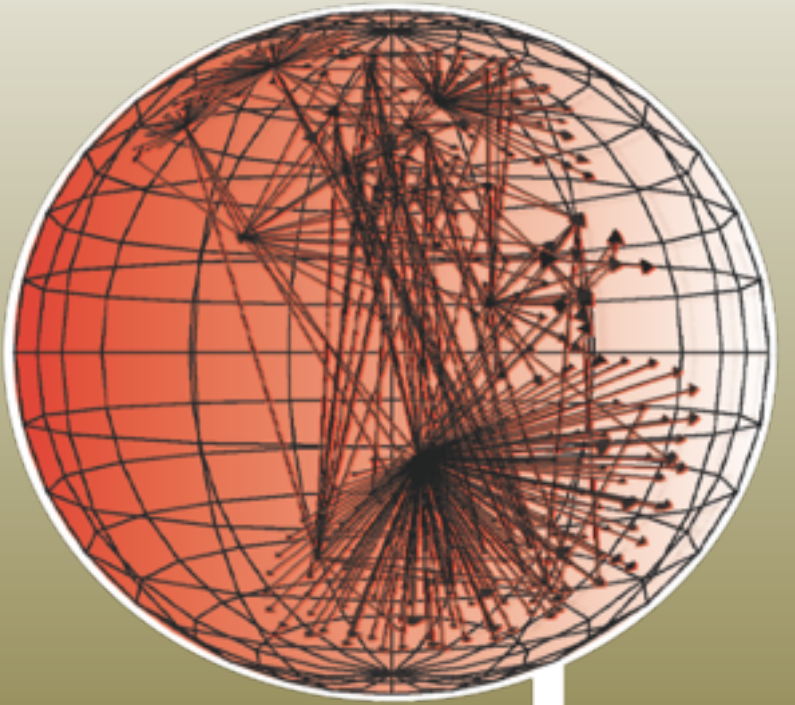
DHS PREDICT project: CAIDA update

PI k claffy, CAIDA
ISI/USC
Marina Del Rey, CA.
28-29 May 2014



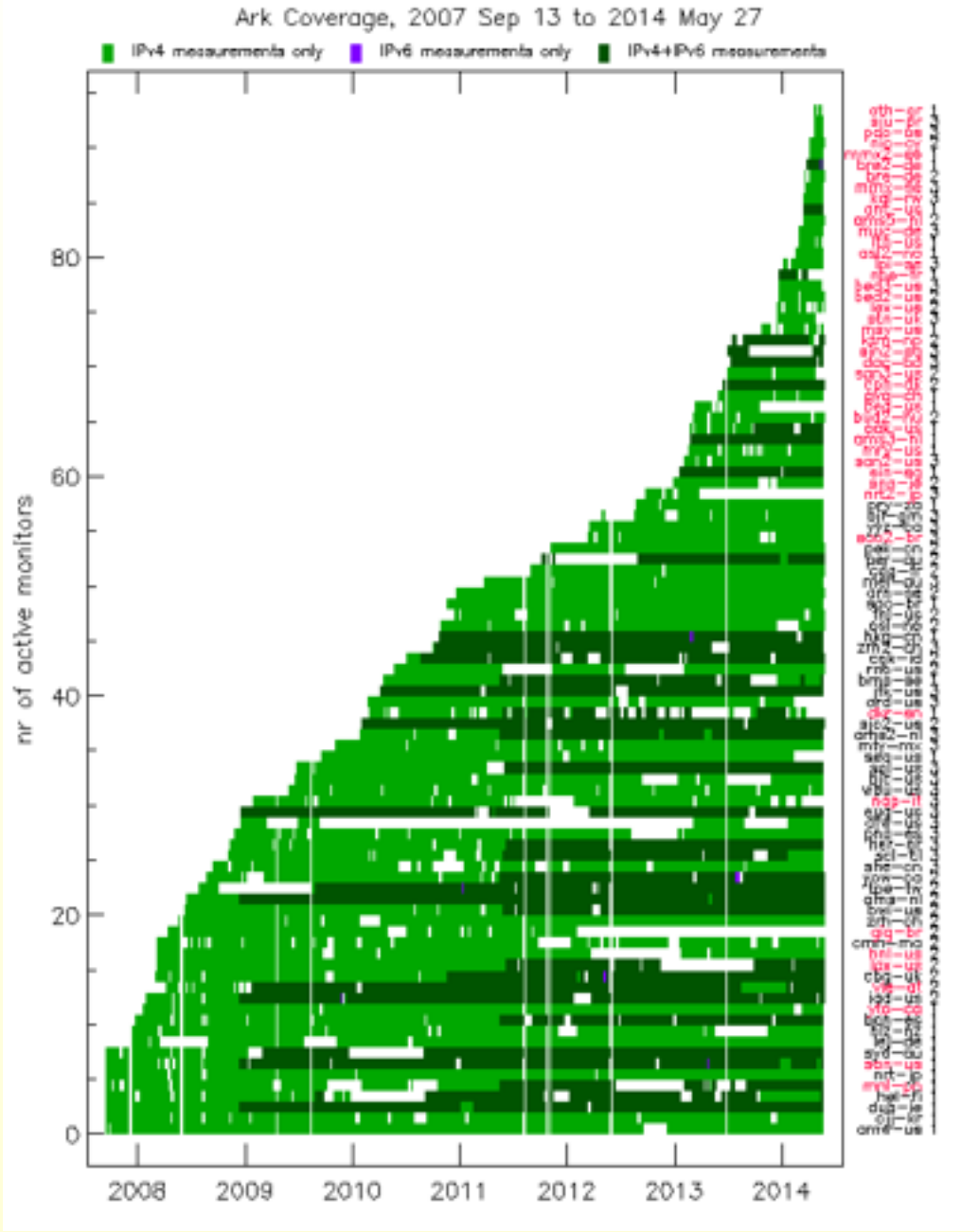
DHS PREDICT project: CAIDA update

- 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
  - continued, fine tuning congestion measurements
- Ark Platform (as of May 2014) - 94 monitors
  - 35 IPv6 enabled
  - 46 Raspberry PIs
- Derived data sets
  - ITDK (April 2014 released soon)
  - AS links, IPv4 daily
  - AS links, IPv6 daily
  - 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

Data Storage

- Continue to use NERSC for the telescope data (free of charge)
- Continue using SDSC Cloud for other data (for fee)
- Continue to expand CAIDA storage
  - Acquired in December:
    - 60 TB disks
    - 10 Gb network card
  - As planned in PREDICT Yr 2

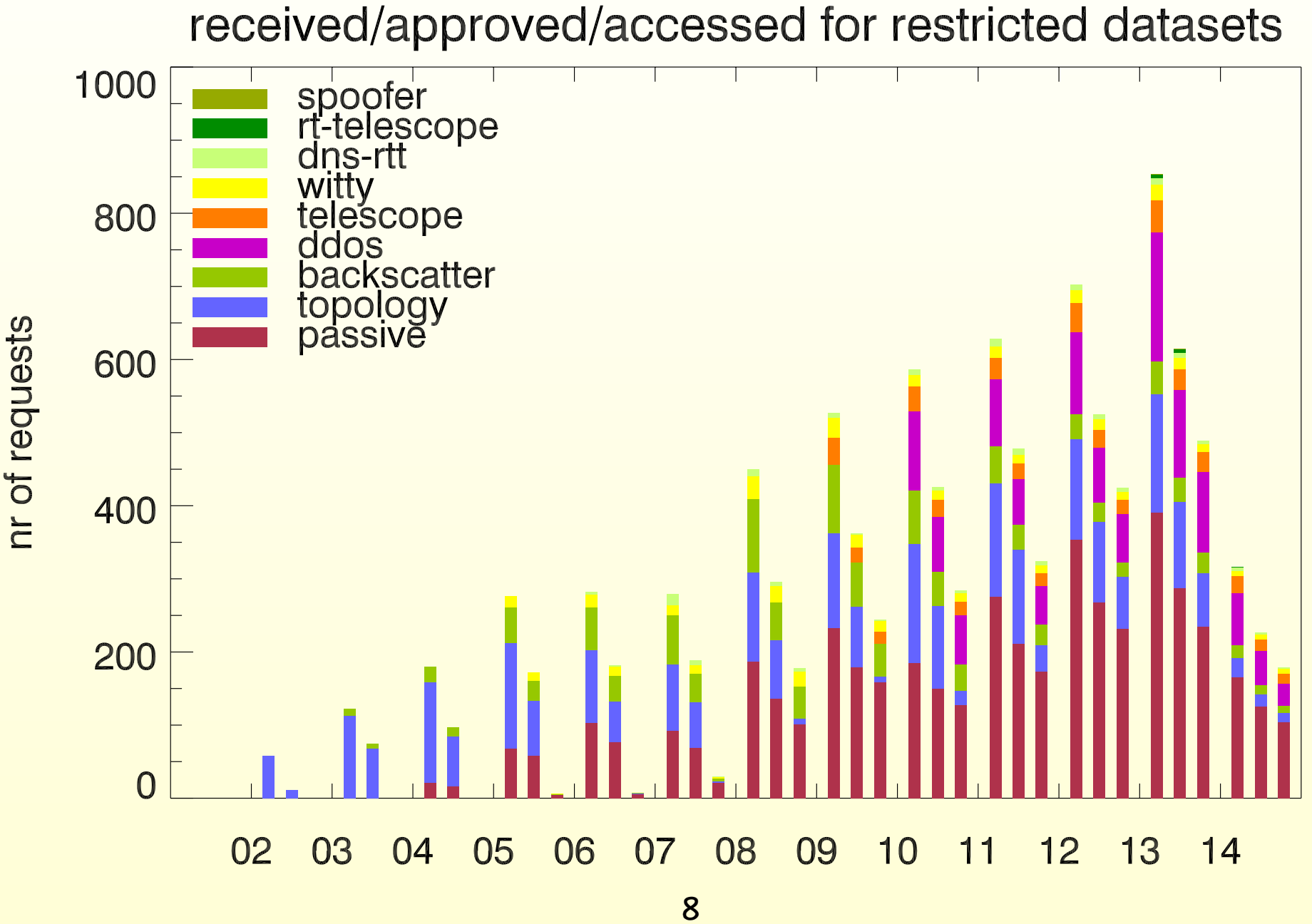
Data Access

- 2013 and 2014 Requests Summarized

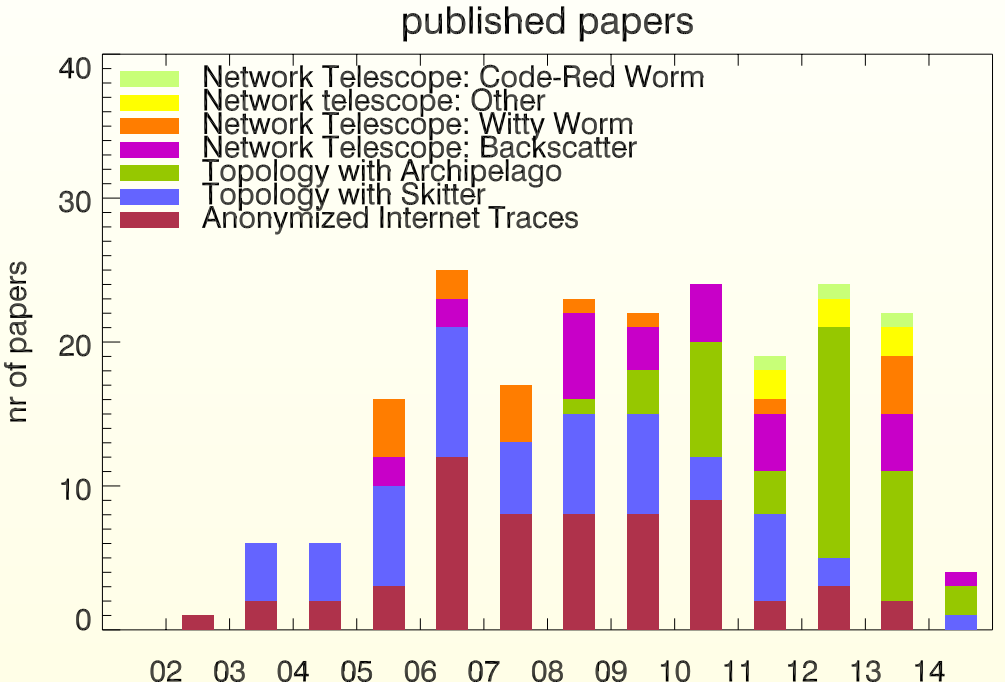
	2013		2014	
Data	**Received**	**Approved**	**Received**	**Approved**
Topology	162	118	27	17
Passive	390	287	165	125
Telescope	44	29	24	16
Witty	21	15	7	7
backscatter	45	33	17	13

- Real Time Telescope Users
  - Zhenxin Zhan (UTSA, TX, US)
  - Kirill Levchenko (UCSD, CA, US)
  - Xiyue Deng (ISI, CA, US)
  - Qiu-Hong Wang (Huazhong Univ. SciTech, CN)
- Pending Requests
  - Tatsuya Mori (Waseda University, Tokyo, JP)
  - Frank Acker (Nova SE Univ, FL, US)
  - Kai-Lung Hui (Hong Kong Univ SciTech)

Restricted Dataset Requests, 2014



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



Number of authors per country (predict=true)

As determined from author affiliations specified in papers.
The count includes authors and co-authors
There are 209 papers with 447 different authors in 32 countries
The average number of authors per paper is 3.1

United States	202	Spain	12	Australia	6	South Africa	2
China	112	Poland	11	Singapore	5	Denmark	1
United Kingdom	61	Belgium	10	Canada	4	Hungary	1
France	53	India	8	Lebanon	3	New Zealand	1
Germany	43	Portugal	7	Korea (South)	3	Romania	1
Italy	39	Argentina	7	Finland	2	Switzerland	1
Japan	15	Taiwan	6	Netherlands	2	Kuwait	1
Greece	12	Tunisia	6	Iran	2	Sweden	1

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

Changed CAIDA data policy (Feb 2014)

- Topology data - 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 NY Resolution”:

make more data easily sharable

(but proceed cautiously

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

Master AUP Revisions

- Removed “no-probe” restriction
  - 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 change of Researcher’s contact information

Propagated changes into CAIDA PREDICT DUA

Recent publications

- k. claffy and D. Clark, **Workshop on Internet Economics (WIE2013) Report**, ACM SIGCOMM Computer Communication Review (CCR), 2014.
- 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, 2014.
- A. Dainotti, C. Squarcella, E. Aben, K. Claffy, M. Chiesa, M. Russo, and A. Pescapè, **Analysis of Country-wide Internet Outages Caused by Censorship**, IEEE/ACM Transactions on Networking, 2014.
- A. Lodhi, N. Larson, A. Dhamdhere, C. Dovrolis, and k. claffy, **Using PeeringDB to Understand the Peering Ecosystem**, ACM SIGCOMM Computer Communication Review (CCR), vol. 44, no. 2, pp. 21--27, Apr 2014.
- A. Lodhi, A. Dhamdhere, and C. Dovrolis, **Open Peering by Internet Transit Providers: Peer Preference or Peer Pressure?**, in IEEE Conference on Computer Communications (INFOCOM), Toronto, Canada, Apr 2014.
- M. Bailey and E. Kenneally, **Cyber-security Research Ethics Dialogue & Strategy (CREDS) Workshop Report**, ACM SIGCOMM Computer Communication Review (CCR), vol. 44, no. 2, pp. 76--79, Apr 2014
- W. de Donato, A. Pescapè, and A. Dainotti, **Traffic Identification Engine: An Open Platform for Traffic Classification**, IEEE Network, Mar 2014.

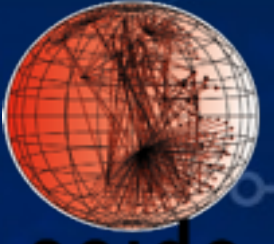
Recent publications

- M. Luckie and k. claffy, **A Second Look at Detecting Third-Party Addresses in Traceroute Traces with the IP Timestamp Option**, in Passive and Active Network Measurement Workshop (PAM), Mar 2014.
- V. Reddyvari~Raja, A. Dhamdhere, A. Scicchitano, S. Shakkottai, k. claffy, and S. Leinen, **Volume-based Transit Pricing: Is 95 The Right Percentile?**, in Passive and Active Network Measurement Workshop (PAM), Mar 2014
- T. Zseby, N. Brownlee, A. King, and k. claffy, **Nightlights: Entropy-based Metrics for Classifying Darkspace Traffic Patterns**, in Passive and Active Network Measurement Workshop (PAM). Mar 2014, PAM 2014.
- T. Zseby, A. King, M. Fomenkov, and k. claffy, **Analysis of Unidirectional IP Traffic to Darkspace with an Educational Data Kit**, Feb 2014.
- A. Elmokashfi and A. Dhamdhere, **Revisiting BGP Churn Growth**, ACM SIGCOMM Computer Communication Review (CCR), vol. 44, no. 1, Jan 2014.
- A. Dainotti, K. Benson, A. King, k. claffy, M. Kallitsis, E. Glatz, and X. Dimitropoulos, **Estimating Internet address space usage through passive measurements**, ACM SIGCOMM Computer Communication Review, Jan 2014.
- A. King, A. Dainotti, B. Huffaker, and k. claffy, **A Coordinated View of the Temporal Evolution of Large-scale Internet Events**", Computing, Jan 2013.

10 Years Later: What Would I have Done Differently? (Or what would I do today..)

DHS PREDICT PI Meeting

PI k claffy, CAIDA/UCSD
ISI/USC
Marina Del Rey, CA.
28-29 May 2014



caida
www.caida.org

UC San Diego

Ancient History (~2004)

KC I need real traffic data to validate my ROI on security technologies!

May the force be with you!

Ten Things I Know About the Internet

[thinking about this question reminded me of my 2008 rant to lawyers, so i revisited it..]

http://www.caida.org/outreach/publications/2008/ten_things/

#1 Legal frameworks are obsolete

1. Updating legal frameworks (from copyright to privacy to wiretapping to common carriage) to accommodate technological advancement requires first updating other legal frameworks (data-sharing) to accommodate empirically grounded research into what we have built, how we use it, and what it costs to sustain.

#2 Most obstacles are not technical

2. Our scientific knowledge about the Internet is weak, and the obstacles to progress are primarily issues of economics, ownership, and trust (EOT), rather than technical.

#3 We know enough to worry

3. Despite the methodological limitations of Internet science today, the few data points available suggest a dire picture.

- e.g., security of naming, addressing, routing

#4 We are not unique

4. [TopTen] The data dearth is not a new problem in the field; many public and private sector efforts have tried and failed to solve it.

- "Unfortunately after four years the PREDICT project has not yet launched, and when it does it will not be able to include data on networks that serve the public, since the legal territory is too muddy for DHS lawyers to navigate while EFF lawsuits have everyone in the U.S. government skittish about acknowledging surveillance of any kind. Even the private networks that PREDICT can serve immediately, such as Internet2 (the research backbone in the U.S. serving a few hundred educational, commercial, government, and international partners) have lamented that the PREDICT framework does not solve their two biggest problems: sketchy legal territory, and fear of RIAA subpoenas and/or lawsuits."

  – *Ten Things Lawyers Should Know About Research*

#5 Lack of data distorts culture

5. Thus the research community finds itself in the absurd situation of not having the fundamental data to do the most basic network research.

- e.g., QoS vs NN, congestion, traffic
- and yet lots of infrastructure spending
  - M-Lab, Bismark, Ark, RIPE Atlas
  - [tho not focused on security]
- culture has responded in two ways
  - under the table, private agreements
  - data-less research (the price of freedom..)

#6 Non-scientists have better data

6. A growing number of segments of society have access to, and use, private network information on individuals for purposes we might not approve.

- WIE2012,WIE2013 workshop reports
  (http://www.caida.org/workshops/wie)
- fervent race to invade our privacy
  as efficiently as possible

#7 Incentives profoundly misaligned

7. Government regulatory agencies have as much reason to be reluctant as providers regarding disclosure of how the Internet is engineered, used, and financed.

#8 Opaqueness breeds uncertainty

8. The opaqueness of the infrastructure to empirical analysis has generated problematic responses from communities earnestly trying to get their jobs done.

- IETF, operators, vendors, ICANN, policy wonks, academics
- IPv6, CGNs, new GTLDs, NN, peering

- little long-term thinking (partially due to funding constraints)
- little sense of Internet history
- no one has oversight for coordination or even articulation of the global picture

#8 (cont)

Academic Internet researchers also operate in a funding environment that does not promote tackling 10-year problems, nor are they equipped to navigate the conflict of interests between the university and the providers of network data. Providers either legally cannot or are reluctant to share data without restrictions on what can be published about their network, and universities have rules limiting such restrictions. And so federal agencies funding research continue to spend millions of R&D dollars per year developing lots of technology, even legal technology, to promote data retention and sharing, but the agencies and the taxpayers they represent get little in return. A related problem is that the lack of experience with data sharing in an admittedly quite young field of science means that there is no established code-of-conduct for protecting user privacy and engaging with Institutional Review Boards to navigate ethical issues in Internet measurement research.

#9 It's not that bad (maybe)

9) There is a reason everyone wants connectivity to all the world's knowledge and each other besides its status as the most powerful complex system ever created by man. The Internet's practical promise for individual freedom, democratic engagement, and economic empowerment provides sufficient inspiration for an open, technically literate conversation about how to invest in technologies and policies to support articulated social objectives.

- *"The Internet is Americas most important platform for economic growth, innovation, competition, free expression, and broadband investment and deployment."*
  -FCC-14-61A1.pdf NOTICE OF PROPOSED RULEMAKING, http://www.fcc.gov/document/protecting-and-promoting-open-internet-nprm

#10 Some next steps are obvious

10). All cross policy-technology boundaries

Guideline: architectural innovation in competitive core is rarely successful

e.g., ATM, multicast, routing security, IPv6, DNSSEC, QOS

Maybe more promising are methods that:

leverage cooperation at edge (where innovation is)

…….. how does any of this relate to PREDICT (and the broader issue of data sharing)?

PREDICT: Reality Check *(from Erin)*

- PREDICT.v0
- **Goal: Partnership among government, critical information infrastructure providers, and security development communities (both academic and commercial), to bridge the gap between producers of security-relevant network operations data and technology developers and evaluators who can leverage this data to accelerate the design, production, and evaluation of next-generation cybersecurity solutions.**

- *REALITY: DHS-funded project to support infrastructure for network and computer security research data collection and sharing primarily by a limited number of academic researchers to academic researchers.*

Premise: PREDICT matters

- Everything about the Internet needs data.
  - cybersecurity: what is the impact of heartbleed? Conficker? how many nets allow spoofing?
  - infra. protection: where are vulnerable points?
  - science: how big is this thing anyway? how is it evolving? can we model anything about it? how can we establish scientific reproducibility?
  - policy: are consumers being harmed by policy X?
  - future architecture: what are we building?

1) Figure out what data is needed

- Lack of feedback loop with researchers
- Need equivalent of community workshop every yea
  - previous attempts have not gotten traction
  - need other ways of reaching users
    - surveys at conferences/workshops
      - e.g., http://www.caida.org/projects/ditl/questions/
    - breakout sessions or workshops at conferences
    - annual PREDICT data users meeting
    - active mailing list, web forum
- Catalog data being used at conferences, workshops
- Send reps to conferences, security meetings, *nogs, report back on what data is used or articulated as needed

2) It's all about Transparency (!)

- Open up data catalog as much as possible (now done?)
- Publish older/less risky data sets
- Link data sets to papers/results/blogs derived from them
- Publish PI meeting slides
- Publish usage statistics
- Publish all MOAs/DUAs/IRB applications (or templates)
  - default to transparency: how are people using data?

3) Learn from previous failures (and document how)

- Document what worked and what didn't
  - Acknowledge previous failures and changes
    - [aside: great example ``Rise and Fall of BIND10'' https://ripe68.ripe.net/archives/video/153/]
  - Invite successful data-sharers to speak at meetings
    - e.g., Farsight, Protein Data Bank, Crawdad, OARC
  - Invite struggling data-sharers to speak at meetings
    - e.g., Internet2

4) Leverage investment by other agencies and organizations

- Leverage funding and energy and vision
  - NSF (e.g., data mgt plan)
  - IARPA
  - FCC (MBA)
  - Google (M-Lab)
  - Farsight
  - RIPE Atlas
  - BEREC, e.g., http://berec.europa.eu/eng/news_consultations/ongoing_public_consultations/2098-public-consultations-on-the-draft-berec-report-on-monitoring-quality-of-internet-access-services-in-the-context-of-net-neutrality
- Reuse components from previous projects (e.g., DatCat)
  - “15-minute rule” (for indexing data)

5) Tech Transfer from Industry into Research Community

- matchmaking, e.g., eharmony
- data product rating system, e.g., amaz, yelp
- reviewer rating system, e.g., amaz,yelp
- data brokering, e.g., amaz marketplace
  - standard procedures for acquiring datasets from federation of providers
- tiered access, e.g., amaz prime
- open source tools for data analysis

6) Leverage publishing industry

- Work with academic and science publishers
  - Stamp of approval on journals that require data in papers to be indexed
  - New journal where all papers link to available data
  - i.e., contribute to conditions that will foster a culture shift toward reproducibility and transparency

7) Internal transparency and mgt

- Copious meeting minutes and tracking
- Action tracking via group wiki (and annual analysis of how long things take to get done)
- Track how many times the same conversation happens (and what new comes out of each conversation)
- (again) Publish PI meeting materials, at least internally (and to predict-users)

8) Measure for success

- Integrate issue resolution and metrics of success into contractual SoWs (~done)
- Some metrics are quantitative (data sets, disk space, accounts, papers, students)
- Also need qualitative evaluation of research enabled
  - cybersecurity impacts
  - use by other agencies
  - success stories
- Publish research results in digestible form (inc. infoviz)

9) Catalyzing Goal: Data Set of Year

- Create longitudinal crowd-sourced data set that illuminates an aspect of cyberspace; visualize it over time and space
  - bcp38 compliance
  - open resolver trends
  - mobile performance
  - topology
- Add new such data every year
- Joint venture with NSF and NIST (measurement science)

10) Prepare for rewriting telecom act

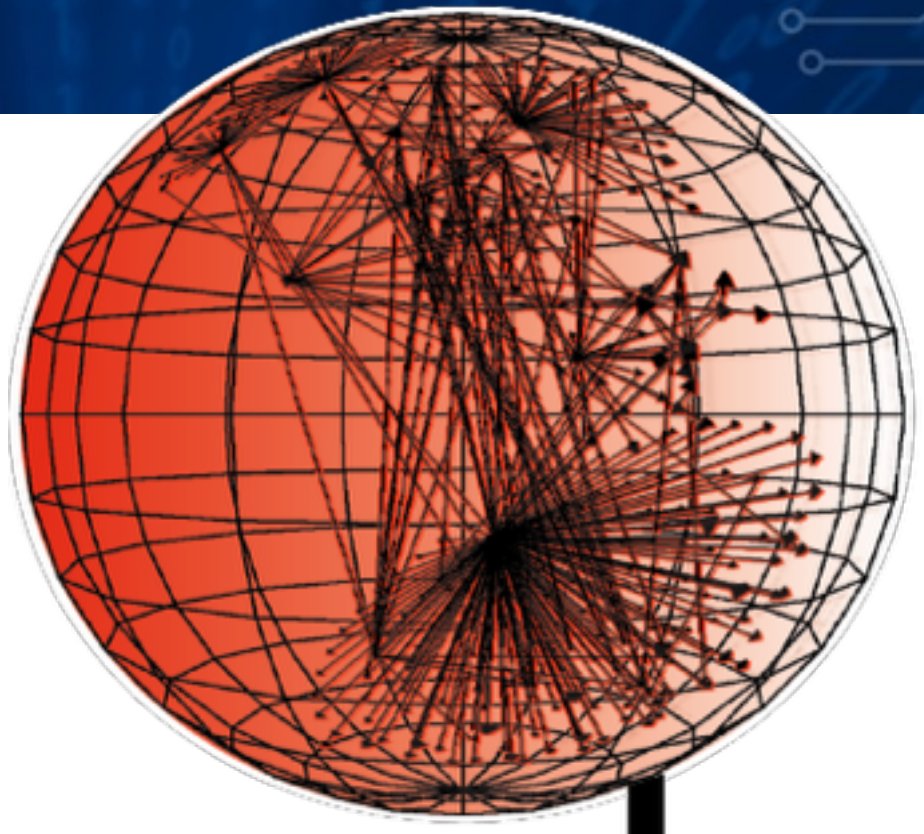
- Among other legislation
- New legislation will happen with or without us
- Allocate real communications policy research group to support public interest in policy evolution
- Add to feedback loop to improve data in PREDICT
- Interagency WG with FCC, FTC, NIST, DHS
- “NITRD” for policy R&D
- Publish in law journals

Contact Information

k claffy

kc@caida.org

http://www.caida.org/



caida

www.caida.org