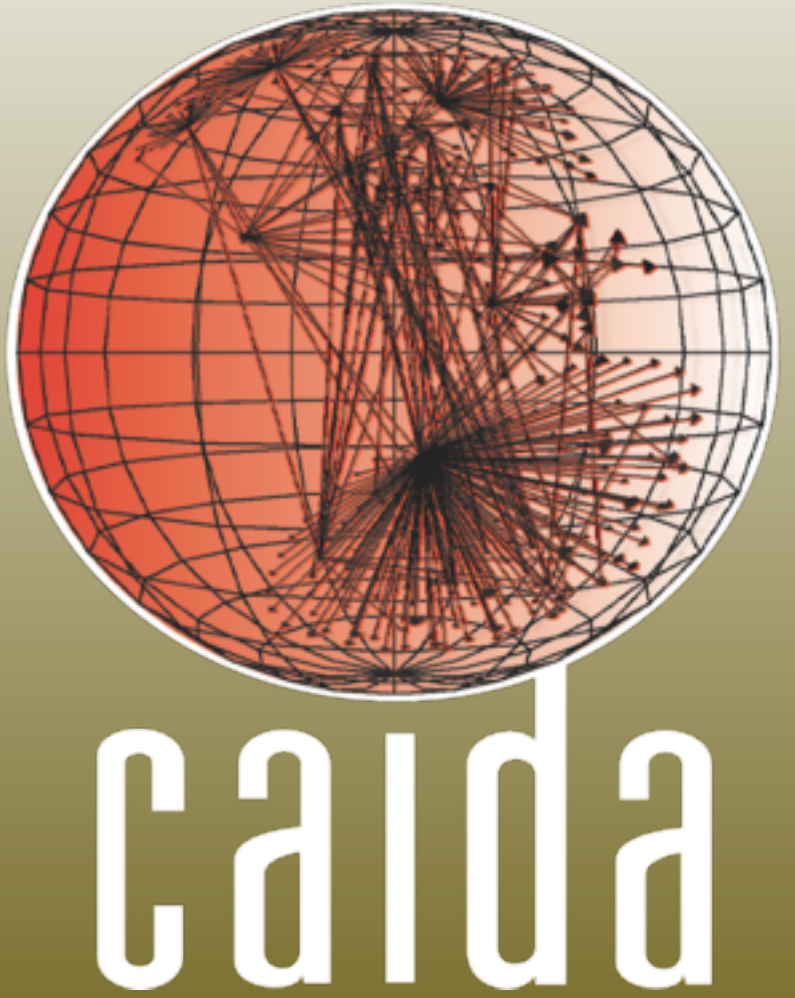
DHS ~~PREDICT~~ IMPACT project: CAIDA update

Bradley Huffaker,
kc claffy, PI
Marina Fomenkov, Co-PI

UCSD, La Jolla, CA
2-3 February 2016

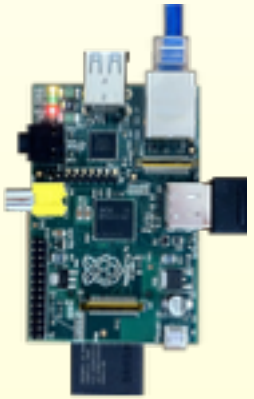


DHS IMPACT project: CAIDA update

- Data collection activities
  - Ongoing measurements
  - Data storage status
  - Data dissemination statistics
  - Aftereffects of changing CAIDA data access policies in 2014
- Other activities
  - Enabling commercial use of CAIDA data
  - Legal issues with providing CAIDA data to foreign military entities
  - Ethical issues in Cyber Research
- Open issues
  - Contributing new data to IMPACT
  - Re-orienting CAIDA data access policies toward IMPACT

Ongoing Measurements: Ark platform

- Concurrent data collection
  - IPv4 and IPv6 topology
  - spoofer
  - targeted congestion measurements
- Ark Platform (as of Jan 2016) - 132 monitors
  - 54 IPv6 enabled
  - 85 Raspberry PIs



=> Accelerated infrastructure growth
All new monitors - Raspberry PIs

Ark Produced Data

- Data sets
  - IPv4 Routed /24
  - ITDK (the last was in August 2015)
  - IPv4 Routed /24 DNS names

=> current - restricted, older than 2 years - publicly available

  - AS links, IPv4 - daily
  - AS links, IPv6 - daily, by monitors
  - IPv6 DNS names
  - AS ranking

=> publicly available

- More to come - access mode TBD
  - spoofer web reports
  - spoofer data
  - congestion data (supporting publications)
  - ...

Ongoing Measurements: Internet Background Radiation

- UCSD Network Telescope
  - 10 TB in December 2014
  - 22.6 TB in December 2015
  - >500 TB archived at NERSC
  - can prepare archived samples on demand
- Gold mine for security-related data sets
- Near-real-time data
  - two months sliding window of IBR traffic
- Archived samples are fine for many research uses
  - anonymization decreases the utility

Ongoing (for now) Measurements: Passive Trace Collection

- Passive infrastructure
  - two monitors with taps and Endace 10GE capture cards on the Equinix link in Chicago
  - **lost** the **San Jose** link in **2014** - it was upgraded to 100GB
  - they still may find a slower link there for us to monitor
- Data
  - Annual “Anonymized Internet Traces” since 2008
  - Our **most popular restricted dataset**
  - about **two thirds** of all requests for restricted datasets

Data storage

- Continue using NERSC for telescope data (free of charge)
- Continue to expand CAIDA storage
  - Tried SDSC Cloud, decided that local storage/ backups are more convenient for us
  - Resources required:
    - Labor maintaining our own hardware
    - Buying new disk shelves
    - Buying new data servers

Data Access

- 2014 - **transitional year**
- Made the following data sets public:
  - Ark IPv4 topology data older than 2 years
    - including raw data and ITDKs
  - All Ark IPv6 topology data
  - Older telescope data:
    - Backscatter data 2001-2008
    - Code-Red and Witty Worm data
    - Sipscan 2011 dataset
    - other archived data
- Announcements
  - February 2014 - blog about **topology** data
  - April 2014 - email to the list of CAIDA data users about changes in access mode for **topology** data
  - **never announced public availability of telescope data**
    - waiting to sort the availability and access issues with PREDICT

Access to publicly available data

- Fast, efficient, very user friendly
  - User info: entirely optional
    - some users still provide their info
  - No hand shake
  - Immediately downloadable
- Simplified AUP - less than a page
  - License
  - Suggested citation format
  - Disclaimer

=> Very popular!

Inferred AS Relationships Dataset - User Info Request

To maintain our funding for providing current and new datasets to the research community, it is very important for us, and our funding agencies, to know what research is done with these datasets.

Please fill out the form below to provide us with this information. All fields are optional. CAIDA will primarily use this information for reporting to our funding agencies. Email addresses will not be redistributed.

If you have any questions or problems using this form, please send email to data-info@caida.org.

Name:
Institution:
Email address:
Relationship to CAIDA:
- I am an academic researcher / faculty advisor / student
- I am a CAIDA sponsor
- I am a U.S. government contractor
- I am from a U.S. government agency
- Other (please specify):

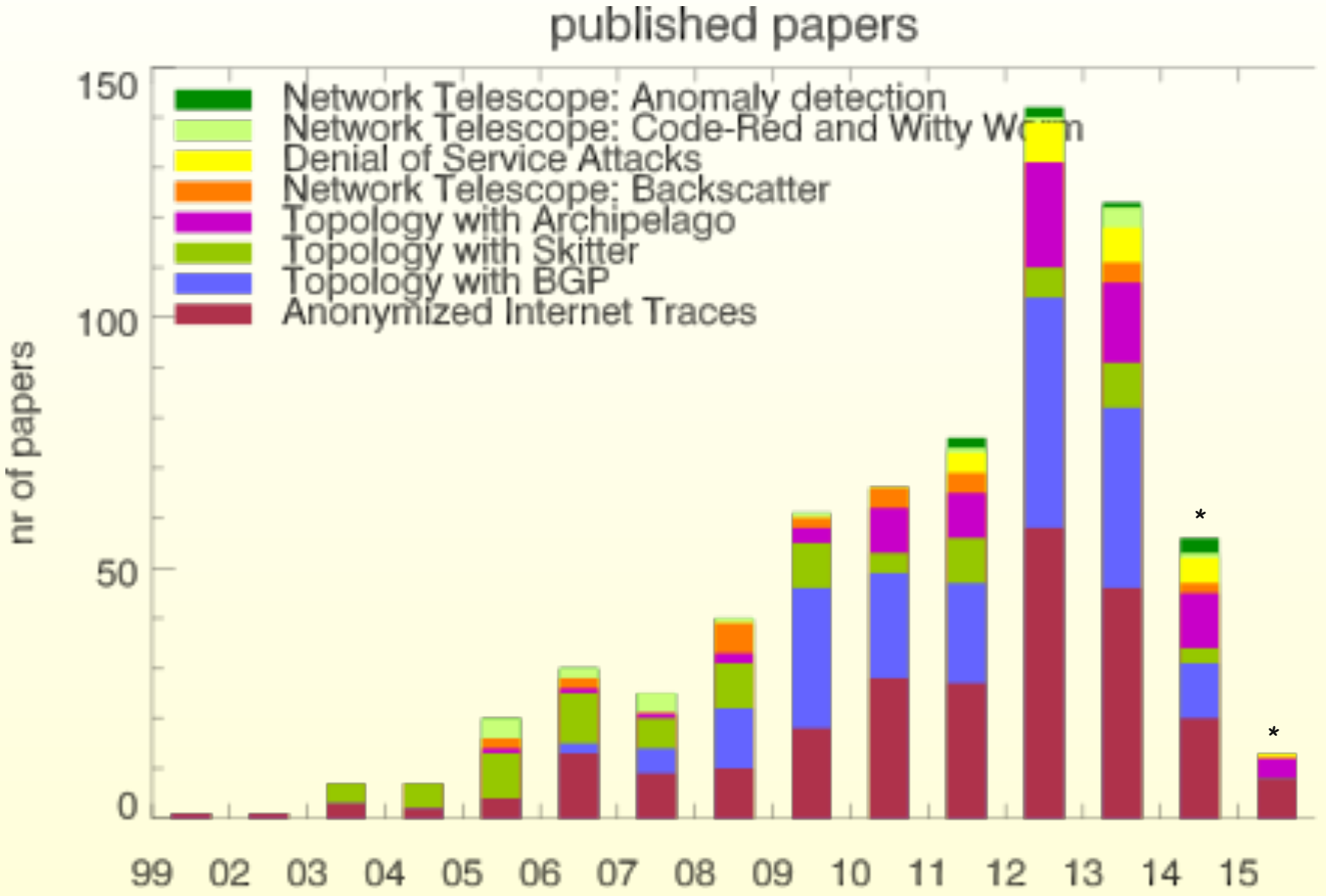
How many students will use the data? (if applicable):

Please let us know what you plan to do with this data:

Subscribe me to data-announce@caida.org (CAIDA data related announcements, e.g. availability of new datasets)

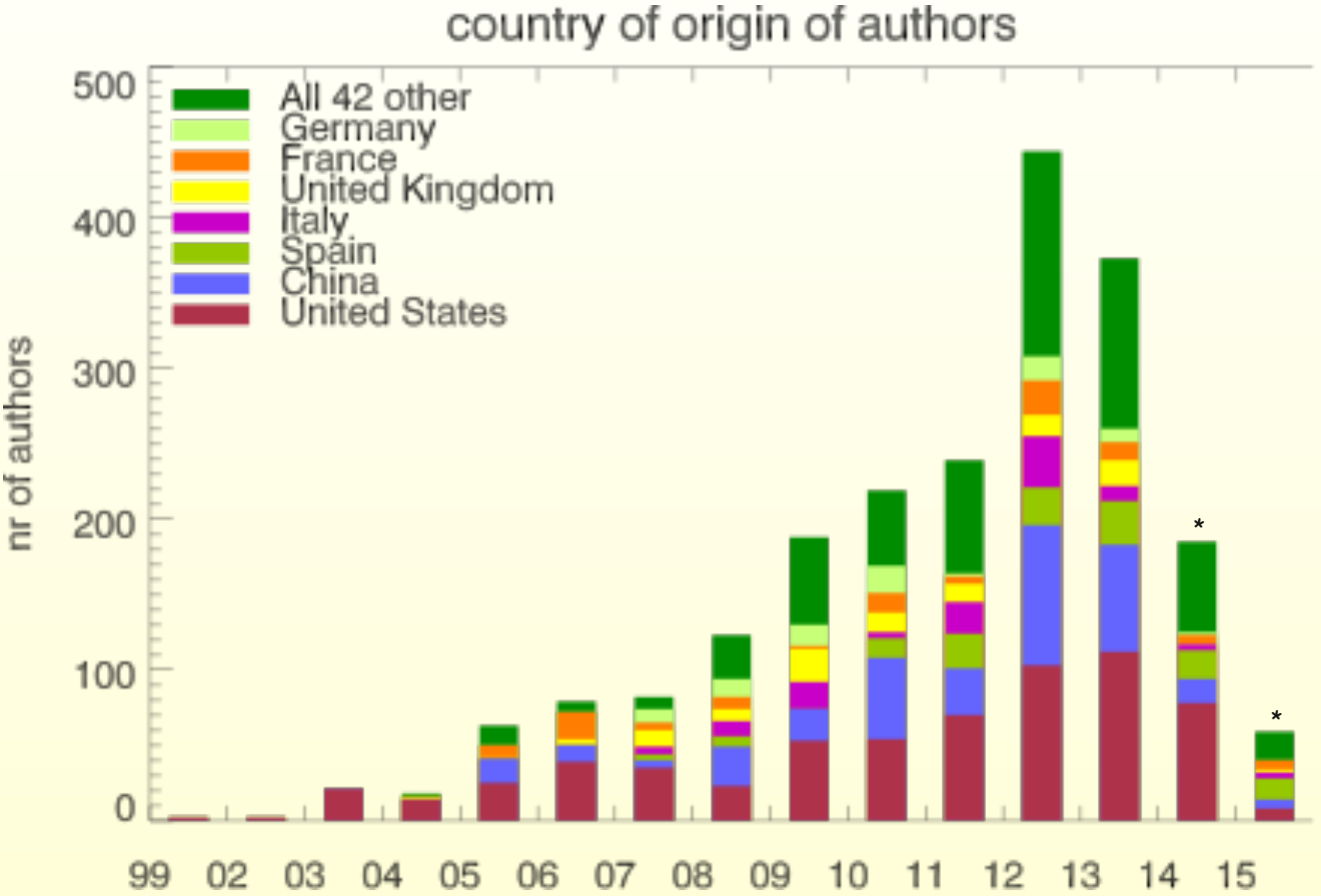
Submit Form and Download Data

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



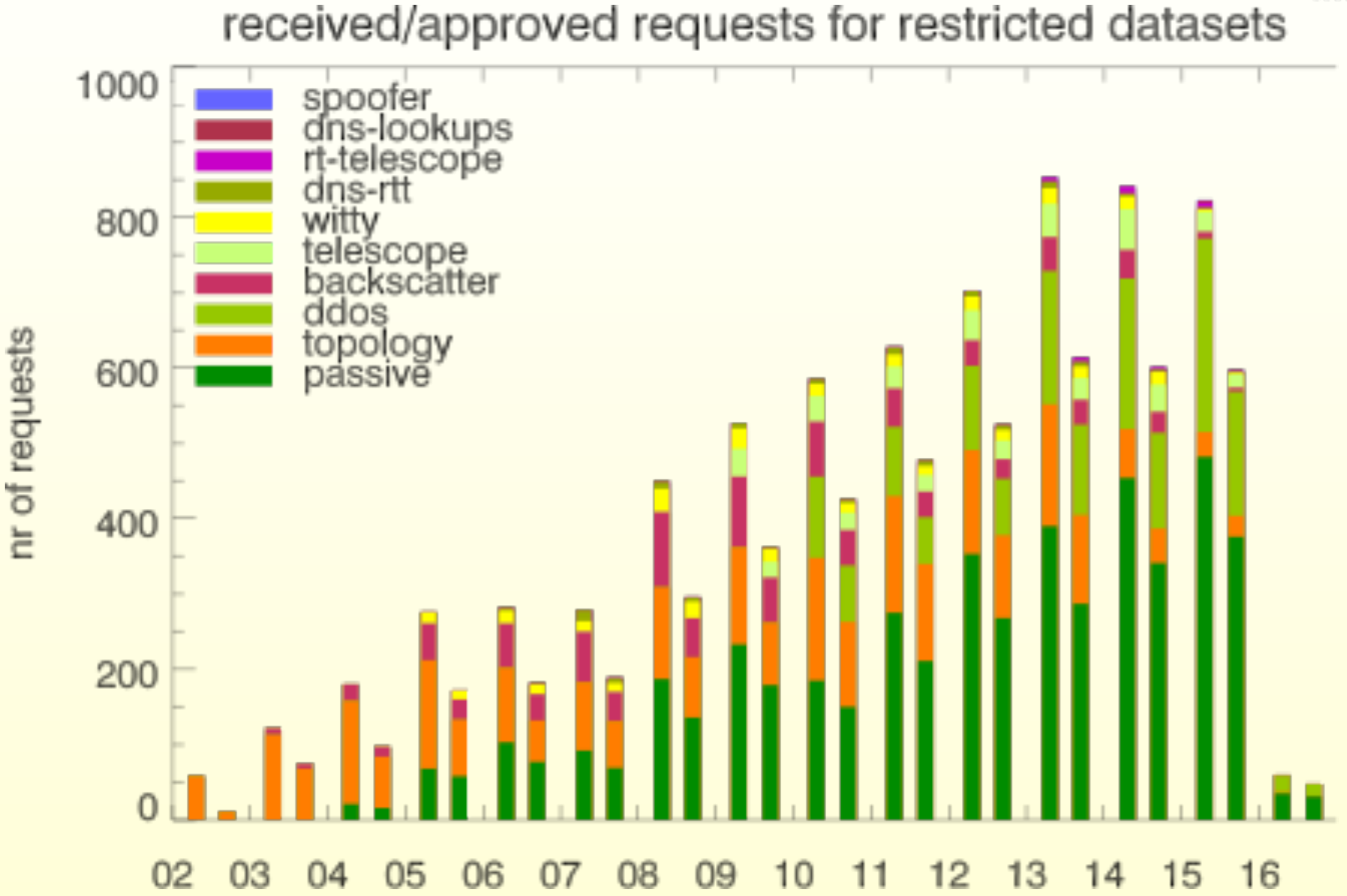
*** only self reported, need to spend a week looking for them.**

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

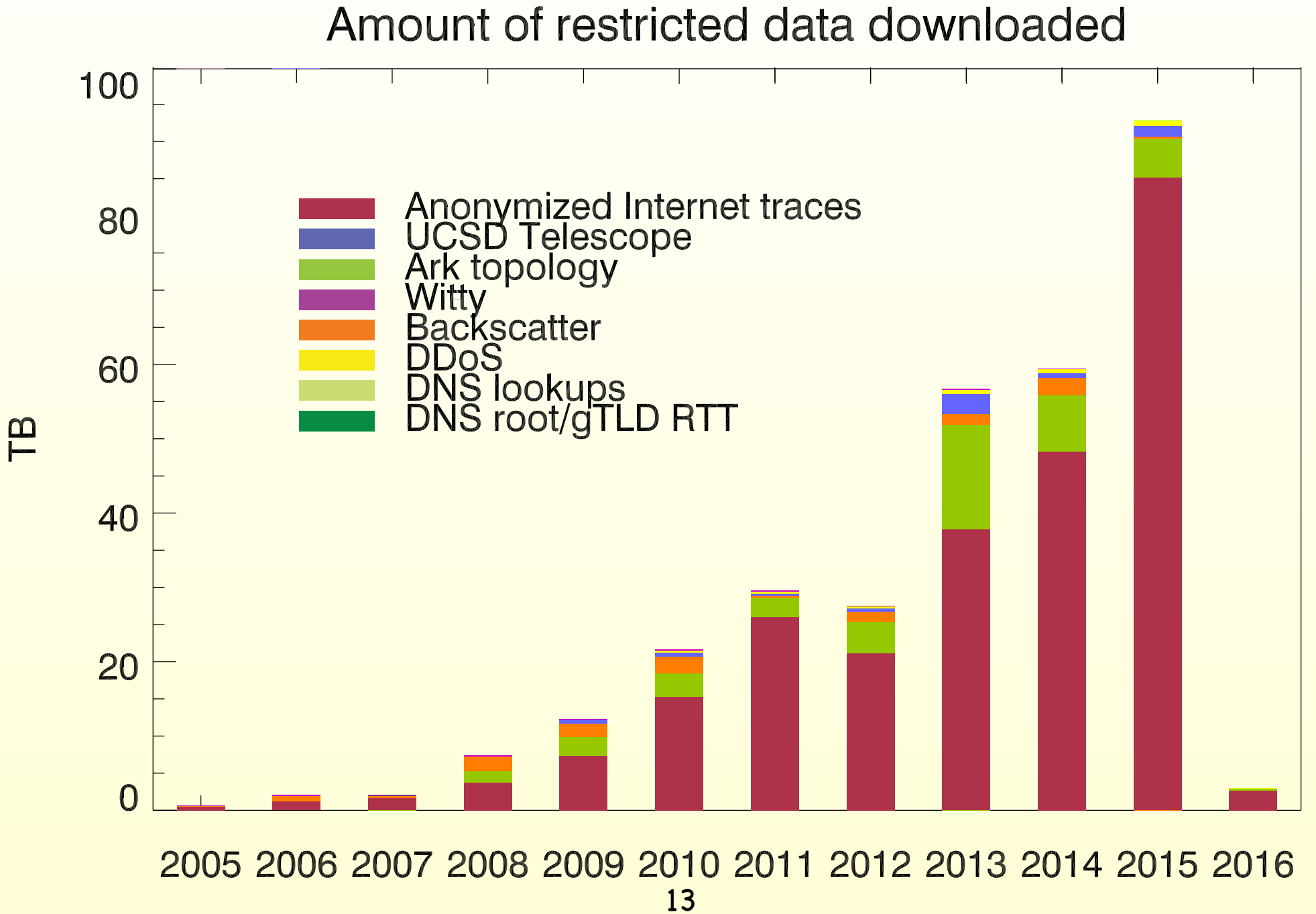


* only self reported, need to spend a week looking for them.

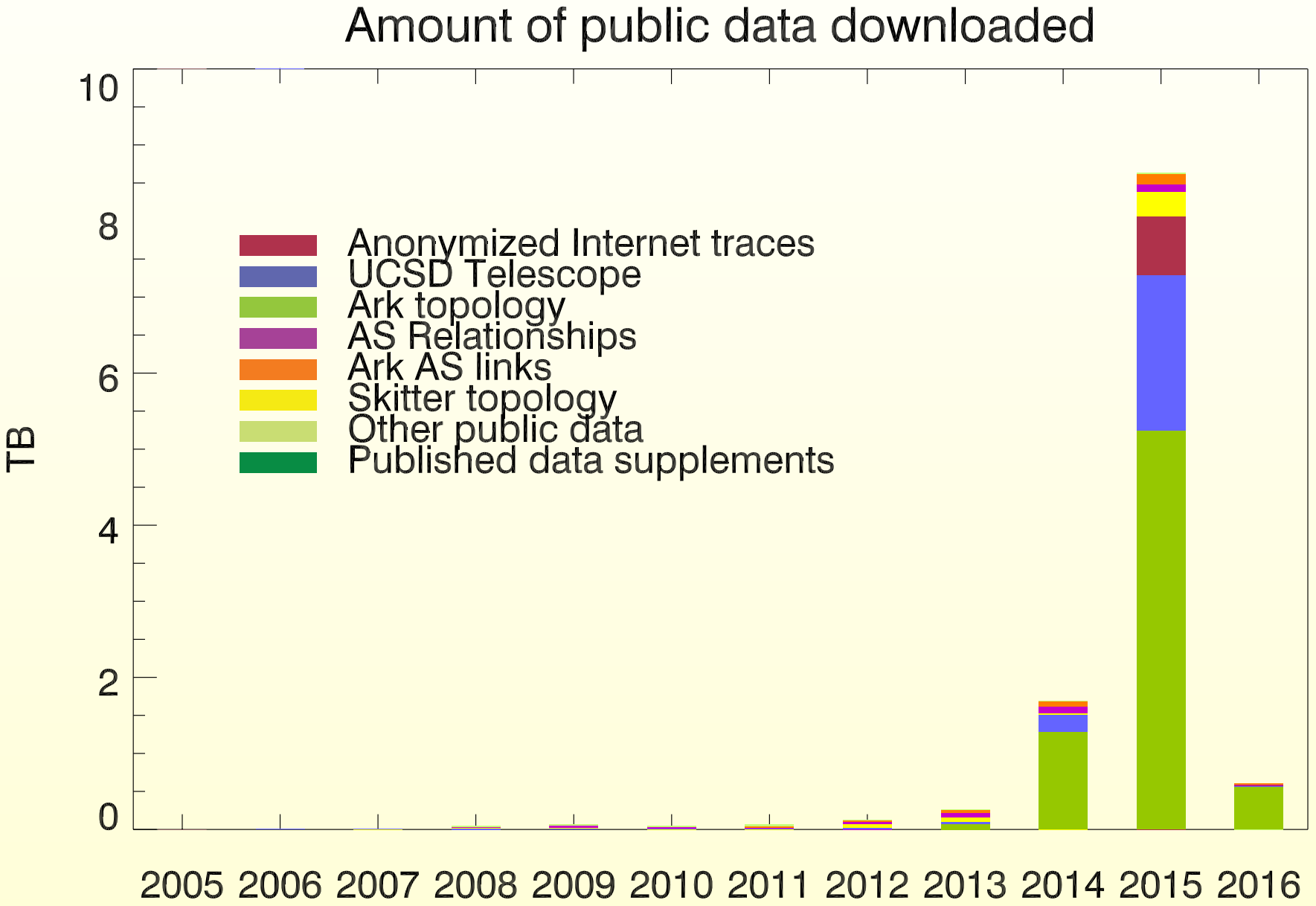
Restricted Dataset Requests



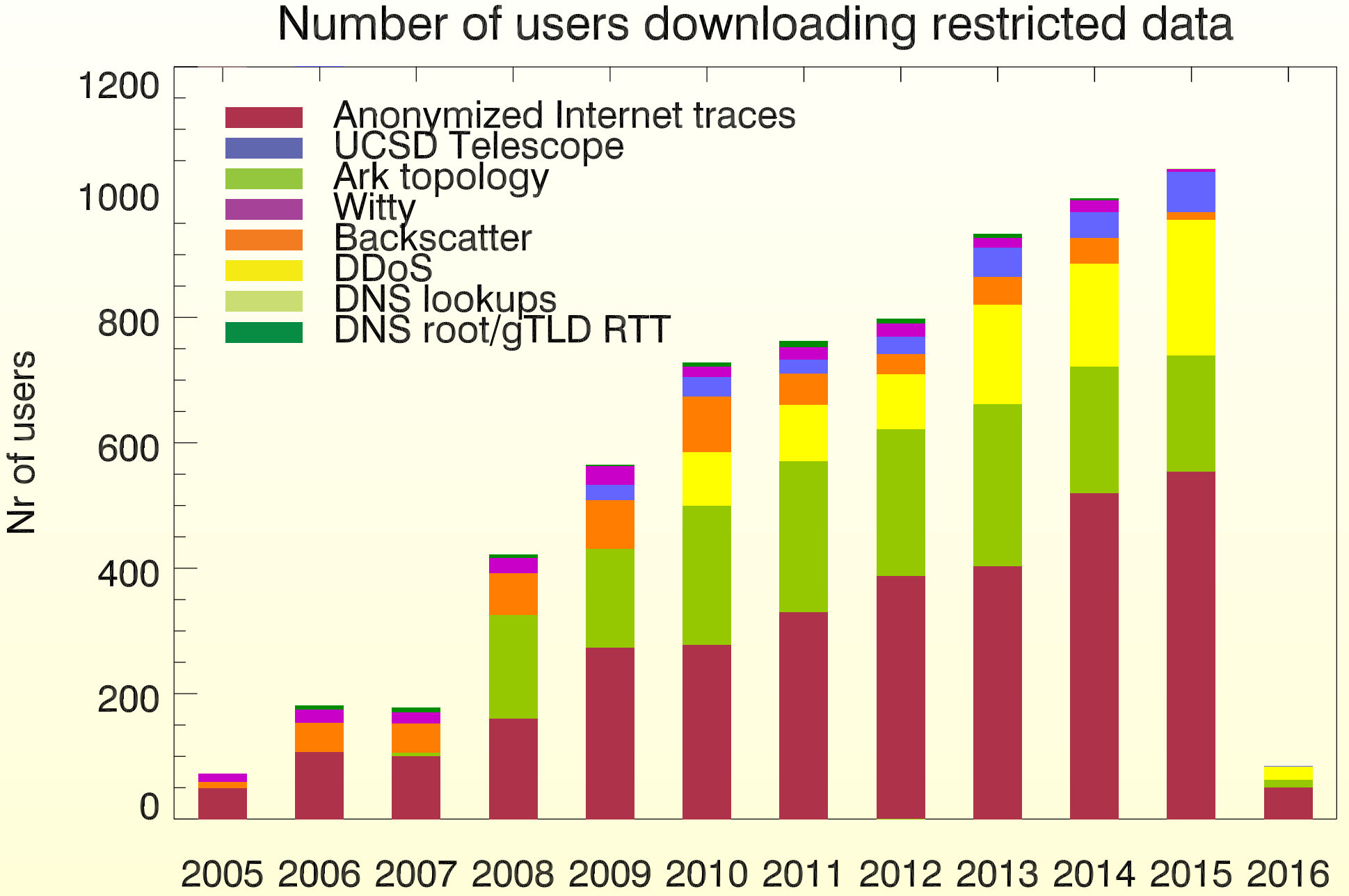
amount restricted data downloaded



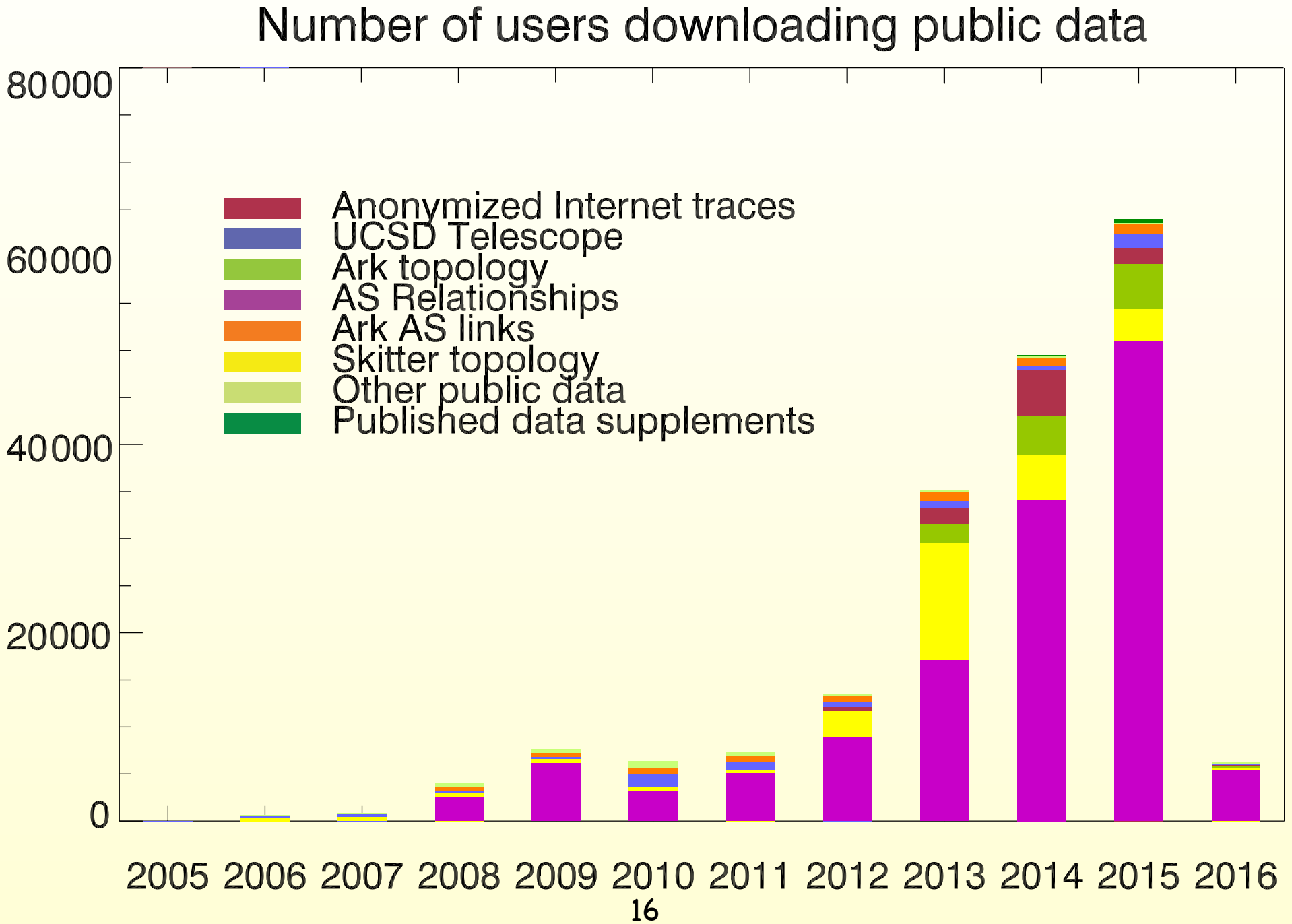
amount public data downloaded



number of users downloading restricted data



number of users downloading public data



Recent publications

- E. Kenneally, “**How to Throw the Race to the Bottom: revisiting Signals for Ethical and Legal Research Using Online Data**”, ACM SIGCAS Computers and Society, Feb 2015
- V. Giotsas, M. Luckie, B. Huffaker, and k. claffy, "**IPv6 AS Relationships, Clique, and Congruence**", in Passive and Active Network Measurement Workshop (PAM), Mar 2015
- R. Beverly, M. Luckie, L. Mosley, and k. claffy, "**Measuring and Characterizing IPv6 Router Availability**", in Passive and Active Network Measurement Workshop (PAM), Mar 2015
- T. Zseby, F. Vázquez, A. King, and k. claffy, "**Teaching Network Security With IP Darkspace Data**", IEEE Transactions on Education, Apr 2015.
- E. Raftopoulos, E. Glatz, X. Dimitropoulos, and A. Dainotti, "**How Dangerous Is Internet Scanning? A Measurement Study of the Aftermath of an Internet-Wide Scan**", in Traffic Monitoring and Analysis Workshop (TMA), Apr 2015

Recent publications

- A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescapè, "**Analysis of a "/0" Stealth Scan from a Botnet**", IEEE/ACM Transactions on Networking, Apr 2015
- E. Kenneally and M. Fomenkov, "**Ethics Research & Development Summary: Cyber-security Research Ethics Decision Support (CREDS) Tool**", in ACM SIGCOMM Workshop on Ethics in Networked Systems Research, Aug 2015
- K. Benson, A. Dainotti, kc claffy, A. Snoeren, and M. Kalitsis, "**Leveraging Internet Background Radiation for Opportunistic Network Analysis**", Internet Measurement Conference (IMC), Oct 2015

Potential New Data Sets

- CAIDA Internet Census 2013
  - category: Address Space Allocation Data
  - sub-category: TBD
  - derived from both active and passive measurements
    - methodology described in *http://www.caida.org/publications/papers/2015/lost_in_space/*
  - classifies /24 address blocks as:
    - IETF-reserved
    - observed as used
    - routed unused (unobserved)
    - unrouted assigned
    - available

  - includes AS announcing in BGP the longest network prefix containing this address block
  - access mode: TBD
- spoofer: project in progress, data sets TBD

Other Security-related New Data Sets

- scanning data, 2008-2015
  - IP, protocol, dst port of scanners
  - hourly rates of packets, flows
  - destinations
  - scanning strategy
- backscatter data
  - attack on Spamhaus 2013 (TCP backscatter)
  - BitTorrent index poisoning attacks
- attacks using DNS open resolvers
- botnet data: ZeroAccess and Sality
  - Command and Control packets

Other Security-related New Data Sets (cont.)

- Details and Methodology in

K. Benson, A. Dainotti, k. claffy, A. Snoeren, and M. Kallitsis, "**Leveraging Internet Background Radiation for Opportunistic Network Analysis**", Internet Measurement Conference, Oct 2015. *http://www.caida.org/publications/papers/2015/leveraging_internet_background_radiation/*

- Access: restricted
  - anonymization would decrease the research utility

Ideas for IMPACT?

- 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 (always) room for more marketing efforts
  - cf. typical advertising strategies: “How did you learn about our product”? (web, newspaper, TV, Facebook, friends…)
- Make CAIDA data accessible exclusively via IMPACT portal?