Designing a Global Measurement Infrastructure to Improve Internet Security

How to get and use real data sets about Internet infrastructure for our ML/AI models

(a little of both talks)

10 Feb 2025

kc claffy, CAIDA/UCSD

caida

CNS-2120399, OAC-2319959, OAC-2131987

Goal of Talk

Understand State of Current Available-for-Research Data about the Global Internet Infrastructure

- Data Types: Active, Routing, Traffic, DNS, Metadata
  - Current State and Operational Challenges
- Data & Tools useful to networking and/or AI researchers
- Invite collaboration on measurement capabilities

Why Should Internet Infrastructure Data Play a Key Role in AI?

First, let us ask ourselves what is the most fundamental difference in this biggest-ever wave of AI?

COMMUNICATIONS OF THE ACM

Explore Topics | Latest Issue | Sign In | Join ACM

OPINION

Artificial Intelligence and Machine Learning

Can Machines Be in Language?

Large language models brought language to machines. Machines are not up to the challenge.

By Peter J. Denning and B. Scot Rousse

Posted Feb 22 2024

https://cacm.acm.org/opinion/can-machines-be-in-language/

"*First, the core ANN is* ***trained on*** *billions of words of text from* ***the Internet*** *to respond to a prompt with a list of most probable next words after the prompt.*"

GPUs may be the gold of AI, but the net is oxygen!

"***GPUs*** *have been called the rare Earth metals — even the* ***gold*** *— of artificial intelligence, because they're foundational* ***for today's generative AI era.***"

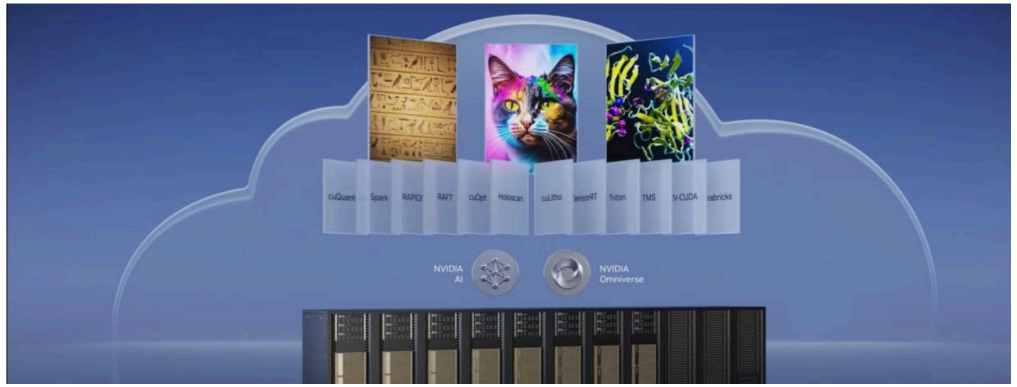
NVIDIA

Home AI Data Center Driving Gaming Pro Graphics Robotics Healthcare Startups

Why GPUs Are Great for AI

Features in chips, systems and software make NVIDIA GPUs ideal for machine learning with performance and efficiency enjoyed by millions.

December 4, 2023 by Rick Merritt



Ask the AI itself:

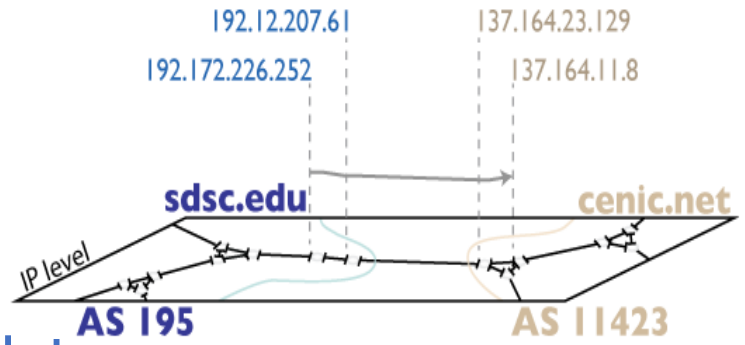
"… if GPUs are the "gold" providing the raw material wealth or the computational power that drives AI, then ***Oxygen*** *(O) symbolizes the* ***fundamental, life-sustaining environment—the Internet—that AI*** *technologies* ***require*** *to function, connect, and flourish."* GPT-4 (after picking `silicon')

The Foundation of … Everything

- And yet… we have **little scientific understanding** of Internet **structure, dynamics, vulnerabilities**
  - Internet security is one of most serious social, economic, and political challenges
  - Yet, … not a discipline

IP level path (traceroute)

```
traceroute to 137.164.11.8 (137.164.11.8), 64 hops max, 52 byte packets
 1  sdsc-rtr                  (192.172.226.252)  13.079 ms  0.285 ms  15.696 ms
 2  mx0-ae7--thor-ae0.sdsc.edu  ( 192.12.207.61)   0.399 ms  0.398 ms   0.361 ms
 3  dc-sdg-agg4--sdsc-1.cenic.net ( 137.164.23.129)   0.901 ms  0.892 ms   0.917 ms
 4  dc-tus-3-agg4-100ge.cenic.net (  137.164.11.8)   2.535 ms  2.503 ms   2.592 ms
```



- But we can (and do) apply decades of expertise to label data
- We can acquire data sets that are global, open, (labeled), relevant!

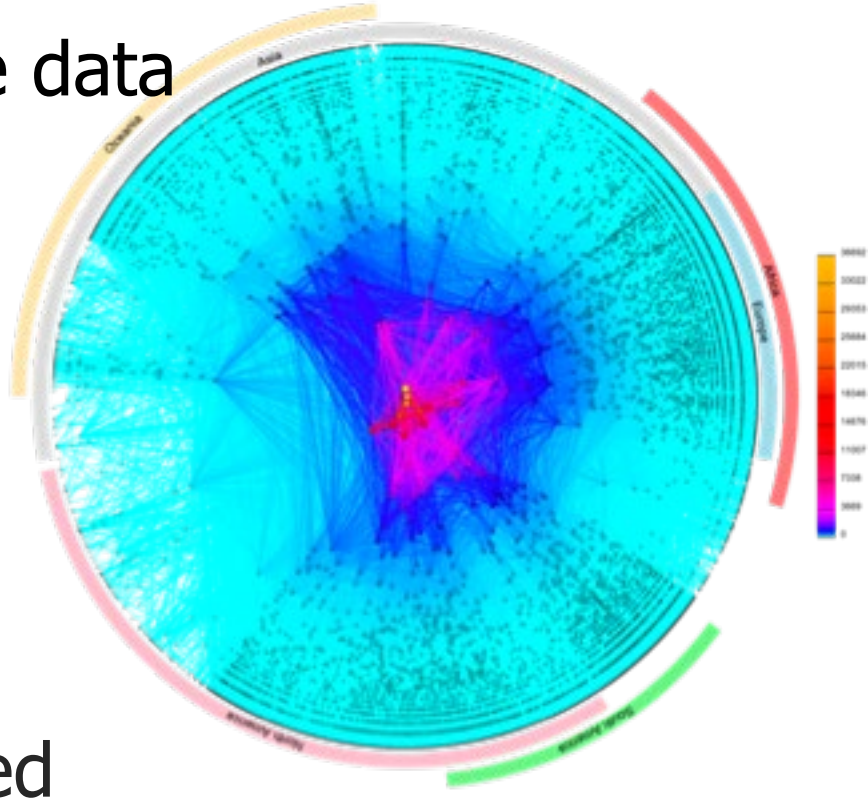
- Compatible w/ national priority to train STEM students
- **Many Internet infrastructure problems are ripe for transformative AI solutions**

NSF 24-553: Improving Undergraduate STEM Education: Computing in Undergraduate Education

BUT.. we need Better and Sustained Data About the Internet

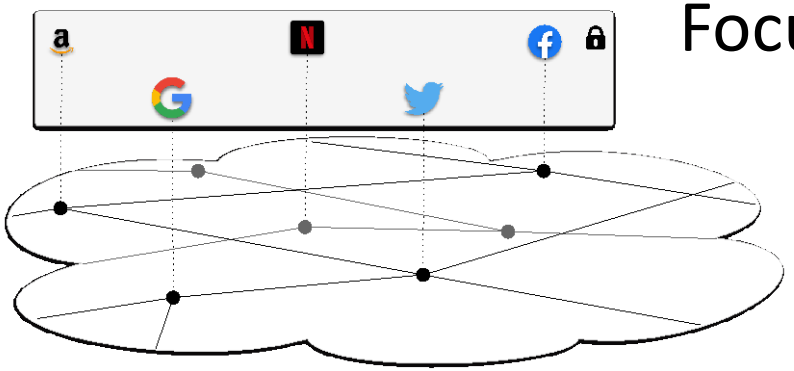
Gaps to transformative AI advances: reduce barriers!

- Even better-labeled and more comprehensive data
- How to find the data
- How to interpret and validate data
- Open LLMs that researchers can train at reasonable cost



Measurement infrastructures, reliable, representative, Internet data sets, and advanced analysis tools are challenging (expensive) to create and sustain

Why Internet Infrastructure Remains Vulnerable



Focus of commercial security: protect the enterprise

The enterprise is the customer.

Who protects the Internet itself?

(Hint: compare it to other critical infrastructures)

Every core transport service has serious, well-known vulnerabilities.

routing, naming (DNS), addressing, encryption key management

High-impact security research requires data-focused infrastructure.

decades of attempts to retrofit security failed to gain traction

now need scientific research that leverages public global measurement

Data Collection Challenges

cost, complexity, misaligned incentives, privacy (PII), commercial sensitivities

Critical Security Flaws At All Layers of IP Transport

Internet Addressing System (IP)

Lack of authentication of source address (Spoofing), DDoS

Internet Routing System (BGP)

Lack of authentication of announcements

Internet Domain Naming System (DNS)

Lack of authentication of IP->hostname mapping (DNSSEC available but complex),Misconfiguration, complexity creates attack surfaces

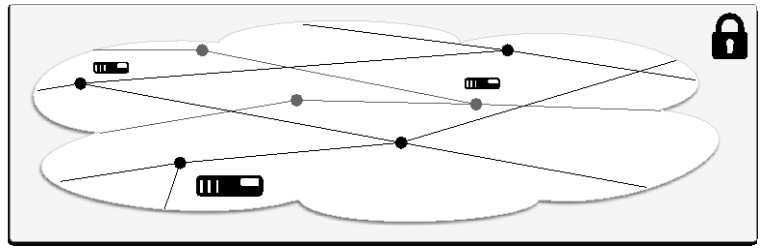
Internet Certificate Authority (CA) System

False certificates

[It’s almost like it was architected to facilitate propagation of misinformation..]

Approach: Global Infrastructure to Advance Security Research

Goal: **Secure the Internet***—a national priority



Approach: **Gather critical data**†—a missing link

Proposed effort: **Design a global measurement infrastructure**†

- Support a range of data collection and research experiments
- Community-driven design

Improve competitiveness of U.S. security research community*

Put data collection, curation and use on **sustainable footing**†

- Framework to handle sensitive data, advance ML/AI methods
- Reduce burden on individual researchers

Improve STEM training, reduce barriers to diversity of participation*

MSRI: Design: GMI3S, OAC-2131987

*Broader Impact
†Intellectual Merit

GMI Design Project: Focal Points

Driven by Internet vulnerabilities and mitigation strategies

"Data needs for Security Internet Infrastructure" Report

https://gmi3s.caida.org/outcomes/documents/vulnerabilities-harms-dataneeds_v2.4.pdf

GMI Design Project: Focal Points

Focused on the following types of data / infrastructure

- Active Measurements
- Interdomain Routing (BGP) Data
- Passive Traffic Capture (unsolicited and two-way/“real”)
- DNS Data (Active and Passive)
- Traffic: One-way (background) and Two-way packet capture
- Supporting infrastructure for data management
- Repeatable practices (AUPs) for access to industry
- Designs for data-driven Internet routing security

Spectrum of Cooperation Needed for Data Collection

more cooperation

more cooperation

Active
- one VP
- many VPs

Routing/BGP
- one VP
- many VPs

Traffic
- one darknet
- one network

DNS/Meta-data
- Ownership of domain names, addresses, networks
- geolocation

many networks

[For each we could discuss: why, state-of-art, challenges, proposed approaches, how we can help, how you can help]

CAIDA Active Measurement Infrastructure: Archipelago

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

ARCHIPELAGO

- **Programmable platform** for global active Internet measurement experiments.
- Teams of distributed nodes
- Ongoing measurements since 2007
  - Longitudinal global topology data
- **On demand measurements**
- **Navigating trade-off** between VP scale and capability

Spectrum of Cooperation

Most Restrictive (Least cooperation)

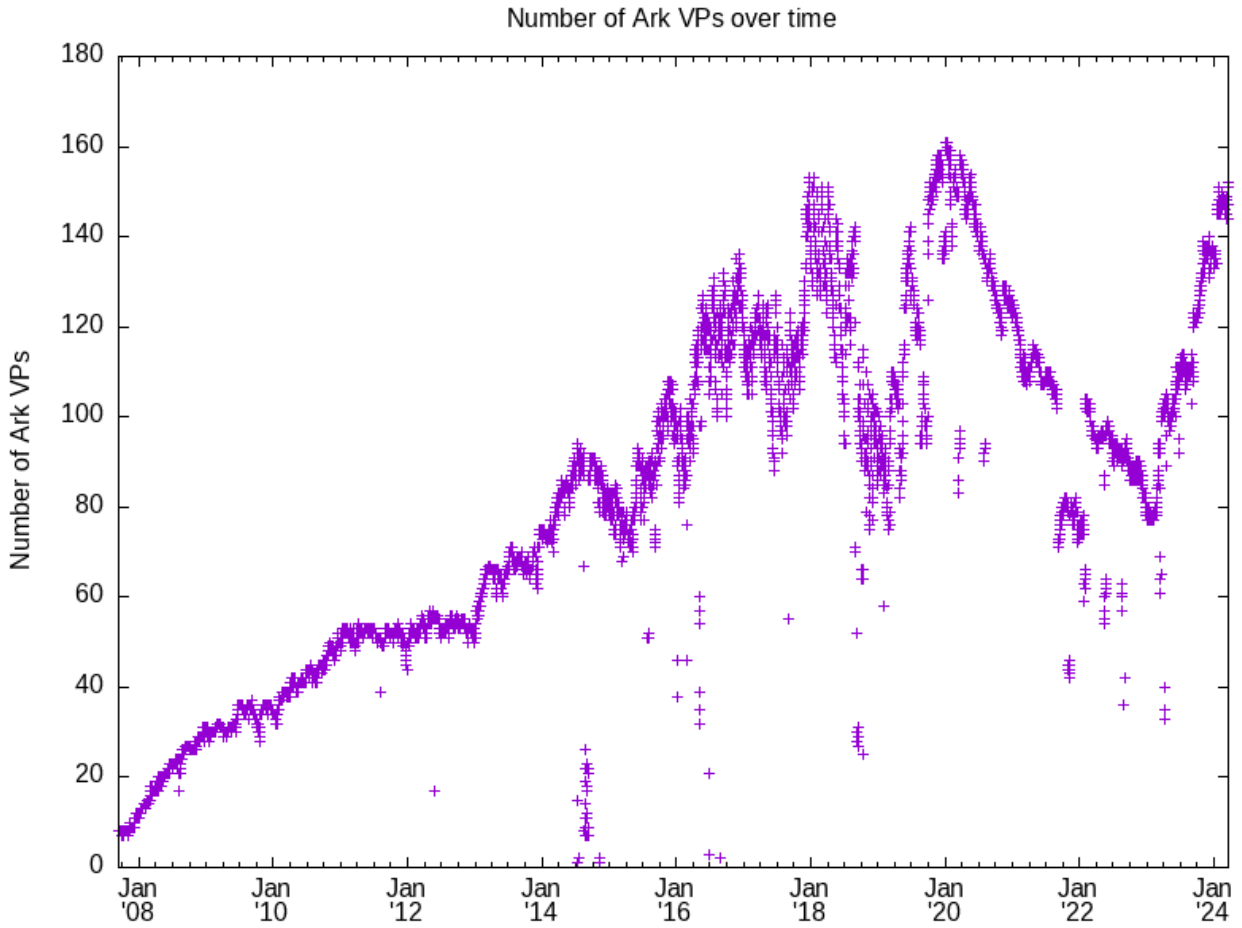
- No access, just use provided data
- API to run tests, send packets, do logic elsewhere / no logic
- VPN access to send packets from probe, do logic elsewhere/no logic
- Domain specific language (DSL) to run tests, send packets, do logic
- Run measurement code in a container on the probe
- Full shell access on the probe

Least Restrictive (Most Cooperation)

Infrastructure Operational Challenges

- Incentivizing people to host VPs
- Automation (DevOps)
- Hardware and software system administration
- Scalable handling of per-VP idiosyncrasies interfering with measurements

Can you see two senior personnel retirements in this graph?



Software Innovation: Providing Programmability

Domain-Specific Language (DSL)

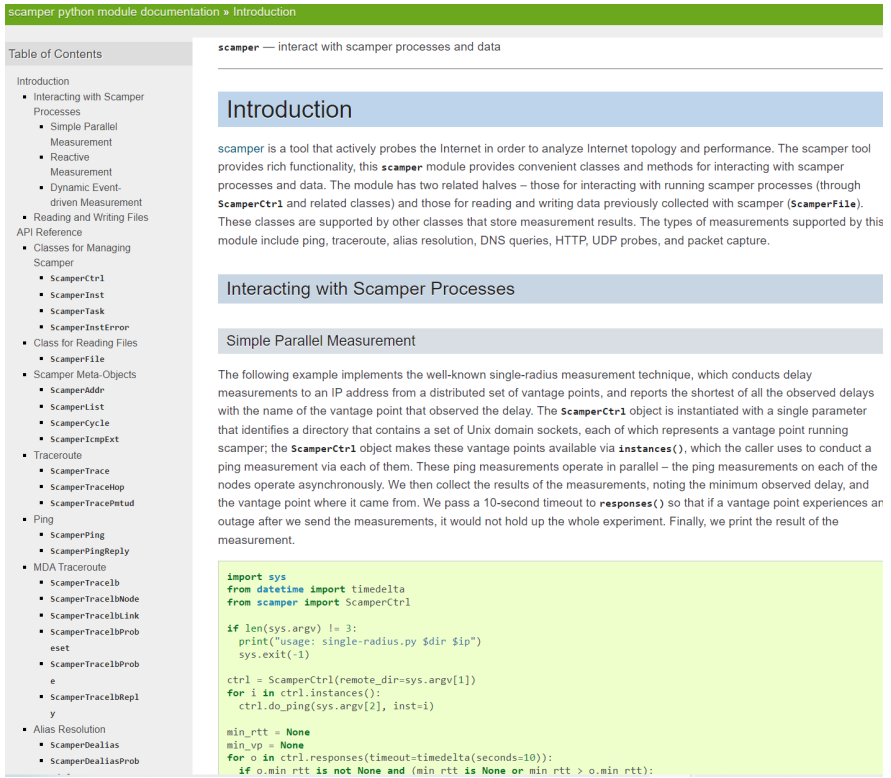
- *Python API to **Scamper*** software on vantage points (VPs)
- Enables parallel measurement from a central location using VPs all over the world.

DSL Functionality

- Rich set of measurement/analysis primitives
- Interfaces to interact with measurement processes

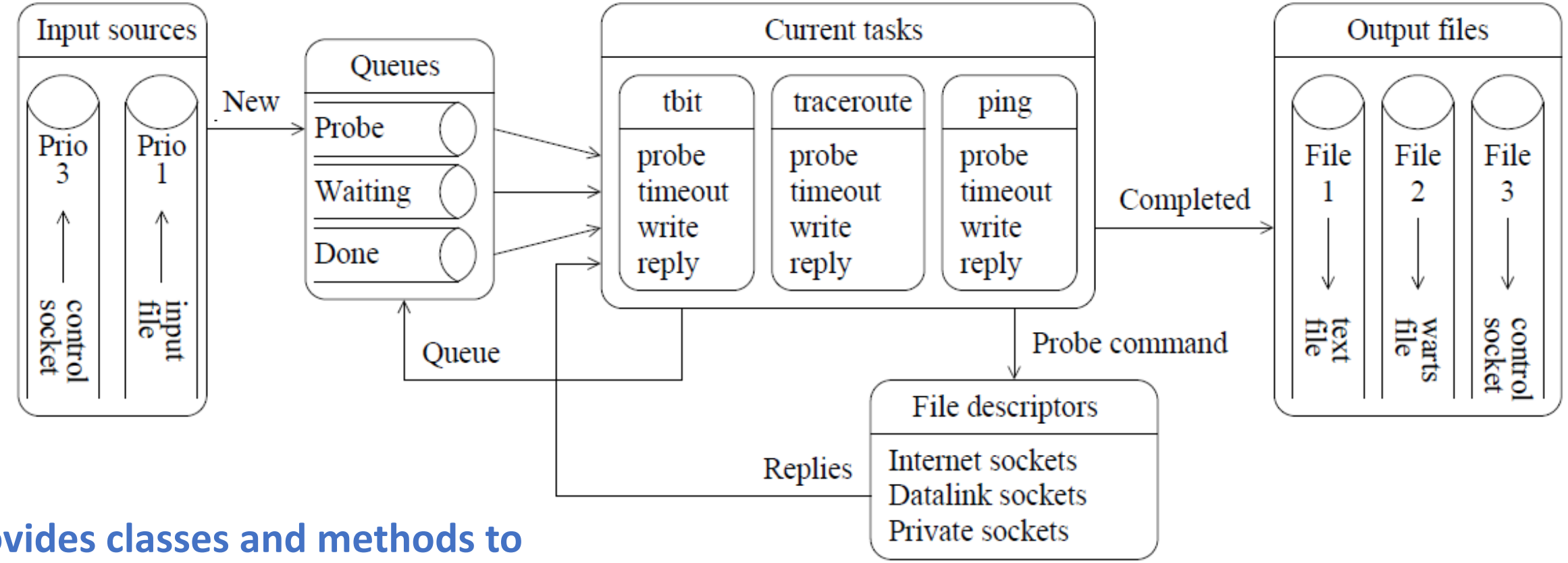
Supported Measurements: Ping, Traceroute, DNS

Queries, HTTP, UDP Probes, Packet Capture, Alias Resolution (which IP addresses belong to same router), ….



https://www.caida.org/catalog/software/scamper/python/

Domain-Specific Language: Concept



Provides classes and methods to

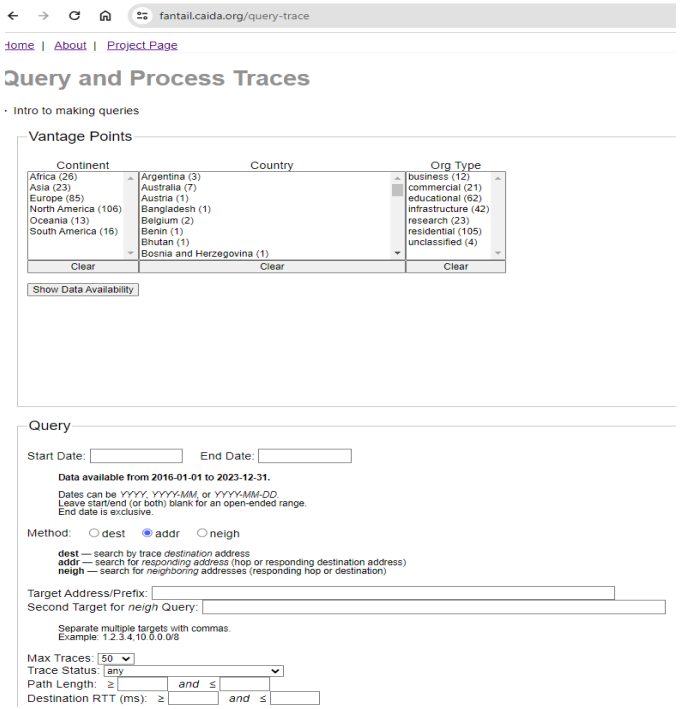
1) **interact with running processes**
2) **Read and write previously collected data**

https://blog.caida.org/best_available_data/2024/01/16/towards-a-domain-specific-language-for-internet-active-measurement/

Other Tools and Data to Support Research

FANTAIL

Comprehensive topology query system designed to search vast archives of raw Internet traceroute data



https://fantail.caida.org/

PERISCOPE

Uniform interface to hundreds of servers with access to thousands of network probing vantage points (monitors) that perform traceroute and BGP queries.
https://www.caida.org/catalog/software/looking-glass-api/
https://www.caida.org/catalog/software/accounts/periscope_request
(Requires authorization with CAIDA's SSO system https://www.caida.org/about/sso/)

DATASETS

https://catalog.caida.org/collection/archipelago

Archipelago (Ark)
CAIDA deploys and maintains a globally distributed measurement platform we call Archipelago (Ark). We grow the infrastructure by...

Ark IPv4 Routed /24 Topology
These are all the Ark IPv4 team-probing data, collected by a globally distributed set of Archipelago (Ark) monitors. IPv4 Routed /24 Topology...

Ark IPv4 Routed /24 AS Links
Data from the IPv4 Routed /24 Topology Dataset are processed by using RouteViews BGP data to identify the Autonomous System (AS) associate...

Ark IPv4 Routed /24 DNS Names
The IPv4 Routed /24 DNS Names Dataset provides fully-qualified domain names for IP addresses seen in the traces of the IPv4 Routed /24 Topolog...

ITDK: Internet Topology Data Kit
Ark-based macroscopic Internet Topology Data Kits (ITDK)

Ark IPv6 Topology
These are all the Ark IPv6 probing data, collected by a globally distributed set of IPv6-enabled Archipelago (Ark) monitors. These data contain...

Ark IPv6 Routed /48 Topology
This dataset contains information useful for studying the IP- and AS-topology of the IPv6 Internet. The goal of these measurements is to...

Ark IPv6 Topology AS Links
Data from the IPv6 Topology Dataset are processed by using RouteViews BGP data to identify the Autonomous System (AS) associated with each...

Ark IPv6 Topology DNS Names
The IPv6 DNS Names Dataset provides fully-qualified domain names for IP addresses seen in the traces of the IPv6 Topology Dataset

How You Can Help

Deploy VPs!

- Hardware (e.g. Raspberry Pi) or Software
- “Archipelago Memorandum of Cooperation Between Hosting Sites and CAIDA” https://www.caida.org/projects/ark/moc/
- Read “Why should my network host an Ark node?”
- Note: We are in a (re)design phase: not all automated!

ARCHIPELAGO
Internet Measurement Infrastructure
http://www.caida.org/projects/ark/

caida
www.caida.org

Why should my network host an Ark Node?

When your network hosts a measurement node that participates in CAIDA's Archipelago (Ark) infrastructure, it broadens the view of the global Internet for the network research community. Network researchers use CAIDA topology data to conceive, develop, and test their models and methods. Participating networks agree to our Memorandum of Cooperation (MoC) [1] and install a Raspberry Pi, 1-U server, virtual server, or software container dedicated to Ark measurement.

Once deployed, the Ark node conducts continuous measurements of the routed IPv4 (and IPv6 when the hosting network supports it) address space. Ark aggregates the resulting data on a server at UC San Diego’s Supercomputer Center. Each additional node contributes to – and increases the completeness and accuracy of -- data representing the topological structure of the Internet core.

CAIDA uses these continuous measurements, as well as sophisticated Internet-scale alias resolution methods developed in-house, to build the Internet Topology Data Kit (ITDK) – a heavily annotated router-level graph of the Internet to support data science on Internet topology. CAIDA further annotates each router with its inferred geolocation, and inferred operator, to support sophisticated analyses of the router-level Internet topology by the Internet research community.

As of 2024, the Ark measurement platform has supported the Internet research community for more than 15 years. Recent work has investigated the unintended consequences of submarine cable deployment on Internet routing (PAM'18), persistent interdomain congestion (SIGCOMM'18), the impact on performance and resilience of regional access network topology structure (IMC'21), and methods to automatically extract meaning from Internet router hostnames (CoNEXT'21, IMC'20, IMC'19). As of 2024, CAIDA is enhancing the infrastructure with a domain-specific language to allow researchers to quickly and correctly build and execute experiments.

This range of scientific experiments has successfully demonstrated our vision of a metaphorical distributed measurement “operating system” to support empirical Internet science.

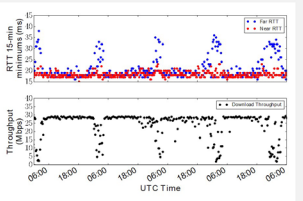
[1] https://www.caida.org/projects/ark/moc/

If interested, please contact
ark-info@caida.org

Figure 1: North America Paths routed through Africa to Brazil

Paths taken from North American Ark nodes to IP address in Brazil that transited Africa after deployment of the SACS Brazil-Angola submarine cable., Unexamined routing configurations induced surprising performance impairments, i.e., the cable resulted in *longer* RTTs for these paths.

Figure 2: Internet interdomain congestion



Time series of latency probes (top panel) and throughput measurements (bottom panel). We used lightweight latency probes to identify interdomain links between networks with evidence of congestion (grey portions). We then conducted throughput measurements to establish the effect of the congestion on performance.

Figure 3: Internet router geolocation

xo.iad02.atlas.cogentco.com
as2828.was14.ip4.gtt.net
te9-2-0d0.cir1.ashburn-va.us.xo.net

vodafone.iad02.atlas.cogentco.com
ae-0.vodafone.asbnva02.us.bb.gin.ntt.net
zayo.vodafone.er2.iad10.us.zip.zayo.com

usqas1-rt002i.i3d.net
interactive.edge1.washington111.level3.net
ce-0-4-0-2.r05.asbnva02.us.ce.gin.ntt.net

level3-as3356.e0-51.switch2.ash1.he.net
ae-1-3510.edge1.washington111.level3.net
level3.ashburn2.ash.seabone.net

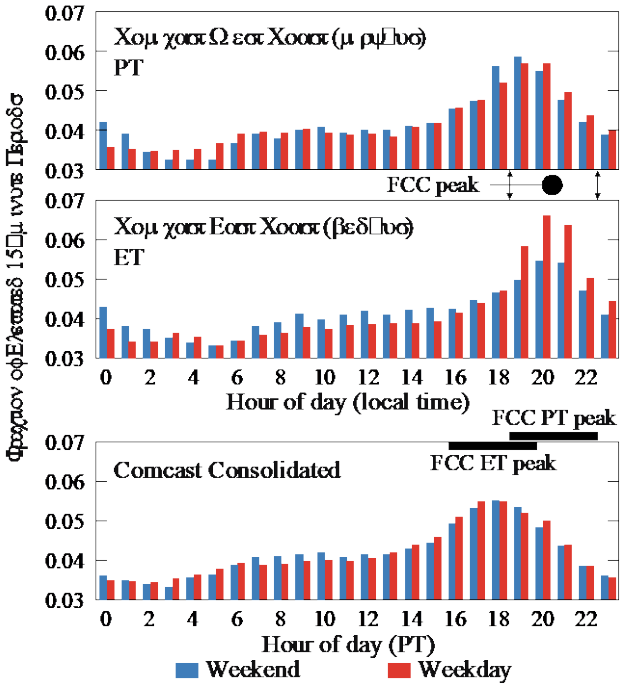
Hostnames of 4 routers located in/near Washington D.C., with router interface names assigned by nine operators. Ark enabled development of a technique to automatically learn the naming convention of each operator using the ITDK as a primary input. We provide the inferred rules, and a public API as a service, to the research community. Note that ash, in red, is an airport in Nashua, NH; our technique learns that the operators used ‘ash’ to mean Ashburn, VA.

UC San Diego SDSC
San Diego Supercomputer Center, University of California San Diego
9500 Gilman Drive, mc0505, La Jolla, CA 92093-0505 | 858-534-5000 | www.caida.org

CAIDA Archipelago Data Usage: Outcomes

Inferring Persistent Interdomain congestion

Dhamdhere et al — SIGCOM '18



Distribution of recurring 15-minute congestion preiods from 2017 as seen from VPs in Comcast

CAIDA/UCSD, MIT, U. Waikato, TU Munich

Illuminating Router Vendor Diversity Within Providers and Along Network Paths

Albakour et al — IMC '23

Router vendor diversity of a path

MPII, Delft U. Tech

Zeroing in on Port 0 Traffic in the Wild

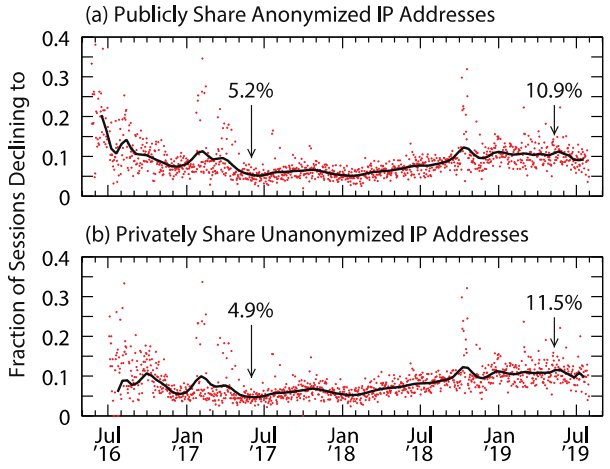
Maghsoudlou et al — PAM '21

Traffic between top 10 (source AS, destination AS) pairs involved in port 0 traffic in the MAWI-short dataset. Max Planck

(sample papers)

Network Hygiene, Incentives, and Regulation: Deployment of Source Address Validation in the Internet

Luckie et al — CCS '19

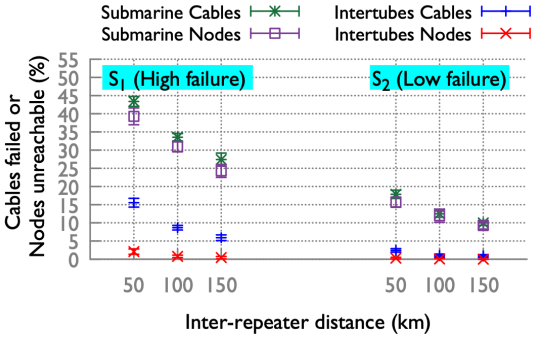


Fraction of Spoofer project tests with the daemonized client per day with sharing restrictions, overtime. The daemonized client was released inMay2016. The Bézier curves hows the trend that the percentage of private tests grew from 5.2% of tests in June 2017, to 10.9% in May 2019.

Waikato, Naval Postgraduate, CAIDA

Solar superstorms: planning for an internet apocalypse

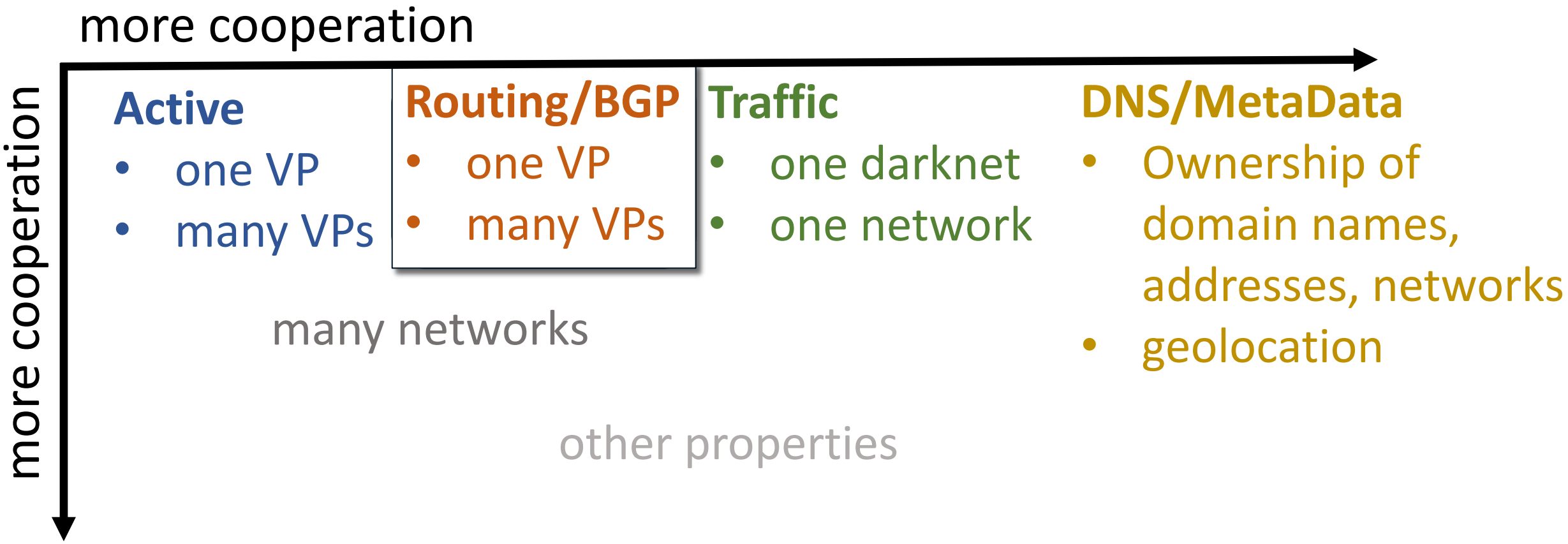
Jyothi et al — IMC '23



Cable and nodes failures under two states of non-uniform repeater failures ((1, (2). In each state, repeaters in a cable are assigned a failure probability based on the highest latitude (!) endpoint of the cable. Three levels of failure are ! > 60, 40 < ! < 60, and ! < 40. Assigned failure probability per repeater in (1 is [1, 0.1, 0.01] and in (2 is [0.1, 0.01, 0.001] across the three levels respectively.

UC Irvine, VMware Research

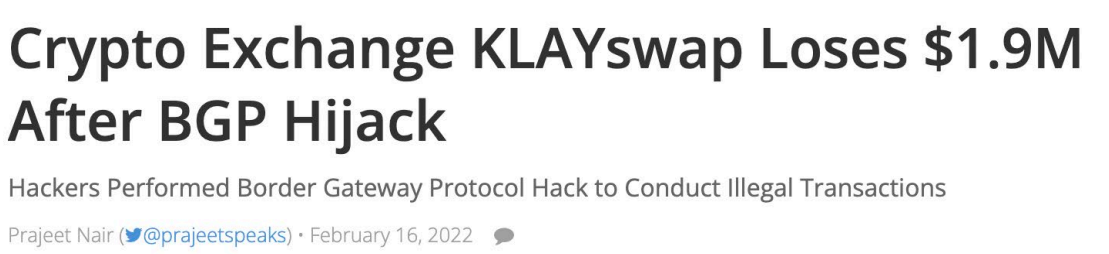
Spectrum of Cooperation Needed for Data Collection



Interdomain (BGP) Routing System Vulnerabilities

- BGP Hijacks (Origin and Path)
- Compromise of RIRs
- Impersonation of address space
- Misuse of AS/prefix revocation
- Malicious use of BGP communities

Unsuccessful attempts to retrofit security into protocol/deployment



Growing Consensus: Transparency will be Critical

Ongoing public measurements will guide the evolution of routing security by enabling

- Evaluation of effectiveness of routing security practices
- Inter-organization auditing and accountability

Current Approaches to Gathering BGP Data

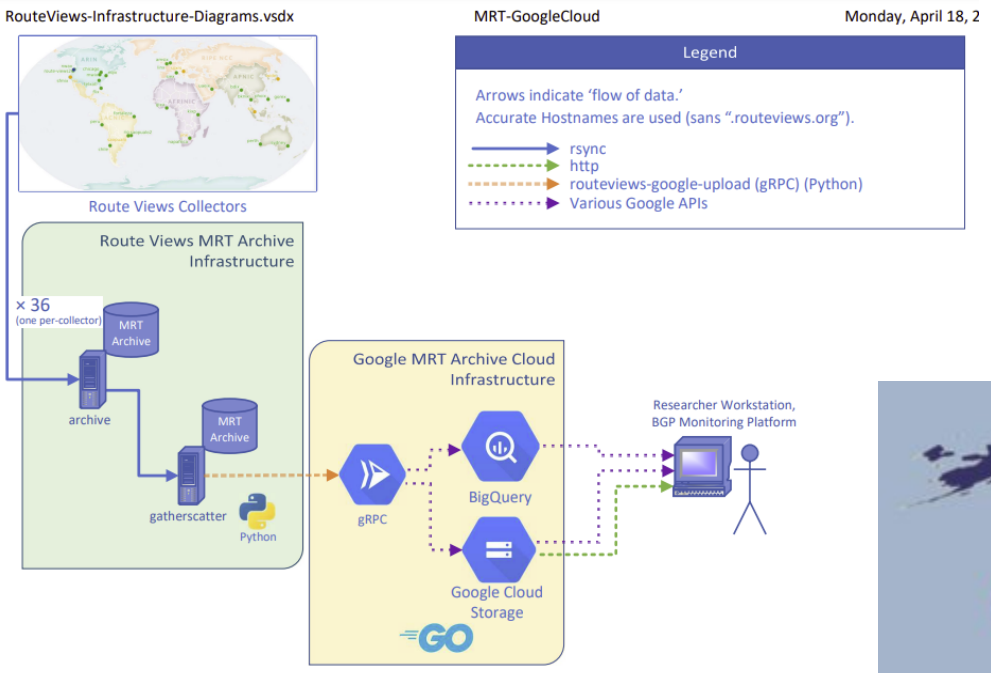
RouteViews Project

https://www.routeviews.org/routeviews/

RIPE RIS

https://ris.ripe.net/

Platforms to obtain real-time and historical BGP information about the global Internet routing system from the perspectives of different backbones and locations



RV Infrastructure

RIS MRT Files

Explore these files for an archive of all data we collect, stored in MRT format.

RIS MRT Files

RIS Live

Monitor BGP messages and detect routing incidents in near-real time with this live feed of RIS data.

RIS Live

RISwhois

Access a whois-style interface to prefix-to-ASN mappings based on the latest RIS data, or download these mappings.

RISwhois

Operational Challenges

Incentivizing infrastructure and peering deployment

Monitoring and managing data quality

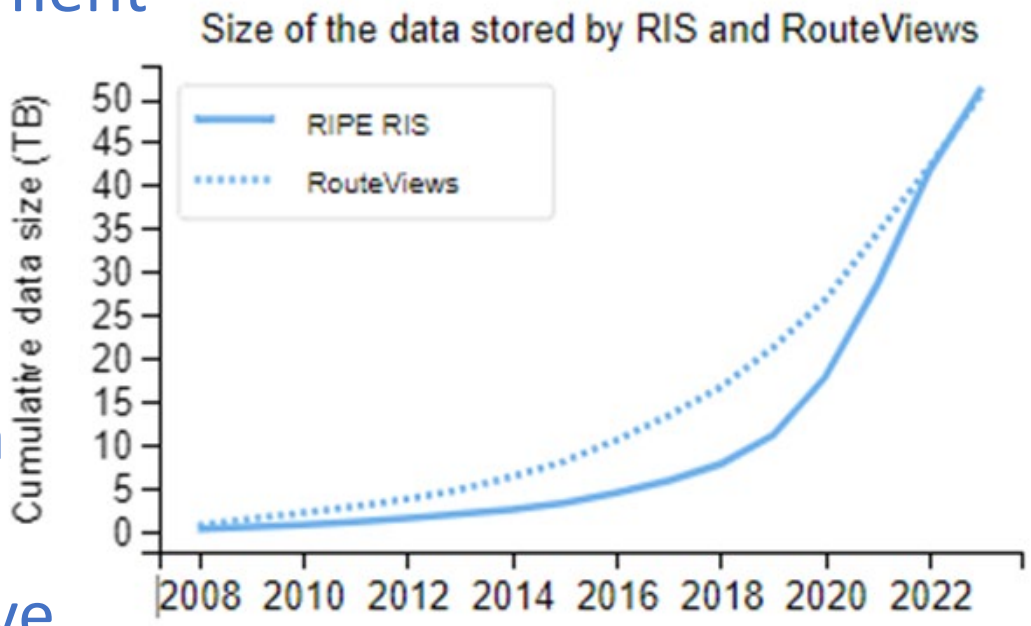
Misconfigurations can pollute archive

Hardware/software administration/automation

Cloud data storage and BQ queries are expensive

Looming challenge: scaling an order of magnitude

(Hijacks can intentionally avoid detection by existing measurement infrastructure)



In 2023, only 1% of the 75k ASes (~8% of transit Ases) export their BGP routes to RIS or RouteViews.

~ 100TB of data

Designed Approach to Scaling BGP Data Collection

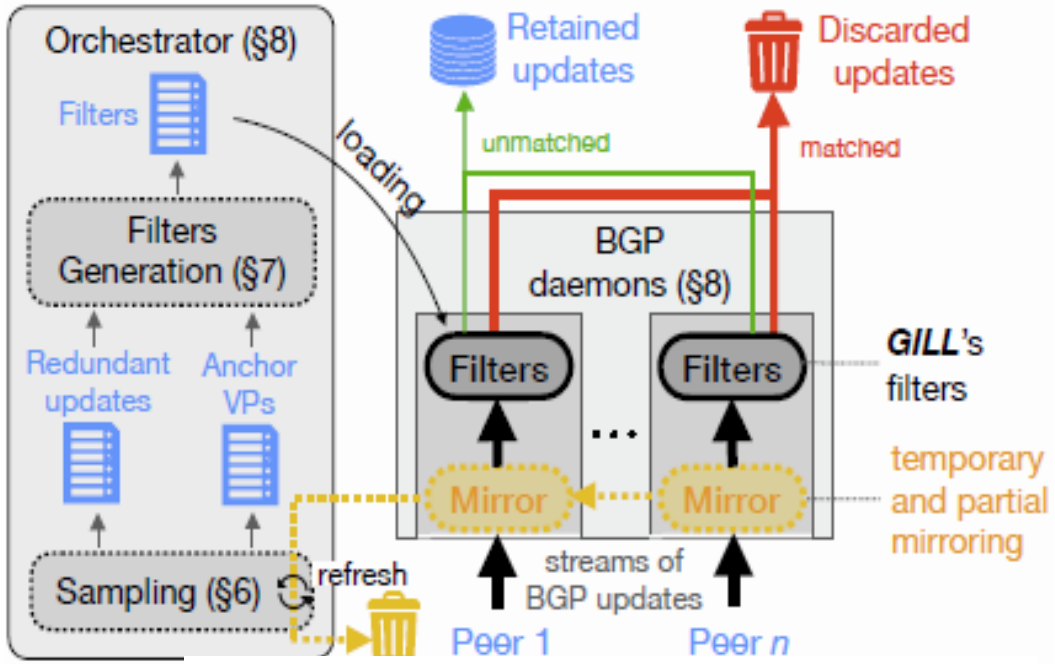
GILL BGP data platform

- U. Strasbourg led, RV/RIS advised
- Architecture to manage collection of routes from at least an order of magnitude more routers than existing platforms
- *Overshoot-and-discard* collection scheme: limits human effort and data volume
- Stores *non-redundant* data
- Requires framework for defining and detecting redundancy

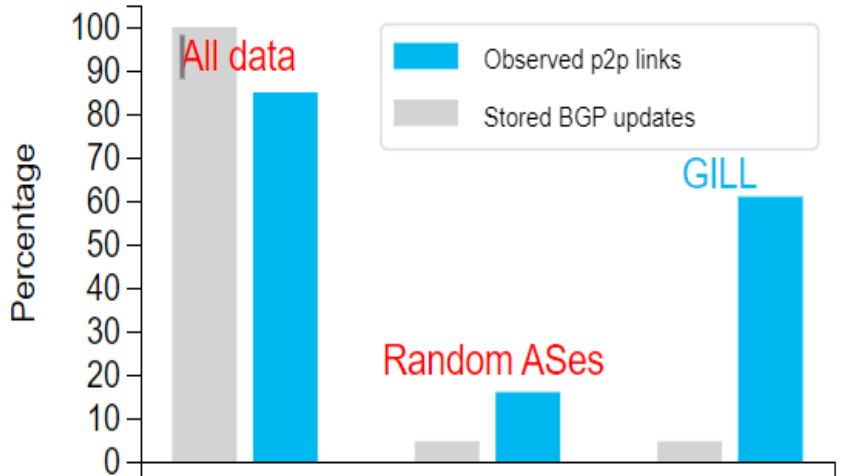
https://bgproutes.quest/

GILL Simulations: if **50%** of ASes peered with GILL, it would store **4.7%** of collected BGP updates, revealing **61%** of p2p links.

GILL's Workflow



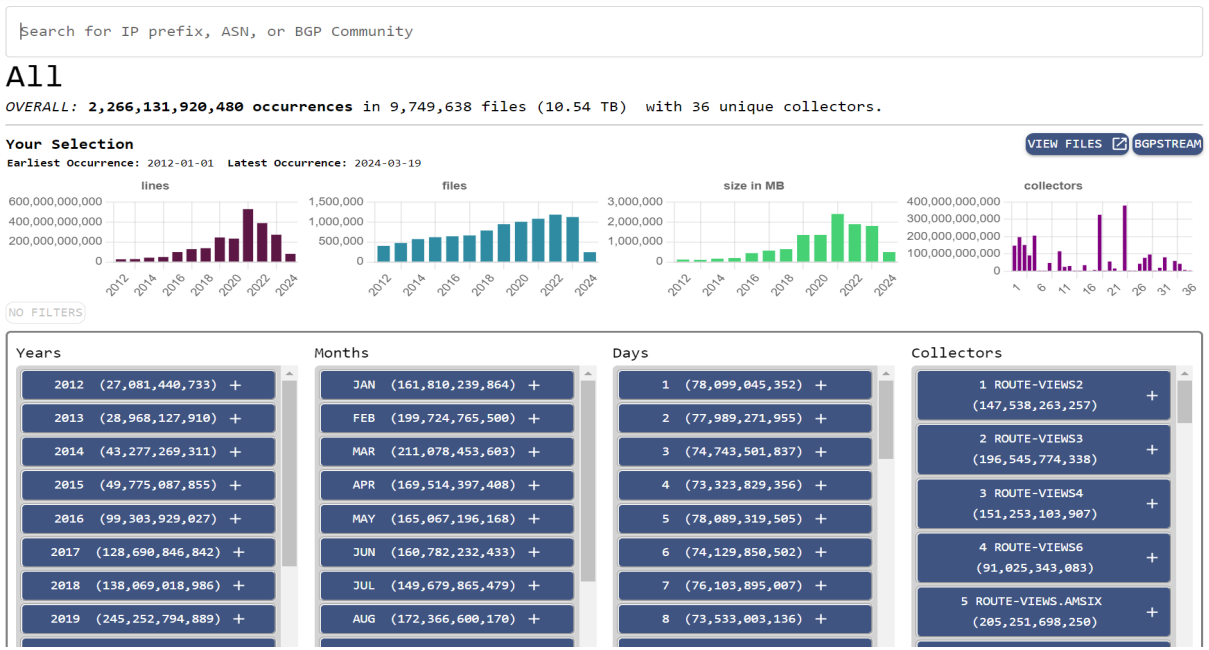
Results of our simulations with GILL and two baselines



CAIDA Tools to Discover and Analyze Routing Data

BGP2GO

Graphical **interface to explore** and compile RV data (MRT files)
http://nids.caida.org:44444/



AS Rank

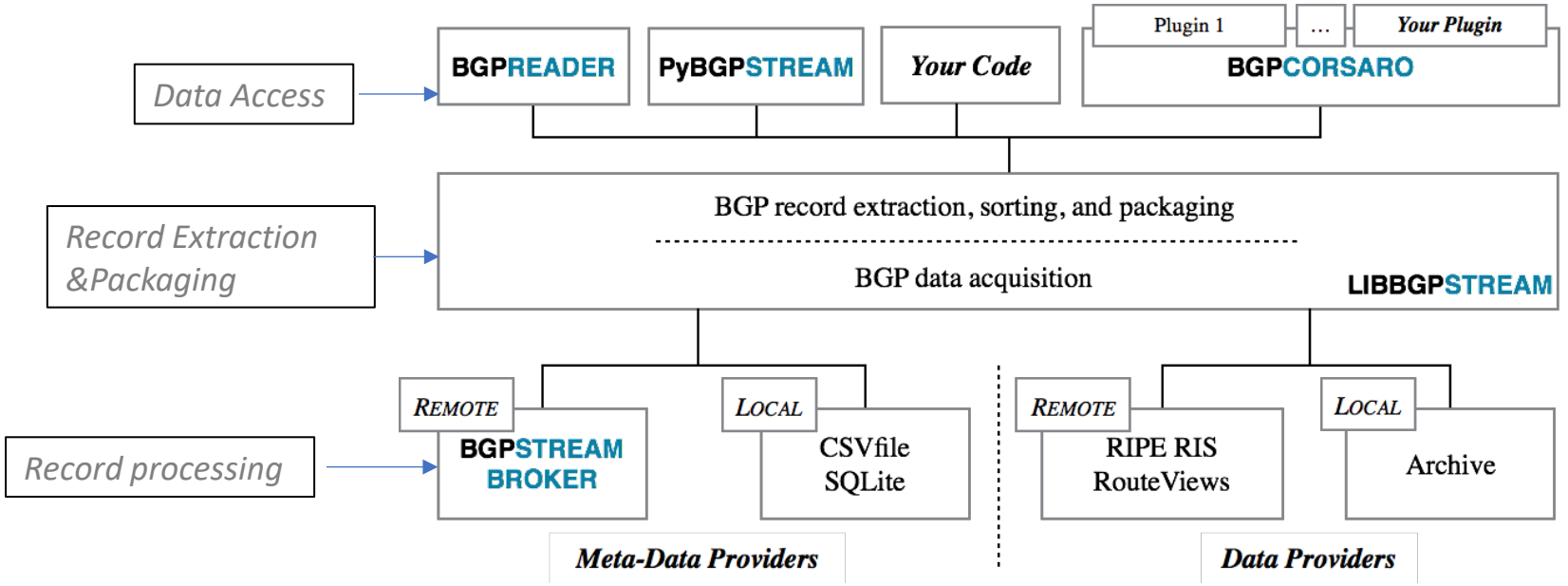
Inferred routing relationships between ASes and orgs

Ranking ASes based on customer cone size, (direct and indirect customers)
https://asrank.caida.org/

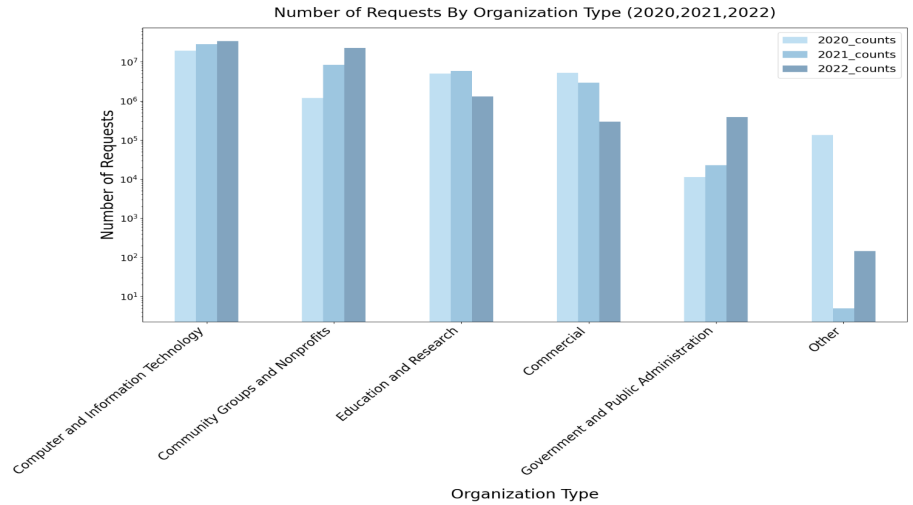
CAIDA Tools to Discover and Analyze Routing Data

BGPSTREAM

Open-source software framework for live and historical **BGP** data **analysis** (relies on RIS/RV)
https://bgpstream.caida.org/



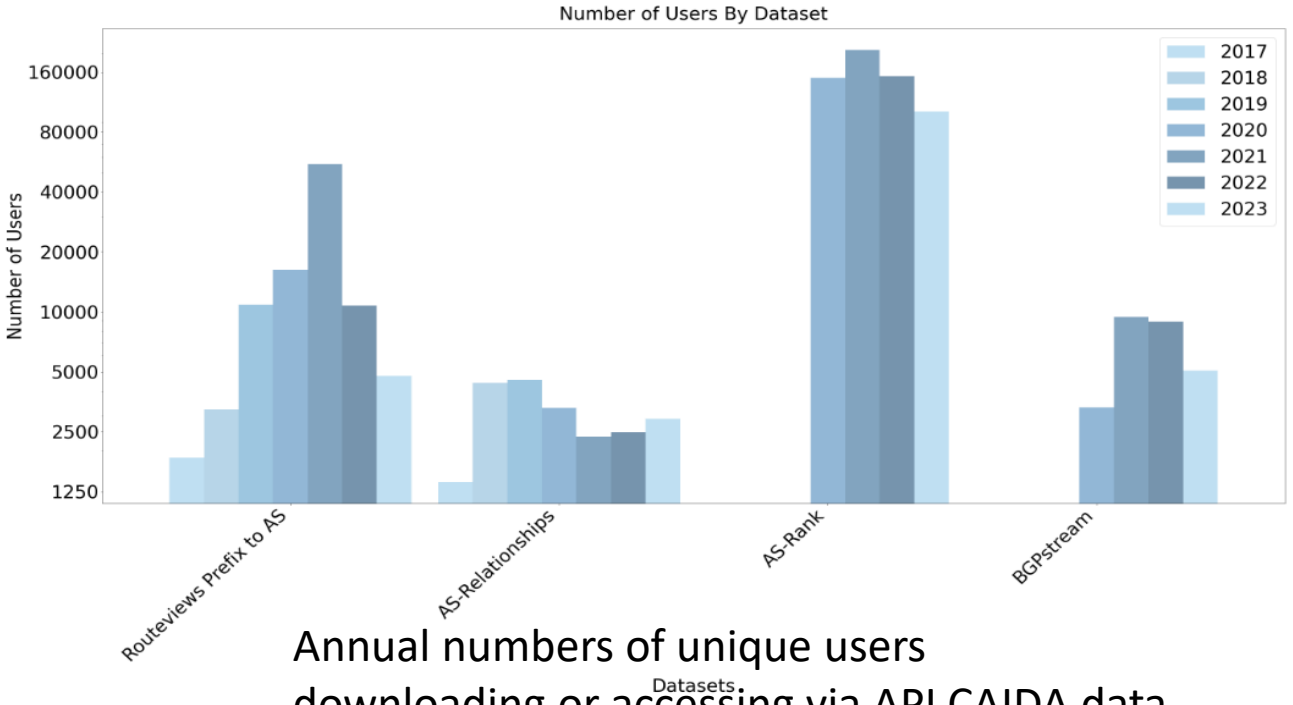
Includes ***bgpview*** **libraries** to facilitate inference of routing tables at **finer temporal granularity** than provided by RouteViews and RIPE RIS.
https://github.com/CAIDA/bgpview



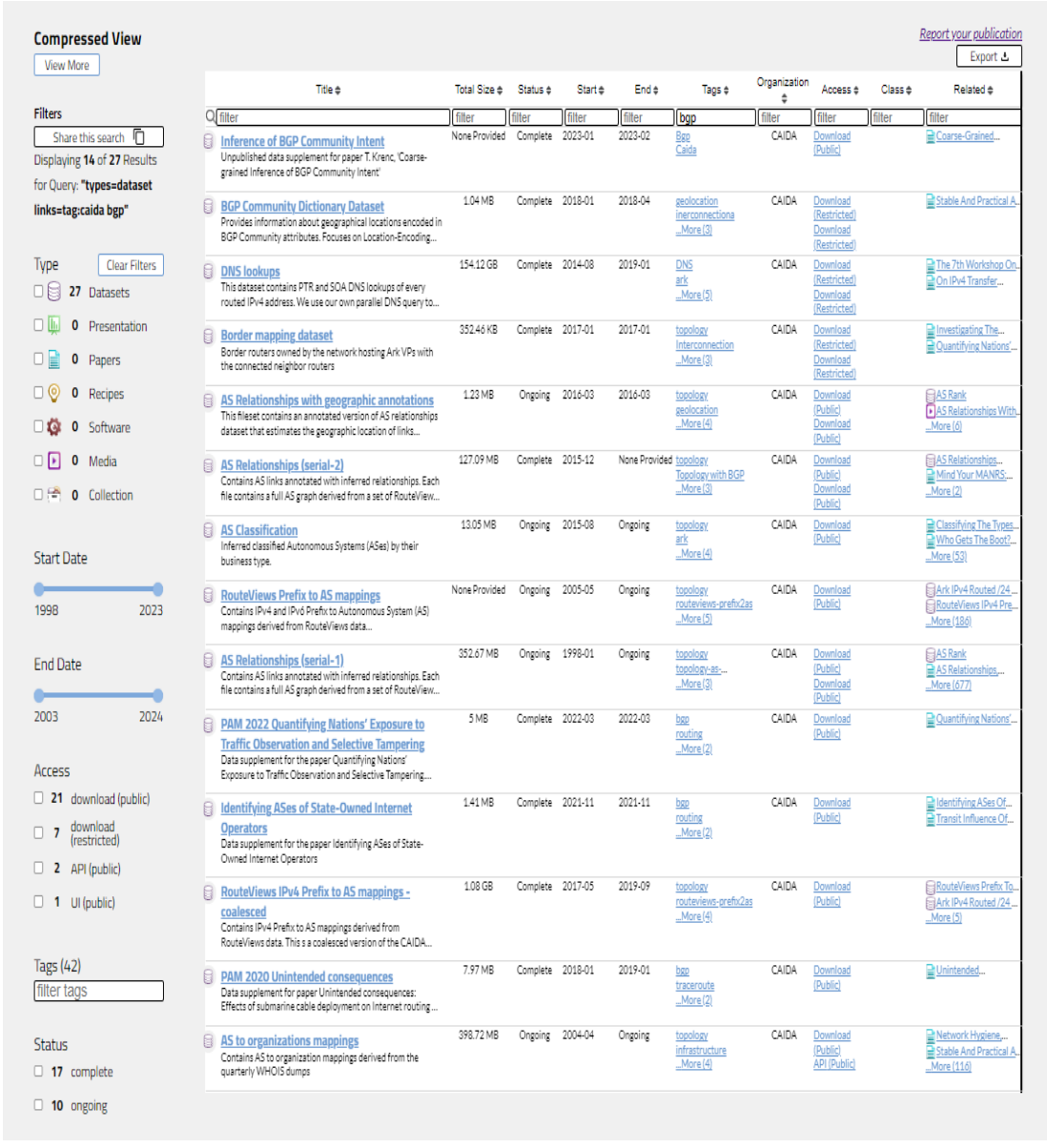
BGP-derived data sets to support research

- Large, labeled data sets
- 10 ongoing datasets
- Popular: AS Relationships, Prefix to AS mapping, AS-to-organization mapping

https://catalog.caida.org/search?query=types%3Ddataset+links%3Dtag%3Acaida++bgp&in_tags=bgp



Annual numbers of unique users downloading or accessing via API CAIDA data



BGP-derived data sets to support research

Request access to BGP2GO

http://nids.caida.org:44444/
https://www.caida.org/catalog/software/accounts/bgp2go_access_request/

BGP-derived data

https://catalog.caida.org/search?query=types%3Ddataset%20links%3Dtag%3Acaida%20bgp

BGP analysis software tools

https://catalog.caida.org/search?query=links%3Dtag%3Acaida+bgp&type=software

Design new data-driven frameworks

A path forward: **Improving Internet routing security**

[Accepted to Journal of Cybersecurity]

- https://arxiv.org/abs/2312.0335
- Demonstrated need for additional data to improve visibility into deployment of routing security practices.[

How You Can Help

Peer with RouteViews:

https://www.routeviews.org/routeviews/index.php/peering-request-form/

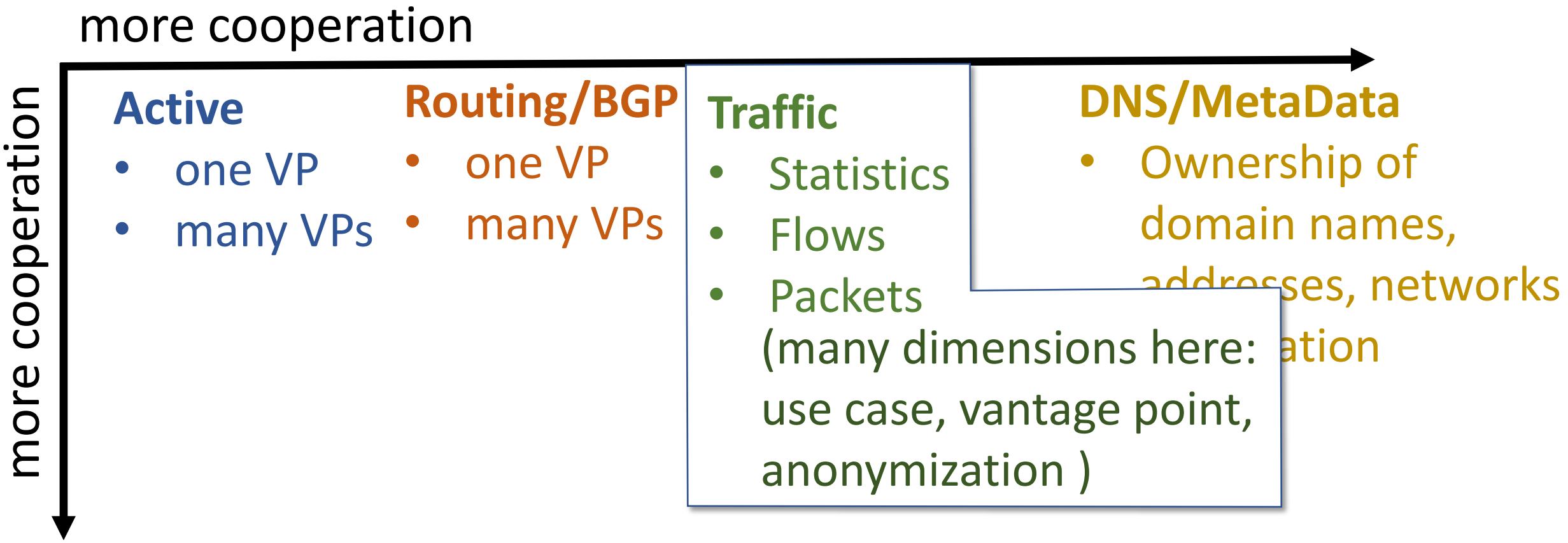
PEER with RIPE RIS:

https://labs.ripe.net/author/emileaben/two-years-of-selective-peering-with-ris/

Peer with GILL (new automated experimental platform!):

https://bgproutes.quest

Spectrum of Cooperation Needed for Data Collection

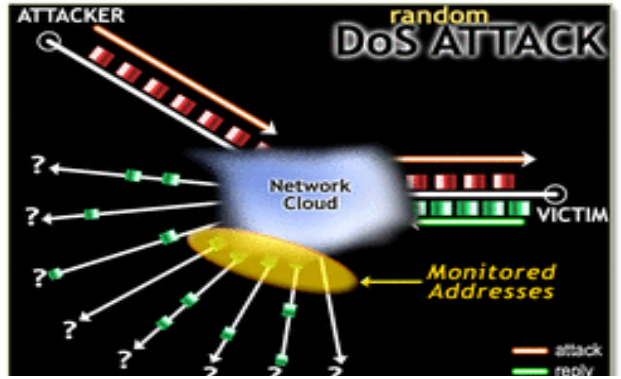


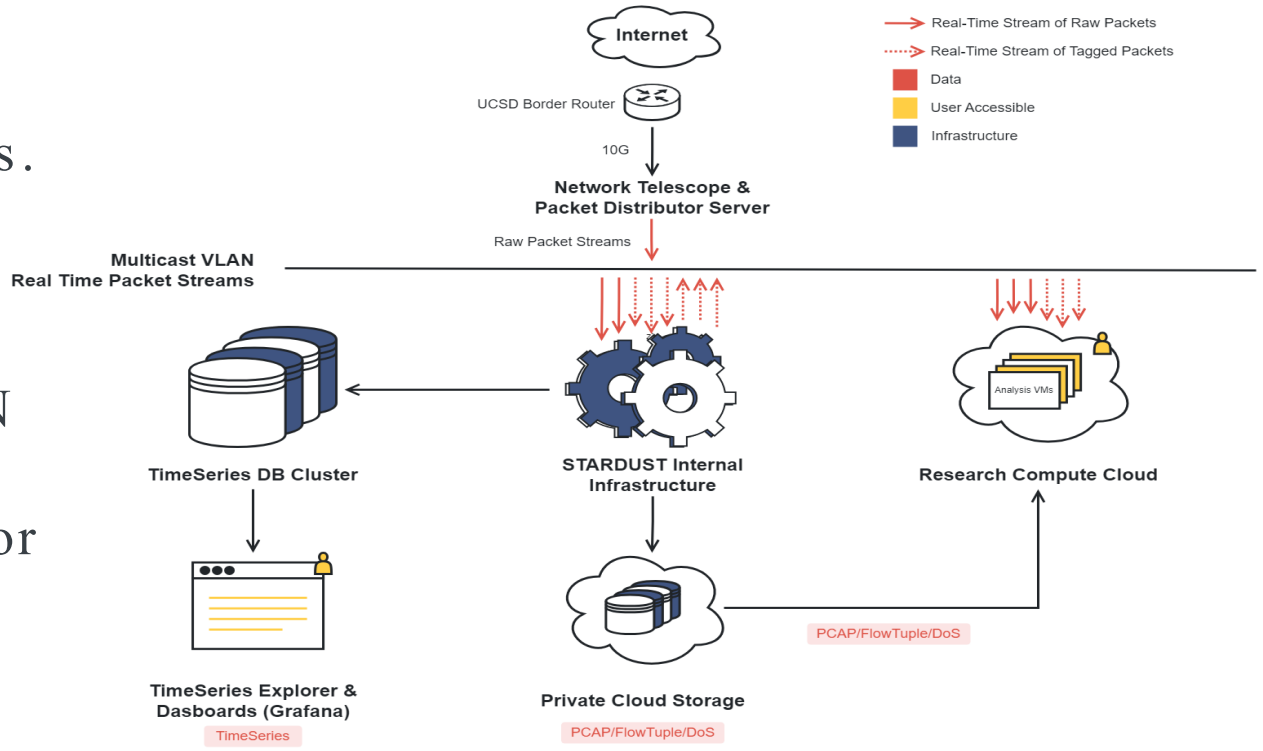
State-of-the-Art: Passive One-Way Traffic Measurement

UCSD Network Telescope

- **Globally routed ~/9 and /10 network -- IP addresses**
- **Continuous view of anomalous unsolicited traffic (Internet Background Radiation, IBR)**

- Users access stream in near real time from VMs.
  - Sensitive and high-volume data
- Raw pcap and flow-level traces (+ RSDoS) at cloud storage (direct or api/software access)
- We label data with geolocation, network, or ASN
- We extract time-series stats (e.g., per-minute count of unique source IPs per country or ASN or protocol port number) --> Grafana dashboard

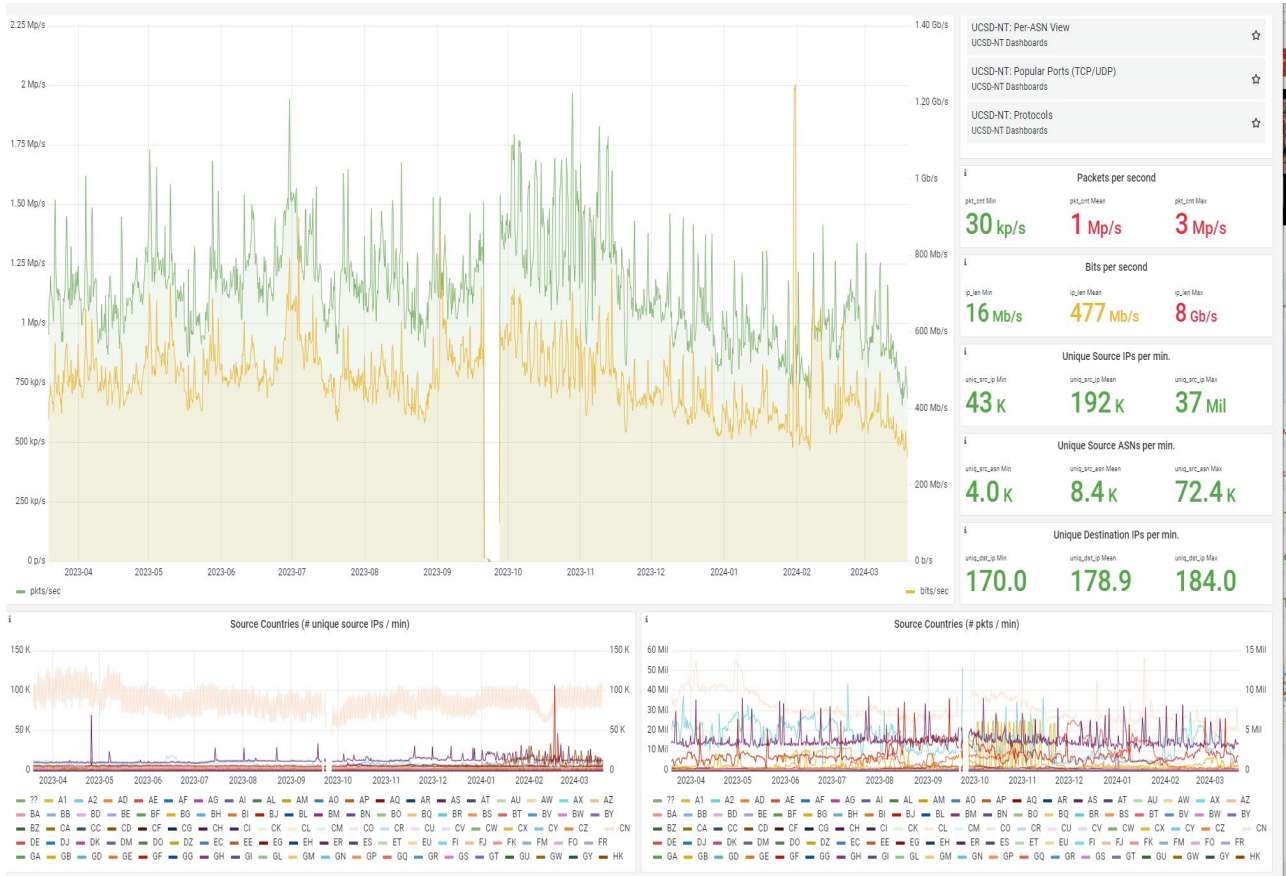




https://www.caida.org/projects/network_telescope/

Operational challenges

- Scaling with traffic growth
- Managing terabytes of data volume
  - The size of 1-hour raw pcap file is ~ 0.5 TB
  - The total size of compressed historical raw pcap data (stored at NERSC) is ~6 PB
- Creating tractable derivative data
  - *i.e, smaller (statistics)*
- Extracting actionable insight



UCSD Telescope traffic data sets to support research

Access to Telescope Data

https://catalog.caida.org/collection/ucsd_telescope_datasets

- Direct stream (on VM)
- Raw historic PCAP files
- Establishing exporter to send a subset of packets received by the telescope directly to collaborators (**expensive, privacy issues)
- Hooked into NSF ACCESS supercomputing facilities at SDSC
- Time-series dashboard of traffic statistics

Designed Approach: STARNOVA

Scalable Technology to Accelerate Research Network Operations Vulnerability Alerts (STARNOVA)
(UCSD PI Ricky Mok)

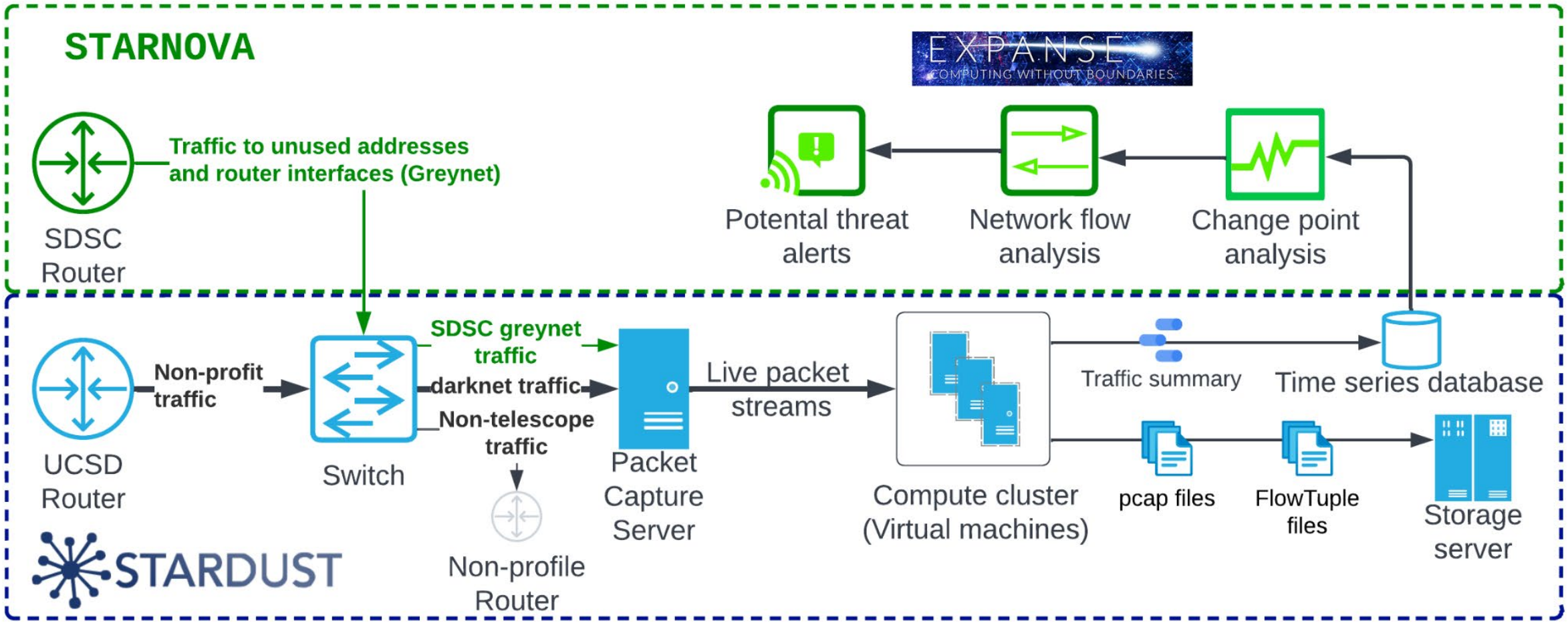
Expanding capability to identify targeted attacks:

- Form greynets – subnets with dark and active IP addresses
- Scale ML-based timeseries analytics using SDSC HPC
- Automate flow analysis to examine intervals flagged by our (more efficient) time-series-statistics anomaly detection method

https://www.caida.org/funding/cici-starnova/

Interested in Your Own STARNOVA Platform?

Helping with Darknet/GreyNet Deployment



Integration of existing UCSD telescope infrastructure and STARNOVA platform

Sharing Traffic Data will always be challenging

Focus on questions we really need to understand

For example:

the **Denial of Service (DOS/DDOS) attack landscape.**

- Is it getting better? Worse?
- Are counter-measures working?
- Would government intervention help?
- If so, what kind?
- How would we know if it helped?

Distributed Denial of Service (DDoS) Attacks

Exploit vulnerable end nodes and the basic packet forwarding function of the Internet to flood an end-node or a region of the network

Image from https://sosafe-awareness.com/glossary/ddos-attacks/

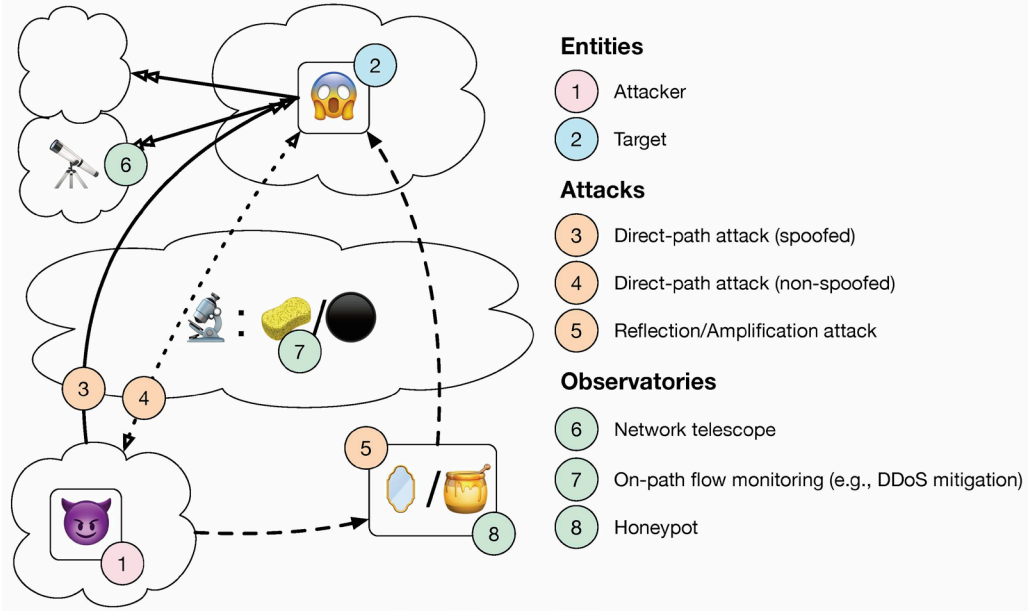
Goal: Can we find any consistent view?

"The Age of DDoScovery: An Empirical Comparison of Industry and Academic DDoS Assessments"

Find and explain discrepancies and similarities between industry and academia observations of attacks

Previous works = cause for concern (even academics don't share data)

17 authors gave it their best..



Three DDoS attack types:

- *Direct-path spoofed (solid line),*
- *direct-path non-spoofed (dotted),*
- *reflection-amplification (spoofed) (dashed).*

Goal: Can we find any consistent view of DDOS?

- Qualitatively compared **9 datasets** (industry &academic)
- Taxonomized data from **24 industry reports** characterizing DDoS in 2022-2023
- New approach to **transparency with industry** by aggregating target information (IPs) from academic sources and sharing with industry,
- Industry players joined it w/ their data sources revealing gaps in visibility, shared results
- Validated industry-reported 2021-2022 drop in spoofed reflection-amplification attacks (during an industry WG effort), but they increased again in 2023
- Proposed self-regulatory advances in transparency involving academic researchers

Attack Type	Observatories Used in This Paper (2019-2023)							Industry Reports (#)
	Network Telescopes		Flow Data		Honeypots			(≈ 2022)
	UCSD	Orion	Netscout	IXP	Hopscotch	AmpPot	NewKid	
Direct-path	▲	▲	▲	▲	n/a	n/a	n/a	▲(5), ▼(0)
Reflection-Ampl.	n/a	n/a	▲	▼	▼	◆	▲	▲(2), ▼(3)

Table 1: Data comparison results: Partially inconsistent views among DDoS data observatories used in this paper measuring decreasing ▼ ($< -10\%$ in 4 years), increasing ▲ ($> 10\%$ in 4 years), and steady ◆ trends of attack types in 2019–2023. The surveyed industry reports from ≈ 2022, which usually compare relative share of attacks, similarly provide inconsistent views. Here, numbers in braces indicate the number of reports out of 24 surveyed reports.

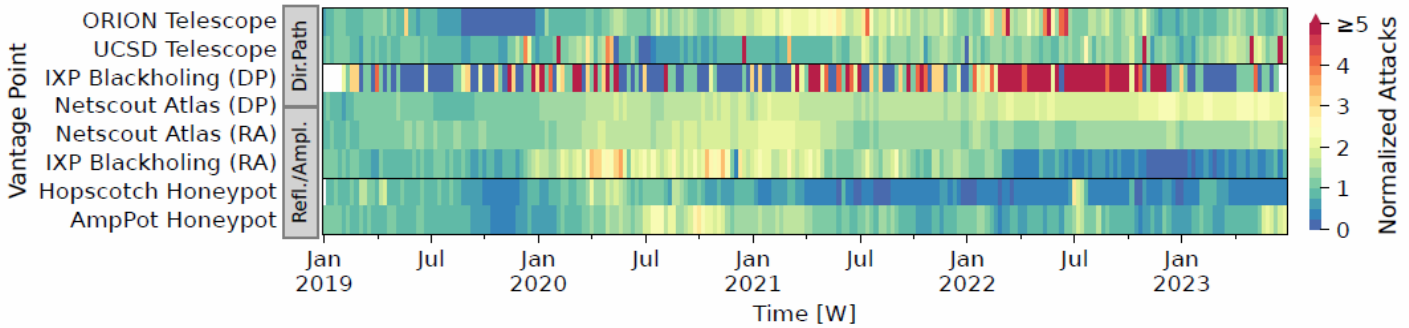
Operational challenges of DDoS Data Sharing

- Vantage Point limitations
  - Location, data sharing constraints
- DDoS reporting from industry is fragmented and scattered
  - Marketing
- No standard definitions, methods
- Data-sharing challenges
  - Even among academics!
  - Industry acknowledged we need improved data governance to facilitate sharing

Table 2: The observatories used in this research vary in collection methods and attack detection strategies. Honeypots use different flow identifiers, see [118]. (Location: **G**eographically & **T**opologically distribution.)

Platform	Type	Attack	Loc.	Coverage	Attack Definition		
					Flow Identifier	Timeout	Threshold
UCSD NT		RSDoS	US	12M IPs	protocol, src IP	300s	$\geq$ 25 pkts, $\geq$ 60s[2]
ORION NT		RSDoS	US	500k IPs	protocol, src IP	300s	$\geq$ 25 pkts, $\geq$ 60s[2]
Netscout (RA)		DP	G/T	proprietary	Hand-craft flow identifiers & thresholds		
Netscout (DP)		RA	G/T	proprietary	Hand-craft flow identifiers & thresholds		
IXP BH (RA) [83]		DP	G/T	proprietary	UDP, ampl. src port		$\geq$ 10 IPs, $>$ 1 Gbps
IXP BH (DP) [83]		RA	G/T	proprietary	TCP		$\geq$ 10 IPs, $>$ 100 Mbps
AmpPot [85]		RA	G/T	$\approx$ 30 IPs	Src IP, src port, dst IP, dst port	60 min	$\geq$ 100 pkts
Hopscotch [167]		RA	G/T	65 IPs	Src IP, dst IP, dst port	15 min	$\geq$ 5 pkts
NewKid [68]		RA	BR	1 IP	Src prefix, dst IP, [dst port][1]	1 min	$\geq$ 5 pkts, [$\geq$ 2 ports][1]

[1] NewKid uses two thresholds, one for mono-(dst port) and for multi-protocol ($\geq$ 2 ports) attacks. [2] See Appendix H for RSDoS inference details.

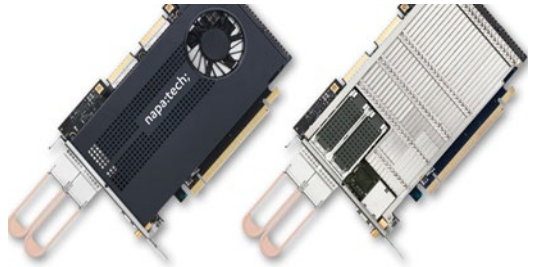


Attack intensity observed at nine vantage points. Observed direct-path (DP) attacks increased in 2022 while reflection-amplification (RA) attacks show the highest intensity towards the end of 2020, and declined thereafter.

Passive Two-Way Traffic Samples

Historical data captured on 10GB commercial backbone link

- Anonymized headers of traces
- Monthly from April 2008 to January 2019
- One-hour traffic capture in pcap format
- Access is provided by request
  https://www.caida.org/catalog/datasets/passive_dataset/

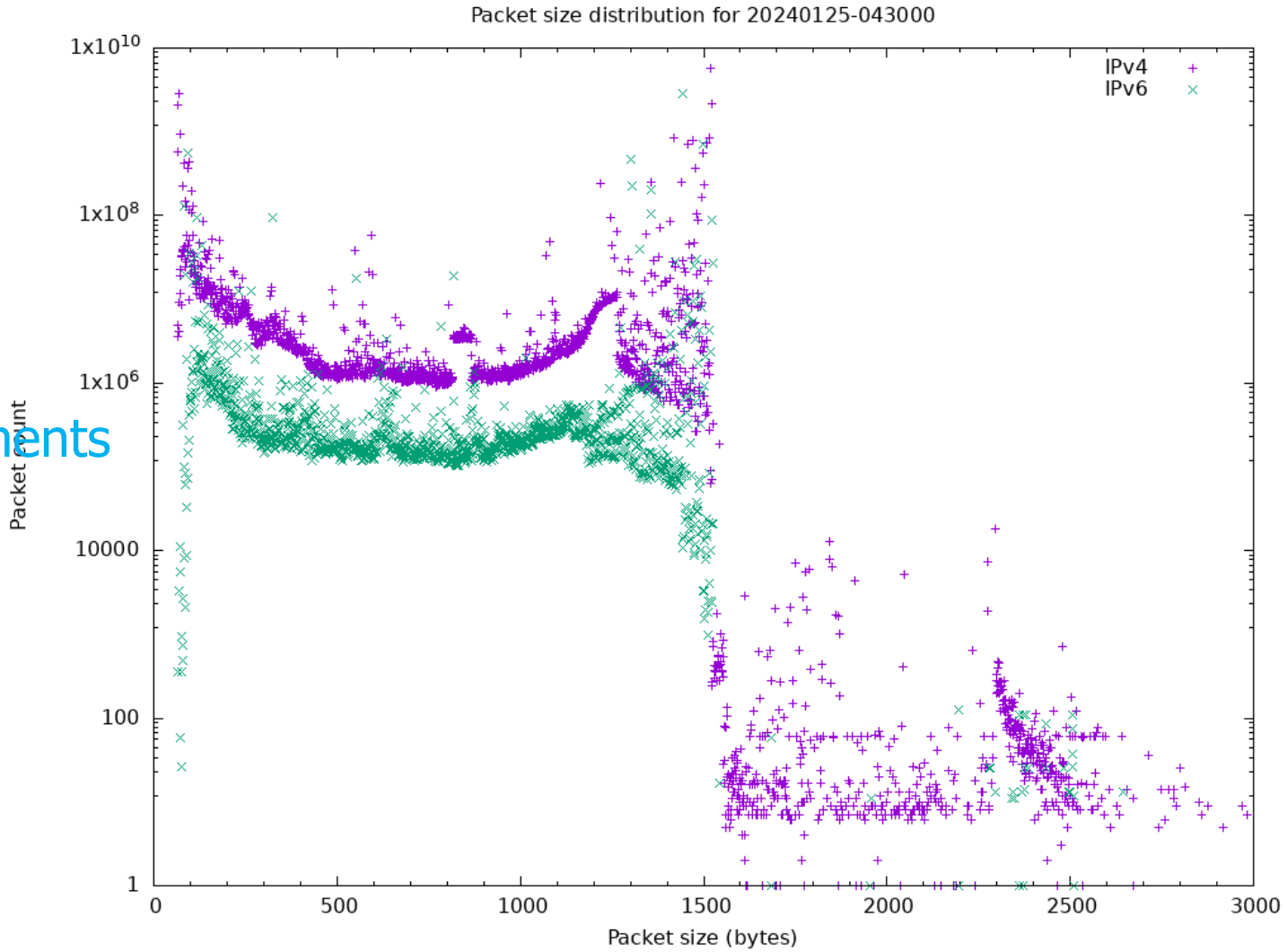


Moved to 100GB backbone link

- Capturing traces since October 2023
- Striping all packet payload after the layer 4 headers (no PI)
- Employing 16 streams -> 16 separate trace files need to be combined into one
- Anonymizing IP with a Crypto-PAn
- Sharing with researchers since early 2024

Operational Challenges

- Fast technological developments
  R&E networks now → 400GB links
- Managing data volume
  1-hour (compressed pcap) = ~ 1TB
- Persuading Vantage Points operators/owners to allow measurements and help with monitor installation
- Getting permission to share data



Proposed Approaches (No pixie dust)

- Brute force
- Well-written legal agreements
- Becoming part-time employee of data owner
- Exploration of privacy technology methods
  - differential privacy

Overcoming Limits of Differential Privacy

"Exploring the Limits of Differential Privacy"

- Overview of DP goals
- Why DP poorly addresses some harms to firms
- (Use synthetic examples)

https://housedemocrats.wa.gov/wylie/2020/01/22/domestic-violence-paid-family-leave-data-privacy-and-transportation/

Limitations:

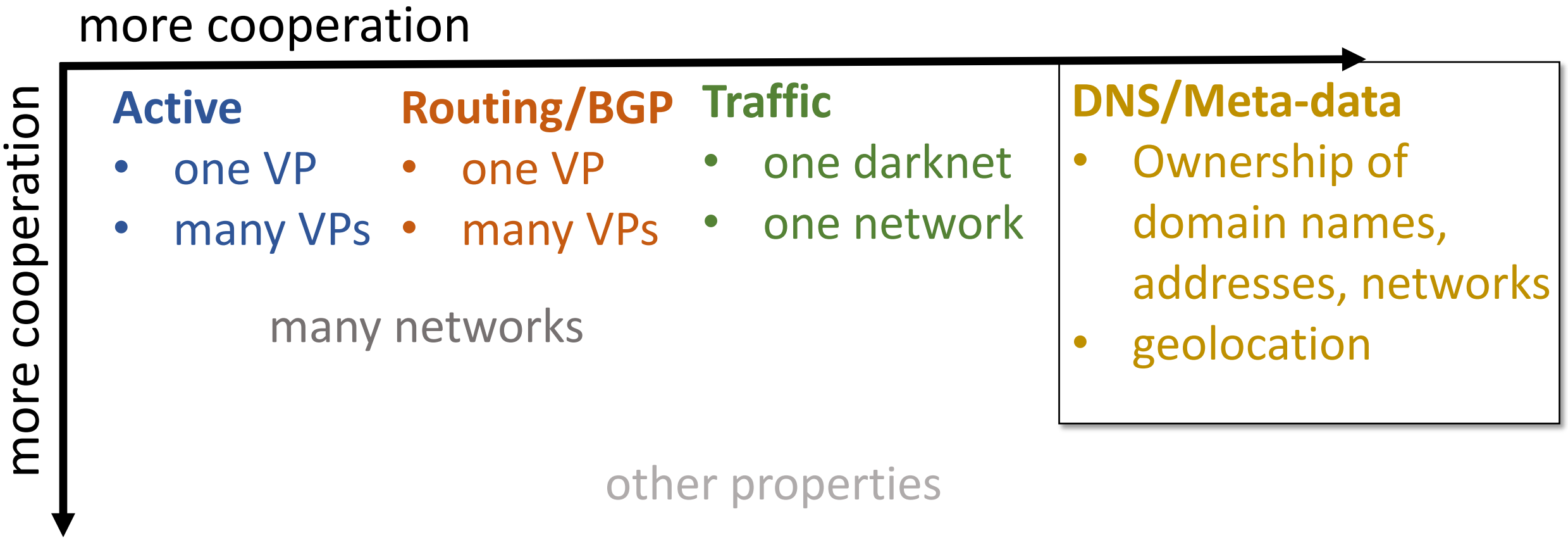
DP is a powerful technology, but not always well-suited to protecting corporate proprietary information while computing aggregate industry-wide statistics

Approaches:

Adding noise to the result of a query *based on pragmatic assessment of harm from that query* can be an important method to reduce potential harm

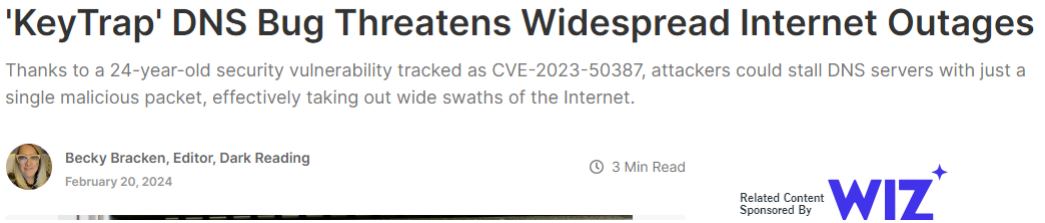
Programmatic assessment of potential harm is extremely challenging

Spectrum of Cooperation Needed for Data Collection



DNS Vulnerabilities

- Service penetration
- Identity theft
- Operational complexity → configuration errors
- User deception (phishing)
- DNS hijacks
- DNS resolution hijacks (many ways)
- Misrouting of DNS queries
- BGP hijacks of DNS resolver/name server
- Malicious name servers
- Cache poisoning



DNS Data Sources Available to Researchers

DZDB https://dzdb.caida.org/

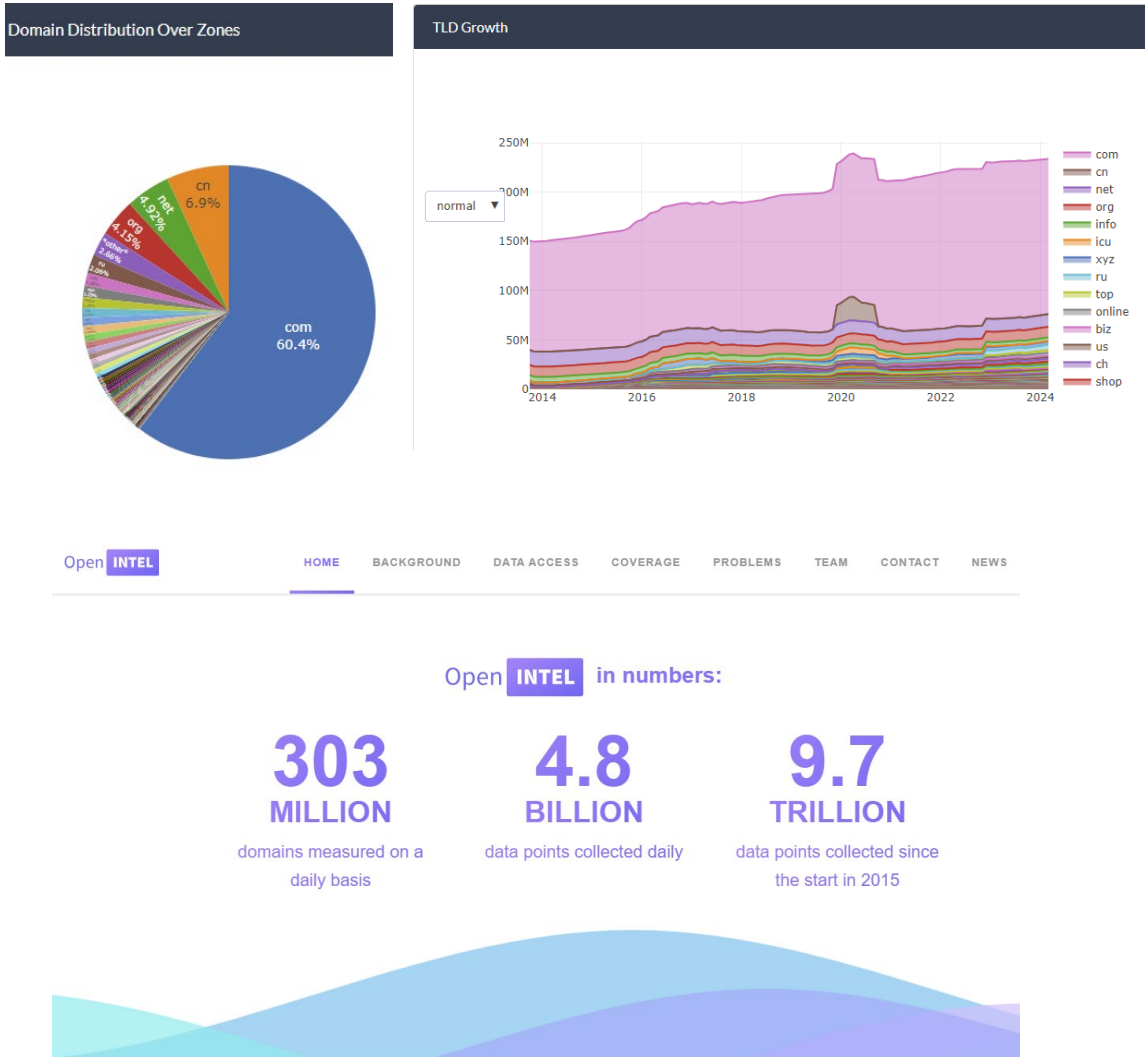
- Database query access to information and historical ~daily zone files provided by Top Level Domains (TLDs)
- Tracks history of a domain, nameserver, and IP records

OpenIntel: https://www.openintel.nl/

- Active DNS measurements of all domains observed in large set of TLDs (using zone files)
- Operated by U. Twente

Other DNS datasets (see CAIDA catalog)

https://catalog.caida.org/search?query=types=dataset%20links=tag:caida%20dns



Operational Challenges

- In addition to hardware/software challenges
- Agreements to access TLD zones files are not seamless
  - Must be updated regularly
  - Zone owners not always responsive
  - Limited to subset of TLDs
- Daily snapshots miss many short-lived domains
    - Which turn out to usually be problematic and thus more important to correlate with other extant domains
    - Represents a “blind spot” for security researchers
- Software licensing challenges when collaborating with industry

New Approach: Really an Old Approach that Died

In September 2004,

VeriSign implemented rapid zone updates, enabling updates to the .COM and .NET zone every 3 minutes

(prior to this VeriSign propagated updates to the .COM and .NET zones every 12 hours)… This data includes domain names, nameservers, IP address additions, deletions and modifications. The proposed service would enable… [access to] updated zone information every five minutes.

....

VeriSign states that the service would be used by recipients to build brand protection and fraud detection services for their customers, and promote security and stability by providing a useful tool to online security companies, ISPs, search engines, financial services companies, and other stakeholders.

VeriSign Application for Registry Service: "Rapid Zone Updates, 2007,
https://www.icann.org/en/system/files/files/memo-dns-update-service.pdf

New Approach: Rapid Zone Updates

Have proposed establishing a "DNS zone of trust" with USG-controlled TLDs supporting rapid zone updates: .gov, .edu. Us.

In process of data-driven risk/benefit analysis

Community design of infrastructure to sharing

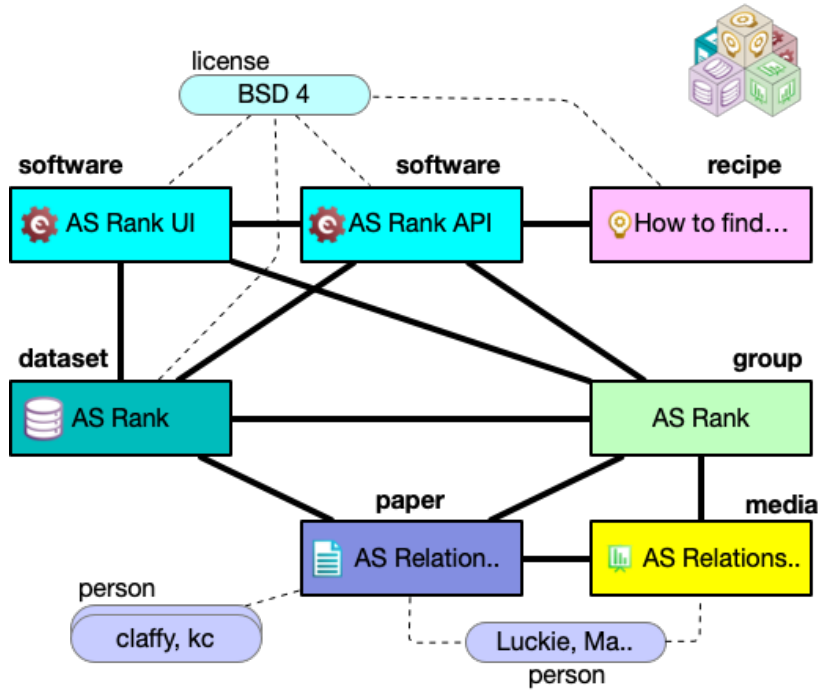
How You Can Help

If you have contacts at trust-conscious TLD registry operators, let us know

To Find What Data You Need: CAIDA Resource Catalog

Currently contains: 143 Datasets, 4000+ Papers, 612 Presentations, 34 Media, 42 Software , 39 Recipes entries as well as 10 Collections

- Search through a library of Publications, Datasets, Software, Recipes, Media to view descriptions, metadata, other resources, and links (rich context)

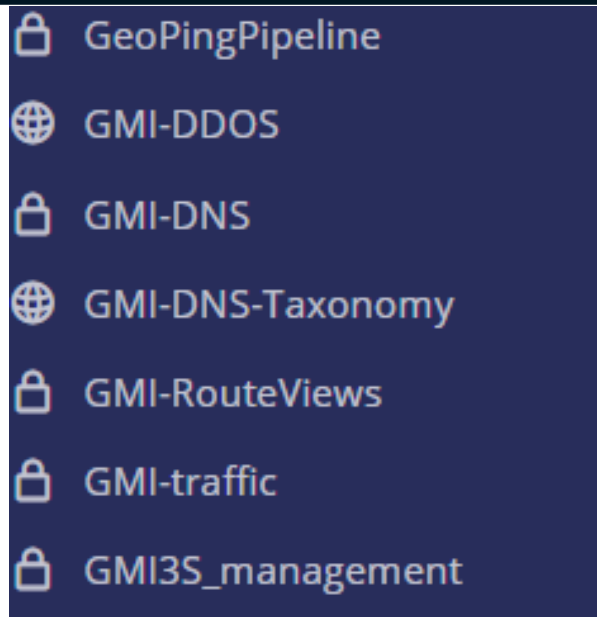


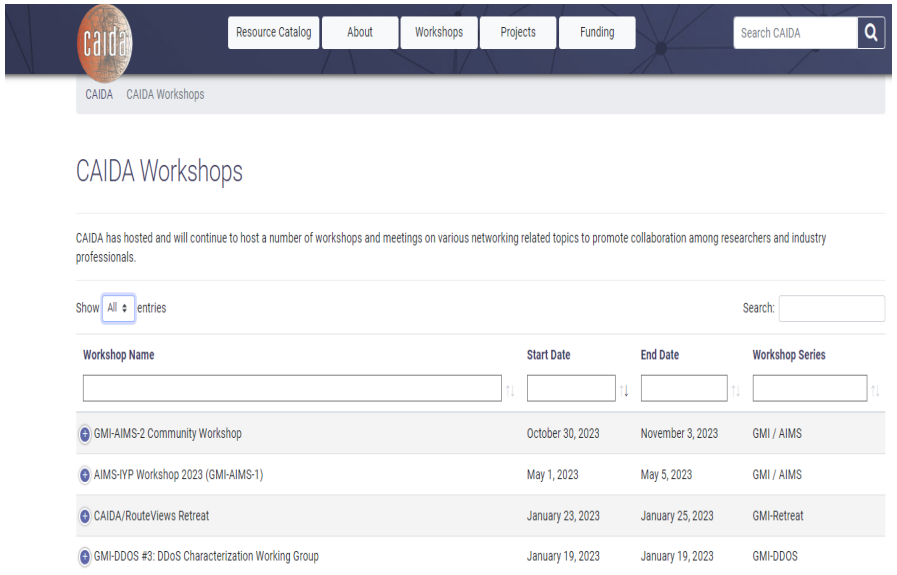
Internet Data Science for Cybersecurity Curriculum

- A data-science framing for conversations about the role of Internet measurement and data science in a range of public policy issues, with an emphasis on cybersecurity
- Focuses on the Internet as a data transport service, and vulnerabilities specific to interdomain routing (BGP), naming (Domain Name System), and certificate management
- Online syllabus slides, assignments, reading materials
  - https://cseweb.ucsd.edu/classes/wi23/cse291-e/syllabus.html
- Initiated working group to integrate Internet data sets into UC San Diego Data Science Machine Learning Platform (DSMLP) https://blink.ucsd.edu/faculty/instruction/tech-guide/dsmlp/

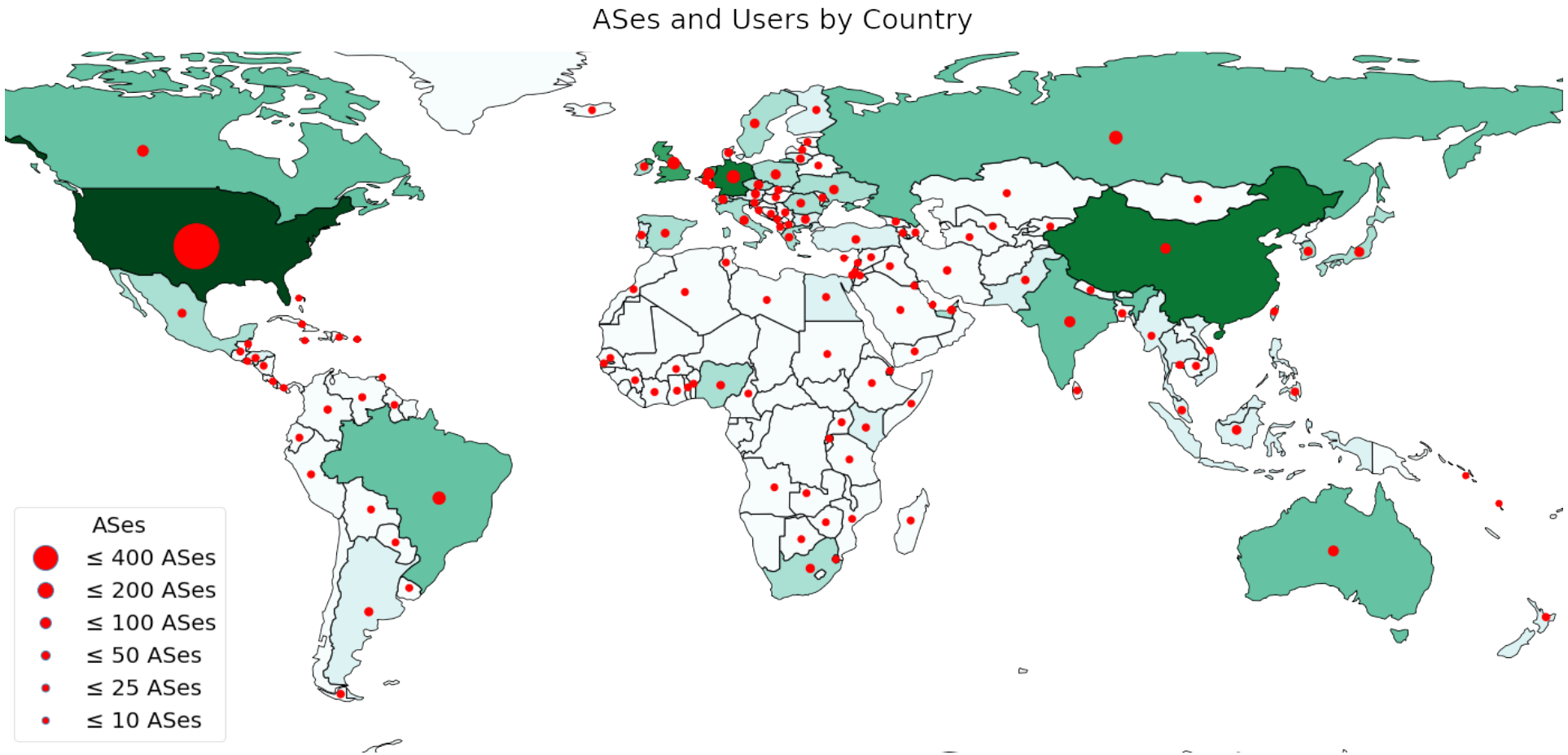
Workgroup Meetings & Workshops

- Created per-data-type working groups
- Conducting weekly/monthly/quarterly meetings
- Multiple working groups with industry partners
- MatterMost (slack) communication in between
- Two advisory committees: industry and academic
- Workshops: May & October 2023
- Next AIMS workshop -- June 2024
  https://www.caida.org/workshops/



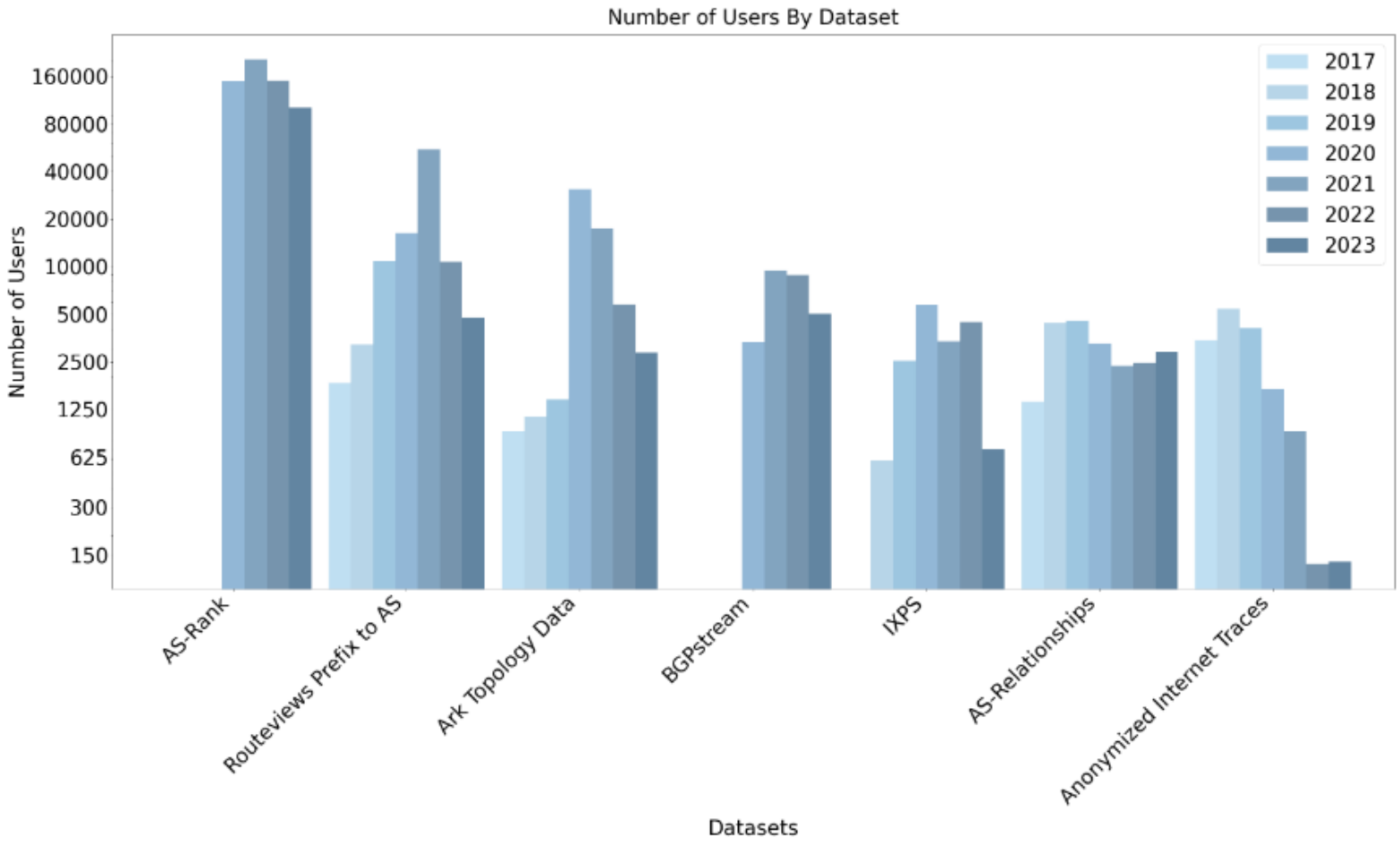


Data Sharing Statistics



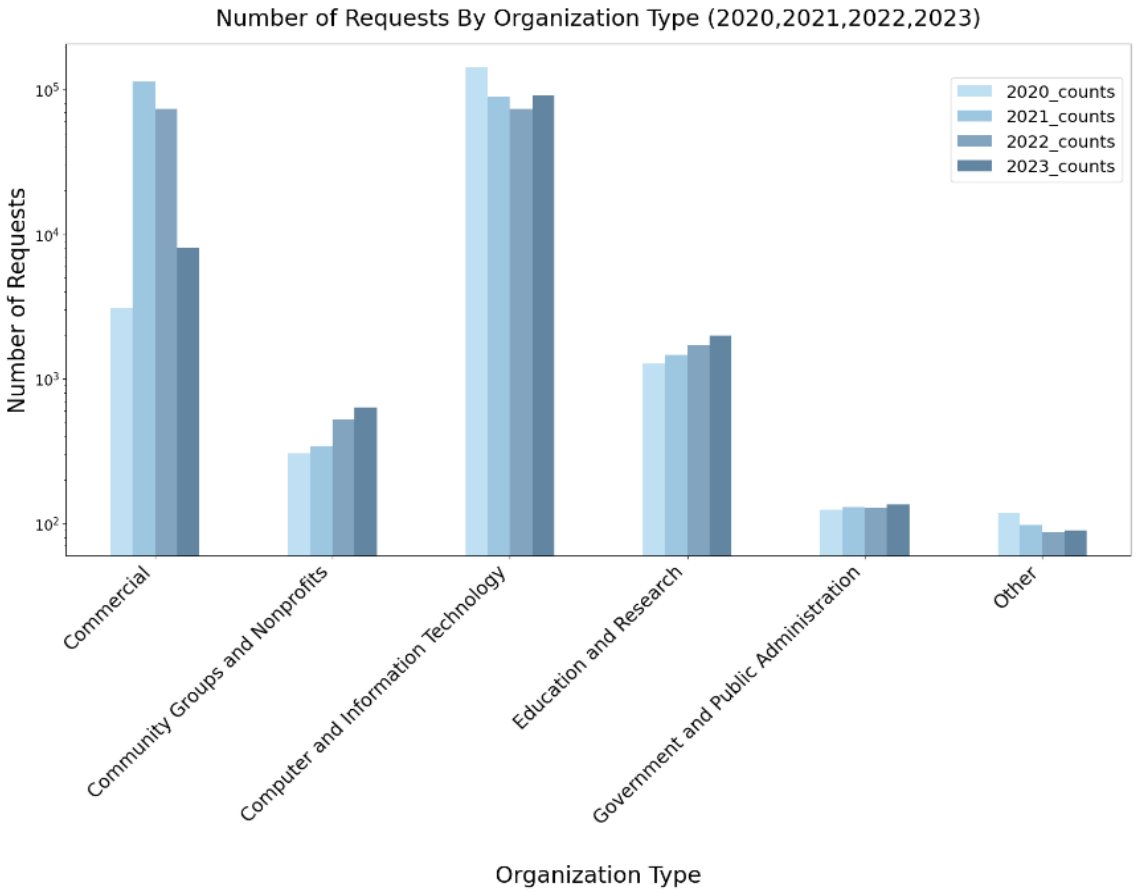
Unique **users downloading CAIDA data** and corresponding ASes aggregated by country.

Data Sharing Statistics

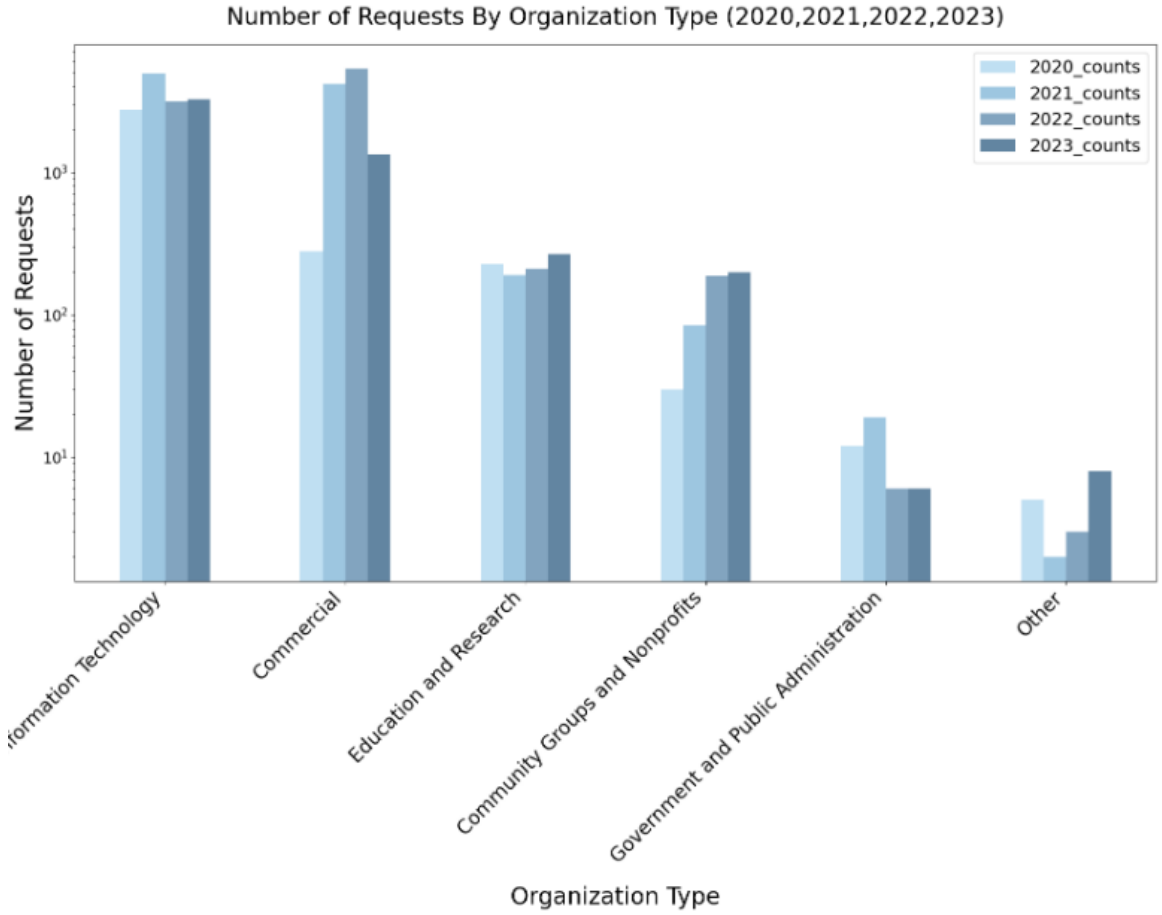


Unique users downloading CAIDA data downloaded annually.
AS Rank and BGPStream unique users calculated based on API access.

API users by type of organization

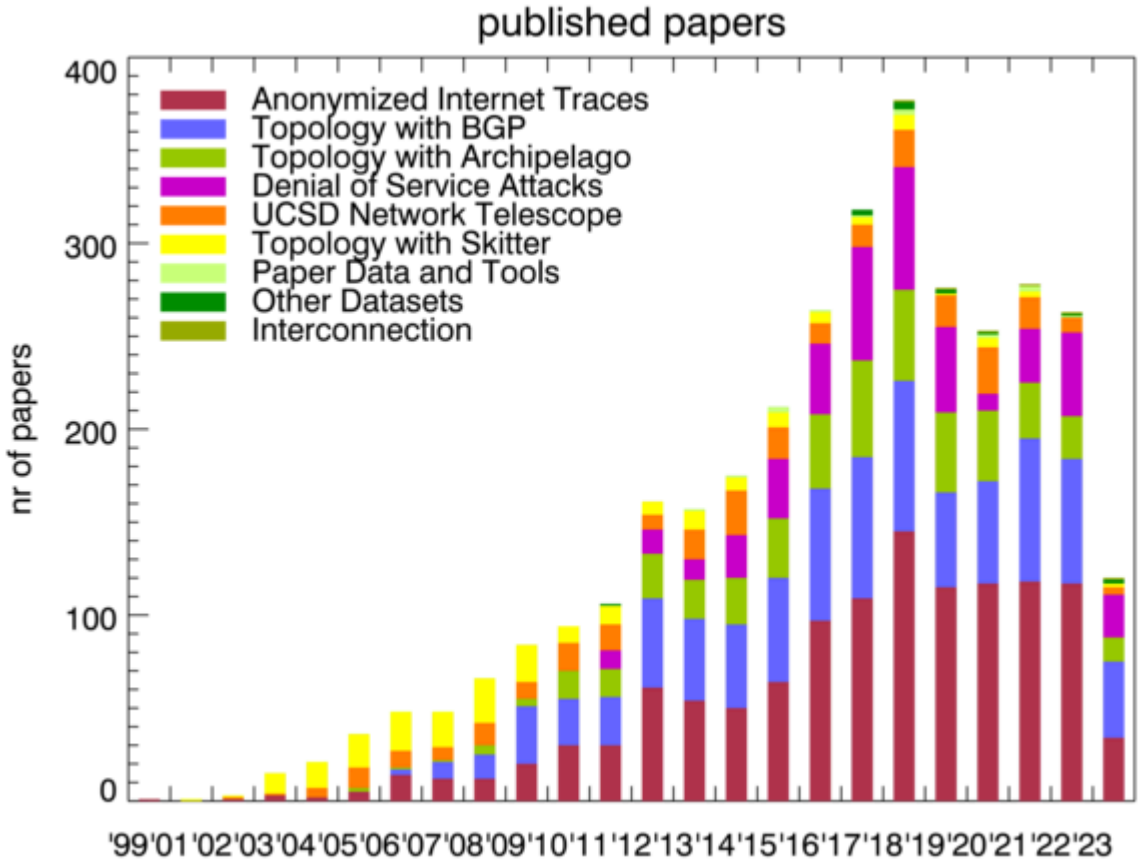


AS Rank requests received annually

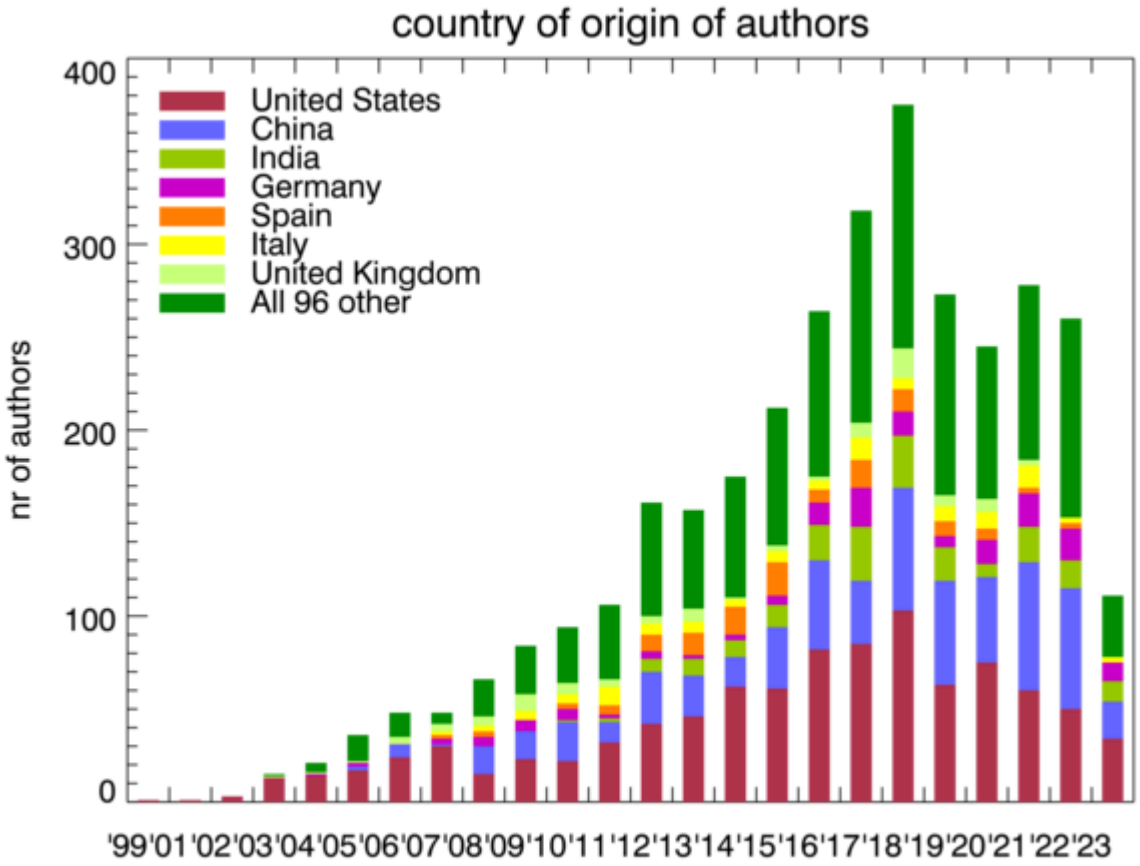


BGPStream requests received annually

Non-CAIDA Publications Using CAIDA Data

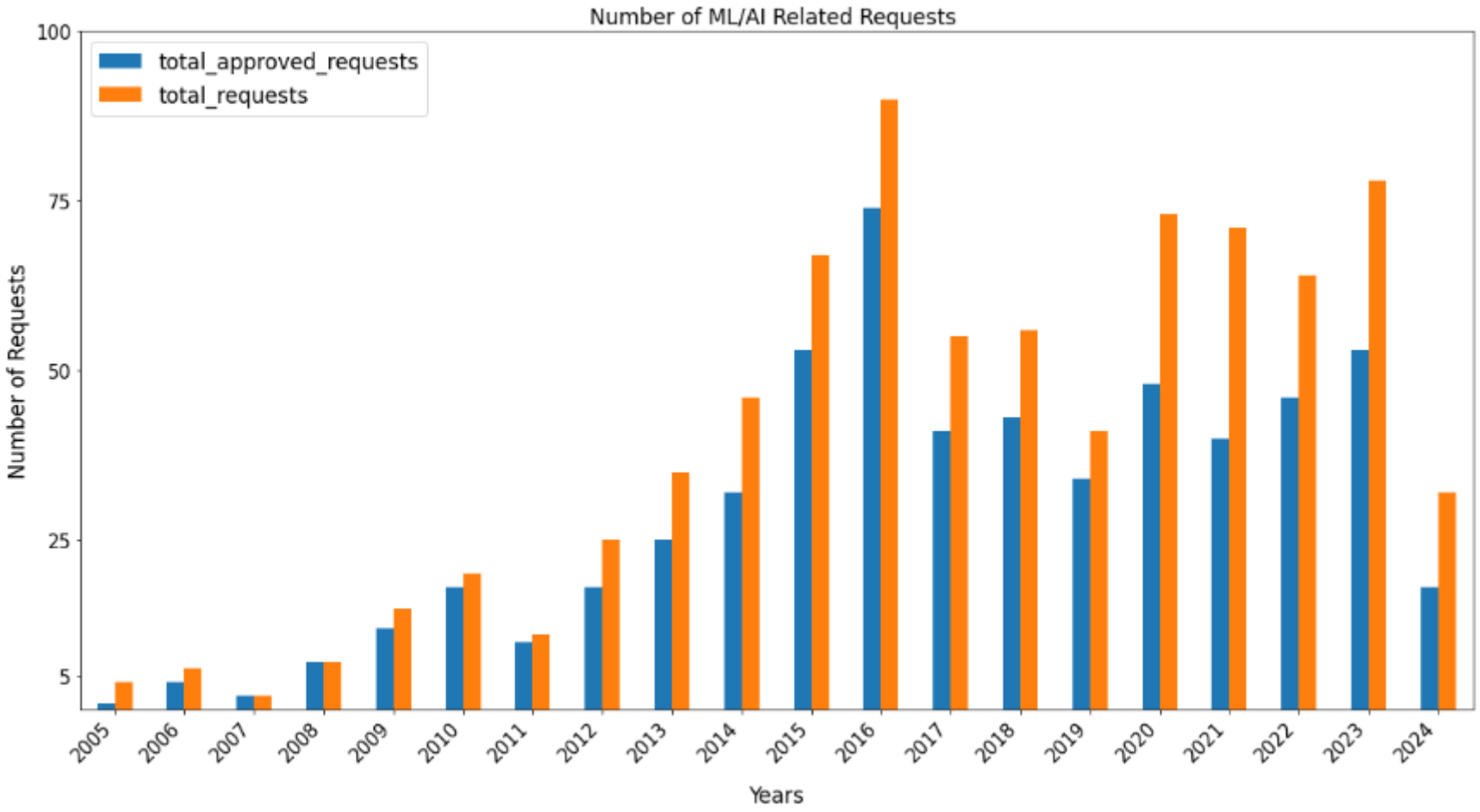


Non-CAIDA publications using CAIDA data (lower bound) As of today we know of > 3500 publications that used CAIDA data



Country of affiliation of first author

ML/AI Related Requests for Restricted Data



Number of requests and approvals each year.

ML/AI Related Requests for Restricted Data

Passive: unsolicited and two-way traffic

- Use open-source LLMs to identify DDoS attack flows in the network in a fine-grained manner.
- Create/compare AI/ML model/algorithms that detect (DoS and other) attacks, classify traffic
- Worm detection and prevention using deep convolutional neural network guided self-attention mechanism
- Optimize networks through AI techniques. Replicate real traffic inside AI simulated network
- Teach network engineers how to use LSTM for anomaly detection and prediction in network traffic
- Apply AI techniques to create a graph model based on available network parameters

Active measurements: Ark topology data

- Neo4j course on graph technology use for analytics: use LLM and RAG to enhance inferred graph
- Refine input locations fed into the IPMap geolocation tool using LLMs, to improve geolocation.
- Build AI model that can extract geo hints from hostnames
- Exploratory ML-based analysis for IP representation learning

LLM Applications for Internet Transport Layer Research

1. **Knowledge sharing** analyzing research papers, threat reports, and security advisories
2. **Identify emerging threats** by analyzing security reports, logs, pcaps, blogs, forums, papers
3. **Infrastructure property classification (network ownership, type, relationships, security attributes)** -- extracting from textual descriptions, measurements, hostnames, topology data, clustering/classification analysis… “ChatBGP” (-- U. Strasbourg project)
4. **Infer semantic meanings of hostnames** by analyzing the context in which they are used , including website content, metadata, and network relationships
5. **Improve geolocation accuracy** by analyzing contextual information associated with round-trip times, topology, routing paths

SDSC coming to the rescue (with LLM infrastructure..)

Please come too!

Summary of Wishlist

How we can help your communities:

1. Data and tools (including for AI/ML..)
2. Curriculum materials
3. Resources to use them
4. Guidance (including connecting to operational experts)

How you can help:

1. **Deploy measurement VPs** (software, peering, our hardware, your hardware)
2. Connect data sets to your infrastructure (step 1: **link to catalog.caida.org**)
3. **Feedback on data** set you use/need

Commercial Collaborators/Supporters

Akamai

iconectiv®

SFMIX

kentik®

CORE

AMPRNet

catnic®

fastly®

IIJ
Internet Initiative Japan

ExpressVPN

CLOUDFLARE®

bgp.tools

NETSCOUT®

GREY NOISE

VERISIGN®

amsix

Google
BigQuery

JSDELIVR