DHS PREDICT project: CAIDA update

Marina Fomenkov, Co-PI
CAIDA

Washington D.C.
9-10 January 2014

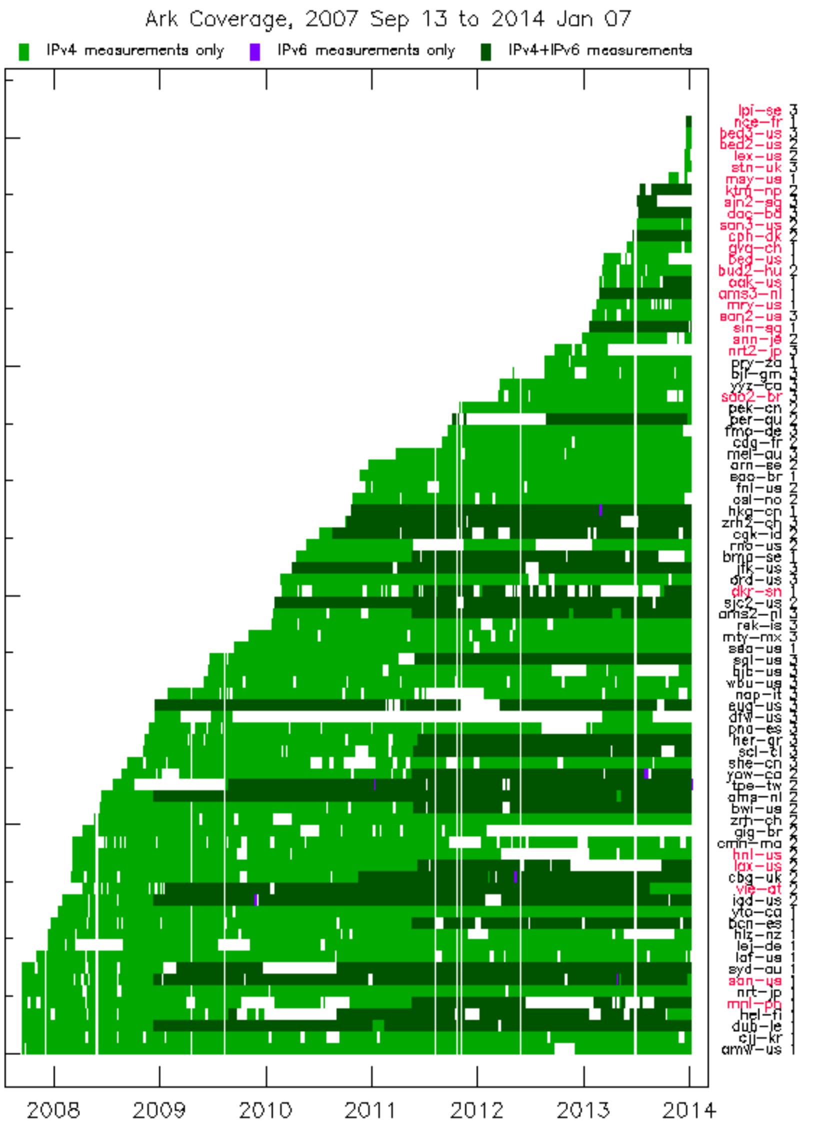
DHS PREDICT project: CAIDA update

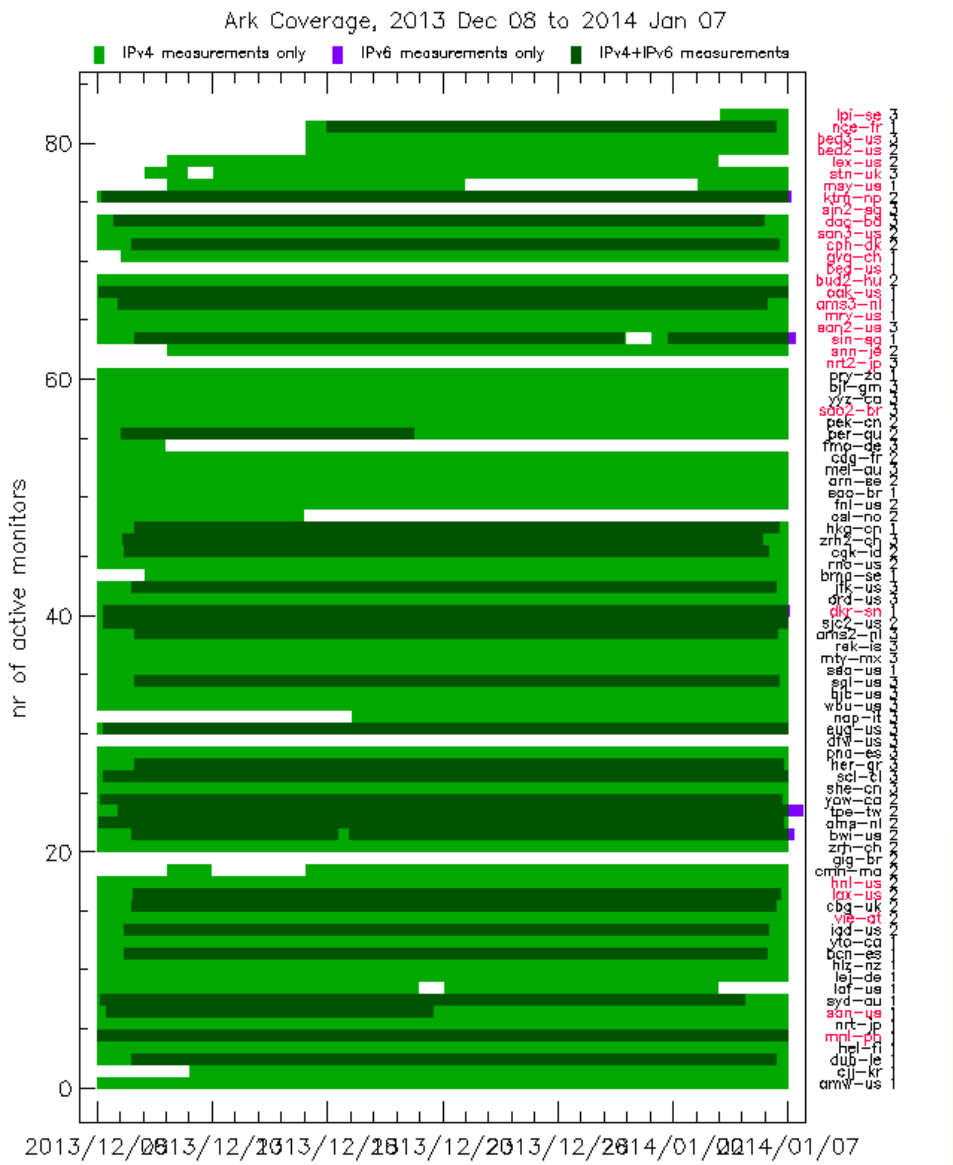
- Data collection activities
  - Ongoing measurements
  - Data storage status
  - Data dissemination statistics
- Other activities
  - New AUP for publicly downloadable data
  - AUP revisions
  - CREDS Workshop report
- Open issues

Ark Measurements

- Concurrent ongoing data collection
  - IPv4 topology
  - IPv6 topology
  - spoofer
  - initiated congestion measurements
- Ark Platform (as of Jan 2014) - 83 monitors
  - 35 IPv6 enabled
  - 29 Raspberry PIs - accelerated deployment
- Derived data sets
  - ITDK (the last was in July 2013)
  - AS links, IPv4 - daily
  - AS links, IPv6 - daily, by monitors
  - AS relationships
  - AS ranking

Ark Data Coverage - Sept 2007 to Present

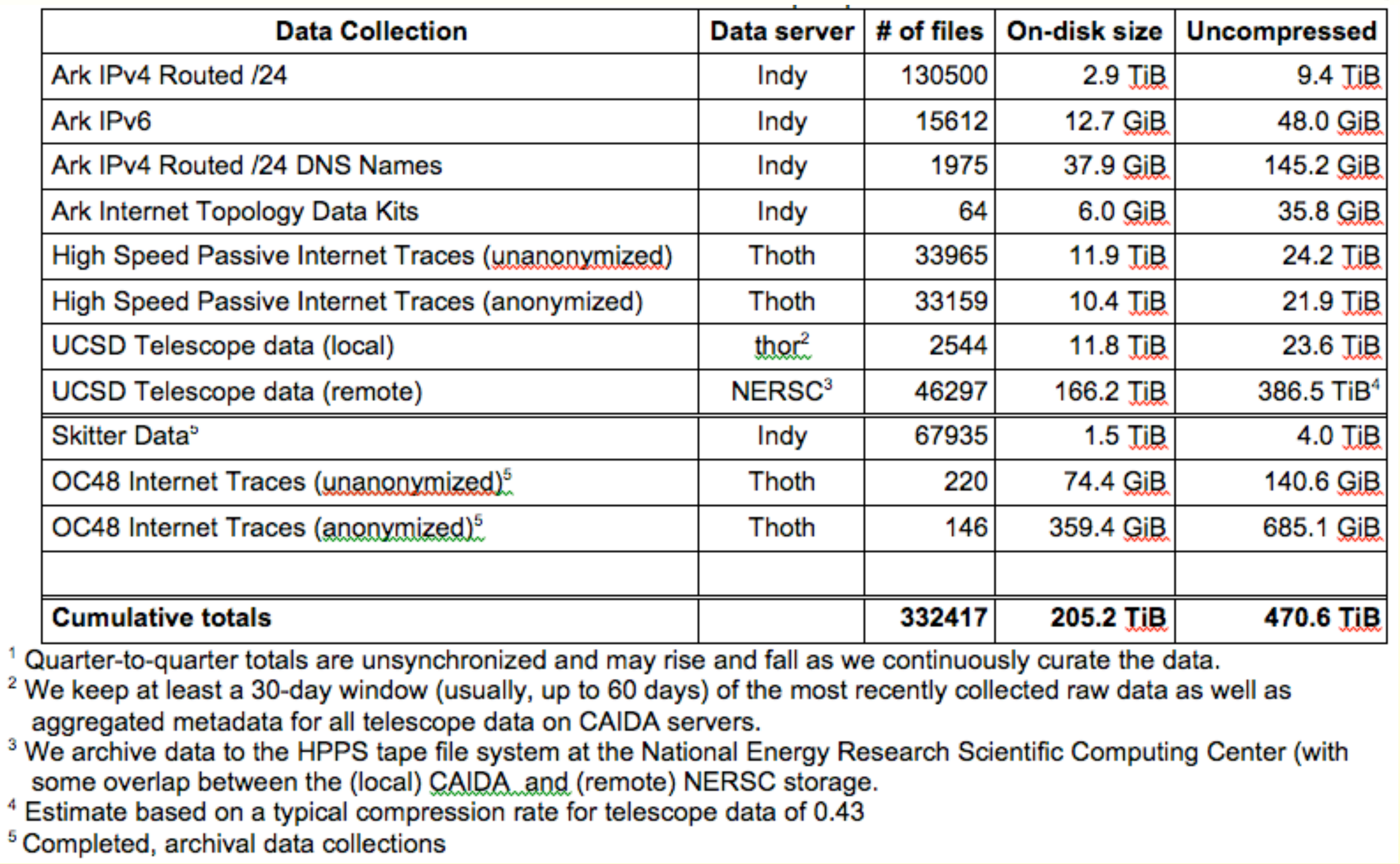




Ongoing Measurements: Internet Background Radiation

- UCSD Network Telescope
  - ~3-4 TB per month
- Stream processing
  - compress into tuples using Corsaro software
  - display 1-day delayed traffic as interactive graph
- Packaging
  - one-time snapshots of interesting events
  - educational data kits
  - industry evaluation samples

Dataset Collection and Size Update



Data Collection	Data server	# of files	On-disk size	Uncompressed
Ark IPv4 Routed /24	Indy	130500	2.9 TiB	9.4 TiB
Ark IPv6	Indy	15612	12.7 GiB	48.0 GiB
Ark IPv4 Routed /24 DNS Names	Indy	1975	37.9 GiB	145.2 GiB
Ark Internet Topology Data Kits	Indy	64	6.0 GiB	35.8 GiB
High Speed Passive Internet Traces (unanonymized)	Thoth	33965	11.9 TiB	24.2 TiB
High Speed Passive Internet Traces (anonymized)	Thoth	33159	10.4 TiB	21.9 TiB
UCSD Telescope data (local)	thor[2]	2544	11.8 TiB	23.6 TiB
UCSD Telescope data (remote)	NERSC[3]	46297	166.2 TiB	386.5 TiB[4]
Skitter Data[5]	Indy	67935	1.5 TiB	4.0 TiB
OC48 Internet Traces (unanonymized)[5]	Thoth	220	74.4 GiB	140.6 GiB
OC48 Internet Traces (anonymized)[5]	Thoth	146	359.4 GiB	685.1 GiB
Cumulative totals		**332417**	**205.2 TiB**	**470.6 TiB**

[1] Quarter-to-quarter totals are unsynchronized and may rise and fall as we continuously curate the data.

[2] We keep at least a 30-day window (usually, up to 60 days) of the most recently collected raw data as well as aggregated metadata for all telescope data on CAIDA servers.

[3] We archive data to the HPPS tape file system at the National Energy Research Scientific Computing Center (with some overlap between the (local) CAIDA and (remote) NERSC storage.

[4] Estimate based on a typical compression rate for telescope data of 0.43

[5] Completed, archival data collections

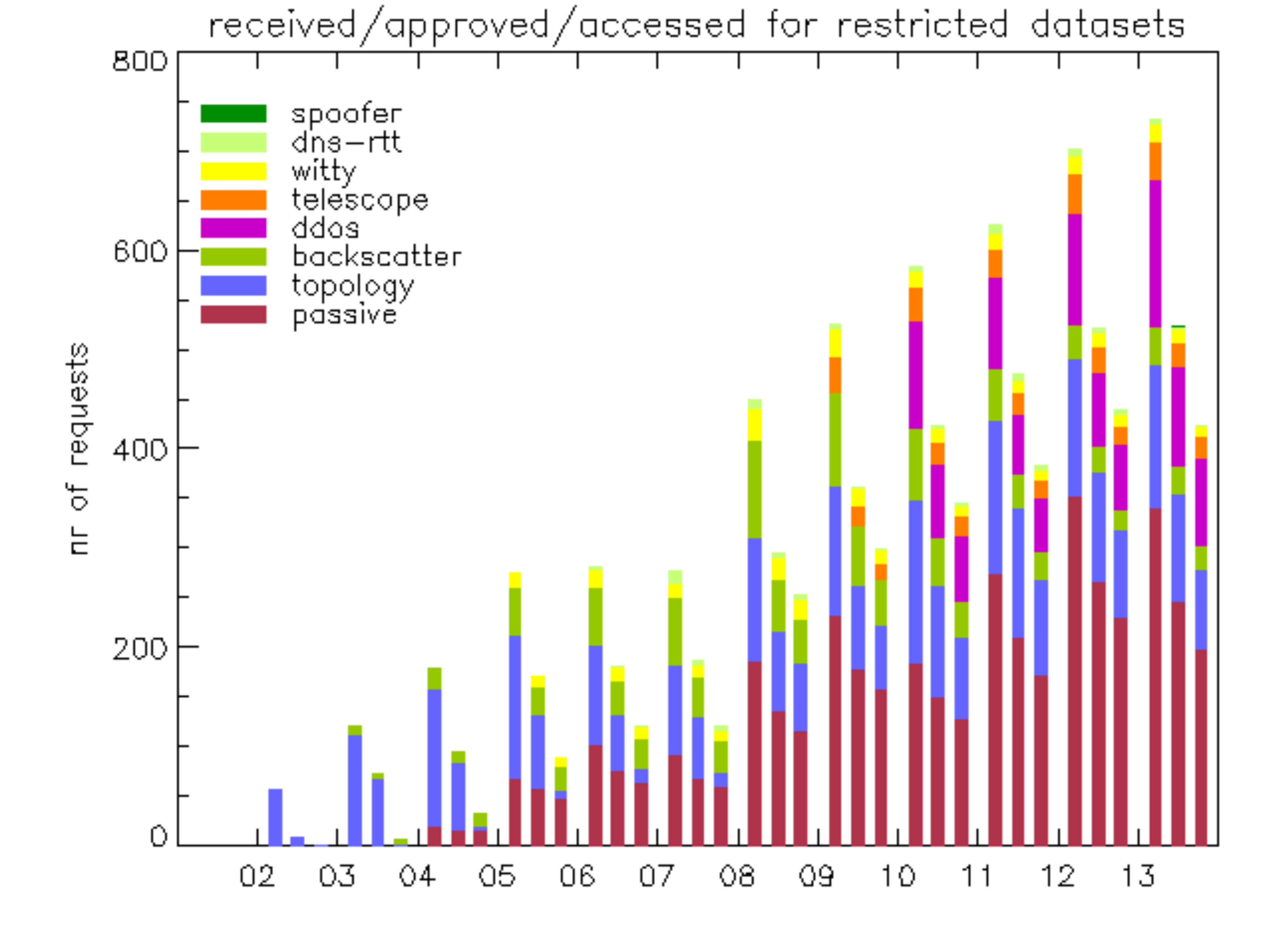
Data storage

- Continue to use NERSC for the telescope data (free of charge)
- Continue to use SDSC Cloud for other data (for a fee)
- Continue to expand CAIDA storage
  - Acquired in December:
    - 60 TB disks
    - 10 Gb network card
  - As planned in PREDICT Yr 2
- Submitted two CRI proposals to NSF
  - New funding to support and grow our measurement infrastructures

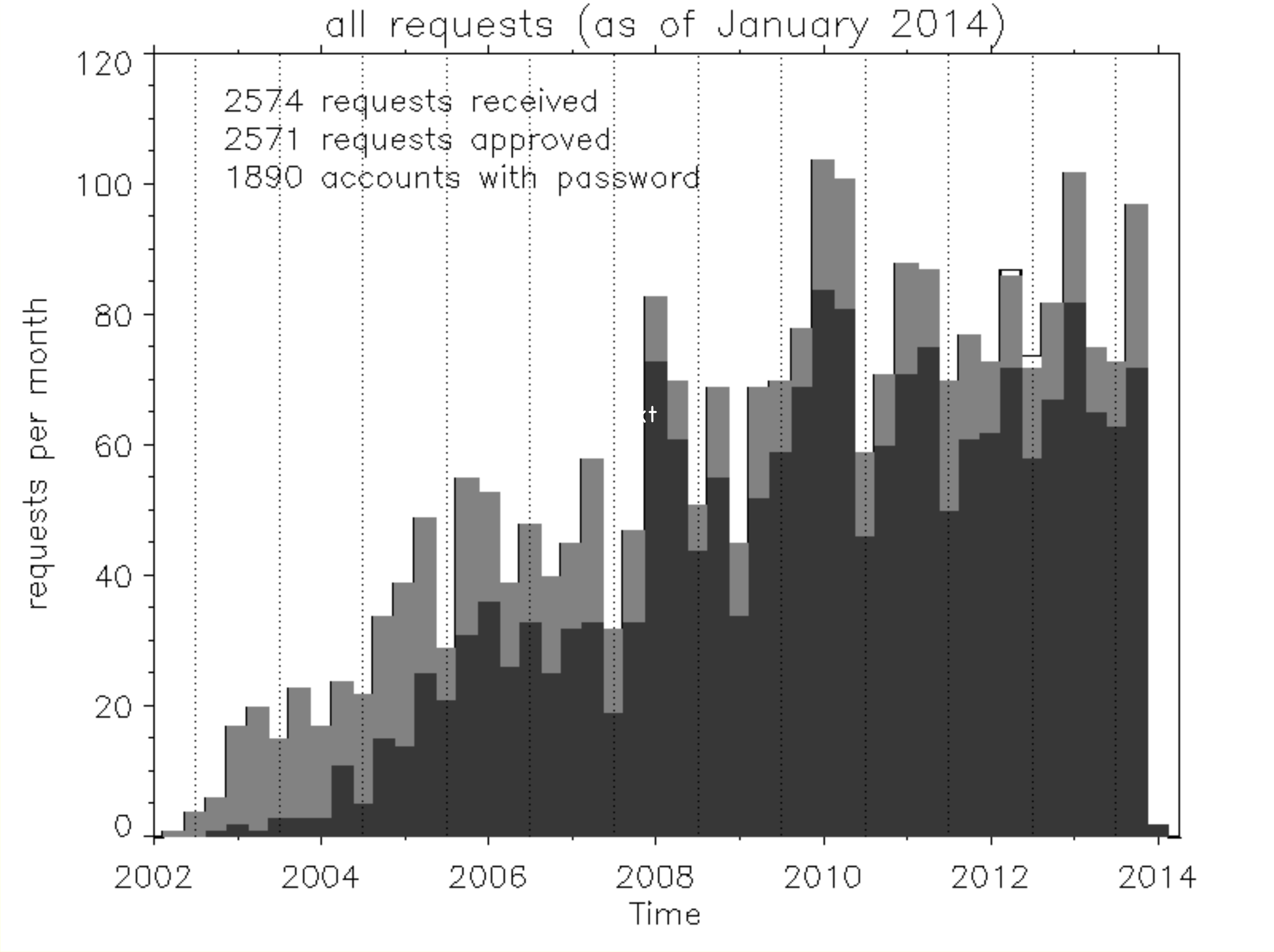
Data Access

- Topology data
  - 162/118 requests received/approved for topology data in 2013 from all over the world
  - usually give access within two working days
- UCSD Network Telescope data
  - 44/29 requests received/approved for telescope data in 2013, also of worldwide origins
  - advertised real-time telescope dataset in October in data overview and semi-annual email to users
  - six requests for realtime telescope, approved three:
    - Kirill Levchenko (UCSD, US); Tatsuya Mori (Waseda Univ., Tokyo, JP); Xiyue Deng (ISI, US)
    - Only Kirill actually has an account
    - three requests still being processed: 2 from China, 1 from US
    - Working with NICT, Japan to accommodate their request
    - Concerns about hardware resource availability

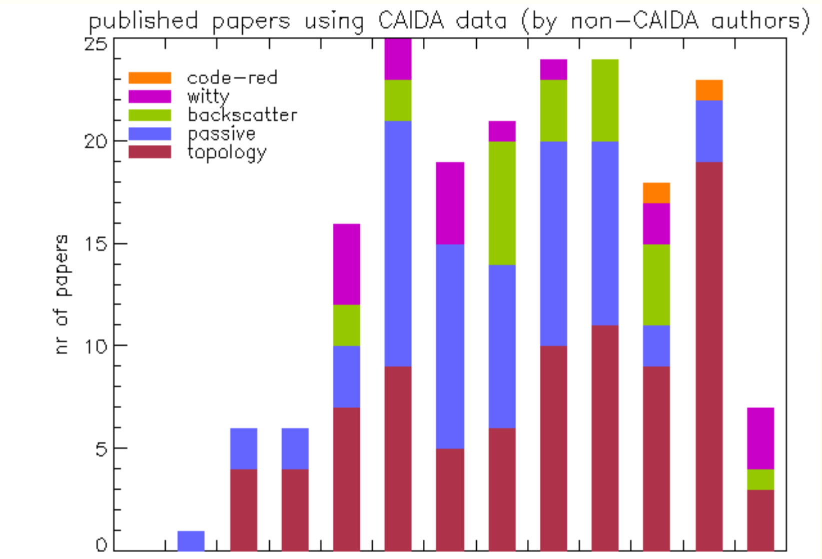
Restricted Dataset Requests, 2013

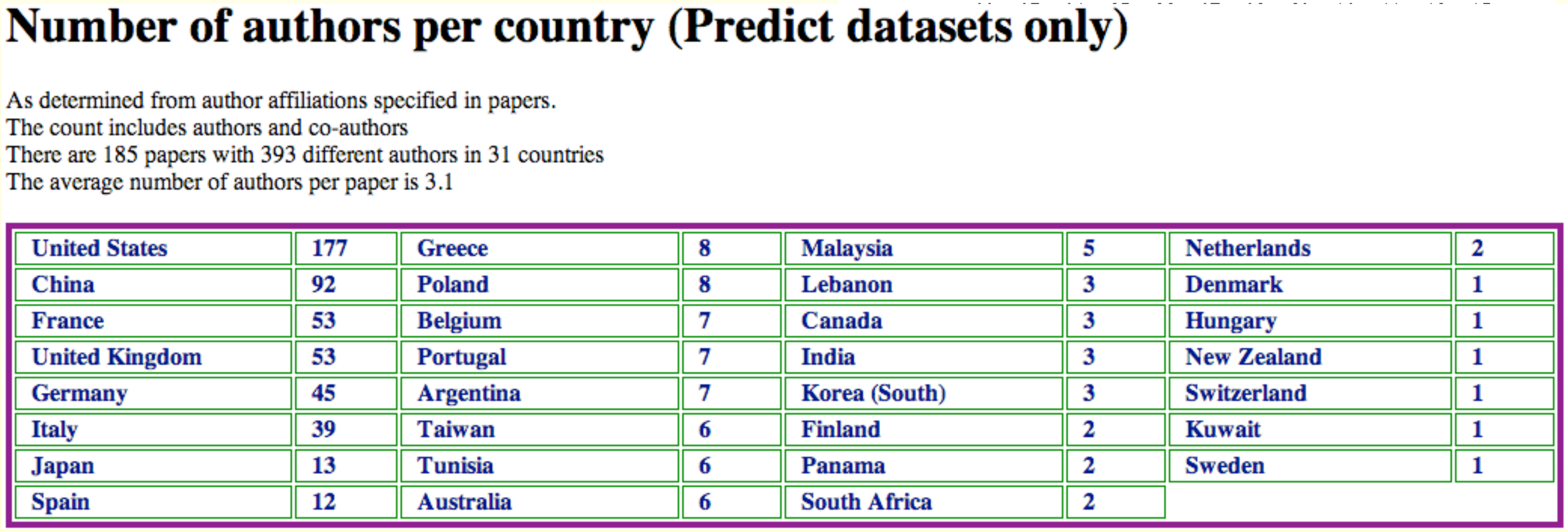


PREDICT Requests 2002-2013



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





Number of authors per country (Predict datasets only)

As determined from author affiliations specified in papers.
The count includes authors and co-authors
There are 185 papers with 393 different authors in 31 countries
The average number of authors per paper is 3.1

United States	177	Greece	8	Malaysia	5	Netherlands	2
China	92	Poland	8	Lebanon	3	Denmark	1
France	53	Belgium	7	Canada	3	Hungary	1
United Kingdom	53	Portugal	7	India	3	New Zealand	1
Germany	45	Argentina	7	Korea (South)	3	Switzerland	1
Italy	39	Taiwan	6	Finland	2	Kuwait	1
Japan	13	Tunisia	6	Panama	2	Sweden	1
Spain	12	Australia	6	South Africa	2		

Other Activities: public AUP

- Publicly accessible data sets:
  - AS IPv4 links
  - AS IPv6 links
  - Statistics of passive traces
  - AS relationships
  - ...
- Access policy
  - User info: optional
  - No hand shake
  - Immediately downloadable

=> Very popular!

- Simplified AUP - less than a page
  - License
  - Suggested citation format
  - Disclaimer

A piece of art - CAIDA public AUP

CAIDA ACCEPTABLE USE AGREEMENT for PUBLICLY ACCESSIBLE DATASETS

The following terms comprise the Acceptable Use Policy and Data License Agreement for all publicly accessible datasets (the “Public Agreement”) made available by the Cooperative Association for Internet Data Analysis (CAIDA), a research unit at the University of California San Diego (UCSD) and governed by The Regents of the Universty of California.

LICENSE. CAIDA’s authorization to access the data grants You a limited, non-exclusive, non-transferable, non-assignable, and terminable license to copy, modify, and use the data in accordance with this Public Agreement. No license is granted for any other purpose and there are no implied licenses in this Agreement. Nothing in this License is intended to limit any rights You may have arising from fair use or due to other limitations on CAIDA’s exclusive rights under copyright law or other applicable laws.

CAIDA has the authority and reserves the right, in its sole discretion, to discontinue further access and use to anyone.

If You create a publication (including web pages, papers published by a third party, and publicly available presentations) using data from this dataset, You should cite the data as follows:

`The CAIDA UCSD` *[DataSet Name]* `-` *[dates used]*,
`http://www.caida.org/data/`*[dataset-URL]*

We encourage You to provide CAIDA with a copy of (or a link to) the publication. We use this information in reports to our funding agencies.

DISCLAIMER OF WARRANTIES. CAIDA USES ITS BEST EFFORTS TO PROVIDE DATA IN ACCORDANCE WITH ETHICAL PRINCIPLES AND SCIENTIFIC INTEGRITY. HOWEVER, THE DATA PROVIDED HEREIN IS ON AN “AS IS” BASIS. NEITHER CAIDA, ITS RESEARCHERS, RESEARCH PARTNERS, LICENSORS, AND DATA PROVIDERS, NOR THE UNIVERSITY OF CALIFORNIA AND ITS TRUSTEES, OFFICERS, EMPLOYEES, AND AGENTS MAKE ANY WARRANTY, EITHER IMPLIED OR EXPRESS, OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, INCLUDING, BUT NOT LIMITED TO, THE ACCURACY, TIMELINESS, COMPLETENESS, RELIABILITY, OR AVAILABILITY OF CAIDA DATA, APPLICATIONS, OR SERVICES ACCESSIBLE THROUGH OR MADE AVAILABLE BY CAIDA.

LIMITATION OF LIABILITY. TO THE EXTENT ALLOWED BY LAW, IN NO EVENT SHALL CAIDA AND THE UNIVERSITY OF CALIFORNIA BE LIABLE TO YOU OR ANY THIRD PARTY FOR ANY INDIRECT, CONSEQUENTIAL, INCIDENTAL, SPECIAL OR PUNITIVE DAMAGES, ARISING FROM YOUR USE OF THE DATA.

If You have any questions about the data or about this Public Agreement, please email `data-info@caida.org`.

Changing CAIDA data policy

- Topology data - make publicly accessible
  - IPv4 raw topology - 2011 and older
  - ITDKs - 2011 and older
  - IPv6 topology
- Telescope data - educational data kits
  - Description and detailed tutorial
  - Will be available shortly
- Passive data - no changes (for now)
  - too sensitive
  - longer “aging” threshold?

CAIDA “2014 New Year Resolution”:

try to make more data easily sharable

(but proceed cautiously)

CAIDA Master AUP vs. PREDICT Non- & Quasi- Restricted Access Policy

- Used beta-portal version for comparison
- Many similarities (obviously)
  - Researcher’s info required
  - Consent for public disclosure
  - Publication info required
  - ...
- But: many (irreconcilable?) differences
  - License formulation
  - Handling modifications to AUP
  - Termination provisions (absent in PREDICT?)
  - Distribution conditions
  - Disclaimer, indemnification, ...
- For now: CAIDA data will be in either Unrestricted or Restricted class

Master AUP Revisions

- Removed “no-probe” restriction
  - will insert it as a special case attachment if necessary
- Removed anonymization requirement for networks and organization
  - there are restrictions on PII
  - focus on restrictions for publication/disclosure, rather than for research analyses
- Added provisions to deal with the change of Researcher’s contact information

Need to propagate these changes into CAIDA PREDICT DUA (work in progress)

CREDS Workshop

- Dissemination
  - Published as CAIDA technical report (October 2013)
  - Submitted to CCR
- Main themes
  - Brave New World - Ethical Research Amidst Expanding Opportunities
  - Checking Our Collective Assumptions - Risks and Benefits at the Frontline of ICT Research
  - Teaching Researchers to Fish - Tools to Implement Ethics Principles and Applications
- Discussions
  - Carna Botnet Data - are they ethical to use?
    - The majority said “yes”
  - Path forward
    - Leadership
    - Awareness and necessary education

Open Issues

- Are we expected to give portal feedback at this meeting?
  - Is such “real-time feedback mode” optimal?
  - What about the feedback given at the August meeting?
- Is there enough time to act on our feedback before January 15?
- Why are CAIDA Skitter and OC48 data still shown as “Restricted” (in beta)?
- Is login required to search the catalog?

Open Issues (cont.)

- MOA Ph III issues:
  - Asymmetrical indemnification clause
  - Need to update CAIDA DUA provisions
  - Need more time for a legal review
  - USCD had a 2 week holiday break
- Access problems
  - Access is given for 12 months
  - New data set appears
    - Same Category
    - Same Sub-category
  - Should a new request be required??

PREDICT Outcomes

- The whole CAIDA Data Collection
  - known worldwide
    - but it took years to develop and promote
    - the easier access - the higher popularity
  - interdisciplinary impact
    - NIST Workshop on Large-Scale Networks
- CAIDA data sharing activities
  - full time Data Administrator
- All our legal documents
  - PREDICT stimulated this work
- All Ethics Research
  - high community value - but how to quantify?

Metrics of Success

- What are the expectations?
  - Another Internet?
  - Another Qualcomm?
  - To get rid of all botnets once and forever?
- “Normalization” issues?
  - How many cybersecurity researchers are there?
  - One Researcher accesses One Dataset and works with it for One Year (or Two...)
  - How many requests would be “good enough”?
- Why “students and papers do not count”?
  - PREDICT is funded by **Science** & Technology Directorate
  - Papers frequently include algorithms and software tools

Ideas for Phase III

- fund meta-data research:
  - ascertain **researchers data needs**
  - access the impact of data age on data usefulness
  - disclosure control policies and methods
- continue experimenting with data access modes
- continue ethics/policy activities
- expand Data Host activities?
- **change metrics**
- there is still room for more marketing efforts
  - cf. typical advertising strategies: “How did you learn about our product”? (web, newspaper, TV, Facebook, friends...)