2017 Cyber Security R&D Showcase and Technical Workshop

July 11 - 13, 2017 I Washington, D.C.

Homeland Security

Science and Technology

Supporting Research and Development of Security Technologies Through Network and Security Data Collection

k claffy / CAIDA/UCSD

13 july 2017

Homeland Security

Science and Technology

Need

Today the "cyber threat" is one of our most serious economic and national security challenges.

But we lack understanding of the structure, dynamics, and vulnerabilities of the global Internet.

Measurement infrastructures, reliable, representative, Internet data sets, and advanced analysis tools are rarely available to researchers and developers.

Need

Researchers require current data on Internet security and stability to study threats:

1. normal and malicious Internet traffic samples
2. baseline topology data
3. real-time changes to topology (instability, hijacks)
4. reachability data
5. compromised hosts
6. observed security vulnerabilities
7. …

Biggest Challenges

Security faces same obstacles as Internet science:

- cost of technology deployment and maintenance
- radically distributed ownership of constituent parts
- operational climate, including legal, privacy, and competitive concerns, that disincents sharing data for all stakeholders:
  - owners of infrastructure
  - owners of data
  - collectors of data
  - distributors of data

Approach

The Department of Homeland Security (DHS) has developed the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) project to provide vetted researchers with current network operational data in a secure and controlled manner that respects the security, privacy, legal, and economic concerns of Internet users and network operators

Approach: IMPACT

Three primary goals of IMPACT are:

1. Develop, implement, and maintain a Web-based portal that catalogs current computer network and operational data and handles data requests.
2. Enable secure access to multiple sources of data collected on the Internet.
3. Facilitate data sharing among IMPACT participants for the purpose of developing new models, technologies and products increasing cyber security capabilities.

More information about IMPACT is available in the [Overview of the IMPACT program](). ([www.impactcybertrust.org/]())

CAIDA role in IMPACT: Enabling R&D

Data activities

- Ongoing measurements
- Data curation and archiving
- Serving data to IMPACT community

Research and Infrastructure activities

- Data collection instruments and infrastructure
- Data analysis methodologies
- Tools for interactive access to archives of data
- Related research activities

Data Collection Infrastructures

Enabling provision of unique, critical data for R&D

- **Ark Platform (as of 20 June 2017)**
  - 181 monitors in 60 countries
  - 83 IPv6-enabled in 33 countries
  - 135 Raspberry Pis
- **UCSD Network Telescope**
  - As of June 2017, >1TB/day compressed traffic trace data
  - 33 TB: last full month (May 2017)
  - 323 TB: 2016 (182 TB: 2015)
  - 195 TB: YTD 2017 (as of 6/20/17)
  - 382 TB: last 12 months at NERSC (as of 6/20/17)
  - 998 TB: total archived at NERSC

Ark Platform Data Sets

Publicly Available Data Sets (Most recent 2 years in DHS IMPACT)

- IPv4 Routed /24 Topology, and associated DNS Names Dataset
- IPv4 Prefix Probing Dataset (finer grained temporally)
- Internet Topology Data Kits (ITDK)
- IPv6 Topology and associated DNS Names Dataset
- IPv4 Routed /24 AS Links (September 2007 - ongoing)
- IPv6 AS Links (December 2008 - ongoing)
- AS Relationships (also via as-rank.caida.org)
- AS Classification
- AS to Organization
- AS Facilities Map (**New**)
- Border Mapping Dataset (**New**)
- IPv4 Prefix-Probing Dataset (**New**)

Network Telescope Data Sets

Publicly Available Datasets

- UCSD Real-time Network Telescope Data
- UCSD Telescope Darknet Scanners Dataset
- UCSD Telescope Sipscan Dataset
- Backscatter
- Code Red Dataset
- Patch Tuesday Dataset
- Three Days of Conficker Dataset
- Two Days in 2008 Telescope Dataset
- Witty Worm Dataset

Available Through IMPACT (vetted)

- near real-time darknet traffic data

Recently Added Datasets

Internet maps updated 2/x yr

- Macroscopic Internet Topology Data Kit (ITDK) (http://www.caida.org/data/internet-topology-data-kit/
- IPv4 2013 Census Dataset http://www.caida.org/data/active/ipv4_2013_census_dataset.xml

Survey of active hosts

- UCSD Network Telescope -- Darknet Scanners Dataset http://www.caida.org/data/passive/telescope-darknet-scanners_dataset.xml

Identifying compromised hosts

- AS Border Mapping Dataset http://www.caida.org/data/active/bdrmap_dataset.xml

Identifying transit borders

Probing all routed prefixes

- Prefix Probing http://www.caida.org/data/active/ipv4_prefix_probing_dataset.xml

---

- Spoofer data (coming soon, at https://spoofer.caida.org)
- AS-to-Facilities Mapping (coming soon)

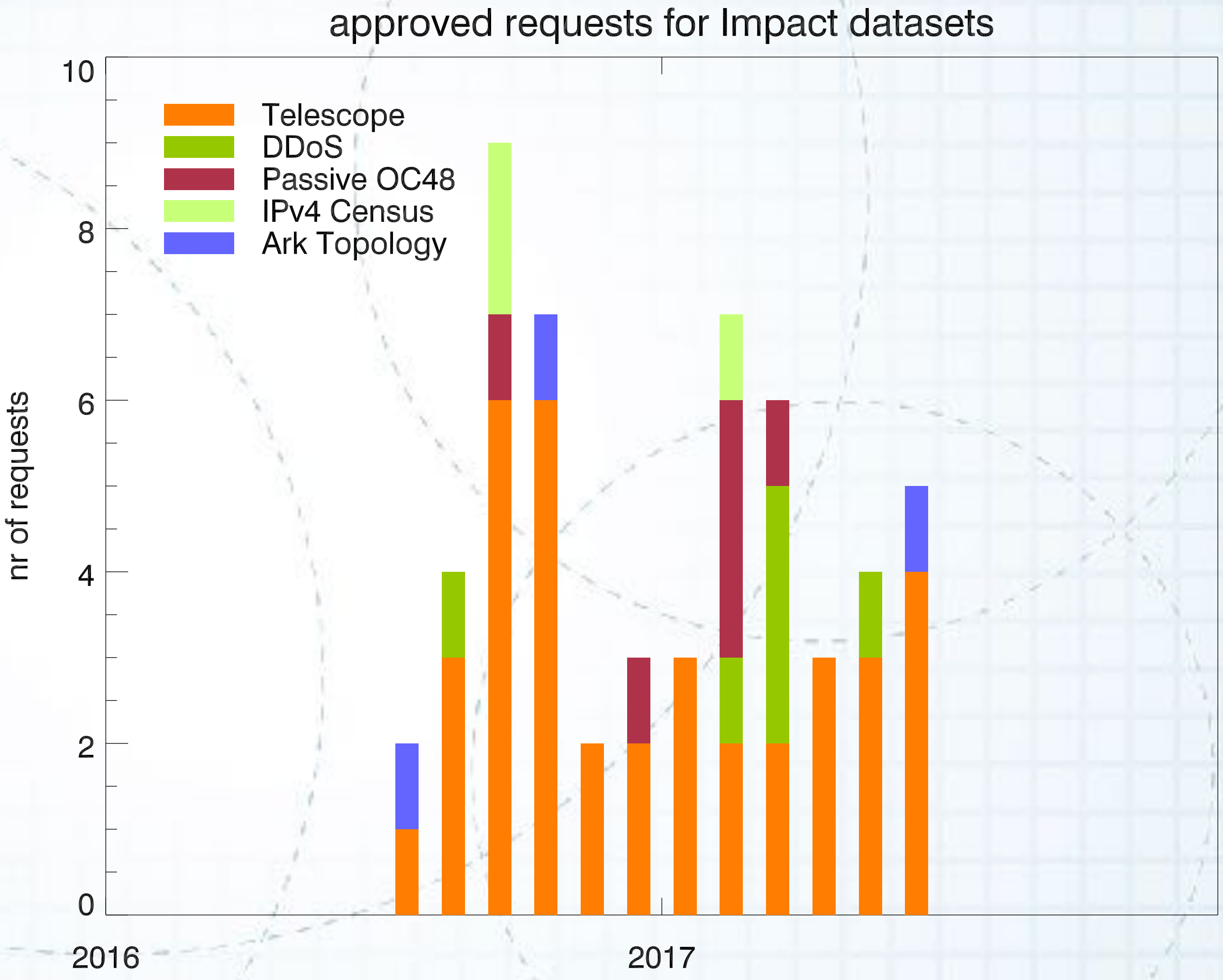
Metrics of success

data-announce mailing list **2,479 members**

Dataset Category	Publications
Denial of Service Attacks	46
Topology with Ark	109
Topology with Skitter	97
UCSD Network Telescope	103

(Up to mid-2016; more publication searching to do…)

Metrics of success



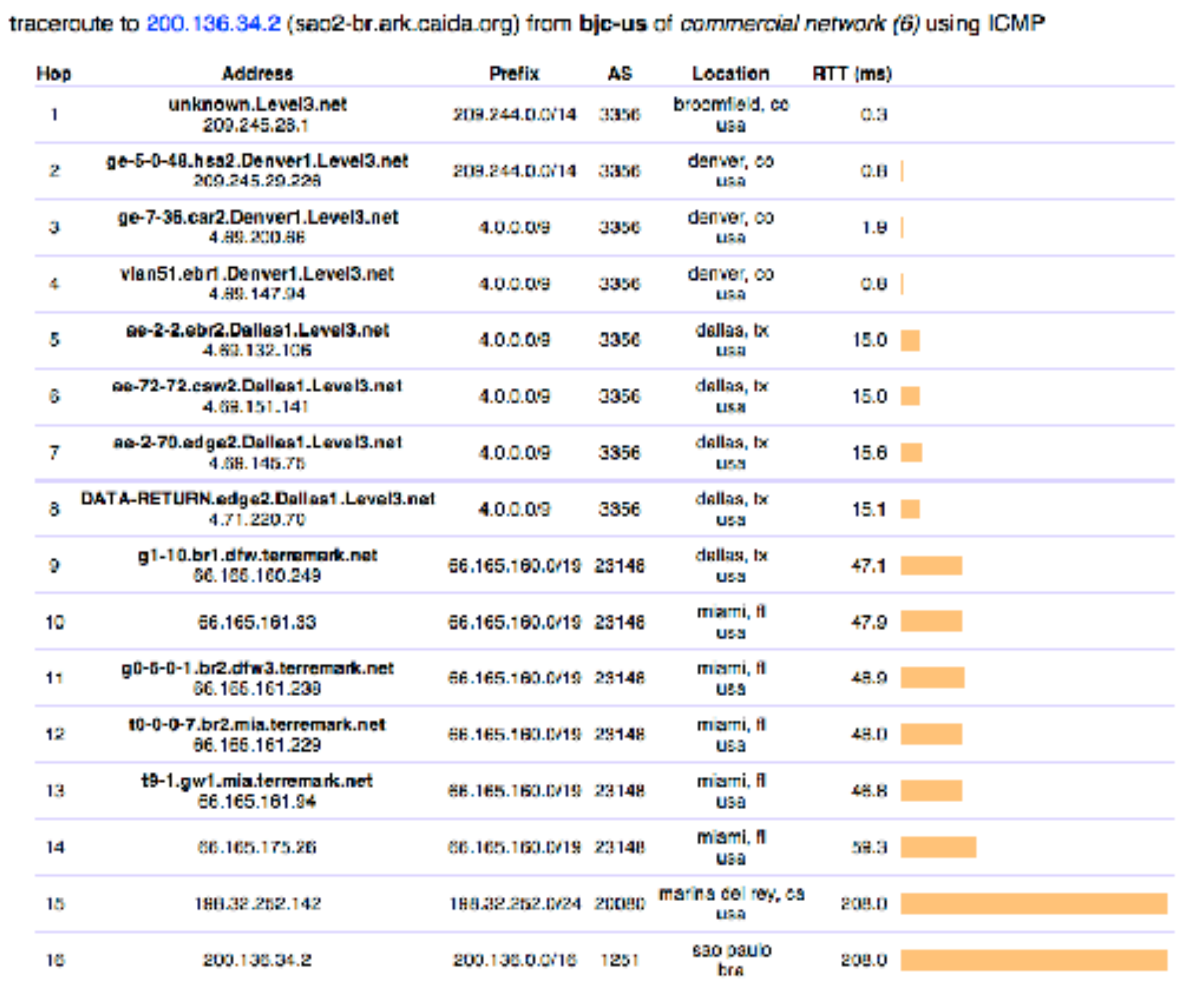
55 approved requests through IMPACT portal (July 2016 - June 2017)

New Tools

- **AS Rank** https://as-rank.caida.org
- ***Vela*** *https://vela.caida.org*
- ***Henya*** *https://www.caida.org/tools/utilities/henya/*
- **Spoofer** https://spoofer.caida.org

Will drill down on these two

GUI access to measurement infrastructure



traceroute to 200.136.34.2 (sao2-br.ark.caida.org) from **bjc-us** of *commercial network (6)* using ICMP

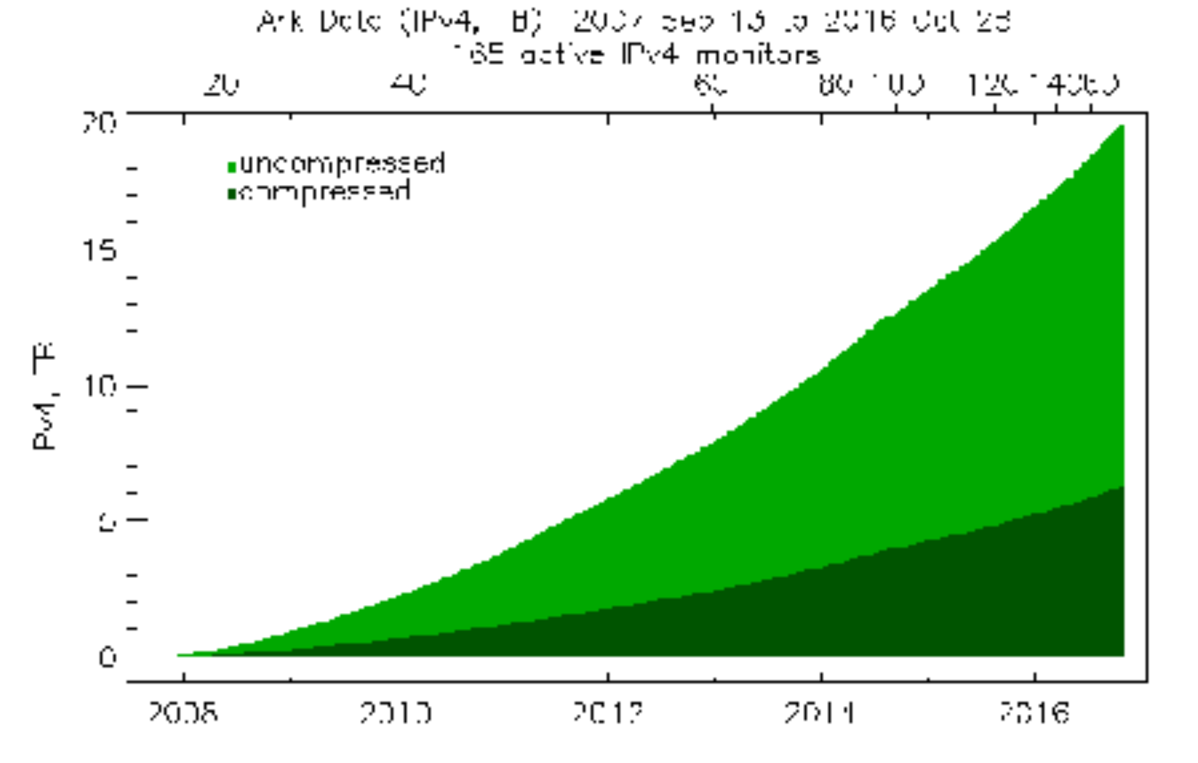
Hop	Address	Prefix	AS	Location	RTT (ms)
1	unknown.Level3.net 209.245.28.1	209.244.0.0/14	3356	broomfield, co usa	0.3
2	ge-5-0-48.hsa2.Denver1.Level3.net 209.245.29.226	209.244.0.0/14	3356	denver, co usa	0.8
3	ge-7-36.car2.Denver1.Level3.net 4.69.200.66	4.0.0.0/8	3356	denver, co usa	1.9
4	vlan51.ebr1.Denver1.Level3.net 4.69.147.94	4.0.0.0/8	3356	denver, co usa	0.8
5	ae-2-2.ebr2.Dallas1.Level3.net 4.69.132.106	4.0.0.0/8	3356	dallas, tx usa	15.0
6	ae-72-72.csw2.Dallas1.Level3.net 4.69.151.141	4.0.0.0/8	3356	dallas, tx usa	15.0
7	ae-2-70.edge2.Dallas1.Level3.net 4.69.145.75	4.0.0.0/8	3356	dallas, tx usa	15.6
8	DATA-RETURN.edge2.Dallas1.Level3.net 4.71.220.70	4.0.0.0/8	3356	dallas, tx usa	15.1
9	g1-10.br1.dfw.terremark.net 66.165.160.249	66.165.160.0/19	23148	dallas, tx usa	47.1
10	66.165.161.33	66.165.160.0/19	23148	miami, fl usa	47.9
11	g0-5-0-1.br2.dfw3.terremark.net 66.165.161.238	66.165.160.0/19	23148	miami, fl usa	48.9
12	t0-0-0-7.br2.mia.terremark.net 66.165.161.229	66.165.160.0/19	23148	miami, fl usa	48.0
13	t9-1.gw1.mia.terremark.net 66.165.161.94	66.165.160.0/19	23148	miami, fl usa	48.8
14	66.165.175.26	66.165.160.0/19	23148	miami, fl usa	58.3
15	198.32.252.142	198.32.252.0/24	20080	marina del rey, ca usa	208.0
16	200.136.34.2	200.136.0.0/16	1251	sao paulo bra	208.0

Vela: On-Demand Topology Measurement Service of CAIDA’s Ark infrastructure

- web interface
  https://vela.caida.org/
- command-line interface

GUI to explore longitudinal data

Henya: Large-Scale Internet Topology Query System



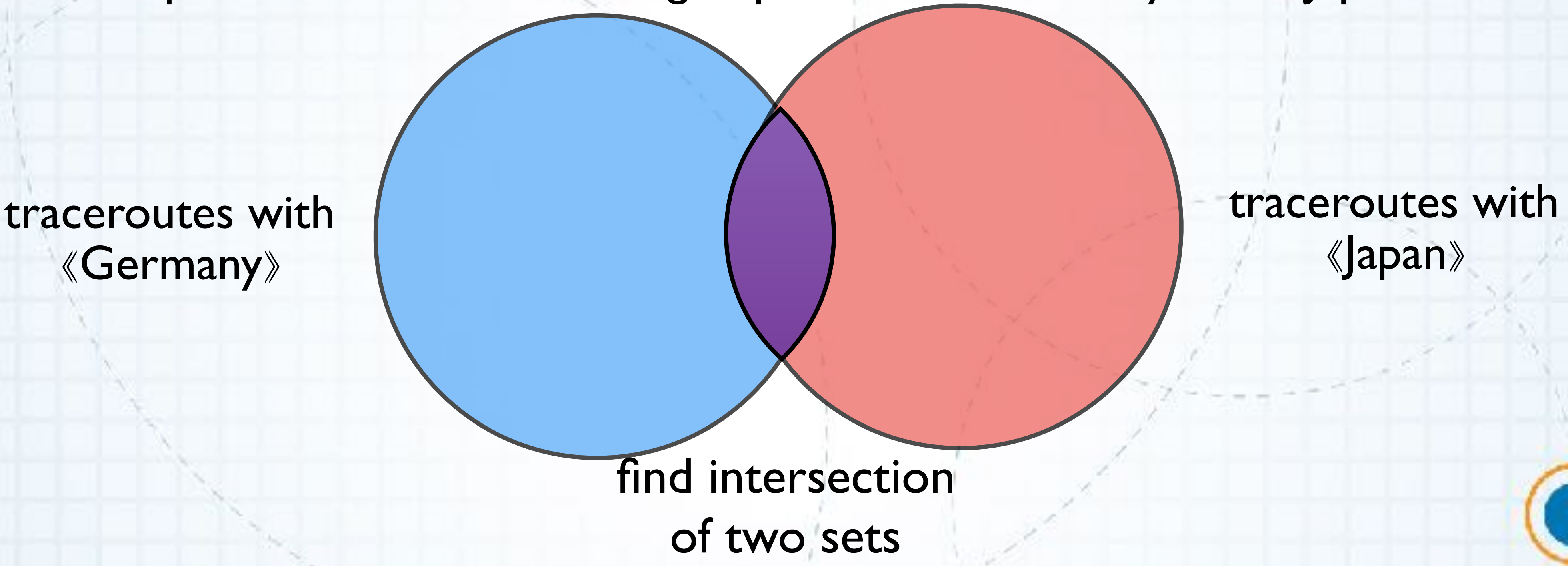
- Access via (same) Vela web interface https://vela.caida.org/
- 9 years of “Routed /24” trace routes
  - 47 billion traces in 20TB of files
  - growing yearly by 10B traces
- 1 year of “Prefix Probing” trace routes
  - growing yearly by 9B traces

Henya Topology Queries

- find occurrences of traceroute path elements
- 《targets》 = IP addreses, prefixes, ASes, or countries
- Queries:
    - traceroutes toward 《targets》
    - traceroutes containing one or more 《targets》
- Parameters:
    - measurement vantage points
    - data collection time periods
    - position of 《targets》 in path
    - hop distance between sets of 《targets》

Henya Topology Query Complexity

- the most complex case:
  - traceroutes containing **two or more** 《*targets*》
  - example: traceroutes containing hops in both 《Germany》 and 《Japan》



- harder: traceroutes with hops in 《Germany or UK or France》 and hops in 《ATT or Level3 network》 and hops in 《Amsterdam Internet Exchange》

Henya Topology Query: Challenges

- large target sets
  - 《Germany》 = 9,906 BGP prefixes = **92M** target IP addresses
  - 《Japan》 = 8,769 BGP prefixes = **154M** target IP addresses
- multiple 《*targets*》 in a single query
  - need the **intersection** of subqueries for 《*targets*$_1$》 and 《*targets*$_2$》 and ...

- these challenges ***poorly met by existing database systems***
  - relational databases not designed/optimized for multi-key searches
    - can't always use column indexes; may need to do table scans on separate columns
  - not a good fit for existing NoSQL databases
    - schema-less document stores (JSON/XML) come the closest

Henya Instructional Video

http://www.caida.org/tools/utilities/henya

(or search “caida henya”)

Vela and Henya Access Policies

- Currently accepting requests for accounts on Vela
- Currently accepting requests for early access to Henya and a subset of total topology dataset.

Benefits of CAIDA’s participation in IMPACT

- support empirically-grounded cybersecurity research
- longitudinal data sets of critical infrastructure
- access to unique data sets and tool chains
- available community of experts to consult with questions
- interactive tools to facilitate exploration of topology data
- enable validation across multiple heterogeneous data

k claffy
CAIDA/UCSD
kc@caida.org
858-534-8333
twitter:@caidaorg

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

caida

UC San Diego

2017 Cyber Security R&D Showcase and Technical Workshop

July 11 - 13, 2017 | Washington, D.C.

Homeland Security

Science and Technology