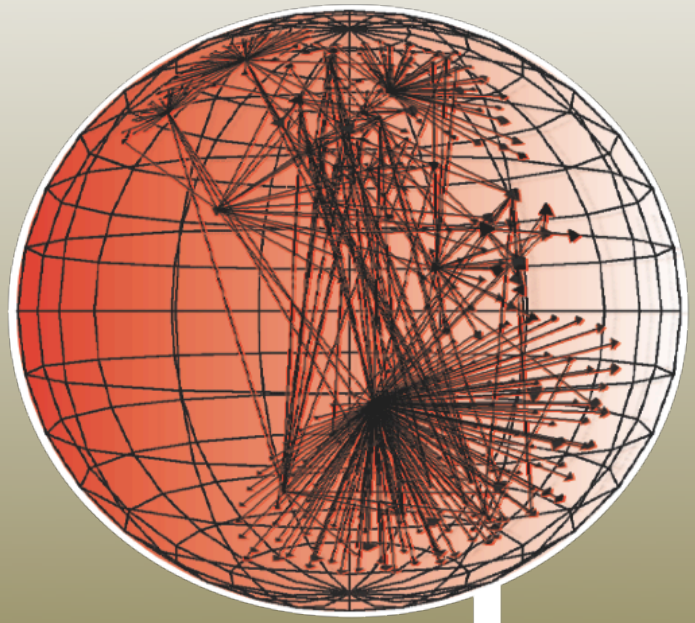
~~Fostering Cybersecurity Research Through Data Sharing~~

~~Data: You Show Me Yours, I'll Show You Mine~~

CAIDA: A Data Sharing Case Study



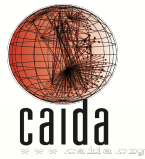
caida

Josh Polterock, CAIDA

NSF Security at the CyberBorder Workshop

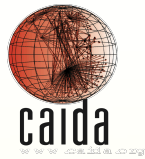
February 23, 2012

Overview



- CAIDA:
- our mission
- conducting research
- building infrastructure
- collecting and curating data
- developing tools
- modeling economics
- informing policy
- coordinating and hosting workshops

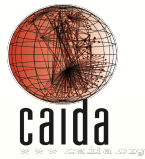
Our Mission



CAIDA The Cooperative Association for Internet Data Analysis (CAIDA) is an independent analysis and research group based at the University of California's San Diego Supercomputer Center. CAIDA investigates both practical and theoretical aspects of the Internet, with particular focus on:

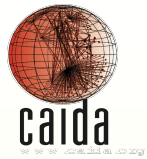
- collection, curation, analysis, visualization, dissemination of sets of the best available Internet data,
- providing macroscopic insight into the behavior of Internet infrastructure worldwide,
- improving the integrity of the field of Internet science,
- improving the integrity of operational Internet measurement and management,
- informing science, technology, and communications public policies.

Conducting Research



- Macroscopic Topology Project
  - IPv4 and IPv6 topology discovery
  - hostname collection
  - Alias Resolution
  - Router to AS assignment
  - dual graph
  - AS relationships
  - IPv4 and IPv6 topology discovery
  - AS Rank
- Internet Interconnection Economics
- Darkspace Analysis
- Named Data Networking
- Modeling Complex Networks
  - Internet evolution
  - Routing complex networks
- DatCat Returns!

Infrastructure



- Archipelago
  - CAIDA's active measurement infrastructure
  - 57 monitors – growing 1 or 2 per month
  - 28 w/ IPv6 connectivity
  - currently used for:
    - Team-probing experiment to collect IPv4 and IPv6 topology
    - alias resolution measurements
    - Spoofer experiment
- Passive Monitors
  - Tier1 10GE backbone link packet header capture

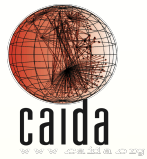
Infrastructure

- UCSD Network Telescope
  - Near real-time access to raw telescope data
  - 2 days of telescope dataset
    http://www.caida.org/data/passive/telescope-2days-2008_dataset.xml
  - 3 days of Conficker dataset
    http://www.caida.org/data/passive/telescope-3days-conficker_dataset.xml
- CAIDA working with SDSC to use cloud storage services (http://cloud.sdsc.edu/)
  - plans to purchase 25 TB of cloud storage for last 60 days of system backups
  - granted 20TB from SDSC Executive Team for scientific data
    - telescope (103TB)
    - packet headers (18.8TB)
    - skitter/ark topology (4TB)

Data collection - passive

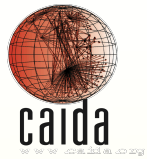
- OC192/10GE backbone: March 2008 - December 2011
  - 18.8 TB compressed, 35.7 TB uncompressed
  - unanonymized: 10.9 TB compressed, 21.3 TB uncompressed
  - anonymized: 7.9 TB compressed, 14.4 TB uncompressed
  - Doing cleanup towards retaining only quarterly traces
  - Completed 2011 Passive Dataset
- Problems:
  - Hardware failures at collection sites: Chicago monitors have been offline since September. We are working with remote hands to troubleshoot.
- Plans:
  - New 2012 annual dataset will start with upcoming trace on 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



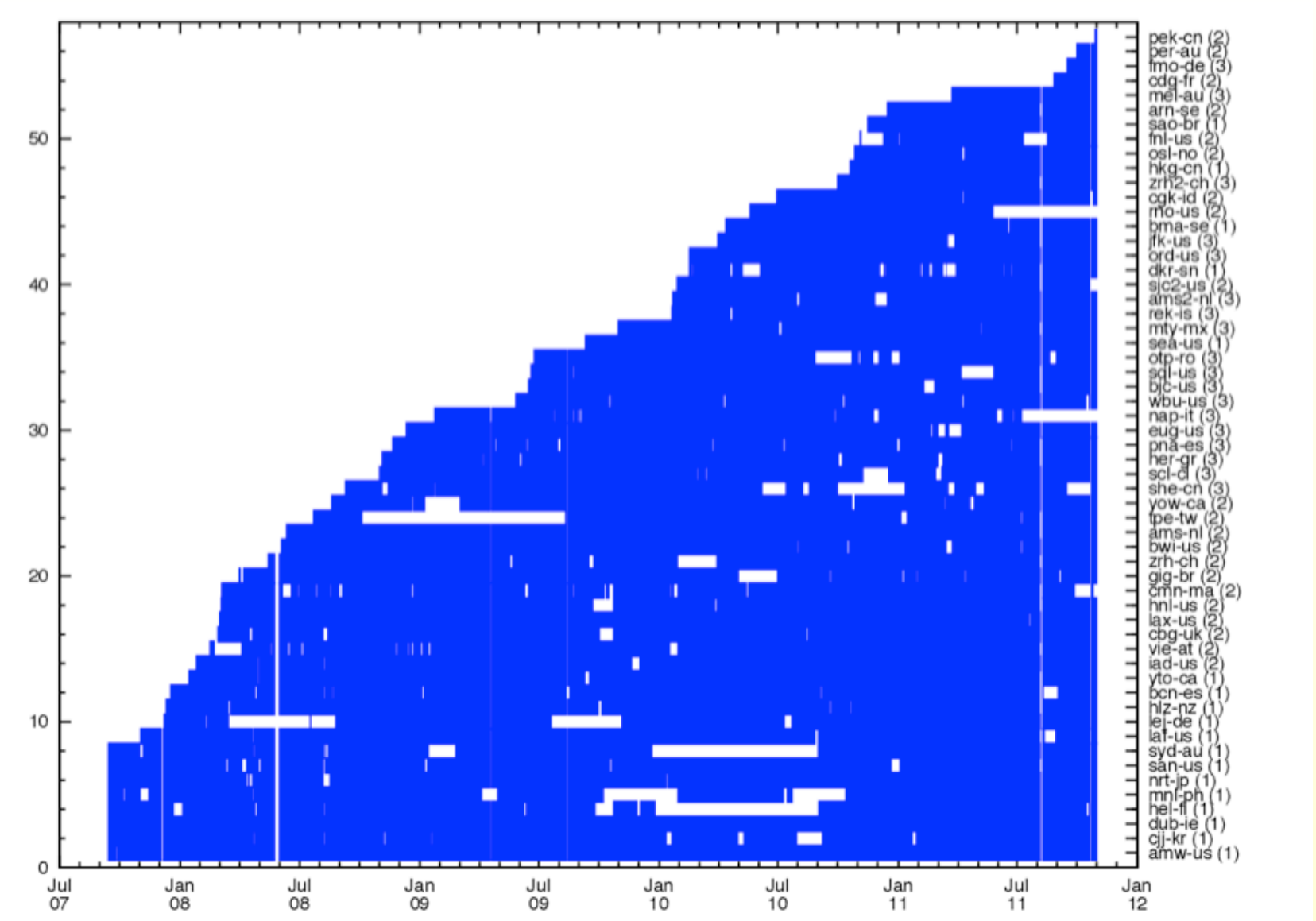
- UCSD telescope:
  - data from most recent 30-days (really five weeks) “live” on disk
    - typically 2.9 TiB compressed, 5.5 TiB uncompressed
  - the previous months - backed up on tape (samqfs)
    - current: 2008/04/12 - 2012/01/12
    - 102 TB (compressed), 192 TB (uncompressed)
    - received new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

Archipelago Monitors and Data

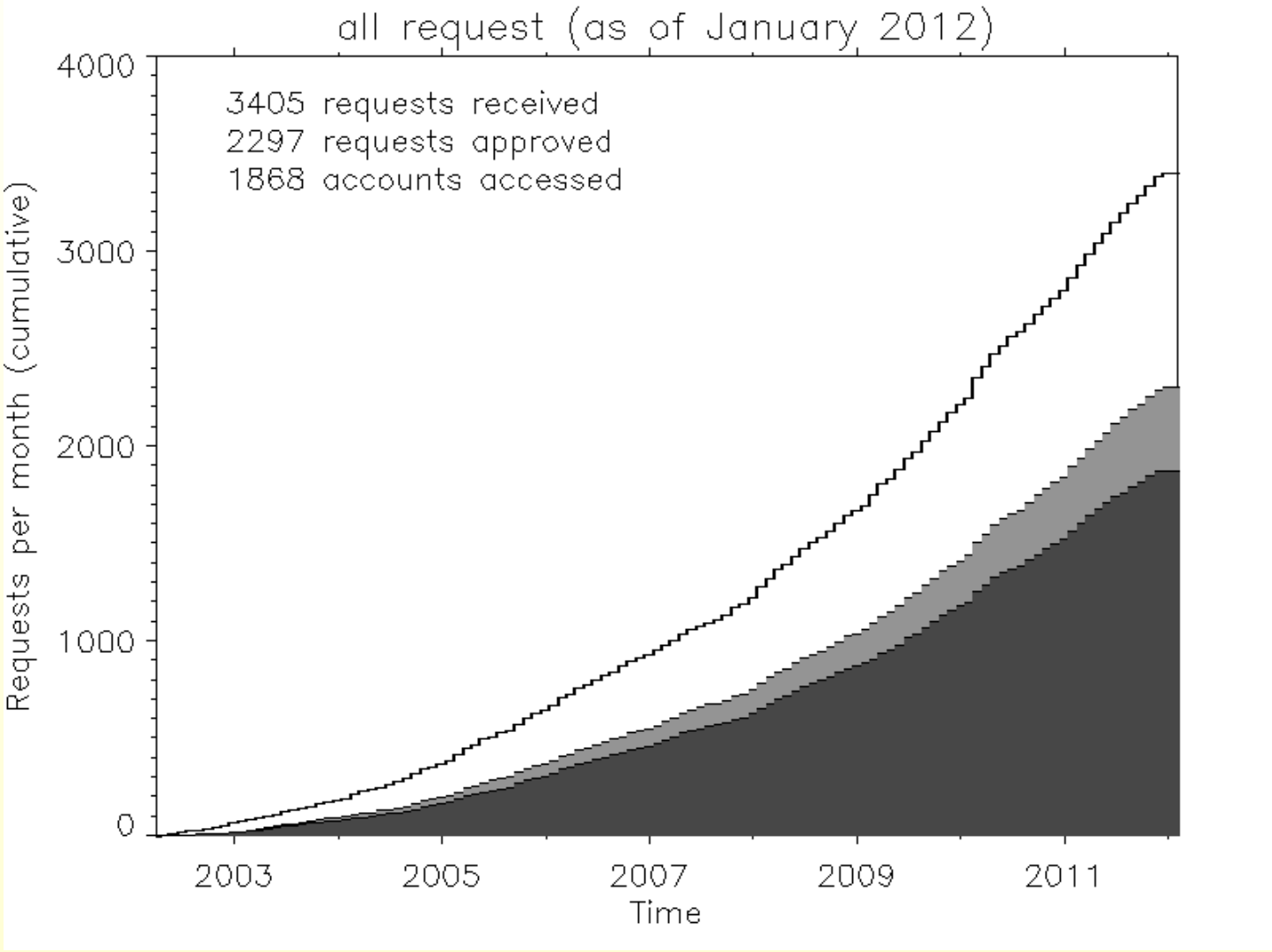


Requests for the data, 2011/2010/2009

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

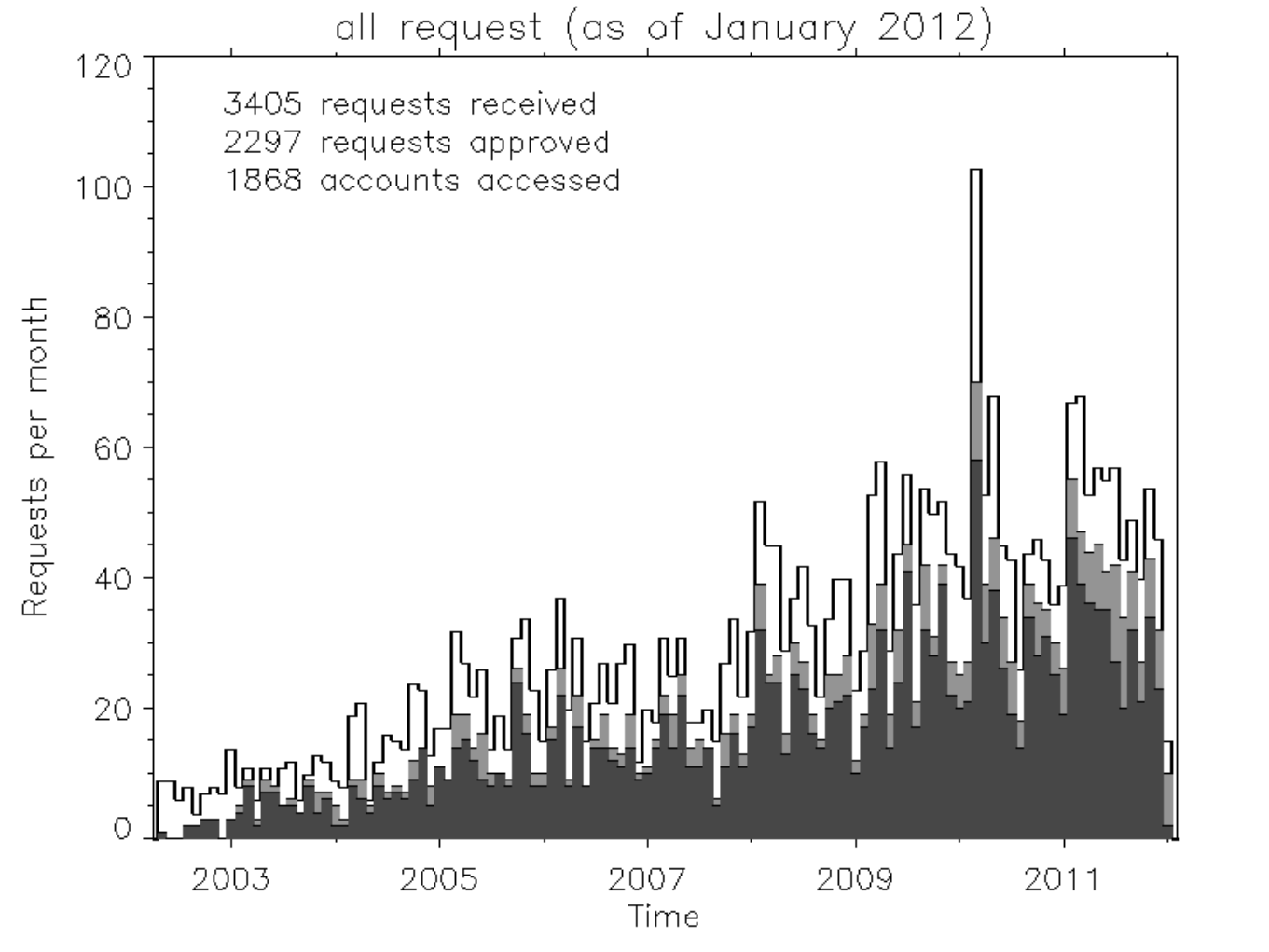
Data request stats

- all requests (cumulative)

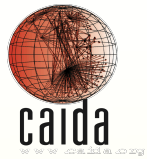


Data request stats (cont)

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

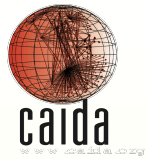


Data Set Popularity



- 1st best - OC192/10GE 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 best - 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

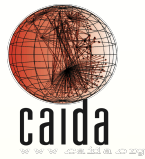


- PREDICT (OC48 traces, topology from skitter, telescope)
- Derived data sets are publicly available (i.e., AS-links)
  - sample use: http://semilattice.net/projects/map-of-the-internet/
- Academics who sign AUP (passive, topology from Ark, telescope)
- Commercial researchers
  - a small sample of data to entice interest
  - join CAIDA, various membership levels are offered

Data statistics - online

- Aggregated, (near) real time
- OC192/10GE 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, ...

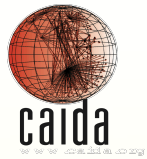
Meta-data for packet traces



- OC192/10GE 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)

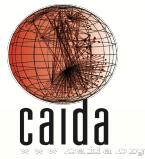
http://www.caida.org/data/passive/trace_stats/

Recent publications



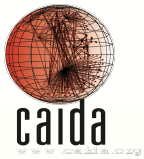
- A. Dainotti, R Amman, E. Aben, kc claffy, *Extracting Benefit from Harm: Using Malware Pollution to Analyze the Impact of Policital and Geophysical Events on the Internet,* ACM SIGCOMM CCR accepted for publication.
- Erin Kenneally *A Refined Ethical Impact Assessment Tool and a Case Study of its Applications,* 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.
- kc claffy, *The 3rd Workshop on Active Internet Measurements (AIMS-3) Report,* ACM SIGCOMM CCR V. 41, p. 37-42, 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

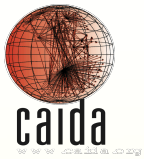
PREDICT Phase II Data Sets



- 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

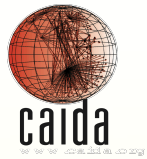
- OC192/10GE backbone: 2007-2011

Updates of CAIDA policies



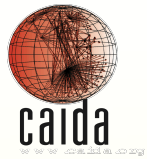
- Telescope data (RTTD)
  - different from previous packaged data
  - simplified and streamlined the AUP language
  - Immediate use by postdoc A. Dainotti and his student
  - analysis of macroscopic events (e.g earthquakes) on the Internet, collaborating with RIPE-NCC on publication.
- ARK hosting sites
  - 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



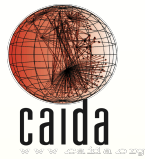
- 4 categories of data - different levels of sensitivity
  - real-time telescope data
  - passive traces
  - active traces
  - derived topology
- Document proliferation
  - 7 data request forms
  - 22 data set web pages
  - 22 README files
- Master AUP 1.0 for all CAIDA data sets
  - Factor out common conditions
  - Remove inconsistencies
  - Sent out to PI list for feedback

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



- what are we protecting?
  - PII (including IP addresses)
  - organization proprietary data
  - Privacy: Individual vs. Organization
- relevant for Best Practice documentation efforts

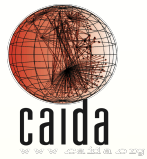
Other activities

- the (disastrous) 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?
- Lots of work on how to better store telescope data; not much to show yet

Other activities

- Hired Ark system administrator, Parisa Nahavandi
- Hired Telescope research programmer, Alistair King
- Hosting Visiting Scholar, Tanja Zseby, working with telescope group
- Hired two postdocs, Matthew Luckie and Alberto Dainotti

Storage Update



Ark IPv4
Total stored data: 1.8 TiB
Total stored no of files: 79600
Total free space: 3.4 TiB (shared with Ark IPv6)
Yesterday growth: 1.7 GiB

Ark IPv6
Total stored data: 2.8 GiB
Total stored no of files: 7444
Total free space: 3.4 TiB (shared with Ark IPv4)
Yesterday growth: 8.2 MiB

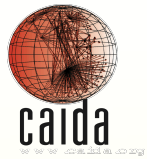
Passive high-speed equinix traces
Total stored data: 7.9 TiB
Total stored no of files: 10094
Total free space: 5.9 TiB
Yesterday growth: 6.0 GiB

Live telescope data (ogma)
Total stored data: 33.9 TiB
Total stored no of files: 8813
Total free space: 3.7 TiB
Yesterday growth: 100.1 GiB

Long-term Telescope storage on tape:
Total stored data: 82.9 TiB
Total stored no of files: 20428
Total free space: N/A
Yesterday growth: N/A

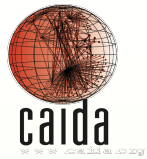
Overall Cummulative Stats
Total stored data: 126.5 TiB
Total stored no of files: 126379
Total free space: 13.0 TiB
Yesterday growth: 108 GiB

Data We Want



- The list below derived in part from k's blogs:
  - "**my second FCC TAC meeting, and its IPv6 promise**" **http://blog.caida.org/best_available_data/2011/04/30/my-second-fcc-tac-meeting-and-its-ipv6-promise/; and**
  - "**data collection and reporting requirements for broadband stimulus recipients**" **http://blog.caida.org/best_available_data/2009/11/12/data-collection-and-reporting-requirements-for-broadband-stimulus-recipients/**
  - interdomain traffic data to look at interdomain traffic patterns
  - financial data
  - peering relationship ground truth
    > BGP data
    > router-level topology ground truth
  - whole maps of ISPs to use as ground truth
  - path performance ground truth
  - maps at various levels
  - anything v6
  - ASes and IP addresses by org
  - actual organization tree

Possible Collaboration

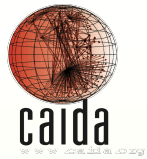


- How the community might interact with CAIDA
- Contribute interdomain traffic statistics to help build a model of interdomain interconnection and dynamics
- AS Rank feedback interface to provide corrections to our AS relationships model

For more information please contact:
info@caida.org

http://www.caida.org/

Sharing Quotes



- "Love only grows by sharing. You can only have more for yourself by giving it away to others." — Brian Tracy

- "And a new philosophy emerged called quantum physics, which suggest that the individual's function is to inform and be informed. You really exist only when you're in a field sharing and exchanging information. You create the realities you inhabit." — Timothy Leary, *Chaos & Cyber Culture*

- "Share your knowledge. It is a way to achieve immortality." — Dalai Lama XIV