

  - (commercial researchers must join caida)

IRB: submitted/accepted, Oct 08

http://www.caida.org/home/about/irb/

How do researchers use the data?

- OC192 backbone: report generator up, traffic classification, performance modeling

- OC48 traces: traffic classification, modeling, monitoring, filtering, generation, locality
http://www.caida.org/data/publications/bydataset/index.xml#OC48

- UCSD telescope: bot/worm monitoring
http://www.caida.org/data/publications/bydataset/index.xml#backscatter

- topology: pkt traceback, marking. DOS defense. topo and routing modeling, discovery, metrics, improvements
http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml

(http://www.caida.org/data/publications/)

how do we use the data?

- OC192 backbone: traffic classification, real time monitor, traffic symmetry, address utilization, other myths

- OC48 traces: traffic classification, modeling, p2p, (also http://www.caida.org/data/realtime/passive/?monitor=sdnap)

- UCSD telescope: traffic classification, real-time monitor, lots of (and not enough..) Conficker analysis

- Topology: annotated Internet mapping http://www.caida.org/research/topology/

(www.caida.org/publications/papers/)

what other data do we seek (|to share)?

- Packet traces: ulonger traces, payload, other sites
- Internet2: better netflow, pkt traces, report gen.

http://www.caida.org/data/realtime/passive/?monitor=sdnap

- UCSD Telescope: unanonymized, payload, real-time

concerns i (still) have about PREDICT

- anonymization situation: we have met enemy
  http://www.caida.org/projects/predict/anonymization/
- policy support: research/position papers
- privacy impact statement: needs repair
- no govt use of data: needs clarification
- no networks that serve public
- metadata catalog
- metrics for success
- community outreach: wikis, blogs, bofs, socialnets
- improved PR