DHS PREDICT project: CAIDA update

Kimberly Claffy, CAIDA

Ann Arbor, MI
31 May - 1 June 2012

DHS PREDICT project: CAIDA update

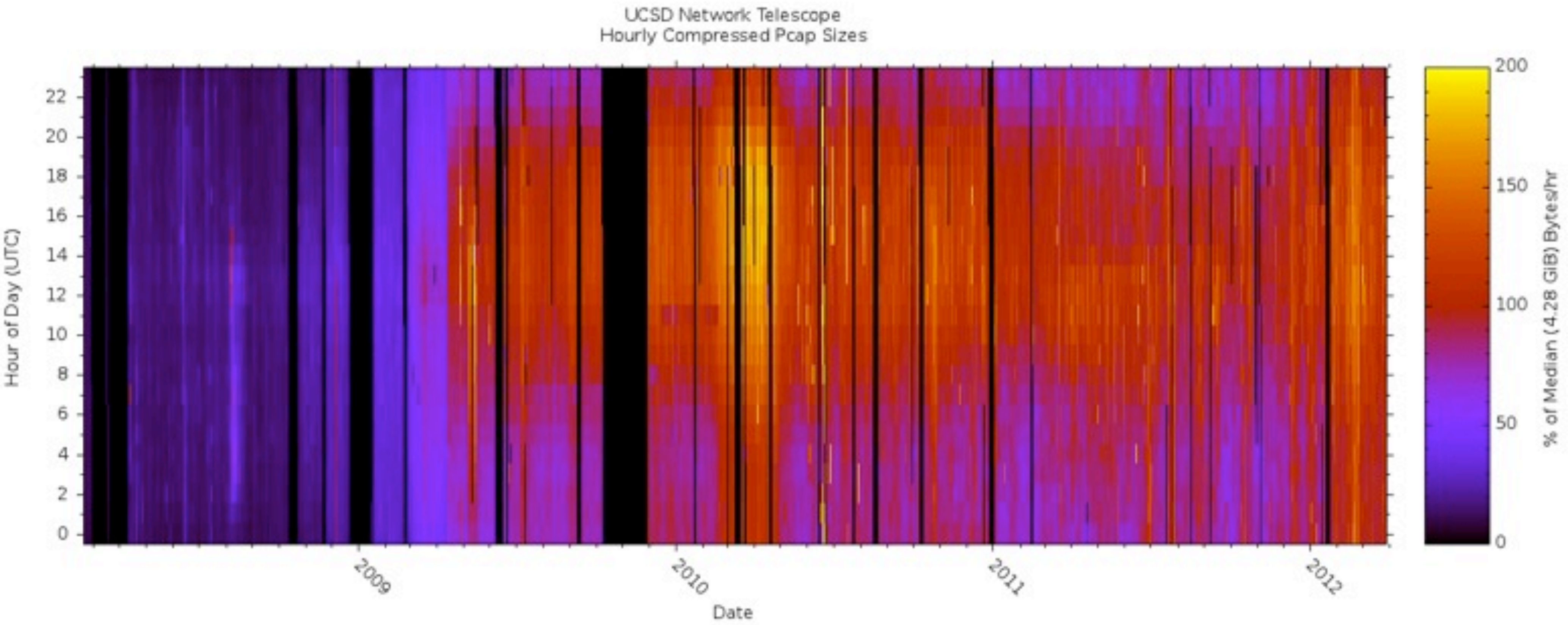
- Data storage status
- Data collection status
- Data set dissemination statistics
- Other activities
- Open issues

Data Storage Status

- CAIDA experimenting with SDSC’s 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)
- Also building our own RAID systems
  - will do capex/opex cost comparison over next 2 years

Data Storage Status

- Transferred 100+ TiB to DOE lab NERS
  - National Energy Research Scientific Computing Center
  - One week to transfer 100TB on 22 March 2012
  - http://blog.caida.org/best_available_data/2012/04/04/targeted-serendipity-the-search-for-storage/



Data collection - passive

- High-speed backbone: March 2008 - May 2012
  - unanonymized: 9.7 TiB compressed, 19.2 TB uncompressed
  - anonymized: 8.7 TB compressed, 17.7 TB uncompressed
- Problems:
  - Hardware failures at collection sites: Chicago monitors offline since September 2011. Replacement hardware recently purchased.
- Status:
  - 2012 data sets online through May
  - took DITL trace, added to 2012 data set
  - set of 'best' quarterly traces completed, but no longer backed up. single copy on spinning disk
  - planning additional traces for IPv6 Launch (June)

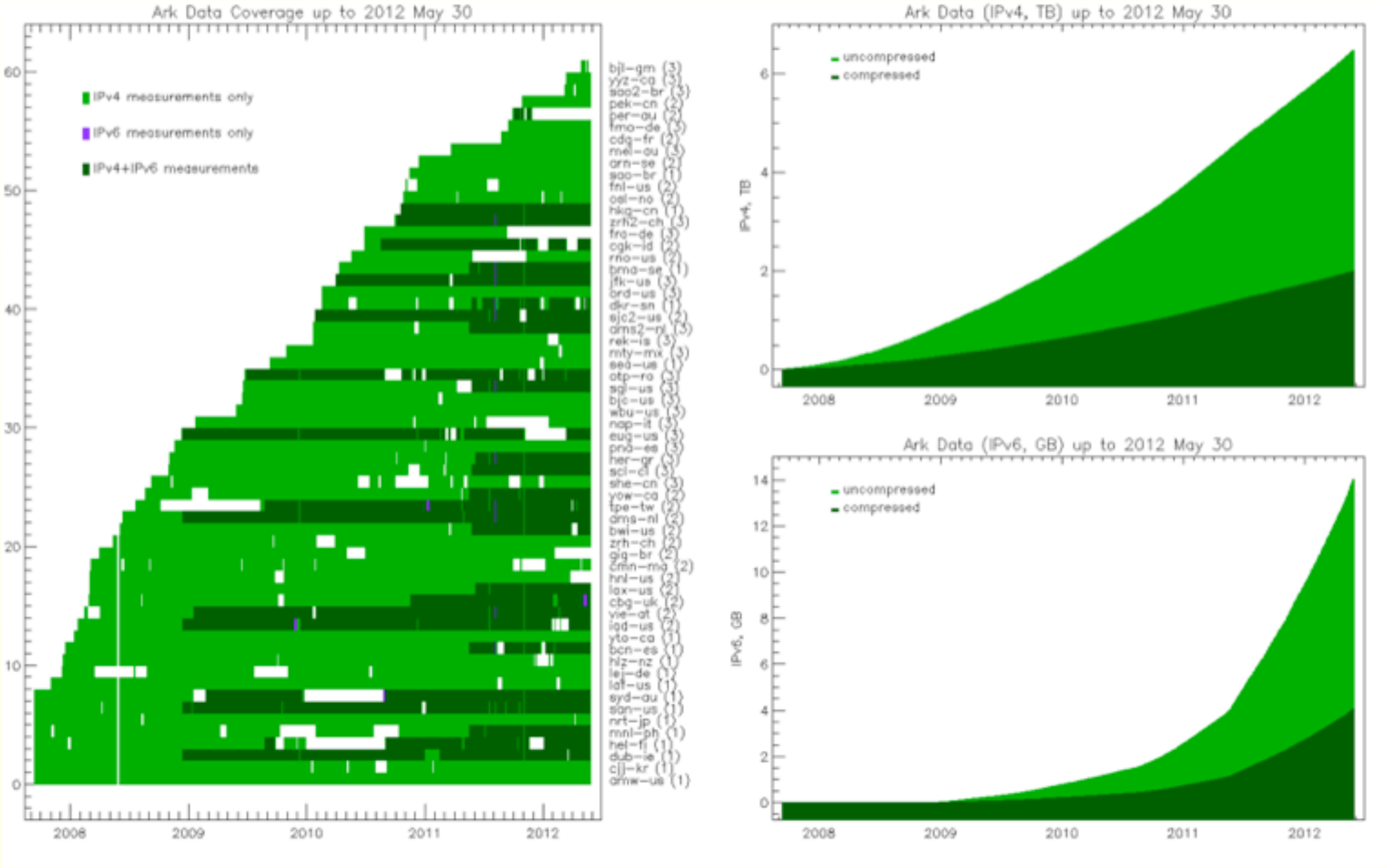
Data collection - passive

- UCSD telescope:
  - 2012 data still ‘live’ on disk (currently ~146 days).
    - 16.92 TiB compressed, 34.28 TiB uncompressed.
  - older data archived to NERSC in April.
    - 105 TiB compressed/encrypted.
  - summaries (Corsaro 8-tuples) stored on local disk
- 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 February 2008
    - skitter ITDK now a public dataset
- current Ark data:
  - IPv4 topology: 2.0 TiB (compressed), 6.5 TiB (uncompressed)
  - IPv6 topology: 4.1 GiB (compressed), 14 GiB (uncompressed)
  - 60 monitors (and growing) in 31 countries, 28 IPv6 capable
- data curation:
  - create derivative data sets
  - aggregate in http://www.caida.org/data/active/internet-topology-data-kit/
    - router-level topologies: nodes and links
    - host names, AS names, geographical info

Archipelago Monitors and Data

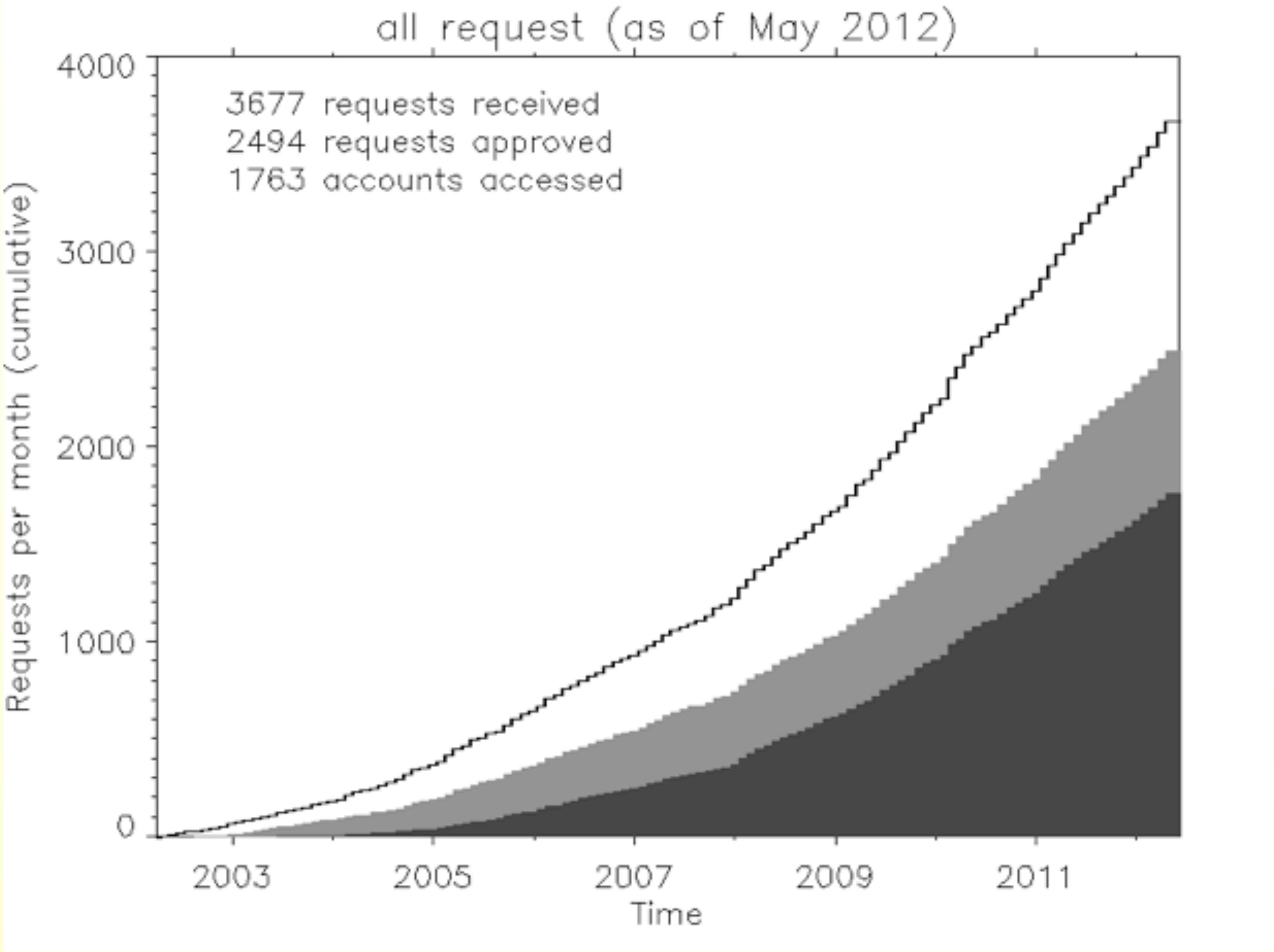


Requests for the data, 2012/2011/2010/2009

Dataset	Requests	Approved	Accessed	Served
Backscatter	15/51/73/95	10/34/47/60	7/28/36/46	Feb 2003
Passive	157/275/185/233	114/211/150/179	98/173/127/157	Feb 2004
Topology	50/155/163/129	40/129/113/83	29/76/73/51	Jul 2004
Witty	5/16/16/27	2/12/13/17	2/10/11/14	Mar 2008
Telescope	12/29/34/37	5/22/23/21	4/18/19/17	Jul 2009
DNS-RTT	1/10/7/7	1/8/5/2	1/6/4/2	Aug 2006
DDoS	47/92/108/NA	324/62/75/NA	28/53/67/NA	Mar 2010
Total	**287/628/586/528**	**206/478/426/362**	**169/364/337/287**	

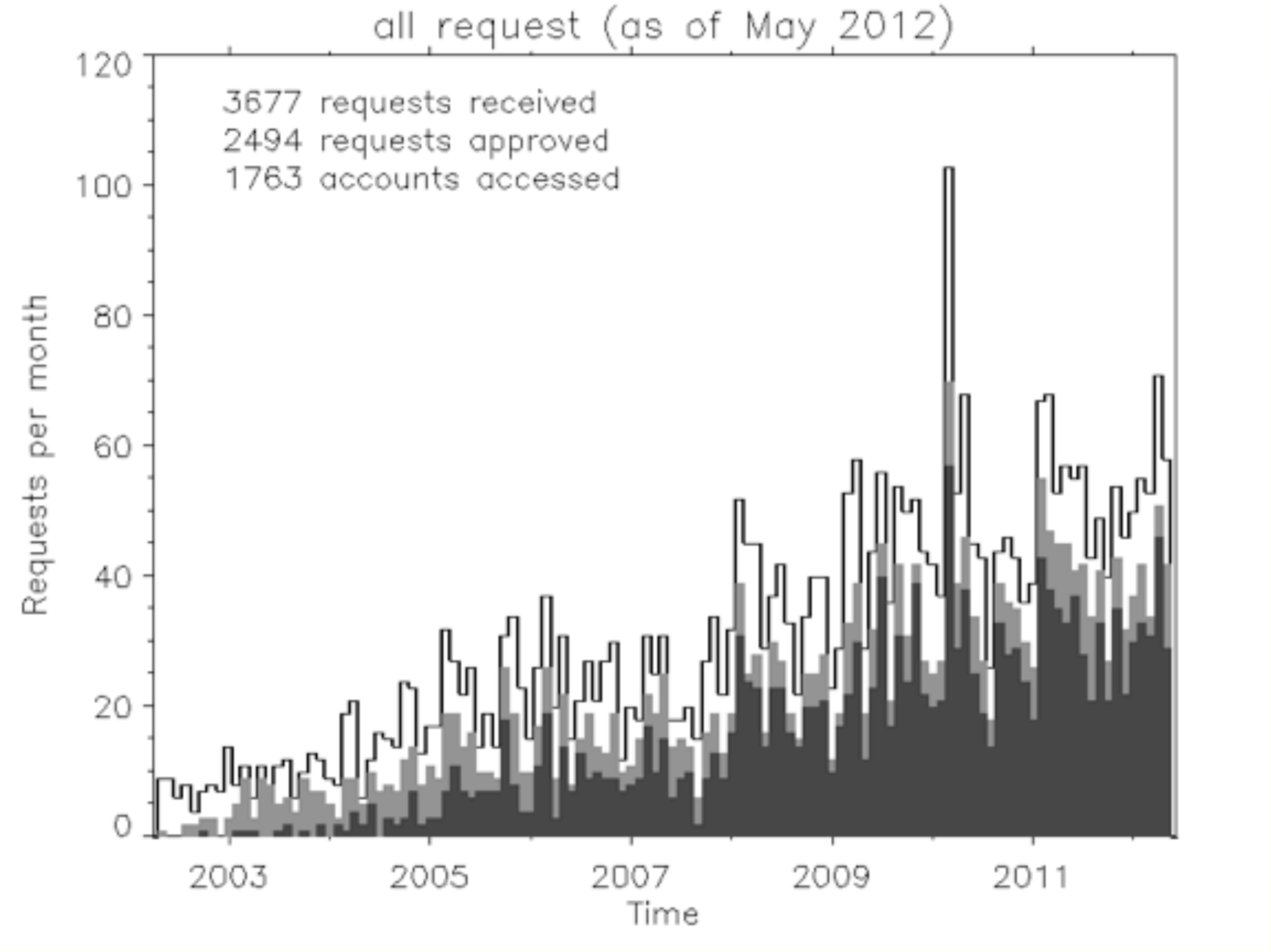
Data request stats

- all requests (cumulative)



Data request stats (cont)

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



Data statistics - online

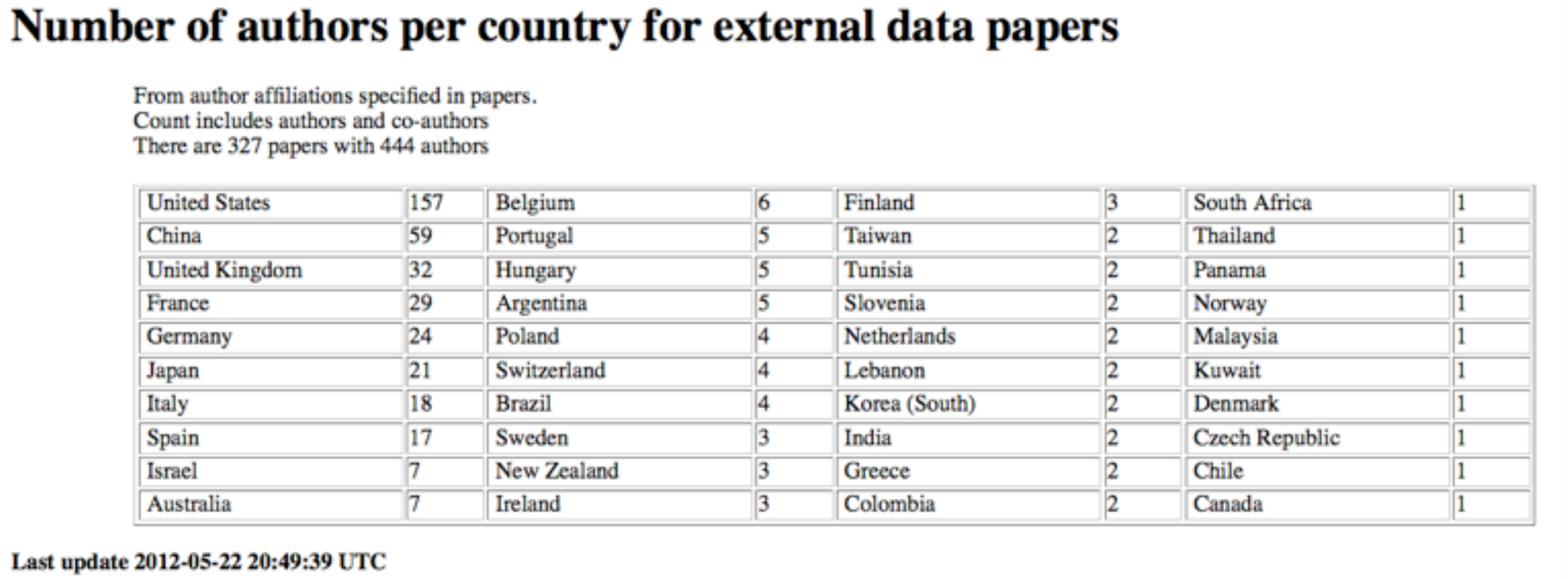
- Report Generator
  - IP-packet-header (traffic) based
  - flows, packet, byte volumes
  - traffic by protocol, port, AS, country, etc
  - http://www.caida.org/data/realtime/passive/?monitor=equinix-sanjose-dirA
- Topology
  - Ark statistics: http://www.caida.org/projects/ark/statistics/index.xml
  - path dispersion (AS and IP), path length distribution, RTT distribution, RTT vs. distance, median RTT per country, ...
- Meta-data for IP packet header data
  - 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/

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
- High-speed backbone link: 2007-2012

non-CAIDA publications using PREDICT-related CAIDA data (that we know of)

total	145
backscatter	21
passive-oc48	56
passive-2007	9
witty	14
itdk	13
skitter	57



Number of authors per country for external data papers

From author affiliations specified in papers.
Count includes authors and co-authors
There are 327 papers with 444 authors

United States	157	Belgium	6	Finland	3	South Africa	1
China	59	Portugal	5	Taiwan	2	Thailand	1
United Kingdom	32	Hungary	5	Tunisia	2	Panama	1
France	29	Argentina	5	Slovenia	2	Norway	1
Germany	24	Poland	4	Netherlands	2	Malaysia	1
Japan	21	Switzerland	4	Lebanon	2	Kuwait	1
Italy	18	Brazil	4	Korea (South)	2	Denmark	1
Spain	17	Sweden	3	India	2	Czech Republic	1
Israel	7	New Zealand	3	Greece	2	Chile	1
Australia	7	Ireland	3	Colombia	2	Canada	1

Last update 2012-05-22 20:49:39 UTC

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 vol. 42, no.1, pp.31--39, Jan 2012.
- kc claffy, *The 4th Workshop on Active Internet Measurements (AIMS-4) Report,* submitted to ACM SIGCOMM CCR.
- kc claffy, *Border Gateway Protocol (BGP) and Traceroute Data Workshop Report",* submitted to ACM SIGCOMM CCR.

Recent publications

- A. Dainotti, A. Pescapè, and K. Claffy, “*Issues and future directions in traffic classification”,* IEEE Network, vol. 26, no.1, pp. 35--40, Jan 2012.
- N. Brownlee, “*One-way Traffic Monitoring with iatmon”,* Passive and Active Network Measurement Workshop (PAM), Vienna, Austria, Mar 2012, PAM 2012.
- K. Keys, Y. Hyun, M. Luckie, and k. claffy, “*Internet-Scale IPv4 Alias Resolution with MIDAR”,* IEEE/ACM Transactions on Networking, 2012.

Recent submissions

- In submission to IMC:
- A. Dhamdhere, M. Luckie, B. Huffaker, K. Claffy, A. Elmokashfi, E. Aben, “*IPv6 Will Be Deployed Any Day Now*”.
- B. Huffaker, M. Fomenkov, K. Claffy, “*Internet Topology Data Comparison*”.
- T. Zseby, N. Brownlee, A. King, K. Claffy, “*Entropy-based Classification of IP Darkspace Events*”. (used Corsaro 8-tuple)
- A. Dainotti, A. King, K. Claffy, F. Papale, A. Pescapè, “*Analysis of a ‘/0’ Stealth Scan from a Botnet*”. (used Corsaro 8-tuple)

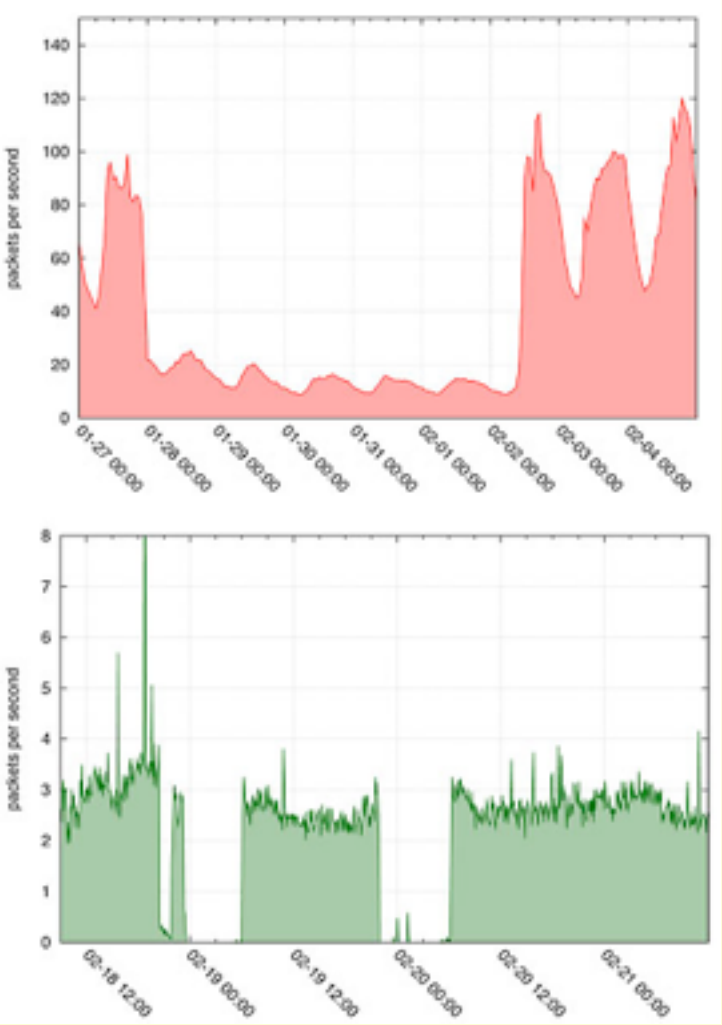
Recent presentations

- k claffy remotely presented "*Extracting Benefit from Harm: Using Malware Pollution to Analyze the Impact of Political and Geophysical Events on the Internet*" at NZNOG, APRICOT http://www.caida.org/publications/presentations/2012/extracting_benefit_from_harm/.

UCSD Press Release

- Internet Censorship Revealed Through the Haze of Malware Pollution
- http://ucsdnews.ucsd.edu/pressreleases/internet_censorship_revealed_through_the_haze_of_malware_pollution/

These graphs show the amount of IBR, or "malware" activity - and the sharp drops related to the actions by the Egyptian government (top) and Libyan government (bottom) in response to the political demonstrations that occurred in early 2011. The data was observed by the UC San Diego Network Telescope in terms of packets per second, or basic messages. The oscillating pattern (better seen in the top image) is due to human, or diurnal, patterns of activity because IBR mostly comes from users' personal computers, not computer servers. Images courtesy of CAIDA/SDSC.



Recent blogs

- kc claffy, *Shutting the phone network off while you're running out of internet protocol numbers* http://blog.caida.org/best_available_data/2012/01/20/shutting-the-phone-network-off-while-youre-running-out-of-internet-protocol-numbers/
- kc claffy, *NASA's recent DNSSEC snafu and the checklist*

  http://blog.caida.org/best_available_data/2012/02/16/nasas-recent-dnssec-snafu-and-the-checklist/
- Josh Polterock, *The Menlo Report and its Companion bring ethical guidelines to ITC research* http://blog.caida.org/best_available_data/2012/02/07/the-menlo-report-and-its-companion-bring-ethical-guidelines-to-itc-research/
- kc claffy, *The 2nd NDN Project Retreat*

  http://blog.caida.org/best_available_data/2012/02/05/the-2nd-ndn-project-retreat/.

Recent blogs

- Josh Polterock, *Internet Censorship Revealed Through the Haze of Malware Pollution* http://blog.caida.org/best_available_data/2012/03/28/internet-censorship-revealed-through-the-haze-of-malware-pollution/
- kc claffy, *Second Workshop on Internet Economics (WIE2011)*

  http://blog.caida.org/best_available_data/2012/03/05/second-workshop-on-internet-economics-wie2011/
- Amogh Dhamdhere, *Twelve Years in the Evolution of the Internet Ecosystem* http://blog.caida.org/best_available_data/2012/04/10/twelve-years-in-the-evolution-of-the-internet-ecosystem/
- Josh Polterock, *Targeted Serendipity: the Search for Storage*

  http://blog.caida.org/best_available_data/2012/04/04/targeted-serendipity-the-search-for-storage/.

CAIDA Master AUA

- Master AUA 1.0 for all CAIDA data sets
  - http://www.caida.org/home/legal/aua/
  - Factors out common conditions
  - Removes inconsistencies
  - Covers multiple categories of data - different levels of sensitivity
    - passive traces
    - active traces and derived topology
- Real-time telescope data: supplemental clauses
- Fixes document proliferation
  - 7 data request forms
  - 22 data set web pages and README files
- (Still) want to discuss having a common AUP on PREDICT portal that meets all PIs' needs

DUST 2012

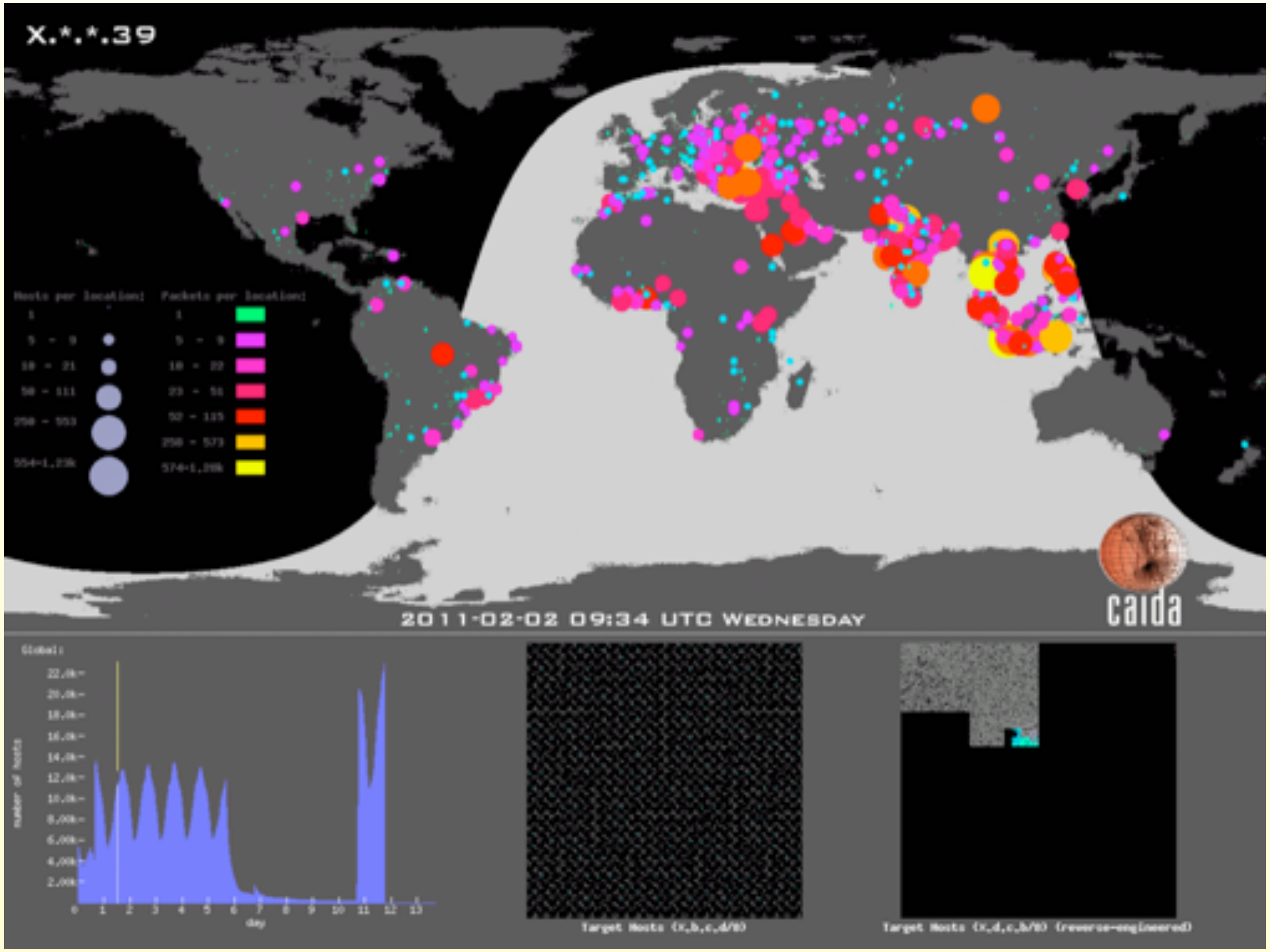
- 1st International Workshop on Darkspace and UnSolicited Traffic Analysis (DUST)
  - 14-15 May 2012 at SDSC, UC San Diego
  - brought together researchers, operators, and analysts interested in unsolicited traffic analysis, especially traffic destined to unassigned (dark) IP address space.
  - 20 presenters
  - workshop report will be submitted to CCR
  - presentations at http://www.caida.org/workshops/dust/1205/

SIPscan

- *“SIPscan: the (IPv4) world scanning itself”*
  - A. Dainotti, A. King, K. Claffy, F. Papale, A. Pescapè
- [demo animation]

Presented at: ISOI, DUST, DANCES, CSE SysNet lunch.

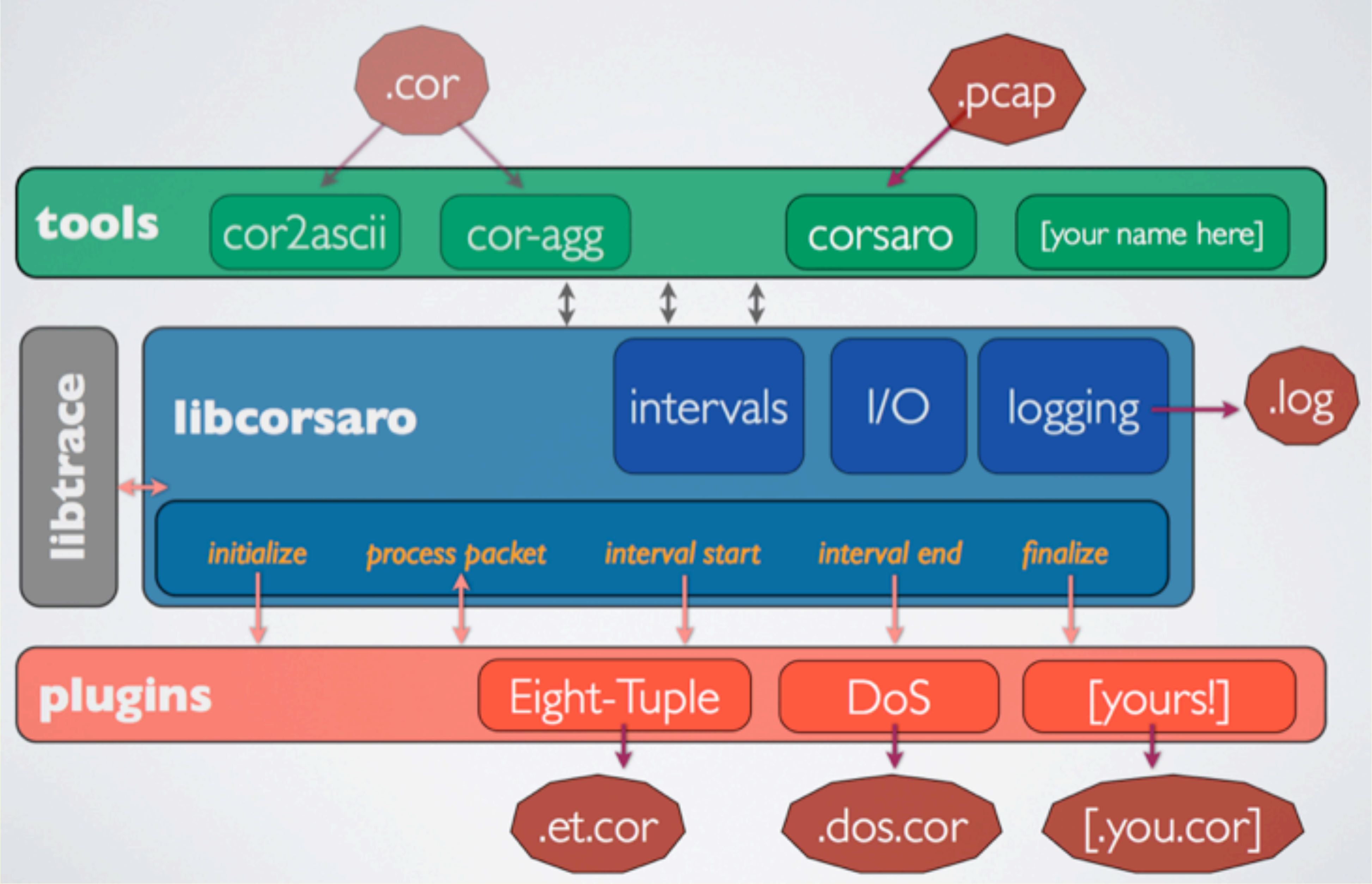
Paper submitted to IMC2012.



Corsaro - Analysis and Indexing Framework

- tools and extensible framework for packet train ad-hoc analysis, plugins, post-processing data management
- aggregates data into intervals (1-min bins)
- plugins
  - Custom 8-tuple flow record balances storage resources and research utility
- http://www.caida.org/publications/presentations/2012/dust_corsaro/

Corsaro - Analysis and Indexing Framework



iatmon - Interarrival Time Monitor

- N. Brownlee, “One-way Traffic Monitoring with iatmon'', PAM 2012
  - reads trace data from file or live interface
  - using WAND’s libtrace library
  - hash table of source addresses
  - writes summary files characterizing sources
  - source available at http://www.caida.org/tools/measurement/iatmon/

CAIDA PREDICT Updates

- Updates to http://www.caida.org/projects/predict/
- Submitted new PREDICT proposal, “Supporting Research and Development of Security Technologies through Network and Security Data Collection”
- Hired new REU to help with PREDICT video (on ASranking)
- Hired recent law school grad to help with “bluebooking” of top ten paper for law journal submission

Menlo Report and Companion Report

- Published: *Ethical Principles Guiding Information and Communication Technology Research: The Menlo Report*.
- Published: *A Companion Report* details the principles and applications more granularly and illustrates their implementation in real and synthetic case studies.
- Published: *The Interaction of the Menlo Report and Revisions to the Common Rule-Comments in Response to the Advanced Notice of Proposed Rulemaking (ANPRM)*.
- See Erin's presentation tomorrow.

Open Issues

- PCC
  - Portal issues: administrative, web site sill unusable
  - UCSD contracting agent still receiving requests before we approve them.
  - Stats on universities who have tried to join PREDICT and given up: why?
- Policy: Commercial use of data
  - OK for govt-funded commercial
  - o/w, CAIDA’s industry evaluation data sets
- Technical: Scalability of real-time telescope
- Strategic/Community-building
  - DUST2013: Day in the Life of Darkspace?