DHS PREDICT project: CAIDA update

- Data collection/infrastructure updates
- Dataset dissemination/statistics
- Research status
- Phase 2 datasets
- Open issues

Data collection

- OC192 backbone: 8.5 TB (3.6 anonymized; 4.9 unanonymized)
  - 2007-2009 data, Jan-Feb 2010
    - monitor is down, cannot transfer more data
  - collect 1 hr trace per mo = 200-250 GB
  - plan to keep a quarterly sample - select the best quality
- UCSD telescope: 3.8 TB on disk (5 weeks window)
  - plan to keep ~ 4TB on samqfs, ~4TB on disk
- OC48 traces: 1.7TB (2004 traces, anonymized, in PREDICT)
- topology: 8 TB
  - old skitter data (in PREDICT): 4 TB
  - new Ark data: IPv4 topology 3.9 TB, IPv6 topology 1.5 TB

Data curation

- OC192 backbone: strip payload/L1/L2, transfer, anonymize, archive
- UCSD telescope: filter out legitimate traffic at the router, 30 days on disk, curate custom data sets upon request
- topology: create derivative data sets, aggregate in ITDK
- need funding to curate/analyze/annotate IPv6 data

Data storage

- Samqfs is about to become expensive
- to keep the current volume = $5k/mo (with overhead)
- Carefully scrutinize old data: keep or discard?
- Upgrade available CAIDA storage: 2 new data servers
- Silicon Mechanics:
  - 48 TB disk space
  - Intel E5620 Quad Core 2.4 GHz processor
  - 6 GB memory
- Ashford Computer:
  - 48 TB disk space
  - Intel X3450 Quad Core 2.66 GHz processor
  - 8 GB memory

how do we serve the data?

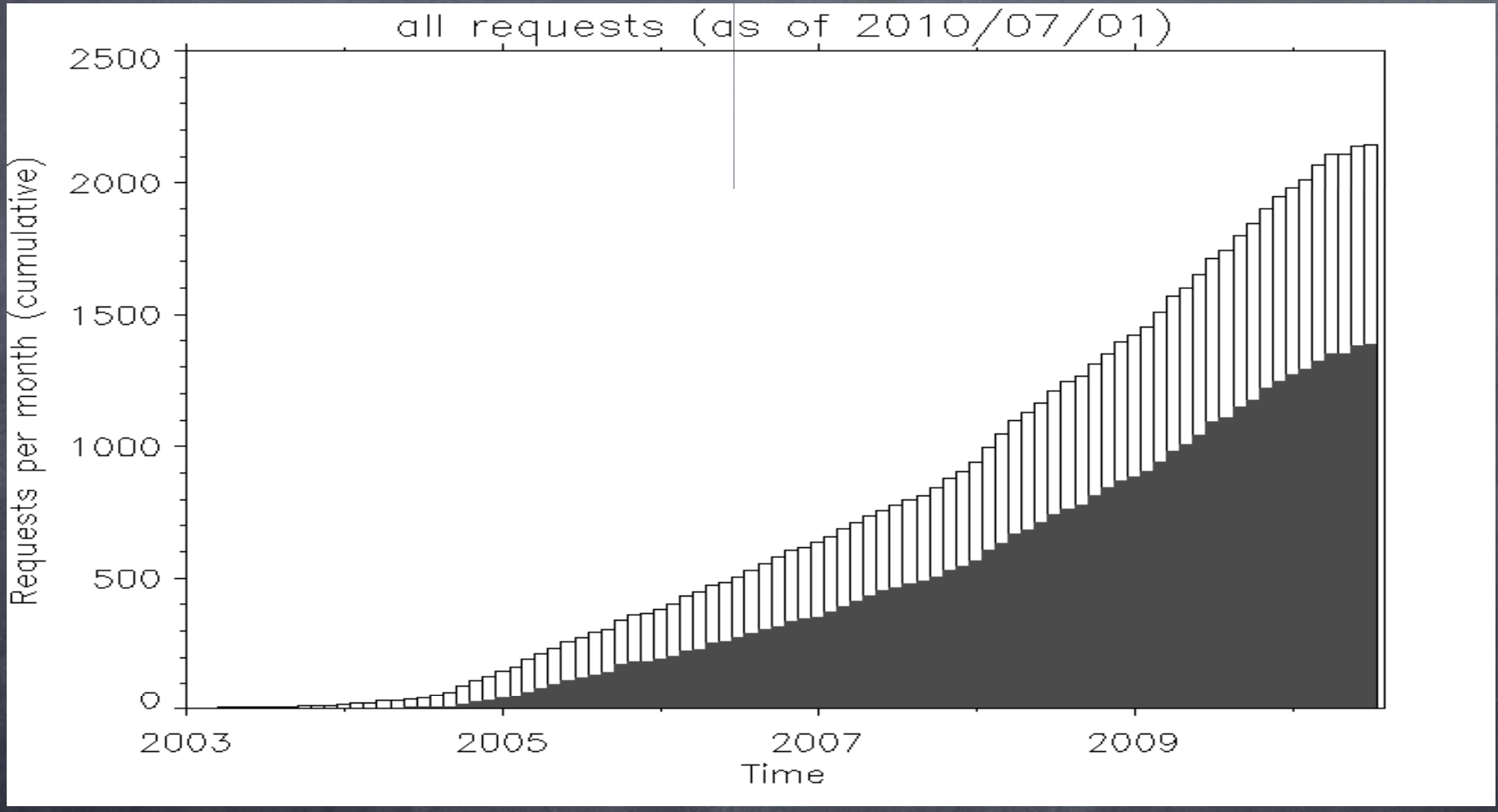
- PREDICT (OC48 traces, topology from skitter, telescope)
- Academics who sign AUP (OC192, topology from Ark, telescope)
- Derived data sets are publicly available (i.e., AS-links)
- Commercial researchers must join CAIDA
- Aggregated statistics online:
- OC192 backbone:
  - report generator: http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA
- topology:
  - Ark statistics: http://www.caida.org/projects/ark/statistics/index.xml
  - For each monitor: path dispersion (AS and IP), path length distribution, RTT distribution, RTT vs. distance, median RTT per country

Requests for the data, 2010/2009

Dataset	Requests	Approved	Accessed	Served since
Backscatter	34/101	23/62	16/45	Feb 2003
Passive	71/242	54/181	41/151	Feb 2004
Topology	82/136	52/90	33/63	Jul 2004
Witty	3/28	3/18	3/14	Mar 2008
Telescope	13/35	9/20	7/16	Jul 2009
DNS-RTT	2/7	1/3	1/3	Aug 2006
	205/549	142/376	101/292	

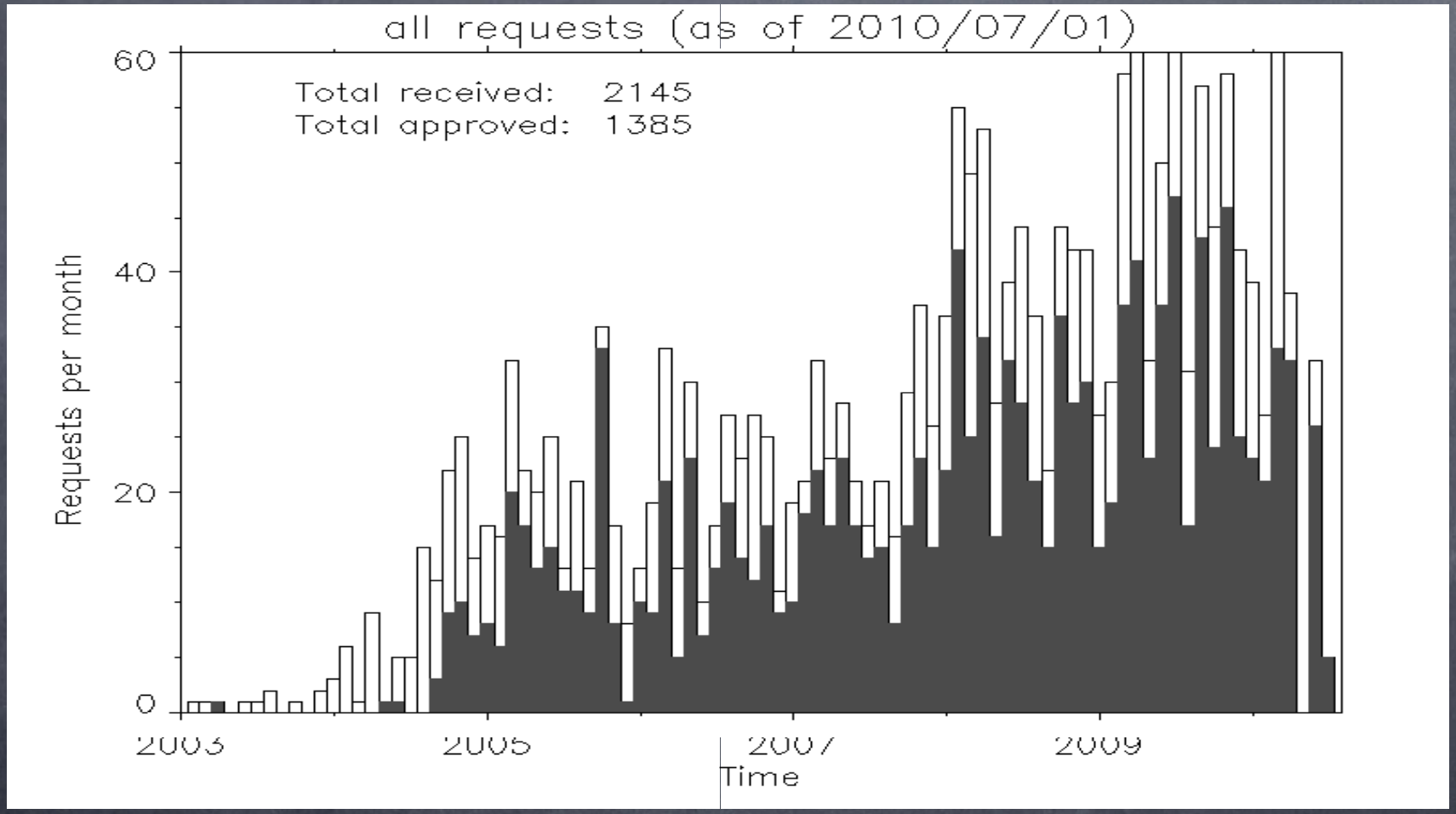
Data request stats

- All requests (cumulative)



Data request stats (cont)

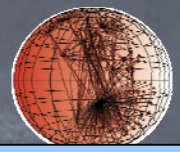
- All requests (monthly)



how many PREDICT requests for our data?

Dataset	Requests	Approved	Accessed
Backscatter	3	3	2
Passive (oc48)	4	3	2
	7	6	4

Community feedback



My student looked at PREDICT and says that without an account, one cannot tell what is in PREDICT and establishing an account involves jumping through many hoops. Further, even with an account, one cannot really tell what data exists until a review board agrees that it is appropriate for the research one is doing. I am sure that there are reasons for this, but this does help attract users. Am I misunderstanding something here?

- The voice of reason?
- Necessary conditions of success:
- Convenience
- Marketing
- Regular updates with newest data

CAIDA new (experimental) marketing strategy

- Commercial requests for data
- Need to join CAIDA = invest money and effort
- How to attract interest?
- Created a “topology data sampler”
  - Old: Jan 2008 for IPv4, Dec 2008 for IPv6
  - Small: 1 cycle
  - Available for evaluation/testing
  - Will join if need more recent/complete data
- Similar “teaser samples” for other kinds of data?
- Useful as educational handouts?

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

CAIDA research using the data

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

- OC192 backbone: traffic classification, real time monitor, traffic symmetry, address utilization, flow sizes, and more…
- UCSD telescope: Conficker analysis, one-way traffic classification, real-time monitor
  - need bodies (interns, students, collaborators)
  - need funding
- Topology: state-of-the-art annotated Internet mapping
  - 48 monitors in 27 countries
  - http://www.caida.org/research/topology/

CAIDA research in data sharing

- Developed and published a data-sharing framework
- Applied this framework to develop risk controls for the UCSD telescope data

- Erin will present

Research results - published

- E. Kenneally and kc claffy, "*Dialing privacy and utility: a proposed data-sharing framework to advance Internet research*", in submission to *IEEE Security & Privacy* special issue, July 2010.
- W. John, M. Dusi, kc claffy, "*Estimating Routing Symmetry on Single Links by Passive Flow Measurements*", published at the 1st International Workshop on TRaffic Analysis and Classification (TRAC).
- B. Huffaker, A. Dhamdhere. M. Fomenkov, kc claffy, "*Toward Topology Dualism: Improving the Accuracy of AS Annotations for Routers*", published in the proceedings of the Passive and Active Measurement Conference (PAM), April 2010 - uses ITDK-like data.

Research results - in submission/preparation

- N. Brownlee, A. Este, kc claffy, "*Internet background radiation: monitoring the packet plague*", submitted to Internet Measurement Conference (IMC) 2010.
- A. Este, M. Zhang, L. Salgarelly, kc claffy, "*The emerging face of Internet traffic*", submitted to Internet Measurement Conference (IMC) 2010.
- A. Dianotti and kc claffy, "Obstacles and challenges to traffic classification", to be submitted to IEEE Networks.
- AIMS-2 workshop report - in the final stage of editing.

Non-CAIDA Research using our data

- OC192 and OC48 traces: traffic classification, performance modeling, monitoring, filtering, generation, locality
  http://www.caida.org/data/publications/bydataset/index.xml#passive
  - 59 publications (47 from data in PREDICT)
- UCSD telescope: Conficker, worm research
  http://www.caida.org/data/publications/bydataset/index.xml#Backscatter
  - 23 publications (all from data that used to be in PREDICT)
- topology: pkt traceback, marking, DOS defense, topo and routing modeling, discovery, metrics, improvements
  http://www.caida.org/data/publications/bydataset/index.xml#Topology
  - 49 publications (39 from data in PREDICT)

Proposed Phase II data sets

- OC192 backbone: 2007-2010
- UCSD telescope: near real time
- topology: newest Ark data
  - IPv4 Routed /24 Topology dataset
  - IPv4 Routed /24 DNS Names dataset
  - IPv6 Routed Topology dataset
  - AS-links, AS relationships - publicly downloadable
- topology: updated ITDK 2010

- Descriptions are on our web site - can be adapted for PREDICT
- More discussion in the afternoon…

Community feedback

I have tons of other data that I can put into PREDICT but it is not real time and may be one-off sometimes. For anyone to put in data for long-enough duration and maintain its quality and currency, I would imagine that they would need incentives. Multiple folks have told me that much of what is there in PREDICT Is not very useful.

- The voice of reason?
- Necessary conditions of success:
- Convenience
- Marketing
- Regular updates with newest data
- … and incentives for Data Providers?

Canonical Data Set Homework

1st best - OC192 and OC48 traces

- popularity: requested 313 times, accessed 192 times (in 2009/2010)
- who used it: 180 .edu, 83 .cn, 36 .uk, 25 .com (since 2004) …
  - and 45 more domains
- feedback: need more, need high-resolution timestamps
- why popular: real data from the real Internet

Canonical Data Set Homework

2nd best - topology data

- popularity: requested 218 times, accessed 96 times (in 2009/2010)
- who used it: 202 .edu, 82 .cn, 31 .uk, 22 .kr, 22 .jp (since 2004) …
  - and 50 more domains
- feedback: need RTT, annotations, more aggregated data
- why popular: real data from the real Internet, interdisciplinary value

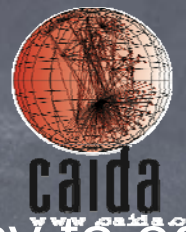
CAIDA newest project

- Claffy and Dunnigan collaborative effort

Open issues in PREDICT

- Improve publicly available meta-data
- Improve the Portal - both “how it looks” and “how it works”
- Review canonical MOAs for public providers - need data from networks that serve public
- Privacy Impact Assessment statement: - needs repair
- Marketing and PR

Suggestions for PREDICT Phase II



- create “todo” list/status updates: accessible, easy to edit - wiki page?
- distribute all documents for review/comments in electronic form: link from wiki?
- develop a set metrics to measure progress: post in the portal, regularly review
- Expand Policy section of PREDICT:
  - post a discussion of Ohm’s paper
  - post a position statement on “government researchers”
  - d1iscuss/post criteria for data eligibility for PREDICT
  - start a public online forum to discuss desirable amendments to ECPA
- active marketing efforts from all team members
  - community outreach: wikis, blogs, bofs, socialnets