MIC - DHS Meeting
CAIDA Research Overview

Bradley Huffaker, CAIDA
(PI kc claffy)
Tokyo, Japan
20 June 2012

Overview

- CAIDA’s mission statement
- global cybersecurity challenges and CAIDA's approach
- measurement infrastructure
- data lifecycle management challenges
- research highlights
- workshops
- japanese related research and collaborations

CAIDA’s Mission Statement

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.

Global Cybersecurity Challenges and CAIDA’s approach

New measurement and data collection technologies and infrastructure to improve situational awareness and understanding of the structure, dynamics and vulnerabilities of the physical and logical topologies of the global Internet.

- internet measurement, analysis and inference techniques
- topology mapping: current annotated Internet topology
- unsolicited traffic analysis to study macropscoic Internet events
- geolocation technology assessment

Global Cybersecurity Challenges and CAIDA’s approach

Provide enabling infrastructure that allows researchers to quickly design and deploy widely distributed cybersecurity research experiments.

- address network science crisis
- facilitate data transfer, curation, archival, and privacy-respecting sharing
- scalabile system management and measurement efficiency
- flexibility in traffic and topology measurement methods
- let researchers spend less time on non-research

Global Cybersecurity Challenges and CAIDA’s approach

Provide best available empirical data sets and analyses relevant to cybersecurity research, future Internet architecture research, and policy/legal frameworks that are persistently behind technological pace of progress.

- best available raw and curated data sets about observable IP-level topology
- provide empirical grounding to parameterize and validate theoretical modeling and applied research efforts
- enable study of clean-slate routing on realistic network topologies
- inform federal communications policy

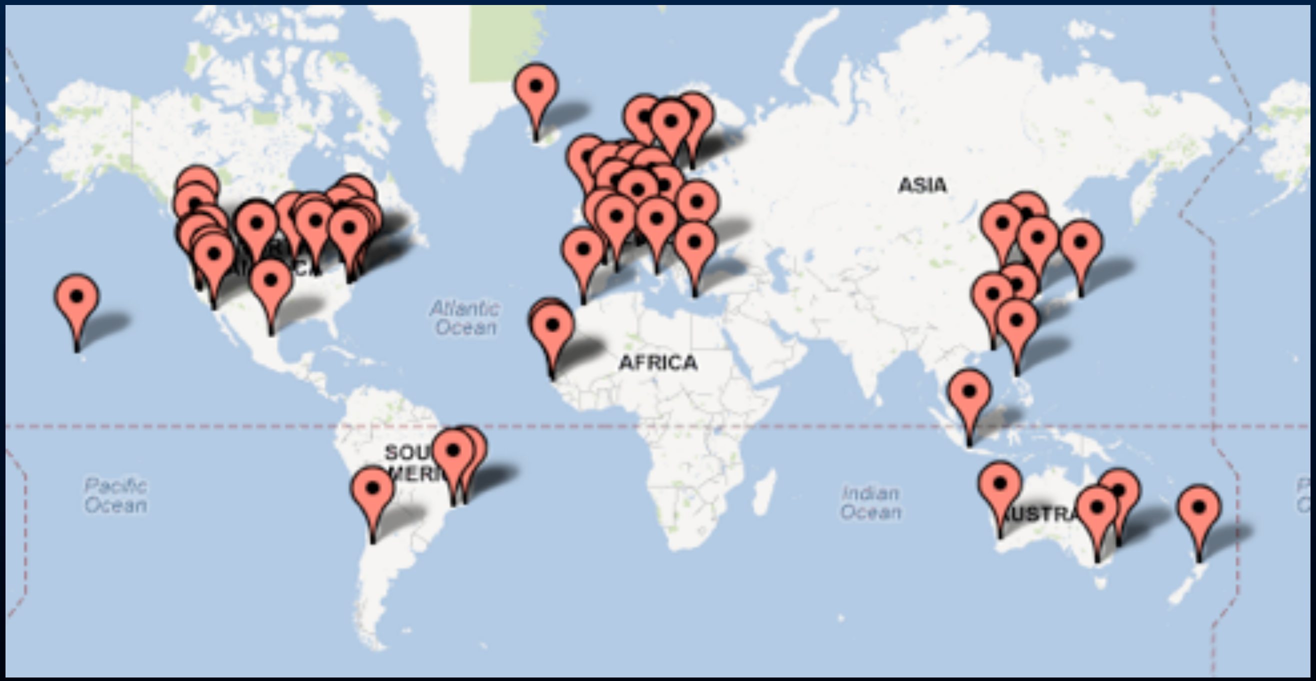
Measurement Infrastructure

- Archipelago (ark)
  - CAIDA active measurement infrastructure
  - supports ongoing topology measurement as well as customized experiments
- UCSD Internet Telescope (darknet)
  - packet capture to largely unused address space (one-way traffic only)
- Passive Trace Capture
  - captures packets on Tier 1 10GE backbone link (two-way traffic)
  - shared anonymized headers only
- HostDB
  - historical record of IP address to DNS hostname database

Archipelago

http://www.caida.org/projects/ark

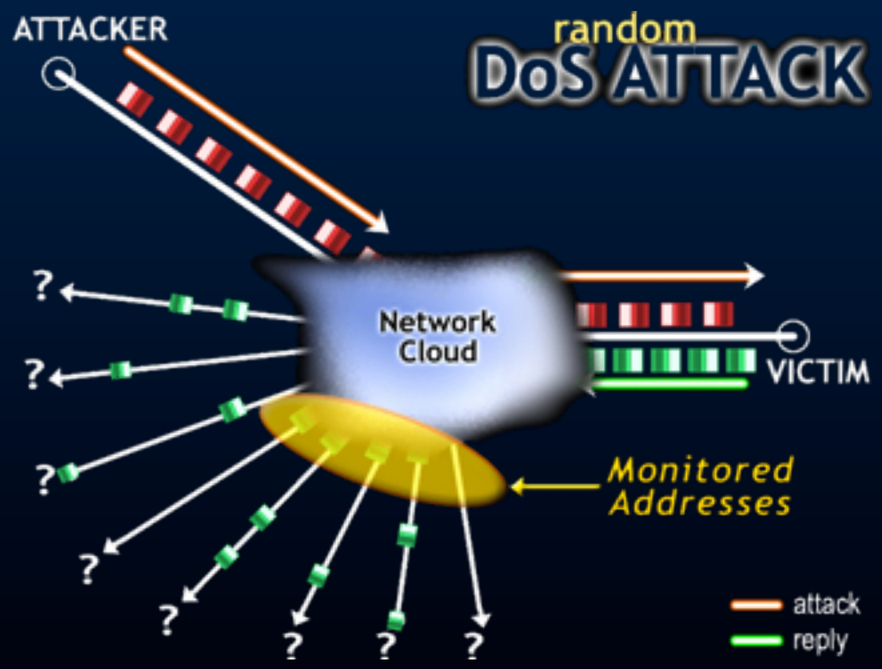
- CAIDA's active measurement infrastructure
- 60 monitors – growing 1 or 2 per month
  - 28 IPv6 capable
  - 31 countries
- current projects
  - team-probing experiment to collect IPv4 and IPv6 topology
  - alias resolution measurements
  - researcher experiments, e.g., spoofer



UCSD Network Telescope

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

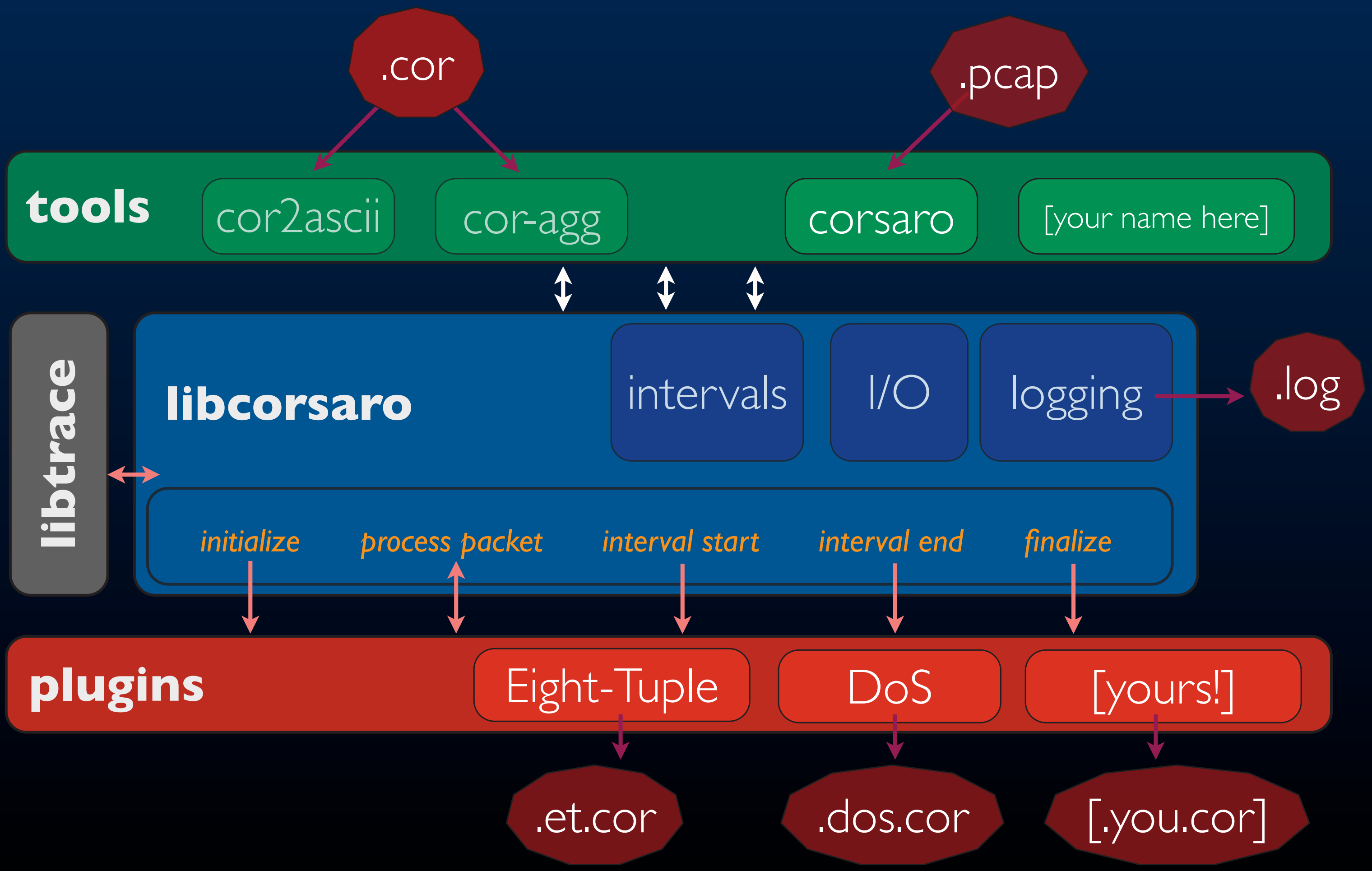
- A portion of the Internet address space that
  - address space that is mostly unused
  - so most traffic it gets is unsolicated
    - botnet, scanning, etc
- **1/256** of the **Internet** address space
- traffic reaching the router is therefore *unsolicited* (Internet background Radiation)
- we collect and analyze



Corsaro - Analysis and Indexing Framework

http://www.caida.org/publications/presentations/2012/dust_corsaro/

- tools and extensible framework for packet train ad-hoc analysis, plugins, post-processing data management
- aggregates data into intervals (1-min bins)
- plug-ins (8-tuple flow records)

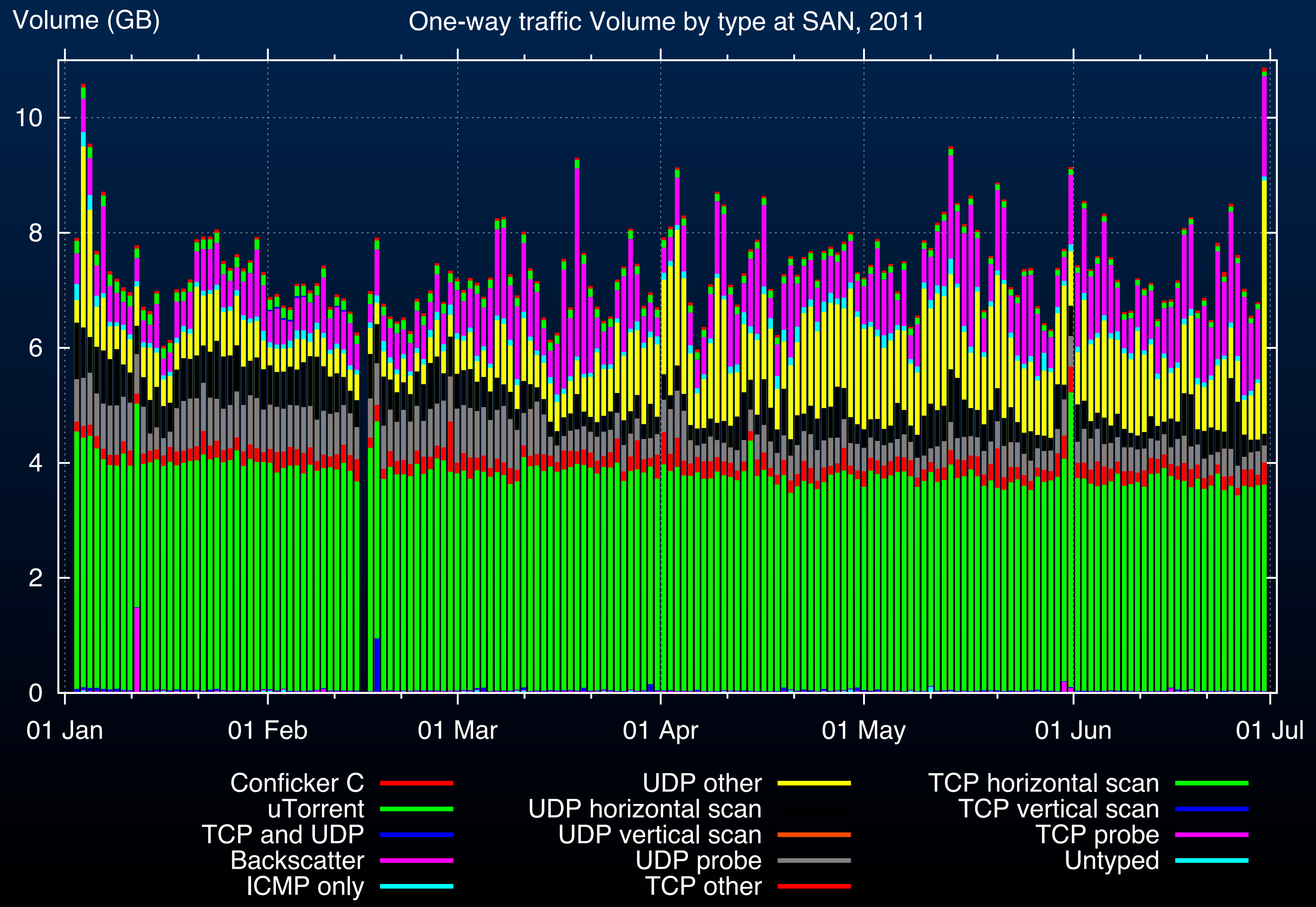


One-way Traffic Monitoring with iatmon

N. Brownlee

Passive and Active Network Measurement Workshop (PAM), Vienna, Austria, Mar 2012, PAM 2012.

- inter-arrival time Monitor
- classify into source types (scan/backscatter/worm)
- analyze inter arrival times
- identifies stealth scans/ sources



CAIDA Data Collections

http://www.caida.org/data/overview/

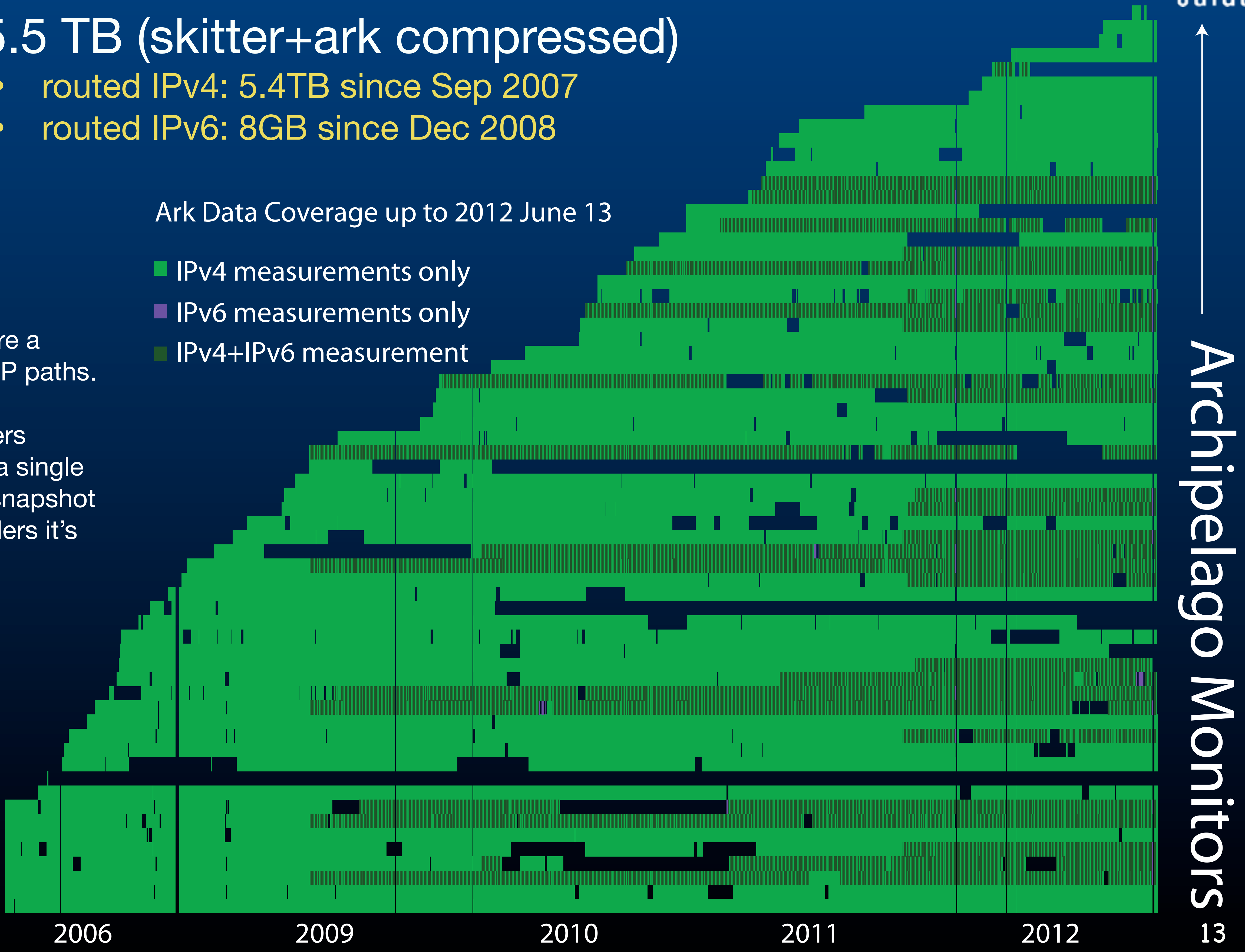
- performance
  - DNS root/gTLD RTT DATA
- security
  - Code Red Worms, Backscatter, DDoS attacks, Witty Worm, Conflicker
- topology
  - AS Links, Prefix to AS, AS Rank, AS Relationships, Archipelago IPv4+IPv6 topology
- topology (processed)
  - Macroscopic Internet Topology Data Kit (ITDK)
- traffic
  - Telescope Data, Telescope (live), Anonymized Internet Traces, Tier 1 packet traces, SDNAP

Archipelago monitors and data

- 5.5 TB (skitter+ark compressed)
  - routed IPv4: 5.4TB since Sep 2007
  - routed IPv6: 8GB since Dec 2008

Raw traces are a collection of IP paths.

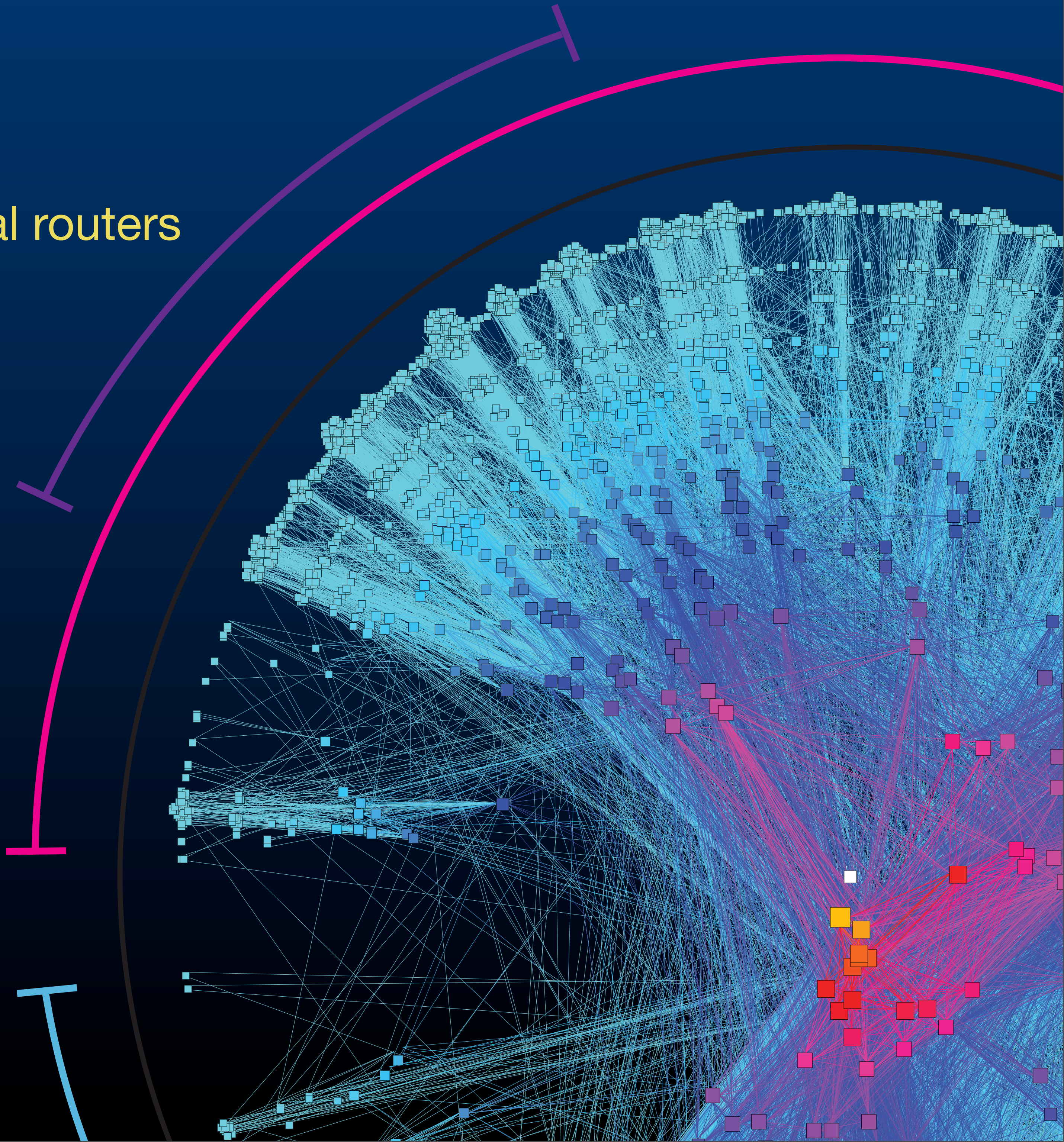
For researchers interested in a single microscope snapshot CAIDA providers it’s ITDK.


Ark Data Coverage up to 2012 June 13

Internet Topology Data Kit (ITDK)

http://www.caida.org/data/active/internet-topology-data-kit/

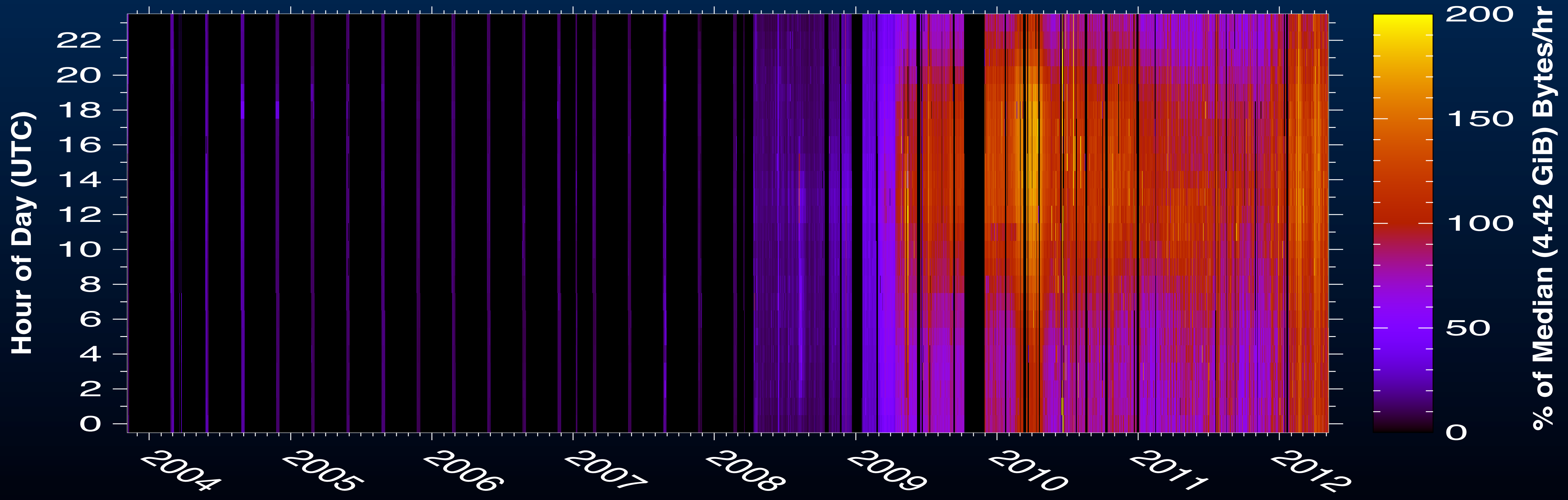
- macroscopic Internet topology snapshot
  - provides a curated, annotated router level topology
- annotations
  - IP address aliased to routers
  - geographic location of individual routers
  - DNS hostnames for IP address
  - router to AS assignments
- multiple snapshots
  - Apr 2010
  - Jul 2010
  - Apr 2011



UCSD Internet Telescope (darknet)

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

- packet capture of unsolicited traffic to unassigned addresses
- 2012 data still live on disk (currently ~146 days)
  - 16.92 TiB compressed, 34.28 TiB uncompressed
- data archived to NERSC in April
  - 105 TiB compressed/encrypted



Data Lifecycle Management Challenges

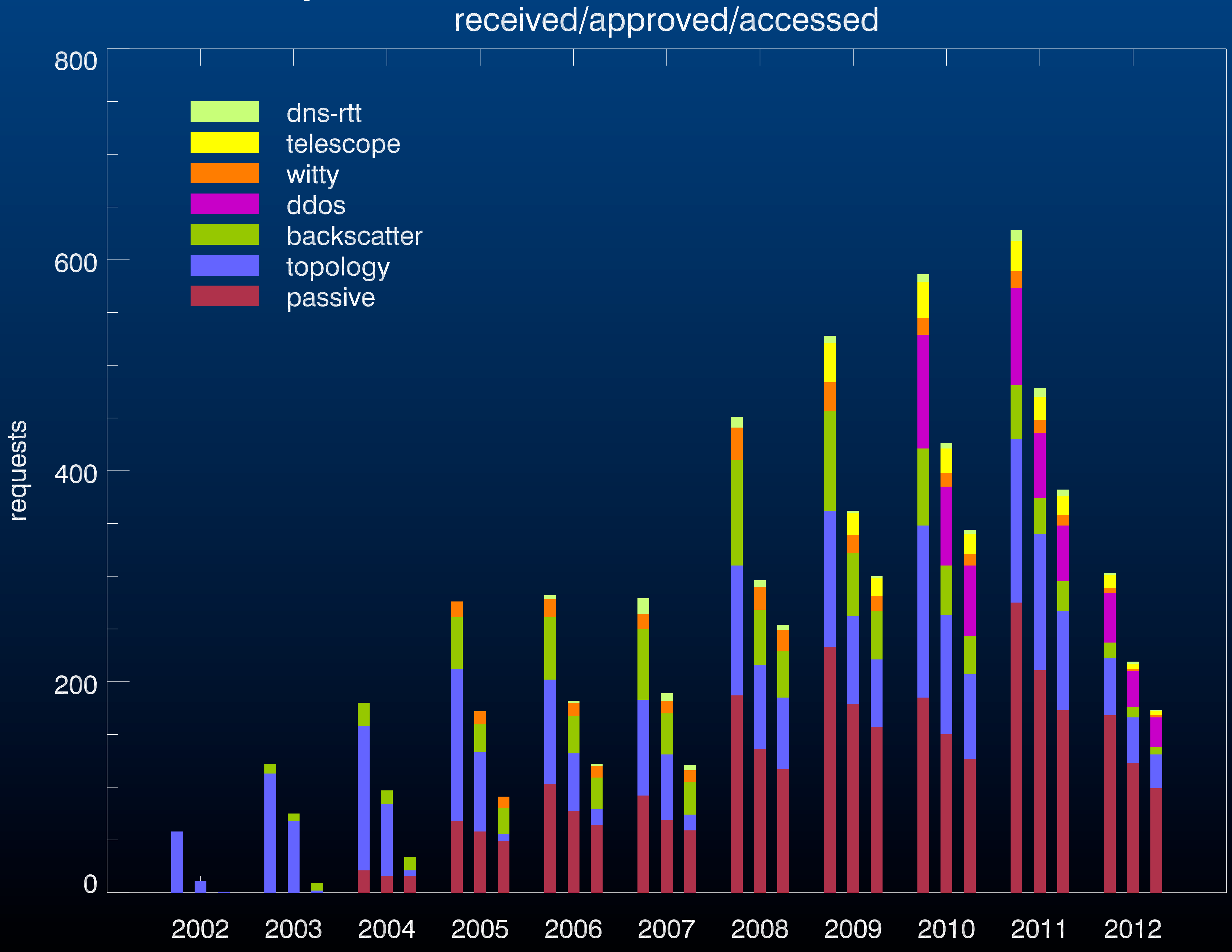
- collection: **technology** and **policy** barriers
- processing: requires **expertise** in data curation
  - many researchers lack time, inclination, or resources
- management
  - moving data from collection points to storage
- storage
  - e.g., CAIDA stores ~110TB (as of 13 Jun 2011)
- sharing
  - disclosure control policies and technologies
  - ethical issues in sharing data
- tracking usage
  - data set dissemination statistics
  - scientific results enabled by data
  - acknowledgement to funding agencies

Data Sharing (AUA)

- What is a AUA (Acceptable Use Agreement)?
    - agreement between CAIDA and an external researcher that specifies restrictions on use of data CAIDA is sharing w the researcher
- Why we need a Acceptable Use Agreement (AUA)?
    - protect misuse of Personally Identifiable Information, e.g., SSNs
    - balance privacy and research utility using CAIDA's framework for ethical data sharing (IEEE Security & Privacy, vol. 8, no. 4, pp. 31--39, Jul 2010.)
- Master AUA 1.0 for all CAIDA data sets

    http://www.caida.org/home/legal/aua/
    - factors out common conditions
    - Removes previous inconsistencies across proliferation of data request forms
    - covers multiple categories of data - different levels of sensitivity
        - passive traces
        - active traces and derived topology
    - supplemental clause for specific datasets (e.g. Real-time telescope)

Data Requests



Data Set Popularity Ranking

- 1st: IP Packet Header traces (10GE/OC192, OC48)
  - requested 850 times, accessed 555 times (since 2009)
  - who used it: 277 .edu, 162 .cn, 45 .uk, 45 .com (since 2004)
    - 57 more domains
    - 1010 total accounts: 270 from U.S.
- 2nd: topology data
  - requested 497 times, accessed 255 times (since 2009)
  - who used it: 265 .edu,125 .cn,45 .uk,32 .kr, 31 .com, 26 .jp (since ‘04)
    - 54 more domains
    - 784 total accounts: 258 from U.S.

Data availability

PREDICT: OC48 traces, topology, telescope

Derived data sets publicly available (i.e., AS-links)

- sample use: http://semilattice.net/projects/map-of-the-internet/

Academics who sign updated AUA

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

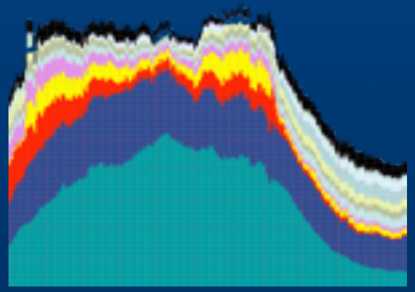
Commercial researchers

- a small sample of ‘industry evaluation’ data to entice interest

join CAIDA, various membership levels are offered

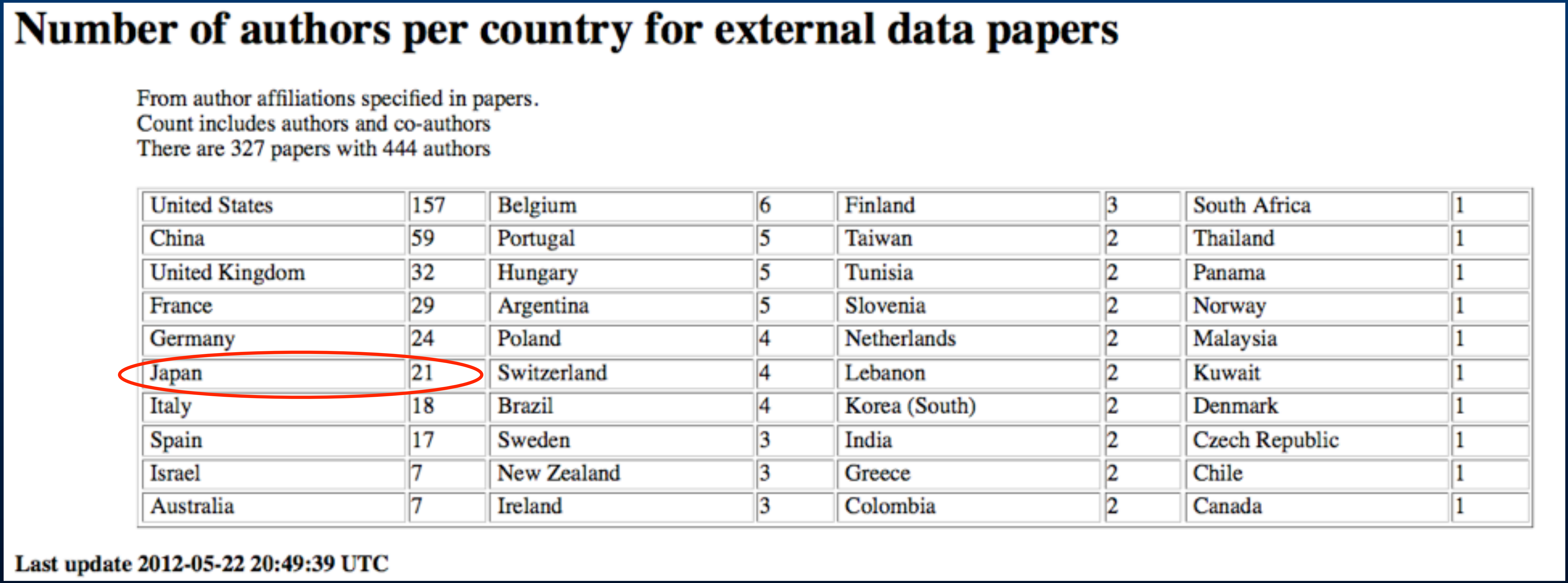
Data statistics - online

- 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/

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



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

Publications using CAIDA data

as of May 2012

- OC192 and OC48 traces: traffic classification, performance modeling, monitoring, filtering, generation, locality

  http://www.caida.org/data/publications/bydataset/index.xml#passive

  - 81 publications (54 from data in PREDICT)

- UCSD telescope: Conficker, worm research

  http://www.caida.org/data/publications/bydataset/index.xml#Backscatter

  - 27 publications (all from data in PREDICT)

- topology: pkt traceback, marking, DOS defense, topology and routing modeling, discovery, metrics, methodology improvements

  http://www.caida.org/data/publications/bydataset/index.xml#Topology

  - 65 publications (53 from data in PREDICT)

Research Highlights

- topology analysis
  - Internet scale router alias resolution
  - comparison between IPv6 and IPv4 topology
  - publicly available macroscopic Internet topology data
- geolocation analysis
  - comparison of public and private geolocation services
- Internet peering analysis
  - inferring AS relationships
  - parameterized computational model
  - AS ranking
- infrastructure security stability
  - analysis of country-wide Internet outages
  - analysis of stealth scan from botnet
- modeling complex networks
  - examining various complex networks through hidden metric spaces
- visualization

Publication review 2011-2012

http://www.caida.org/publications/papers

- topology analysis (2 papers)
- Internet scale router alias resolution (2 papers)
- geolocation analysis (1 papers)
- Internet peering analysis (2 papers)
- infrastructure security stability (2 papers)
- modeling complex networks (4 papers)
- policy and ethical guidance (4 papers)
- measurement methodology (5 papers)
- workshop (4 papers)

Communication via CAIDA's blog

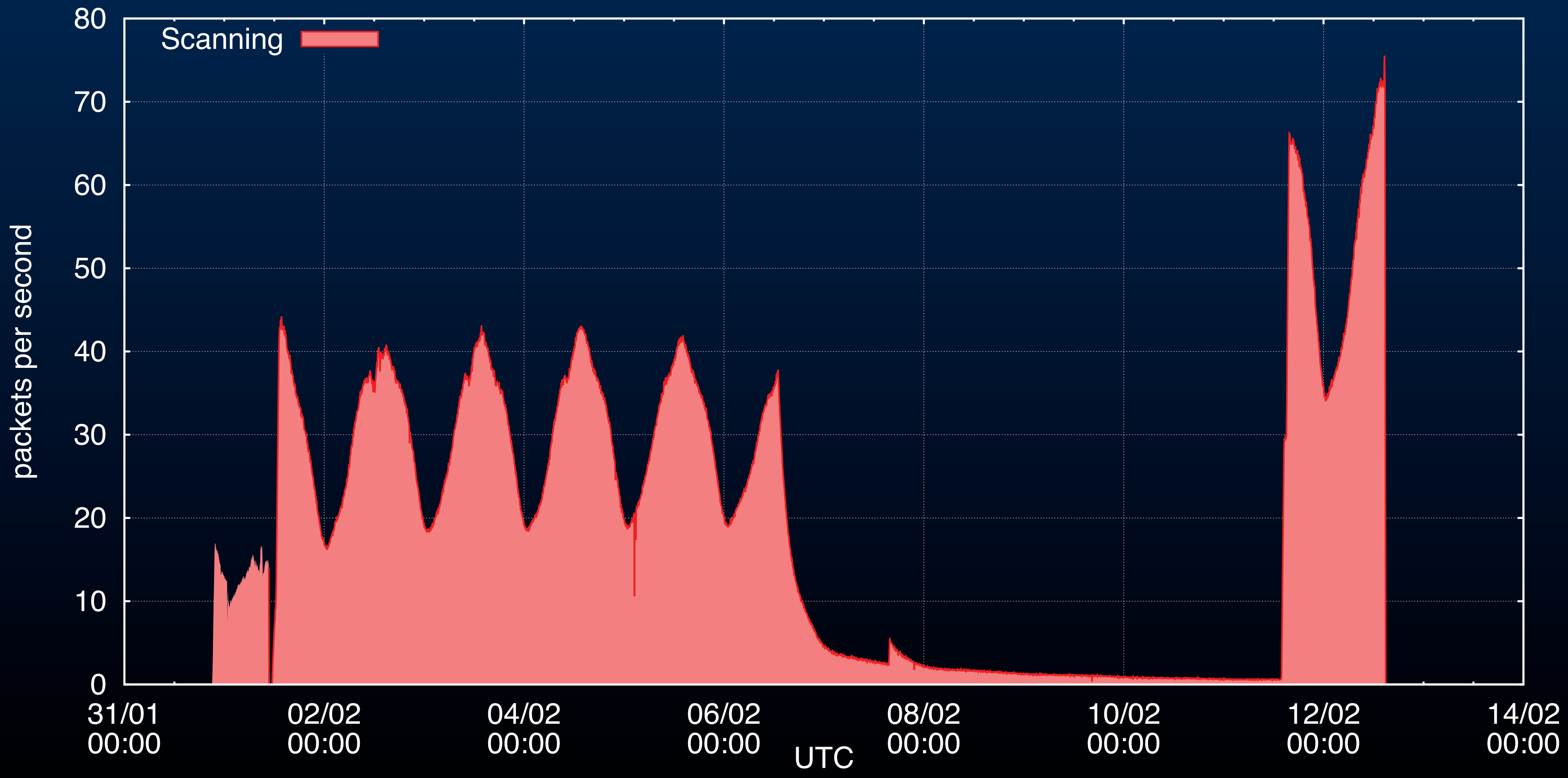
http://blog.caida.org

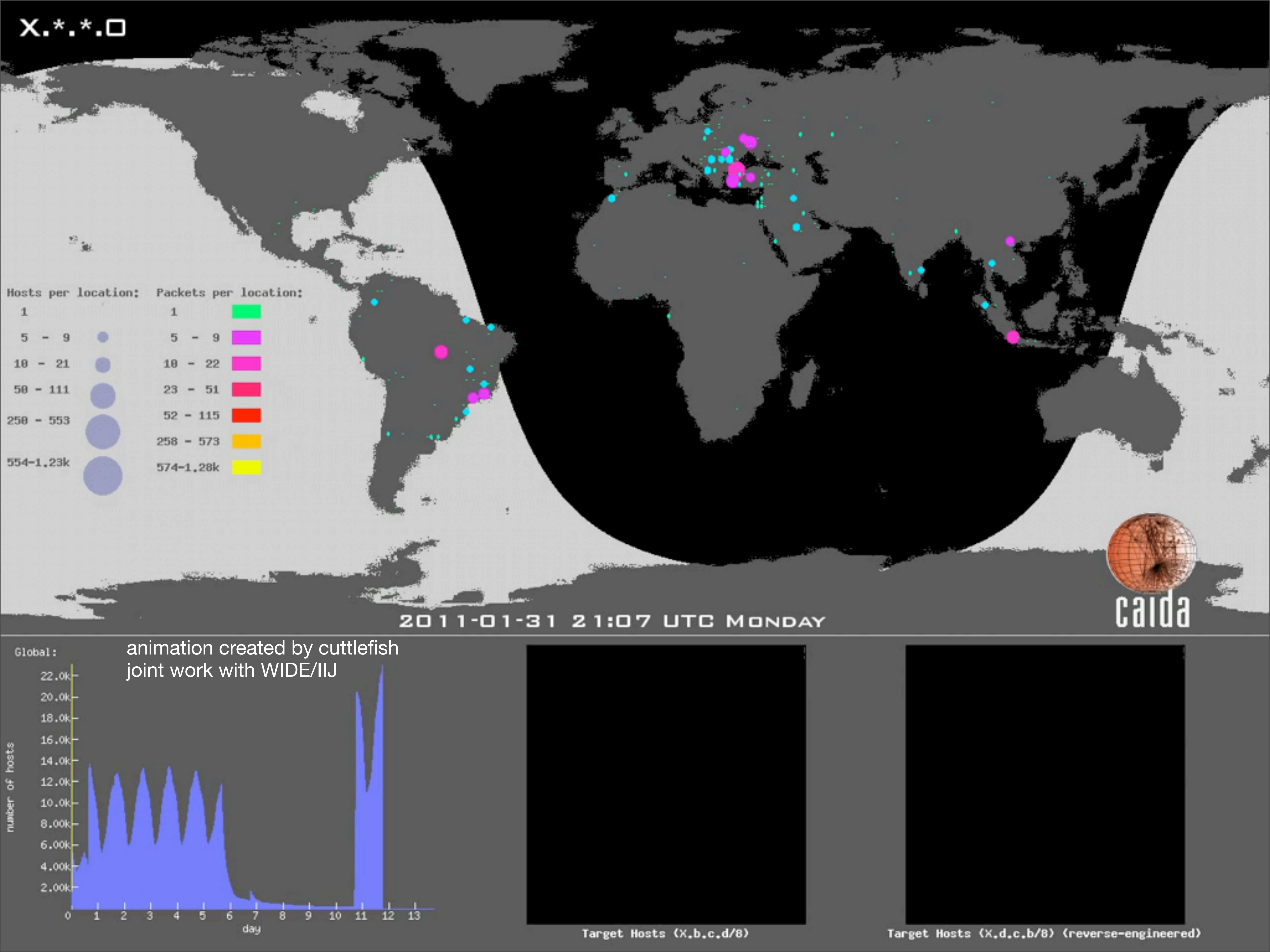
- Matthew Luckie, IPv6: What could be (but isn't yet)
- kc claffy, Shutting the phone network off while you're running out of internet protocol numbers
- kc claffy, NASA's recent DNSSEC snafu and the checklist
- Josh Polterock, The Menlo Report and its Companion bring ethical guidelines to ITC research
- kc claffy, The 2nd NDN Project Retreat
- Josh Polterock, Internet Censorship Revealed Through the Haze of Malware Pollution
- kc claffy, Second Workshop on Internet Economics (WIE2011)
- Amogh Dhamdhere, Twelve Years in the Evolution of the Internet Ecosystem

Analysis of a '/0' Stealth Scan from a Botnet

A. Dainotti, A. King, kc Claffy, F. Papale, A. Pescapè

- a “/0” scan from a botnet (SIPscan)
- observed by the UCSD telescope (a /8 darknet)
    - validated with MAWI traffic traces from WIDE (JP)
- scanning SIP Servers with a specific query on UDP port 5060 and SYNs on TCP port 80





animation created by cuttlefish
joint work with WIDE/IIJ

research

Analysis of Country-wide Internet Outages Caused by Censorship

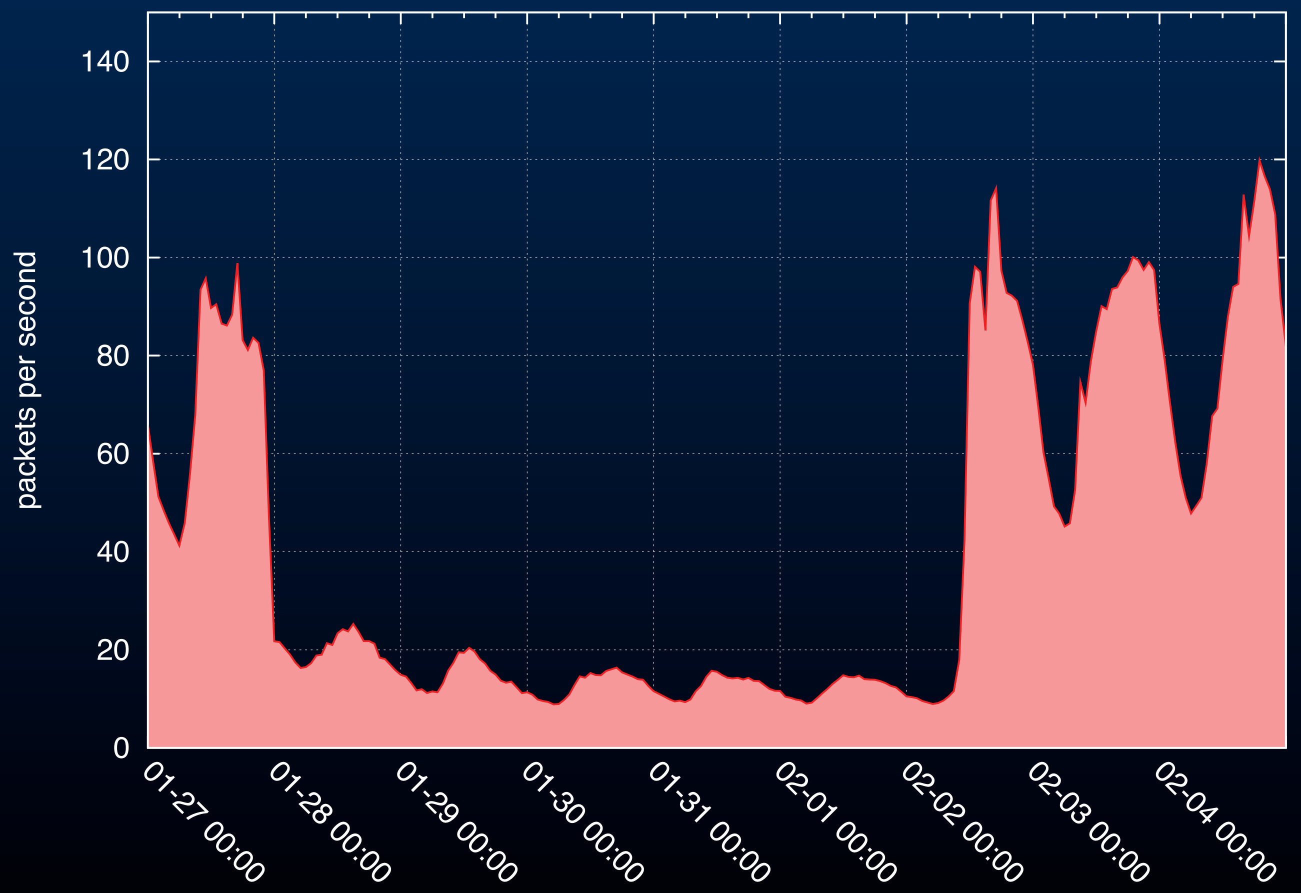
A. Dainotti, C. Squarcella, E. Aben, kc Claffy, M. Chiesa, M. Russo, A. Pescapè

Internet Measurement Conference (IMC), Berlin, Germany, Nov 2011, pp. 1--18, ACM

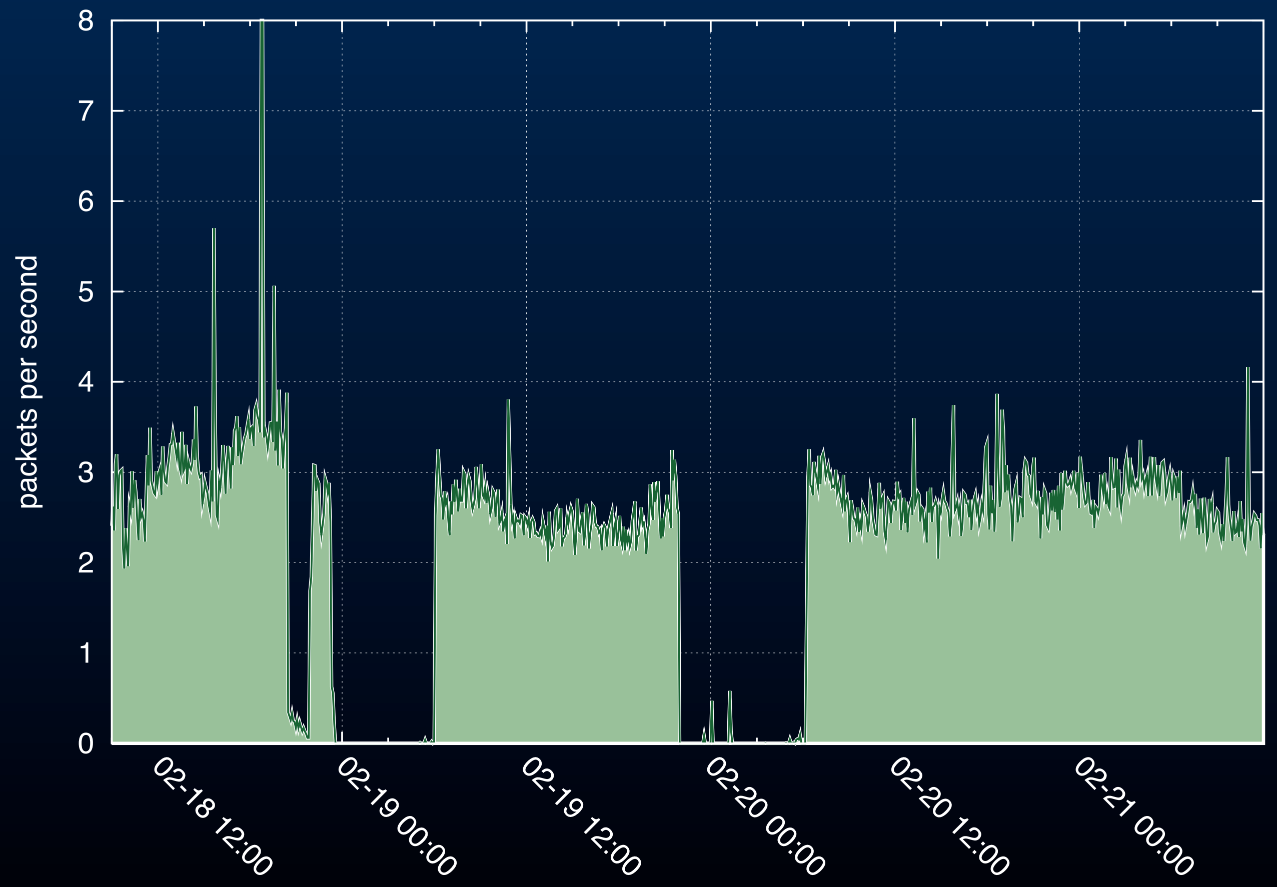
2012 Applied Networking Research Prize (IRTF)

- combining BGP, Archipelago, and UCSD Telescope measurements to analyze country-wide outages during the “Arab Spring”

Egypt



Libya



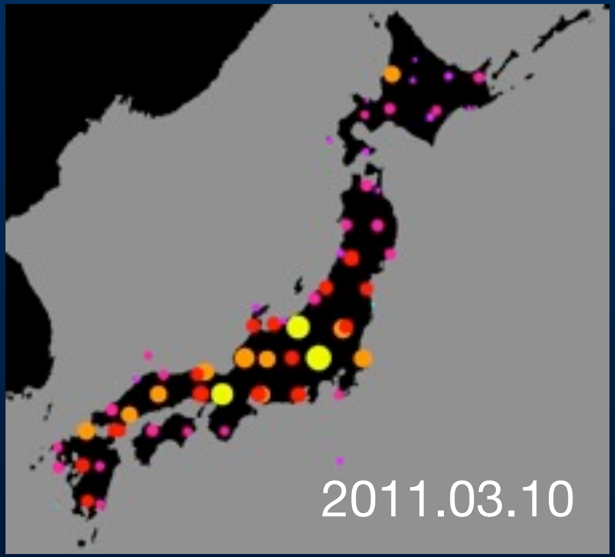
Extracting benefit from harm

A. Dainotti, R. Amman, E. Aben, kc Claffy

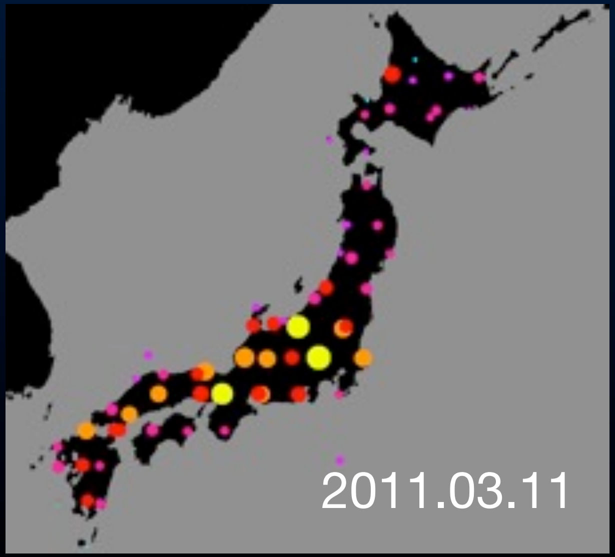
ACM SIGCOMM Computer Communication Review (CCR), vol. 42, no. 1, pp. 31--39, Jan 2012

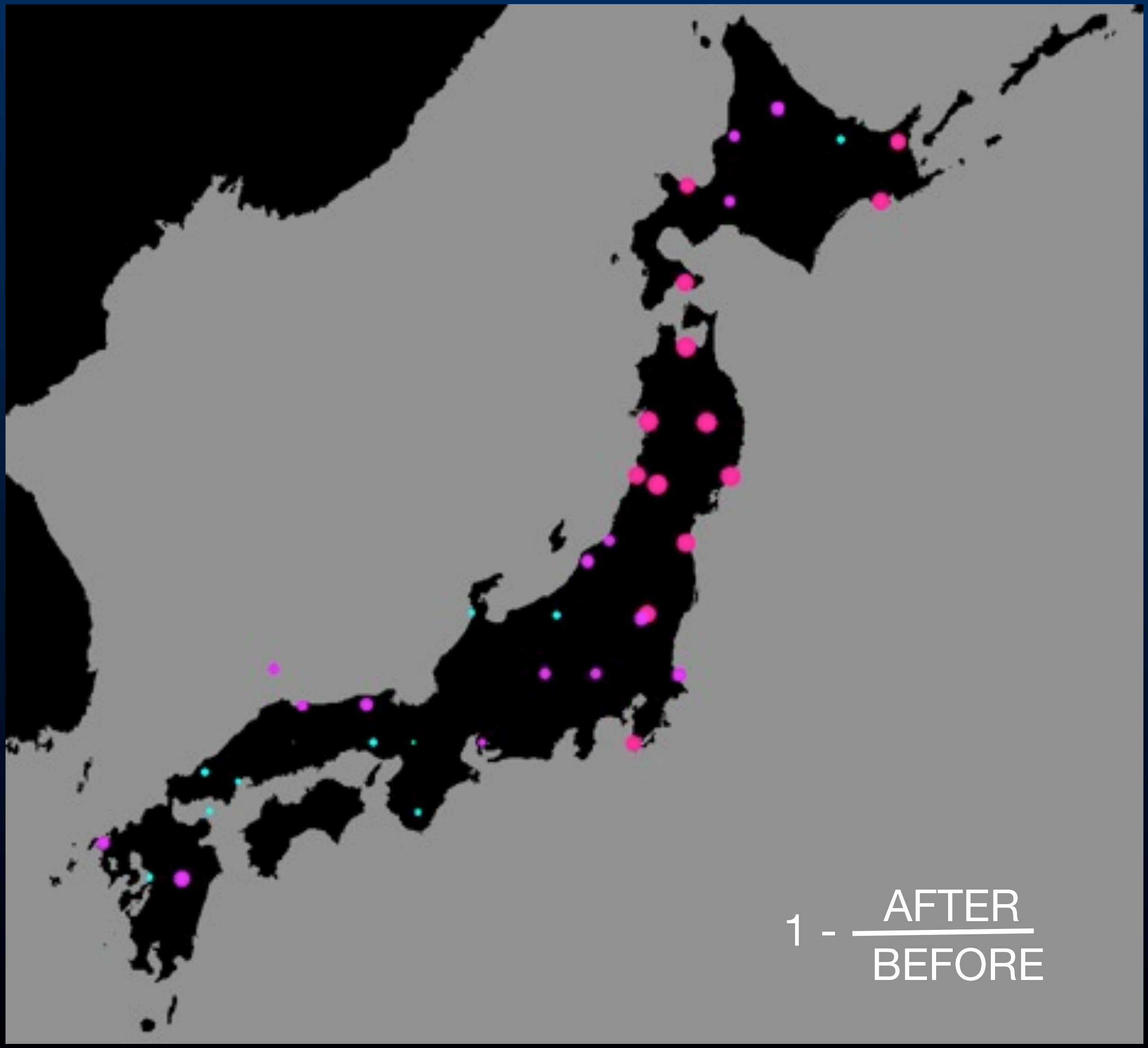
among three best ACM CCR papers 2011-2012

- Effects of Tōhoku Earthquake seen from UCSD Telescope



Number of Unique IP addresses in 24 Hours



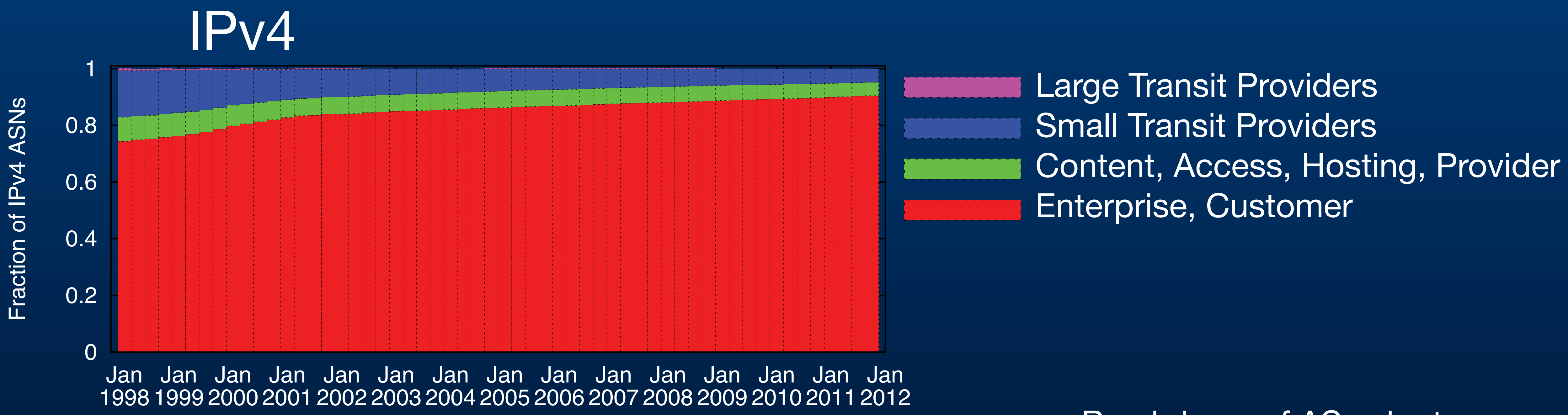


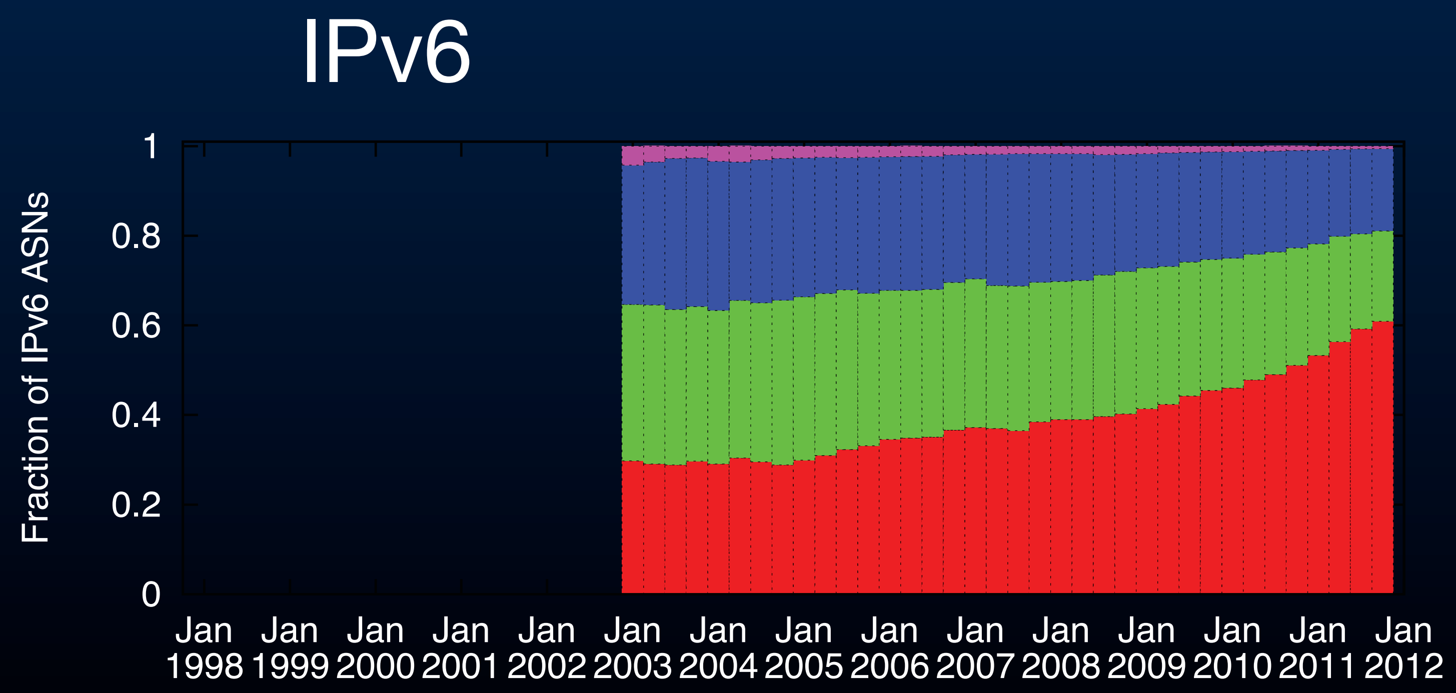
$$1 - \frac{\text{AFTER}}{\text{BEFORE}}$$

precentage change

IPv6 Will Be Deployed Any Day Now

A. Dhamdhere, M. Luckie, B. Huffaker, kc Claffy, A. Elmokashfi, E. Aben



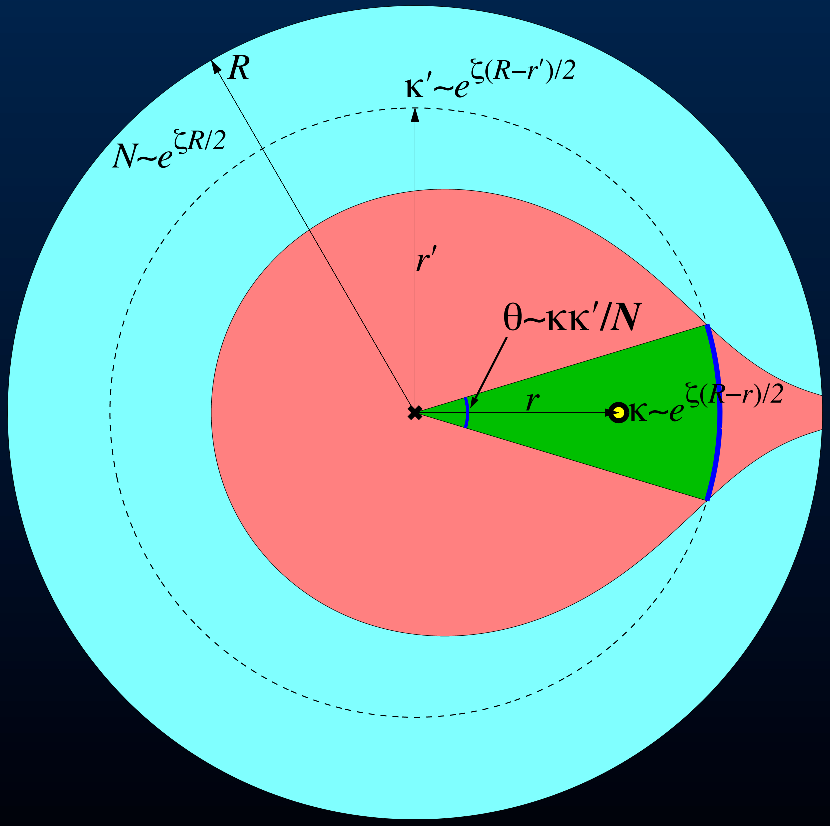


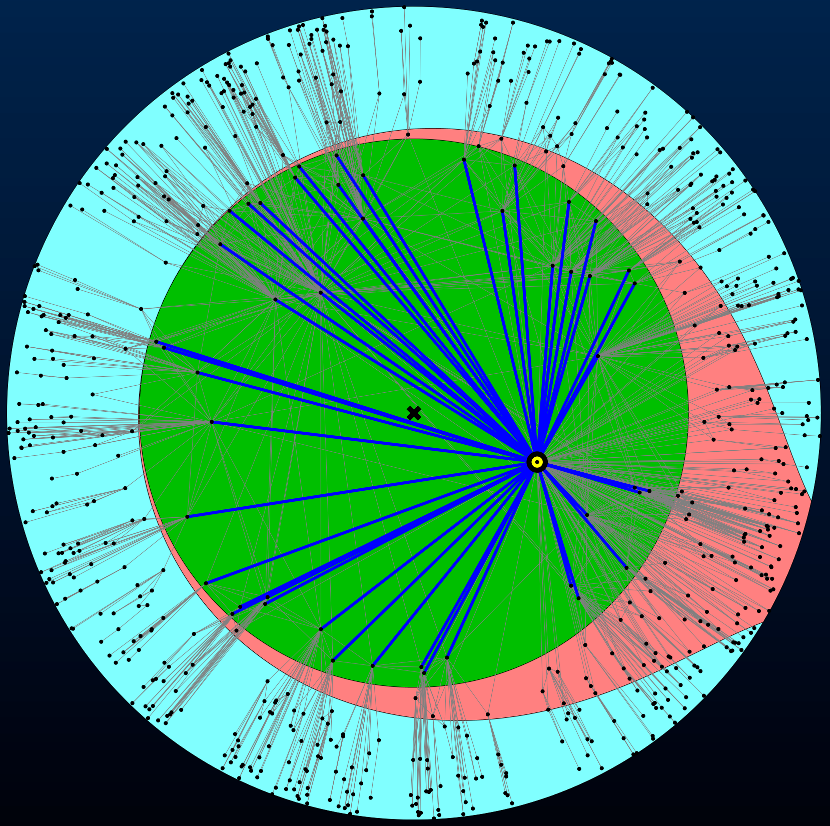
Breakdown of ASes by type.
Over time IPv4 is becoming
more like IPv6.

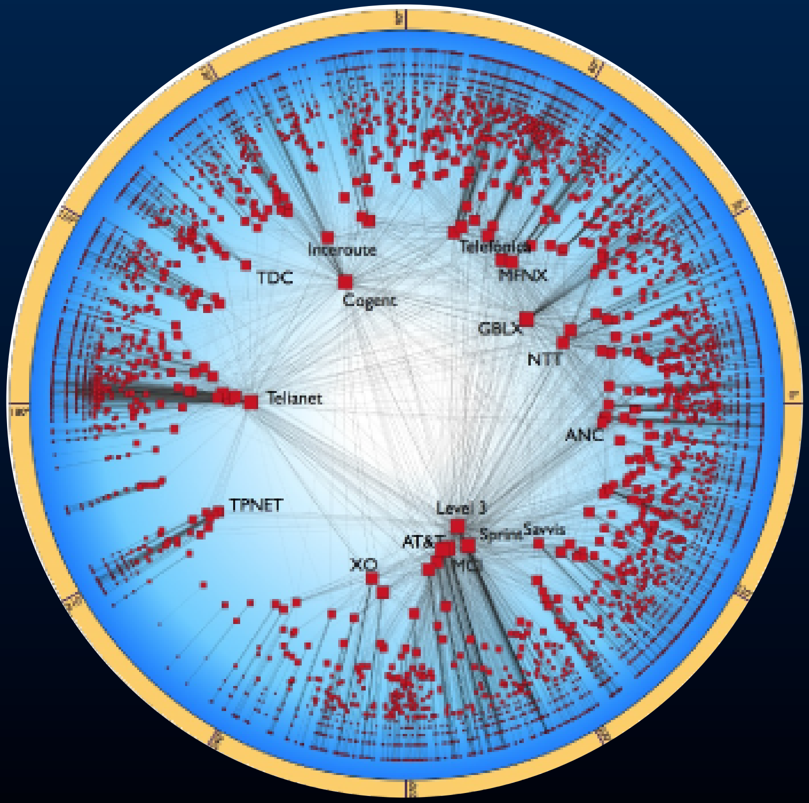
Modeling Complex Networks

D. Krioukov, M. Kitsak

- hidden variables influence structure of network
  - (which nodes are connected)
- variables form a hidden metric space that can be used to enable shortest path routing without global knowledge of topology
  - suggests potential direction toward infinitely scalable Internet routing architecture


R
$\kappa' \sim e^{\zeta(R-r')/2}$
$N \sim e^{\zeta R/2}$
r'
$\theta \sim \kappa\kappa'/N$
r
$\kappa \sim e^{\zeta(R-r)/2}$

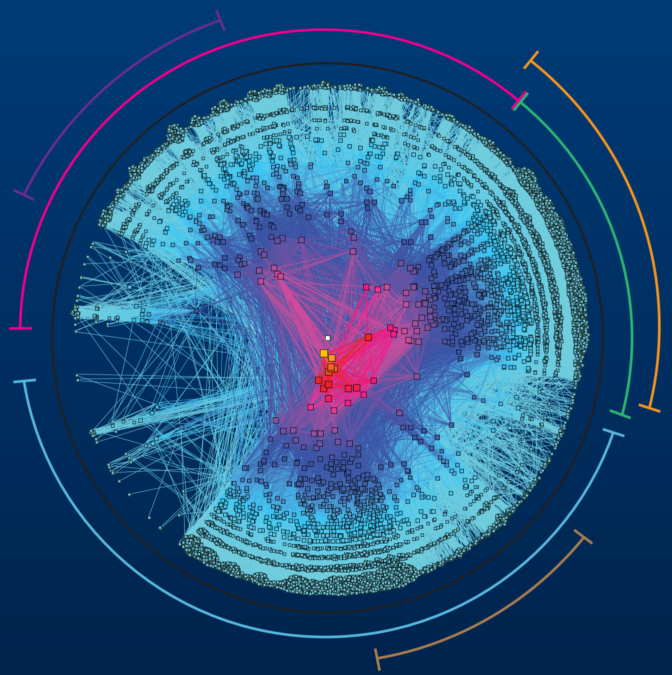




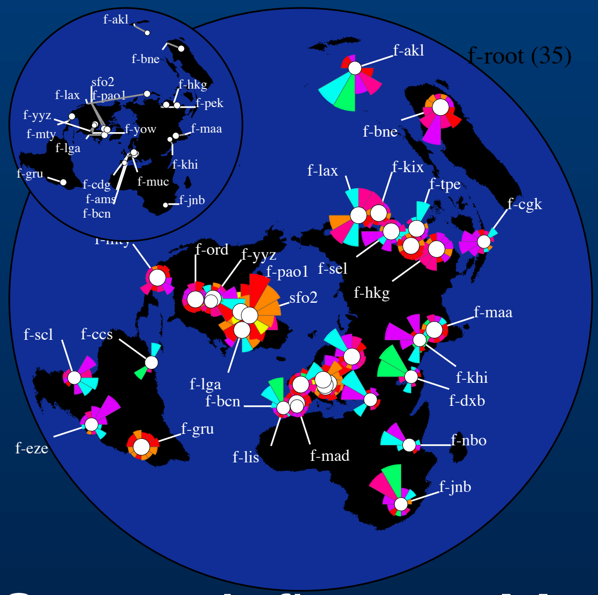
research

Visualization

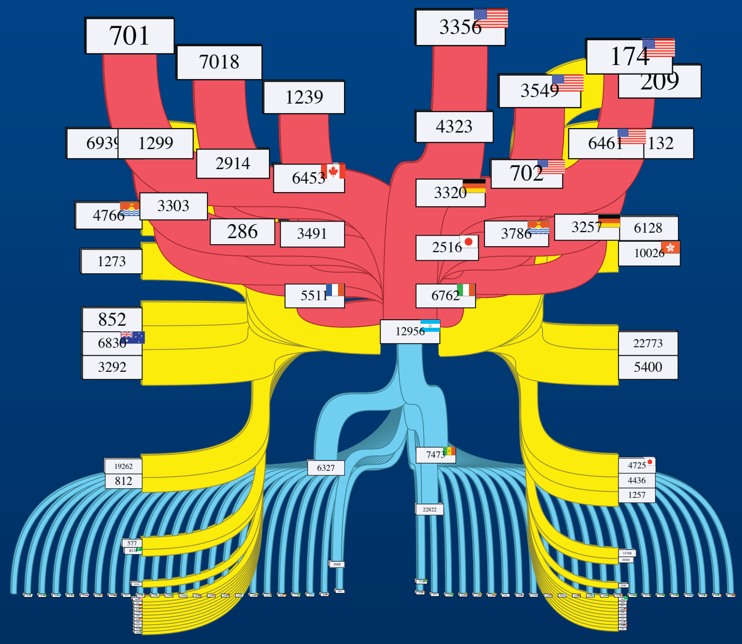
B. Huffaker, Y. Hyun



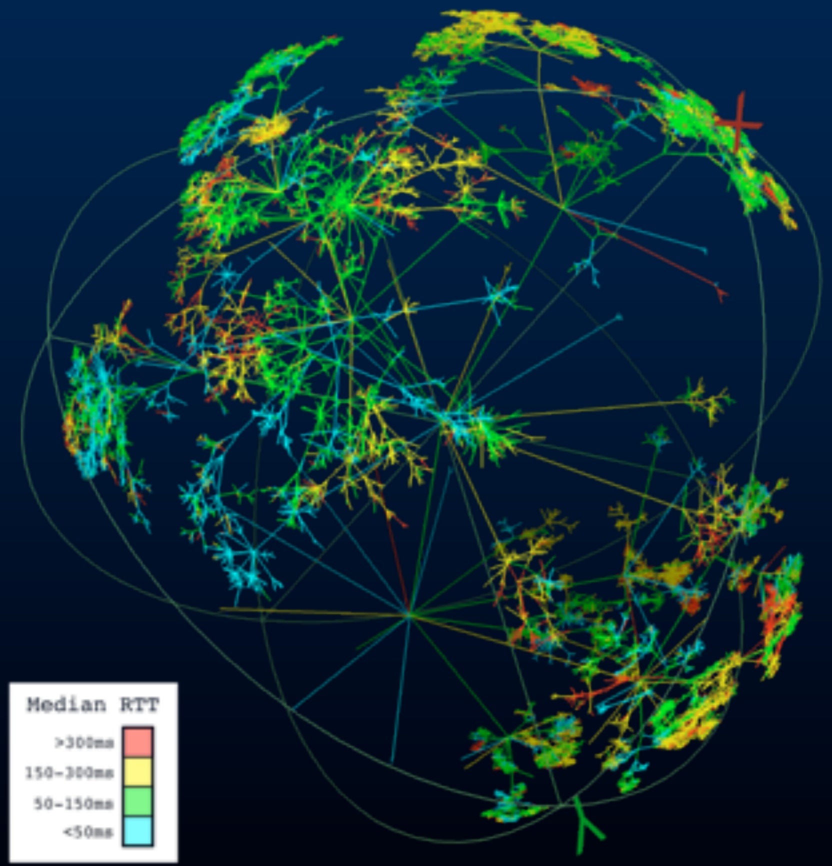
AS Core 2011



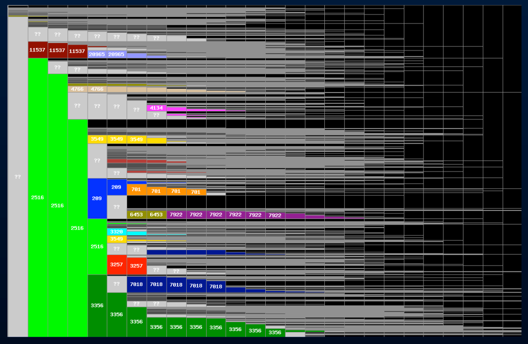
Source Influence Map
F-root Instances



AS Relationships for
Telefonica (12956)



RTT by Country



AS Dispersion

NTT Ark Monitor

RTT per hop from
A-root monitor

Workshops

http://www.caida.org/workshops/

- Active Internet Measurement Systems (AIMS)
  - Feb ’09, Feb ’10, Feb ’11, Feb ‘12
- BGP/Traceroute Workshop
  - Aug ‘11
- Workshop on Internet Economics
  - Sep ’09, Dec ‘11
- Joint workshop with WIDE (Japan) / CASFI (Korea)
  - Aug ’08, Apr ’09, Apr ’10, Dec ’11
- Darkspace and UnSolicited Traffic Analysis (DUST)
  - May ’12

Japanese Related Research and Collaborations

- Infrastructure
  - Ark monitors (Japanese Host)
  - Gulliver Project monitor (CAIDA Host)
- publications

  Identifying IPv6 Network Problems in the Dual-Stack World

  K. Cho, M. Luckie, B. Huffaker
- joint work
  - cuttlefish (diurnal visualization)
- workshops
  - CAIDA/WIDE/CASFI
- related research
  - Tōhoku Earthquake

Questions?

- publications
  http://www.caida.org/publications/papers
- collections
  http://www.caida.org/data/overview/
- workshops
  http://www.caida.org/workshops/

Bradley Huffaker
CAIDA/UCSD
bradley@caida.org