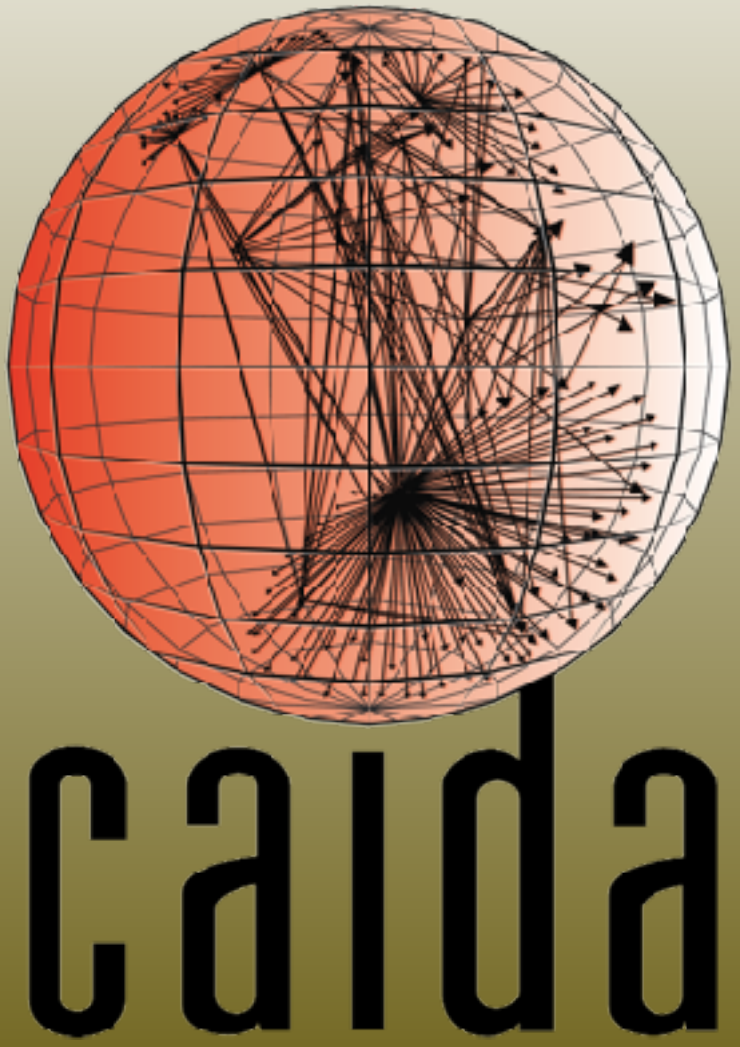
CAIDA update

PI k claffy, CAIDA
MDPnP
Cambridge, MA ,
7-8 May 2018



Needs Addressed

Data to study security and stability-related events and identification of cyber dependencies (TTA#1)

- Data collection and curation
- New datasets supporting real-time incident identification
- Data hosting and dissemination

Analysis system to identify, monitor, and mitigate the infrastructure vulnerabilities (TTA#2)

- Hub for Internet Incident Investigation (HICUBE)
- Web services and visual user interfaces for exploratory data analysis
- Software infrastructure for data storage, query, and transformation

Community outreach and service

- Disseminate information about cyber security data and analytics
- Host professional workshops and meetings

CAIDA as Data Provider

63 CAIDA Datasets served

IMPACT PORTAL

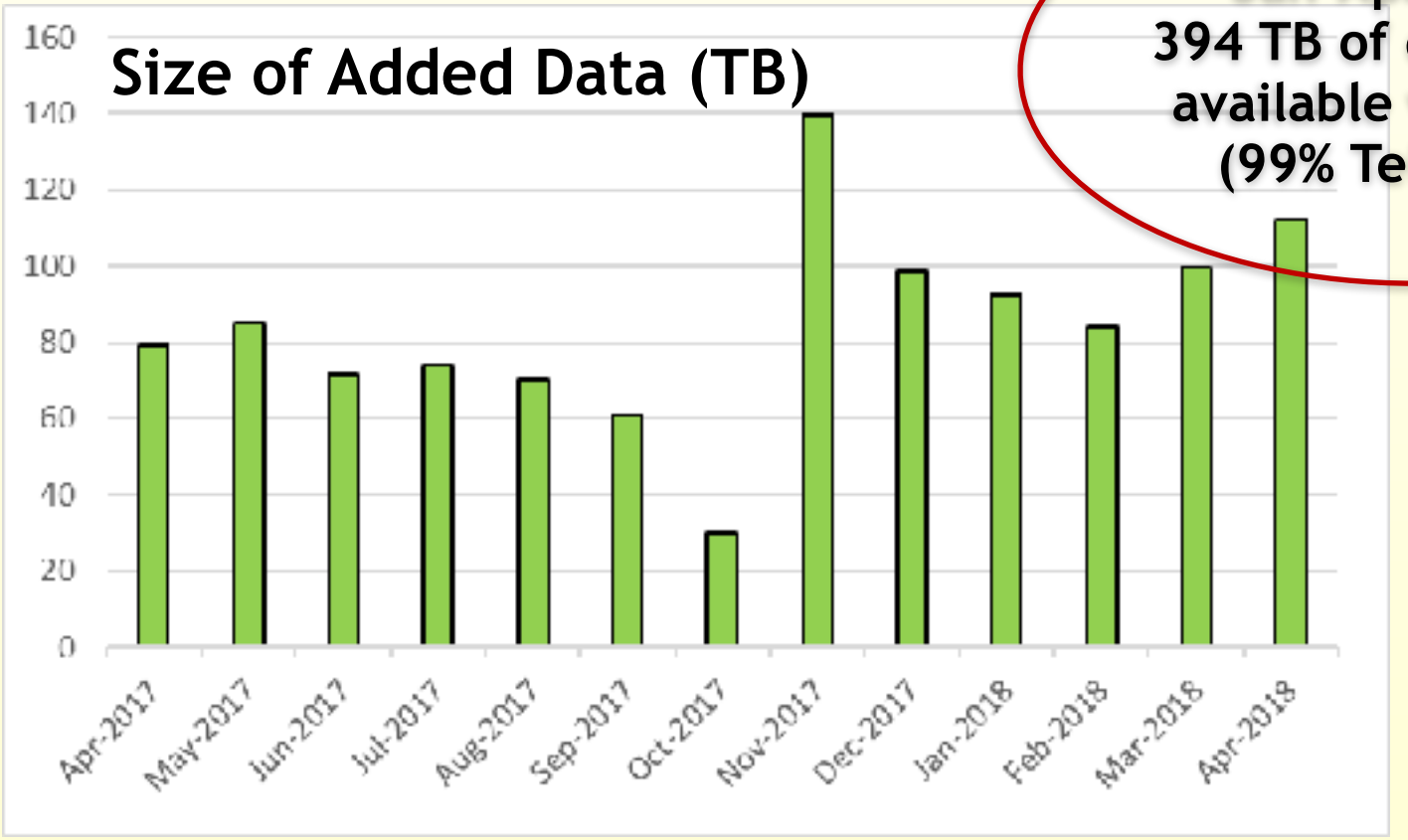
Restricted:	7
Quasi Restricted:	3
Unrestricted:	10
Unrestricted Tool:	1
Restricted Tool:	2

CAIDA PORTAL

Public:	26
(incl. IPv4 Topology older than 1 yr.)	
Paper Supplement:	13
Interactive Graph:	2
Passive traces (R):	3

Ongoing Measurements

- **Internet Topology Data and Metadata (Ark Platform)**
  - Internet Topology Data Kits
- **UCSD Real-time IBR Data (UCSD Network Telescope)**
  - Scanners, malware attacks,
- **Passive Traces (From Equinix-NYC, not in IMPACT)**
  - Backbone links, user traffic, packet headers

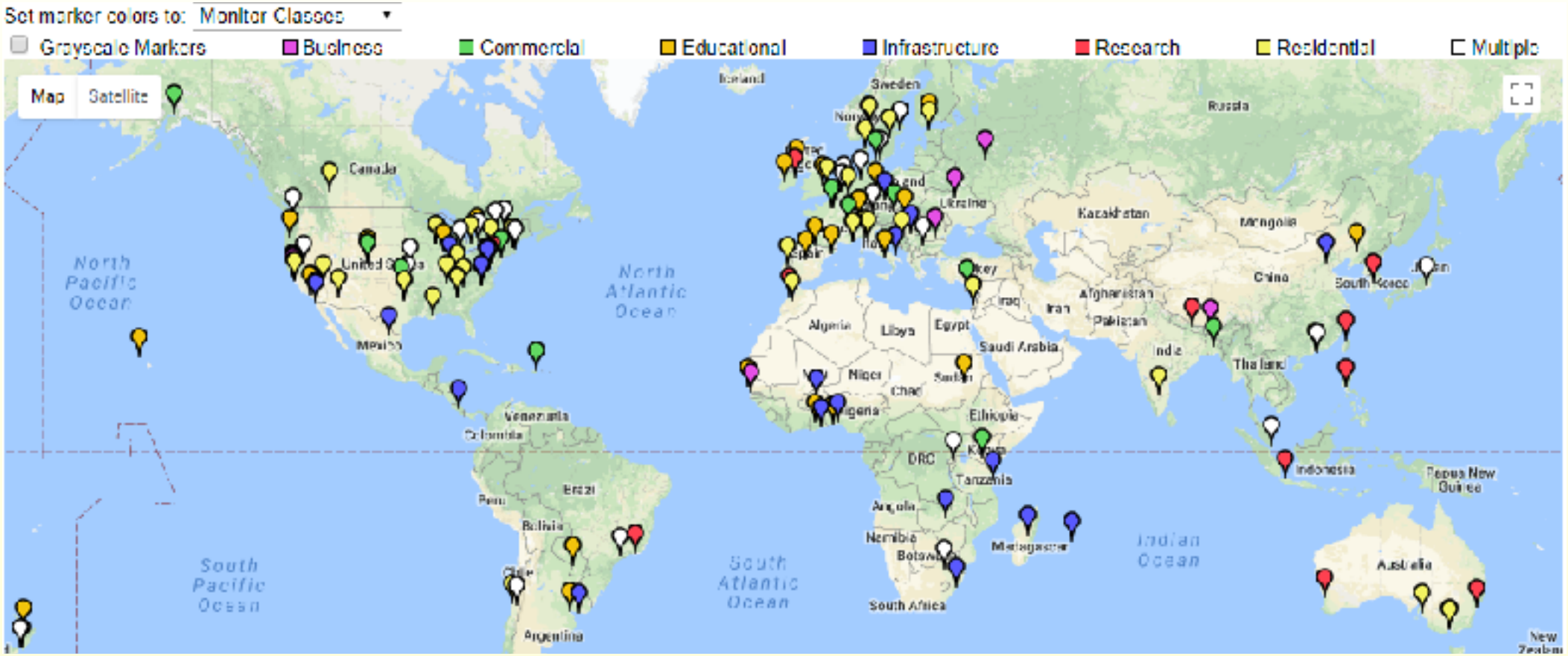


Jan-Apr 2018 :
394 TB of data added available via IMPACT (99% Telescope)

Ark Platform: Topology Measurements

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

New: Data less than 1 year old available only through IMPACT.



- 210 Ark monitors in 64 countries
- 98 IPv6-enabled
- 166 Raspberry PIs

fault-tolerance features

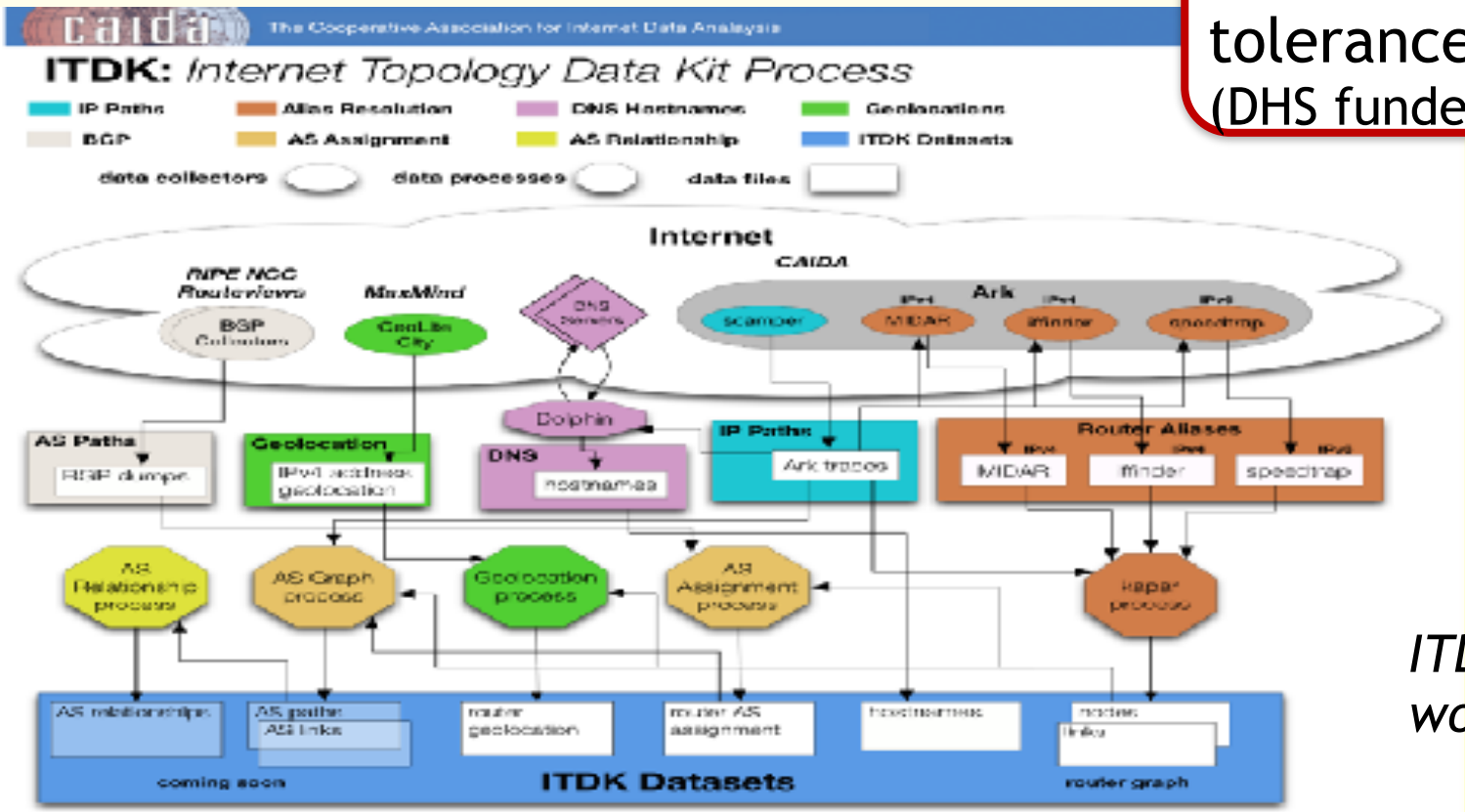
Internet Topology Data Kits

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

2017-08 (served) & 2018-03 (coming soon)

- Traceroutes conducted on Ark and RIPE Atlas
- 2 related IPv4 + IPv6 router-level topologies
- Router-to-AS assignments
- Geographic location of each router
- DNS lookups of all observed IP addresses

Uses new MIDAR code with fault-tolerance features (DHS funded)

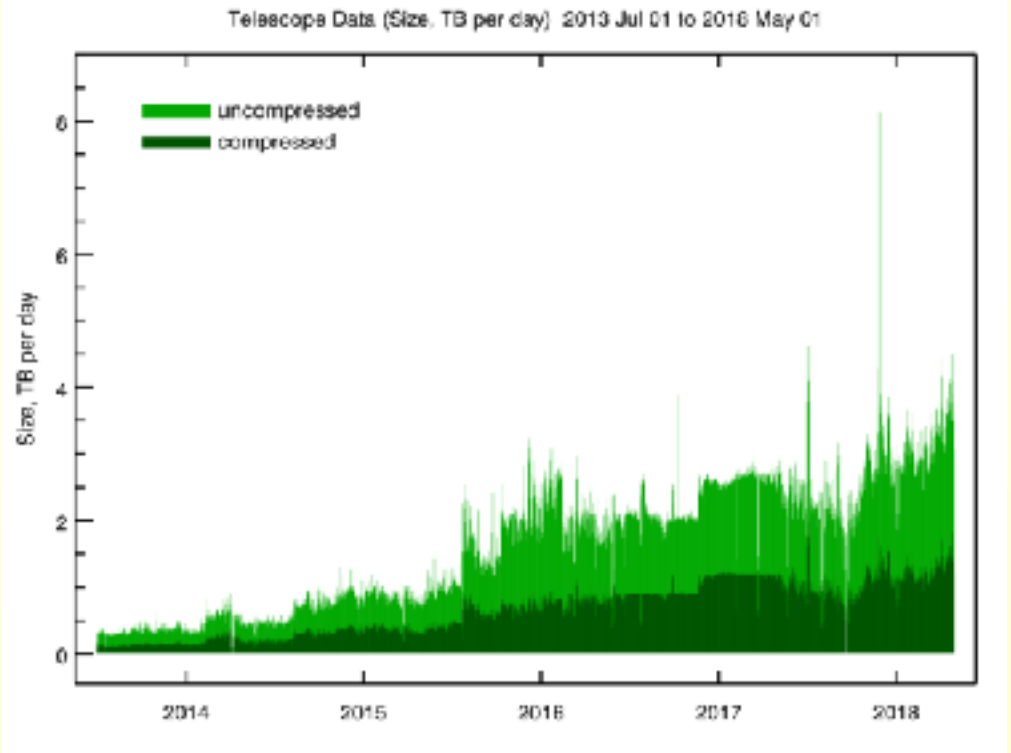


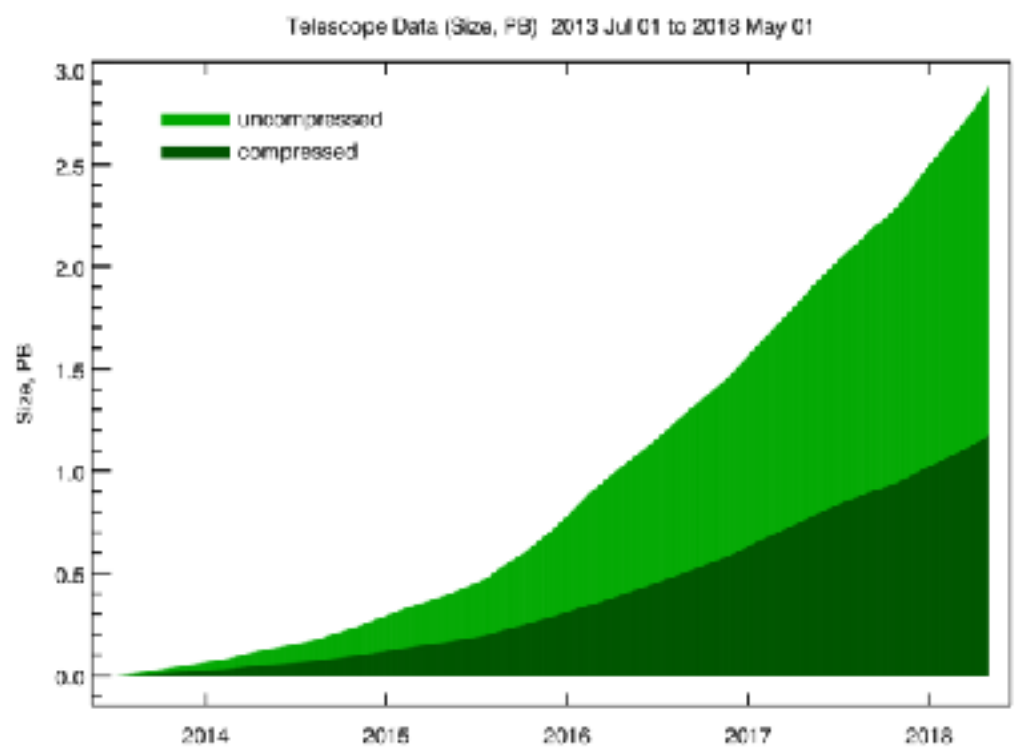
ITDK generation workflow

UCSD Real Time Telescope

http://www.caida.org/data/passive/telescope-near-real-time_dataset.xml

- 391 TB added 1 Jan - 30 Apr 2018
- 1.34 PB compressed data @NERSC (since '08)
- Currently adding ~3-4 TB/day
- 1 Jan - 30 Apr 2018: 4 requests from IMPACT





Anonymized Passive Internet Traces

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

- M]easurements from 2008 to Oct 2016.
- Starting March 2018, data contain anonymized traffic traces from CAIDA's new 10 Gb link Equinix-NYC monitor
- http://www.caida.org/data/monitors/passive-equinix-nyc.xml
- Raw traces taken on Endace DAG cards
- Stripped of payload, anonymized, split into 1-minute chunks
- March and April traces added and served
- Compressed size added March-April 2018: 582GB
- We will collect a single one-hour trace each month
- Still aiming for 100GB monitor in 2019

New: Internet Exchange Point Data

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

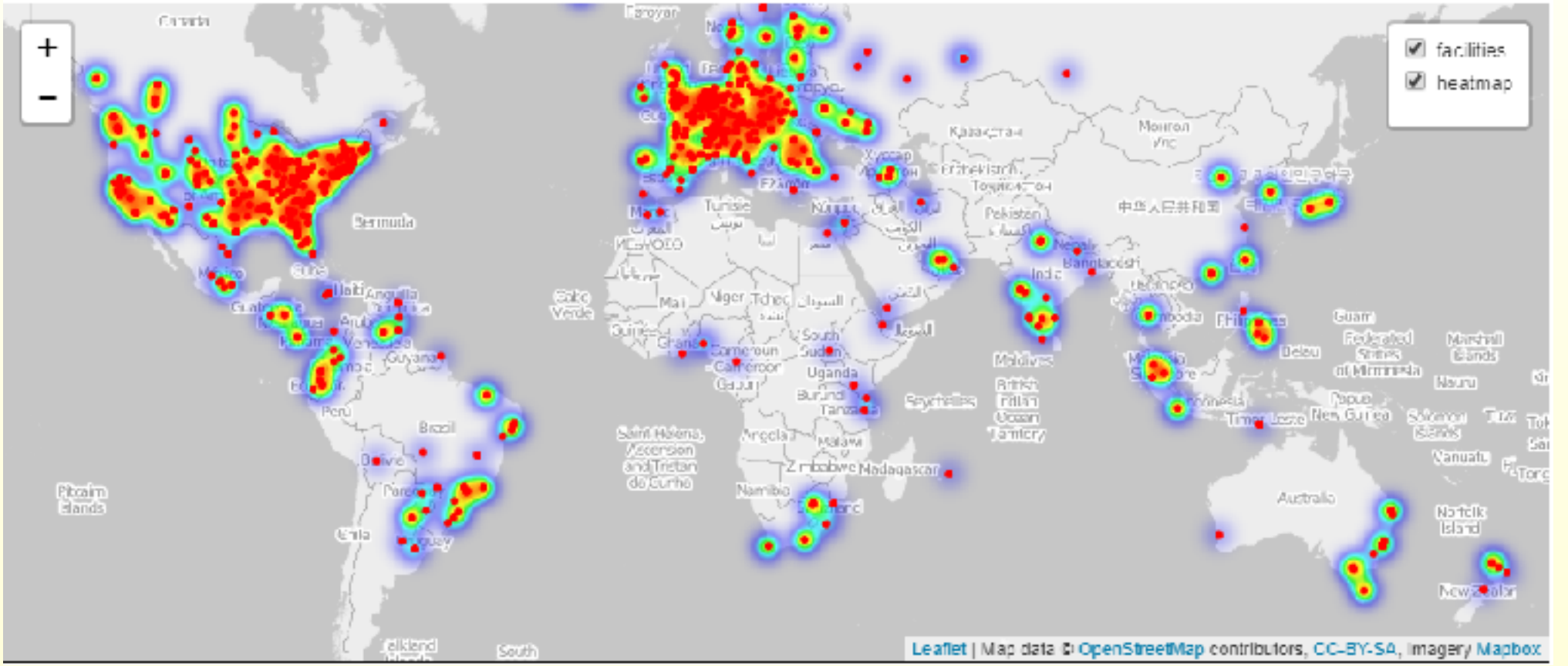
- Info per IXP: *name, source id, geolocation, prefixes*
- Info per facility: *name, source id, geolocation*
- Mapping between facilities and IXPs
- IP addresses used at a given IXP by each member AS
- Combines data from PeeringDB(**PDB**), Hurricane Electric (**HE**) and Packet Clearing House (**PCH**)
- Feb 2018 data: information about 858 IXPs :

SOURCE	COUNT
PDB only	224
HE only	45
PCH only	138
PCH&PDB	68
HE&PDB	75
HE&PCH	26
HE&PCH&PDB	282

New: AS Facilities Mapping

http://www.caida.org/data/as-facilities/

- Mapping of peering interconnections to facilities
- PeeringDB map of ASes to interconnection facilities
- Geonames of cities hosting facilities



Information about geographic locations of interconnection facilities and ASes that interconnect at those facilities

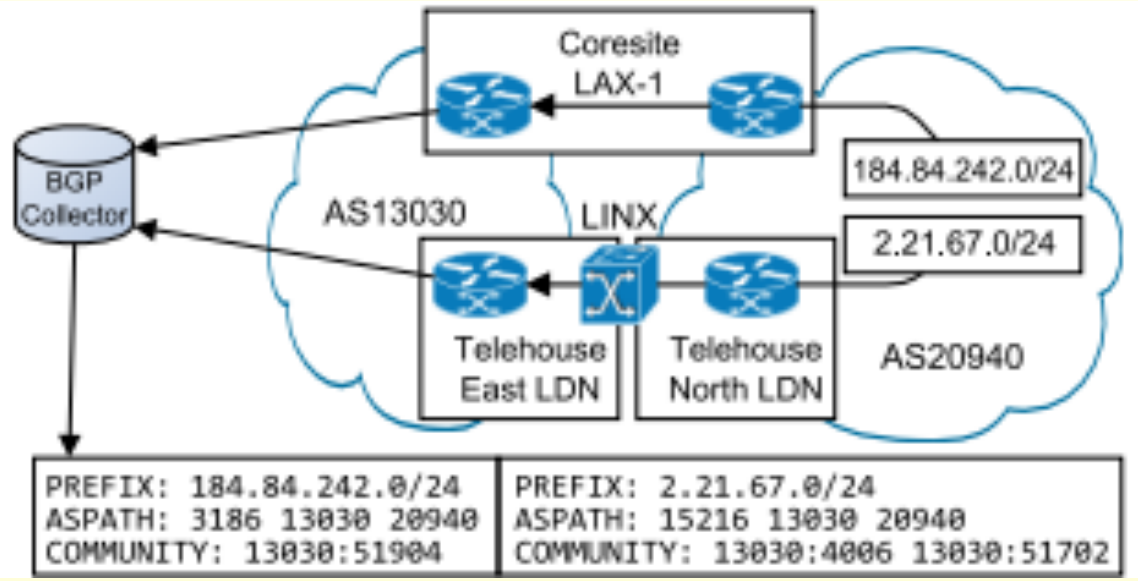
New: BGP Communities Dictionary

http://www.caida.org/data/bgp-communities/

- Focuses on Location-Encoding Ingress Communities
- Method extracts text from remarks sections of IRR records and from ASes's web pages, analyzes w/Natural Language ToolKit
- Provides info on geolocations encoded in BGP community attribute
- Meta-data to infer topological and geographic locations of outages
- Described in Giotsas et al.“Detecting Peering Infrastructure Outages in the Wild”
- (http://www.caida.org/publications/papers/2017/)

Inferring physical locations from BGP Communities, April 2018:

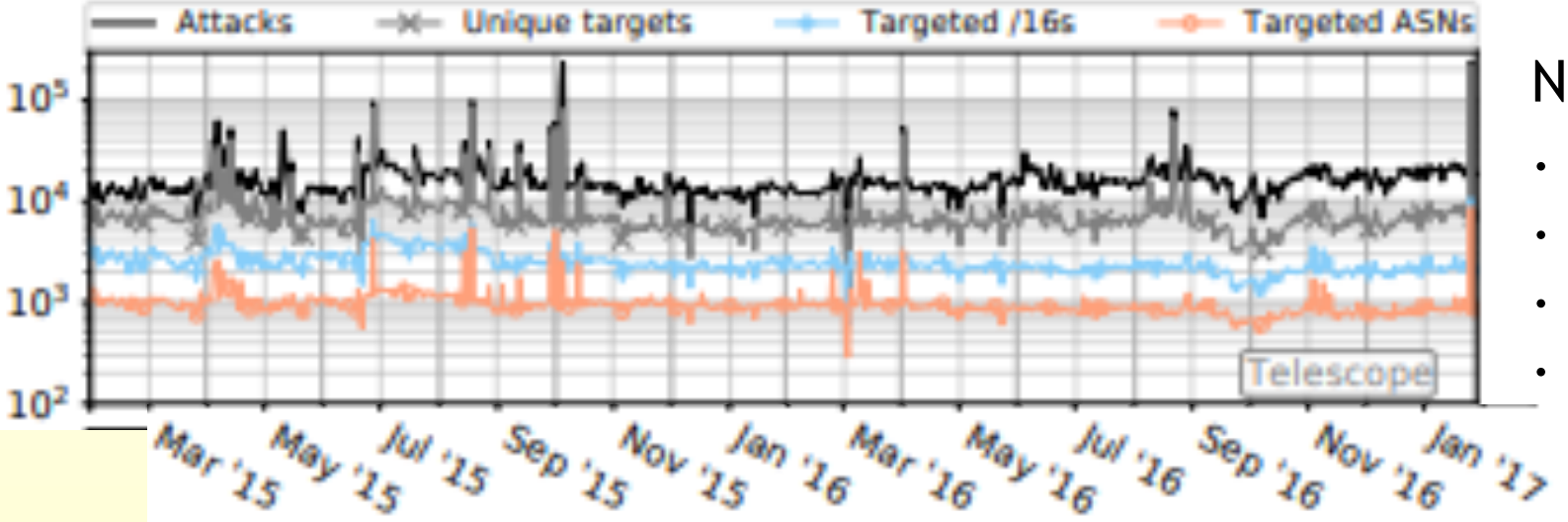
- ✓ *7350 geolocated communities,*
- ✓ *714 AS*
- ✓ *88 Countries*



Randomly and Uniformly Spoofed Denial of Service Metadata (soon)

- Metadata on DoS activity seen by UCSD Network Telescope
- Attack vectors (AVs) extracted from 5-minute intervals
- Geolocation and AS added to completed AVs
- AV has unique combo of target IP address and start time
- Two-year window: March 2015 - February 2017

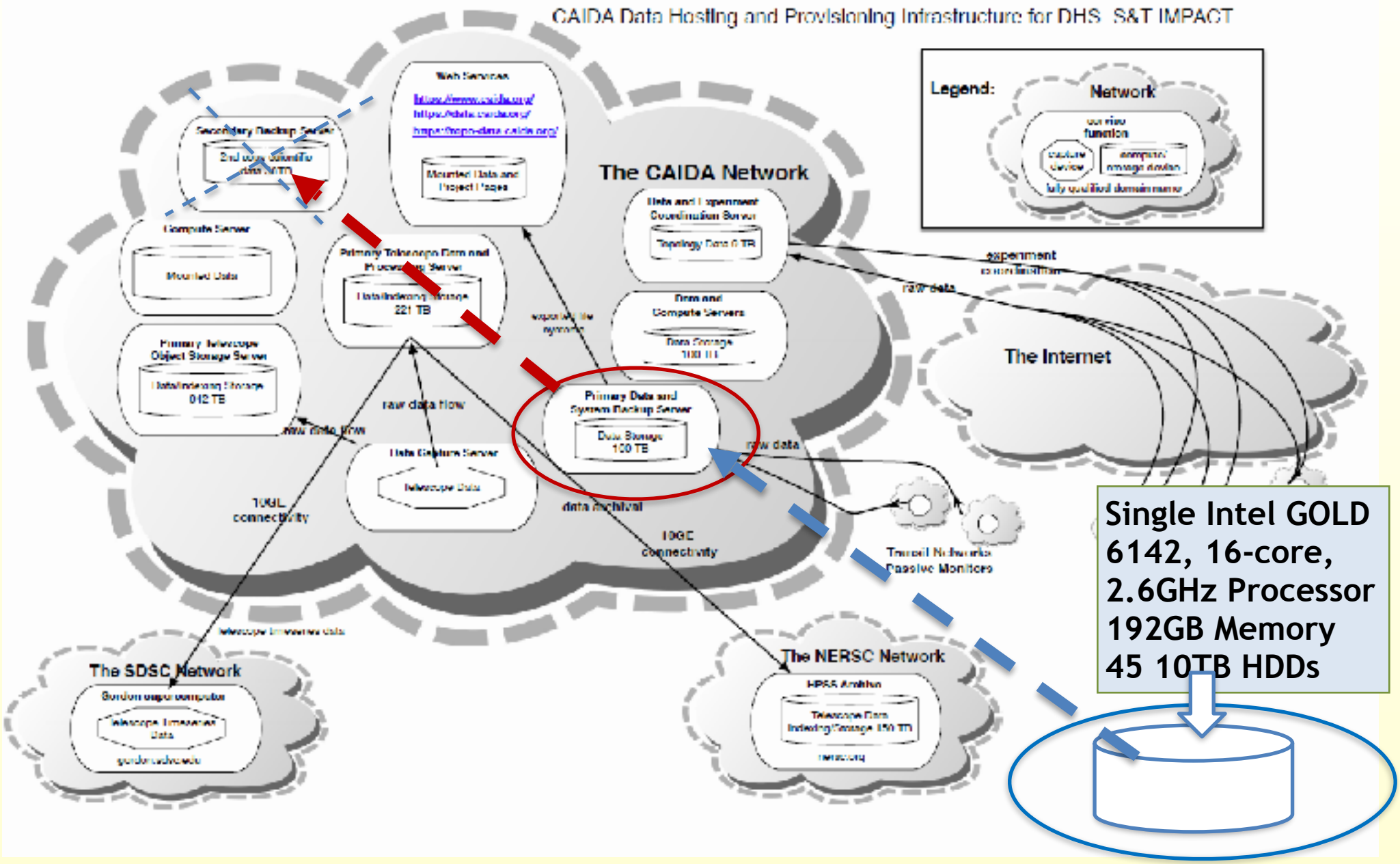
http://www.caida.org/publications/papers/2017/millions_targets_under_attack/



Numbers of targeted
. attacks (black line)
. IP addresses (grey)
. /16 blocks (blue)
. ASNs (orange)

Data Hosting: Infrastructure

Installed and configured new storage server



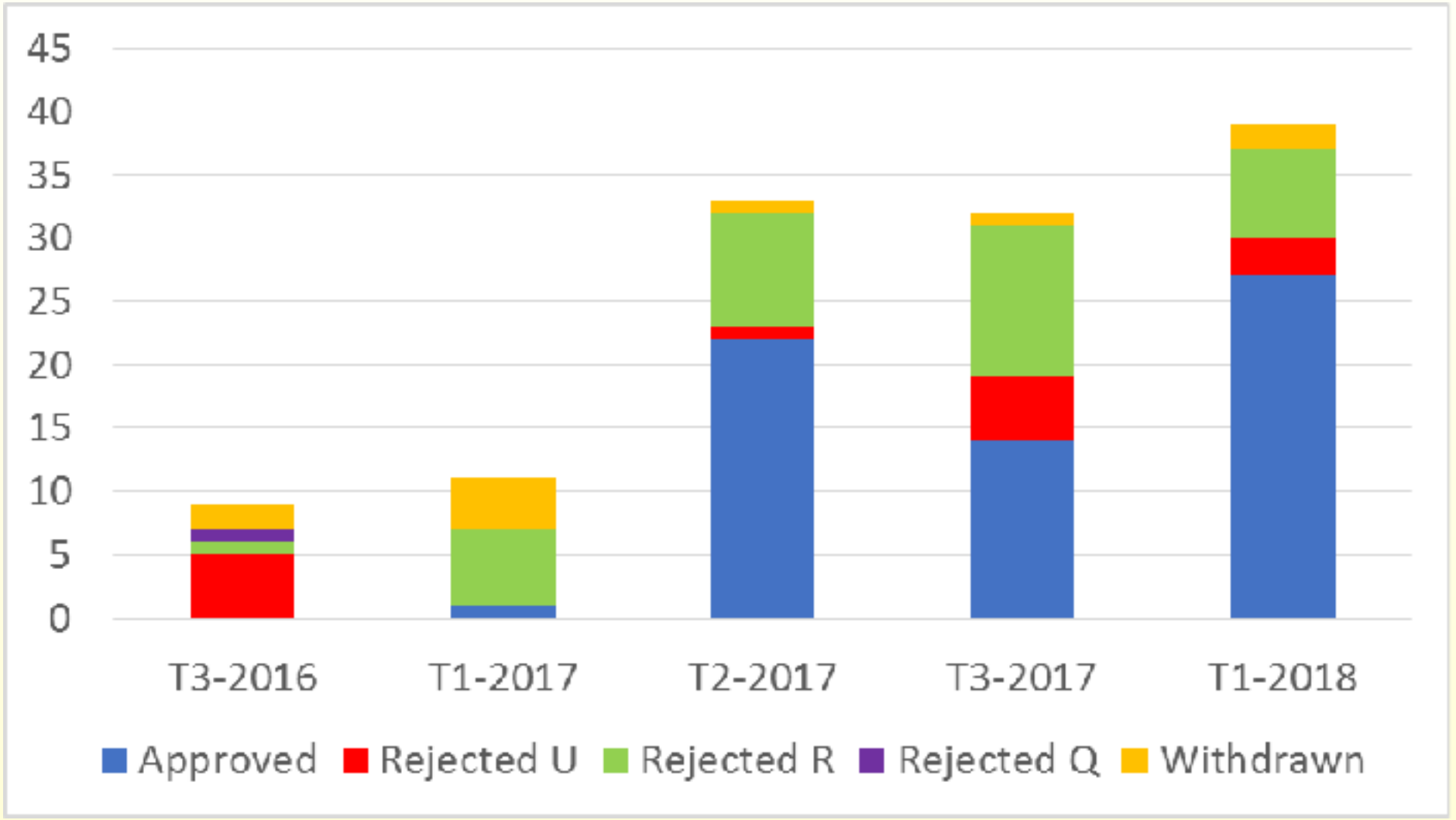
Data Hosting: Size Added

In January-April 2018 we added 39K files (~400 TB)

Data Collection	No. of files	On-disk size	Uncompressed size
Ark IPv4 Routed/24	32222	517 GB	1.6 TB
Ark IPv4 Routed/24 DNS Names	30	2 GB	9 GB
Ark IPv4 Prefix Probing	2812	303 GB	1 TB
Ark Internet Topology Data Kits	14	3 GB	27 GB
UCSD Network Telescope near real-time data	2911	154 TB	391 TB
Passive Traces	759	582 GB	1.2 TB
Cumulative totals	**38748**	**155 TB**	**395 TB**

Data Dissemination: Data Requests

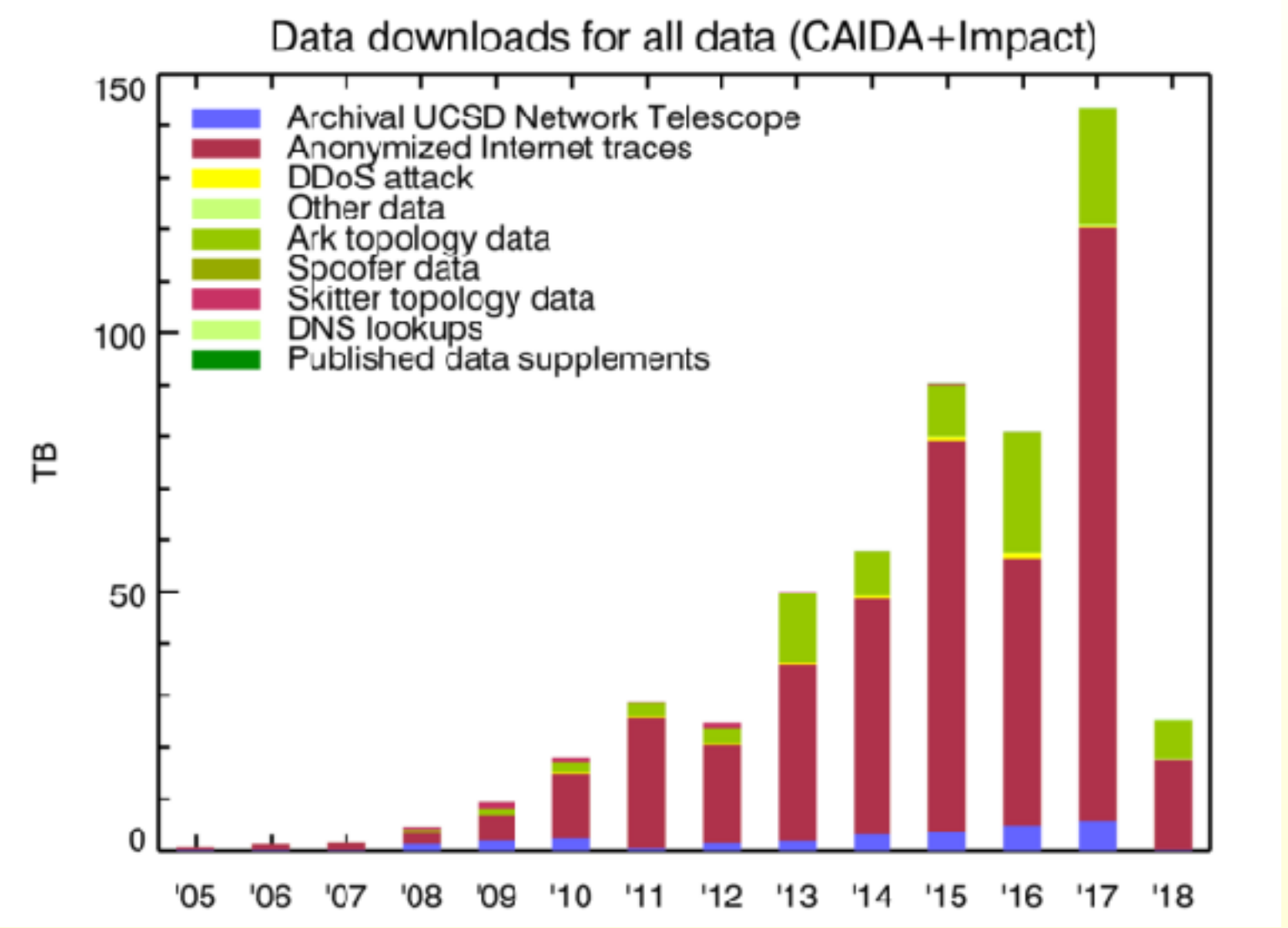
CAIDA Data Request Statistics through IMPACT
September 2016 – April 2018



Requests increasing, still too many rejected requests: discuss

Data Dissemination: Downloads

Two most popular datasets: Topology (IMPACT), Passive Traces (CAIDA)

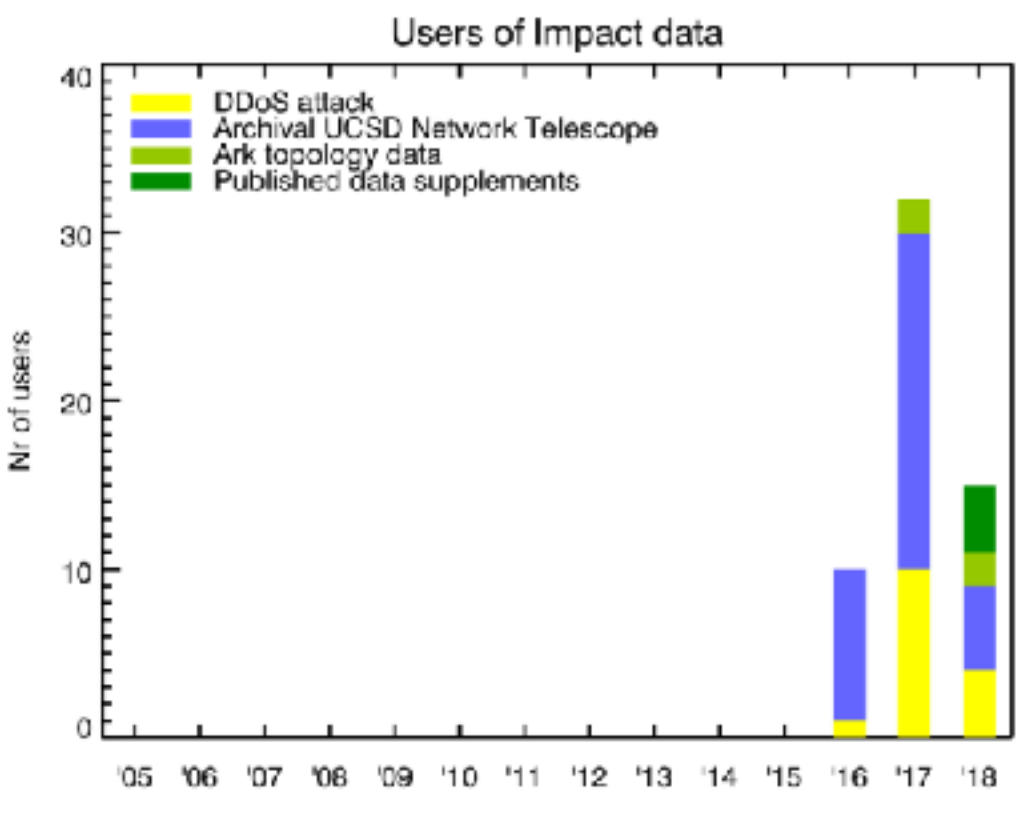


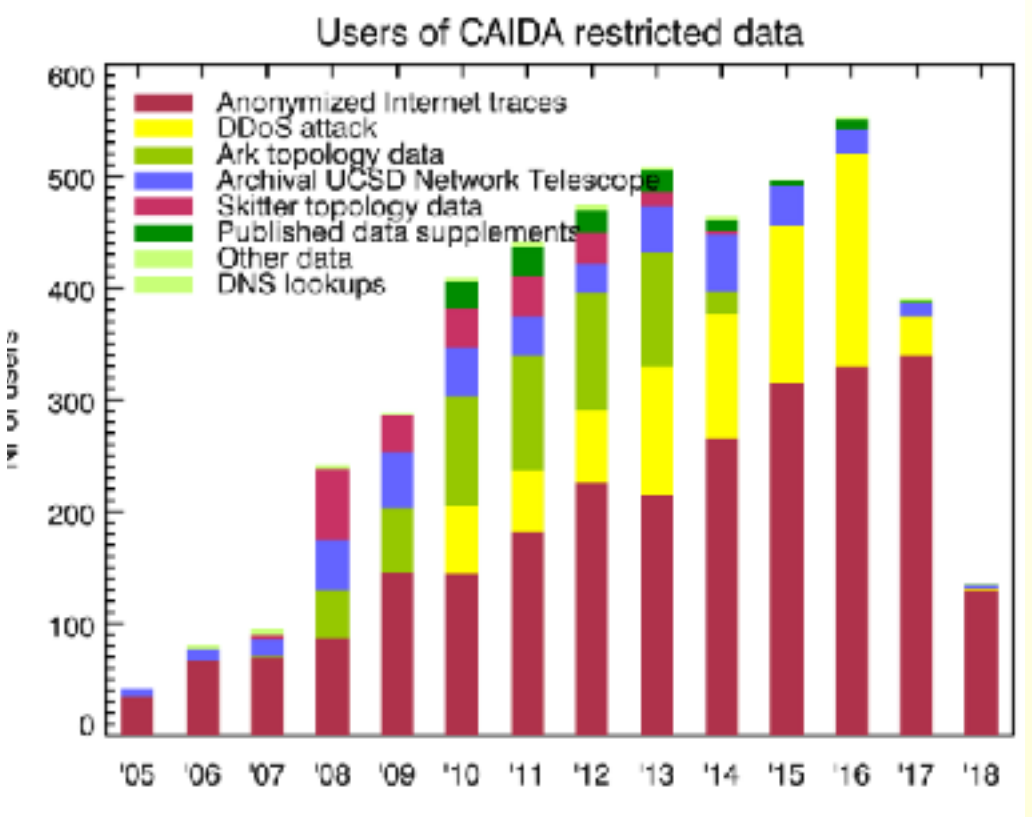
(Plot does not include UCSD (Near-RealTime) Telescope data available only on CAIDA machines)

Data Dissemination: Users

Number of users in first trimester of 2018 consistent with 2017

Passive traces (dark red) are only restricted dataset shared via CAIDA (not allowed in IMPACT)





Facilitating Data Discovery

Reorganized CAIDA Datasets in IMPACT

➢Bundled various datasets for user friendly access

- Combined 5 backscatter sets into “CAIDA UCSD Network Telescope Traffic Samples”
- Combined 14 ITDK datasets into “CAIDA UCSD Ark IPv4 Internet Topology Data Kits”
- Merged IPv4 Routed /24 Topology “Current” and “Older than 2 years” datasets
- Merged IPv4 Routed /24 DNS Names “Current” and “Older than 2 years” datasets
- Combined OC48 data sets

➢Recent ARK topology data (<= 1 year) available only through IMPACT

➢Withdrew old datasets from IMPACT and made them public

- Ark IPv6 topology
- Skitter Internet Topology Data Kits (outdated)
- ITDK’s older than 10 years

Approach 2: CAIDA as a Decision Analytics-as-a-Service Provider

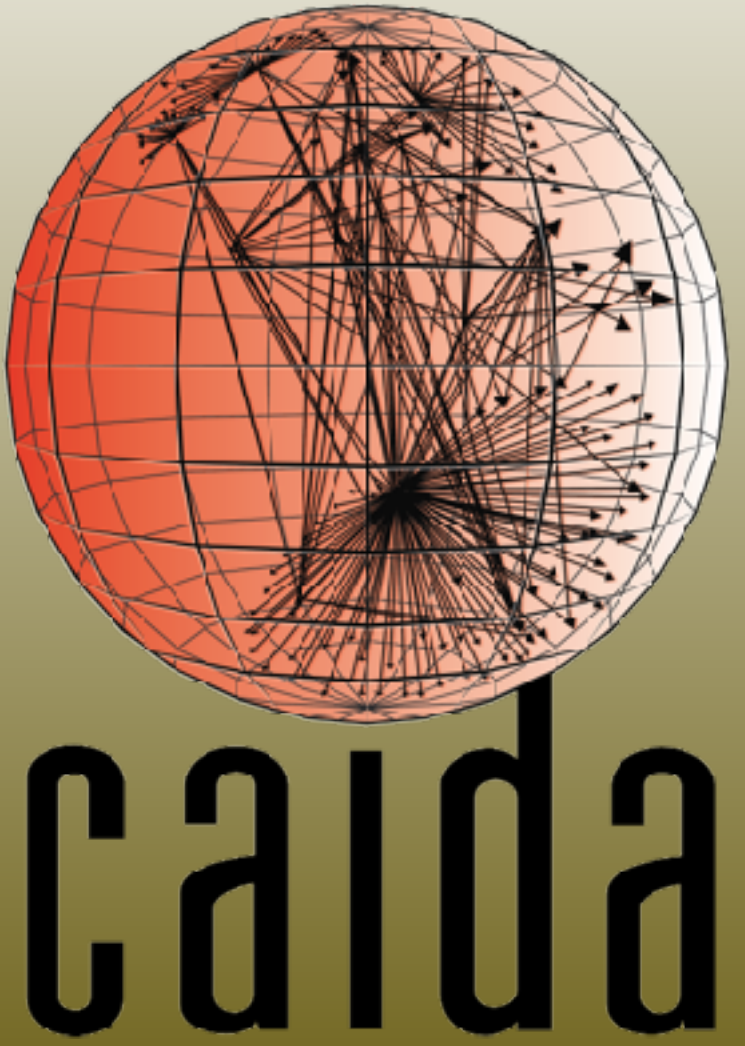
companion talk

Approach 3: Community Outreach

Workshop on Active Internet Measurement (AIMS 2018)

http://www.caida.org/workshops/aims/1803/

- 32 Non-CAIDA participants from 9 countries
- Industry (e.g. Google, Mozilla, NTT, RIPE), Government (e.g. Universities, DHS, FCC)
- Focused on interaction and coordination between different existing measurement infrastructures:
  - Measurements that address current or future public policy needs
  - Infrastructure sharing
  - Infrastructure development issues
  - Data access and sharing
  - Future measurement infrastructure architectures
  - Innovation in scientific discourse to promote reproducibility

Contact Information

PI: k claffy, CAIDA
kc@caida.org
http://www.caida.org/