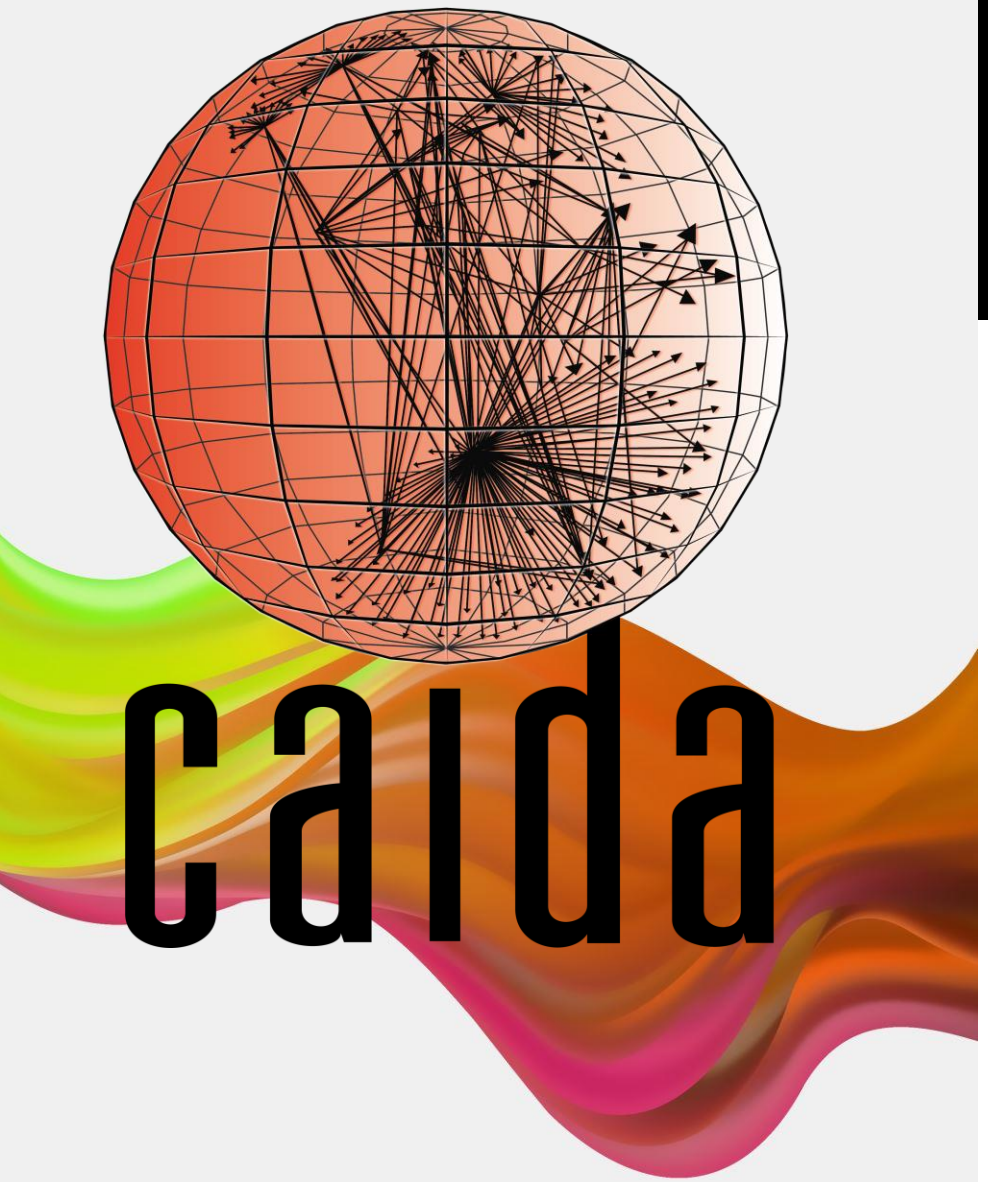


CAIDA'S MEASUREMENT AND DATA INFRASTRUCTURE

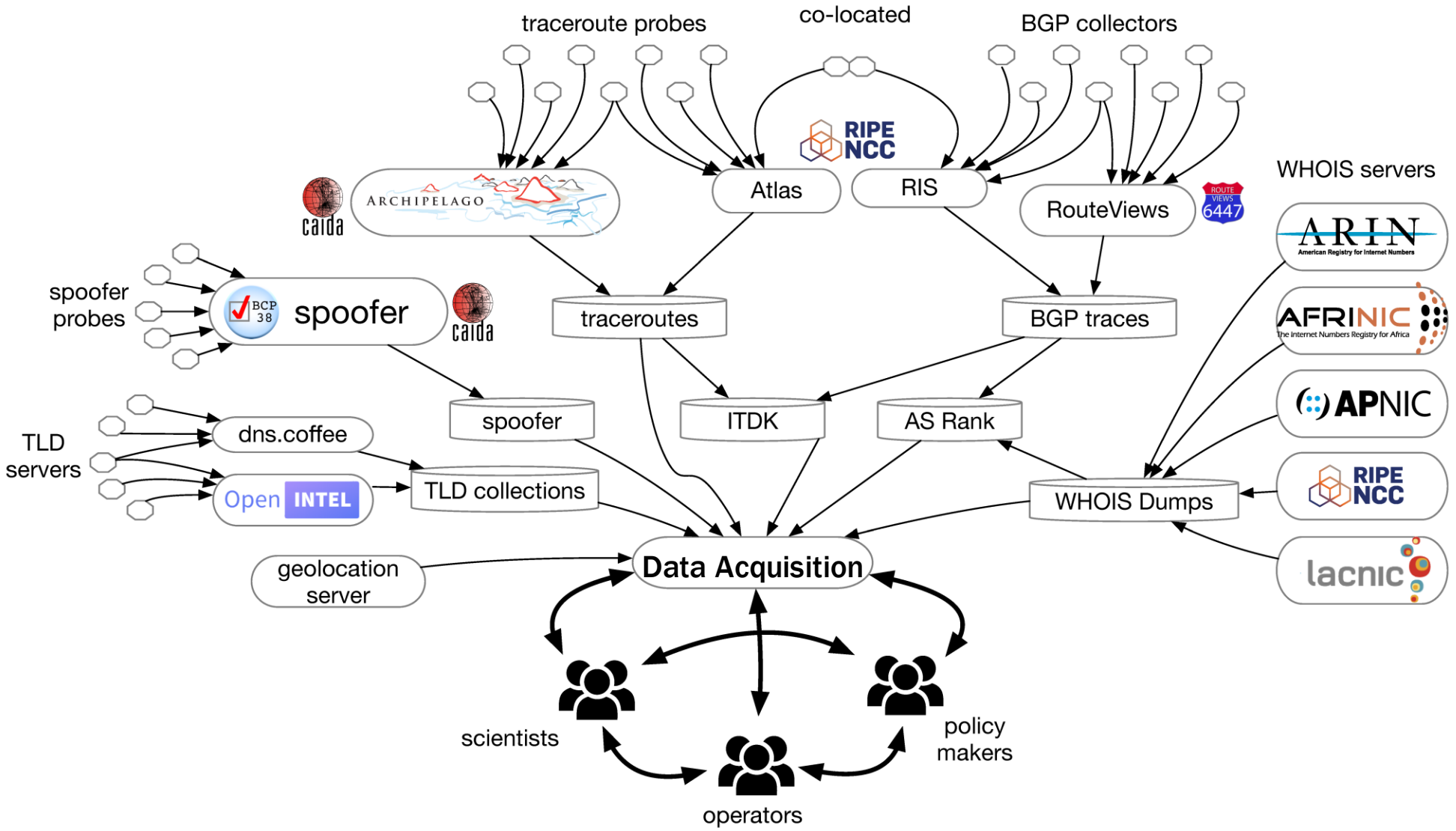
January 2022

kc claffy, CAIDA/UCSD

Bradley Huffaker, CAIDA/UCSD

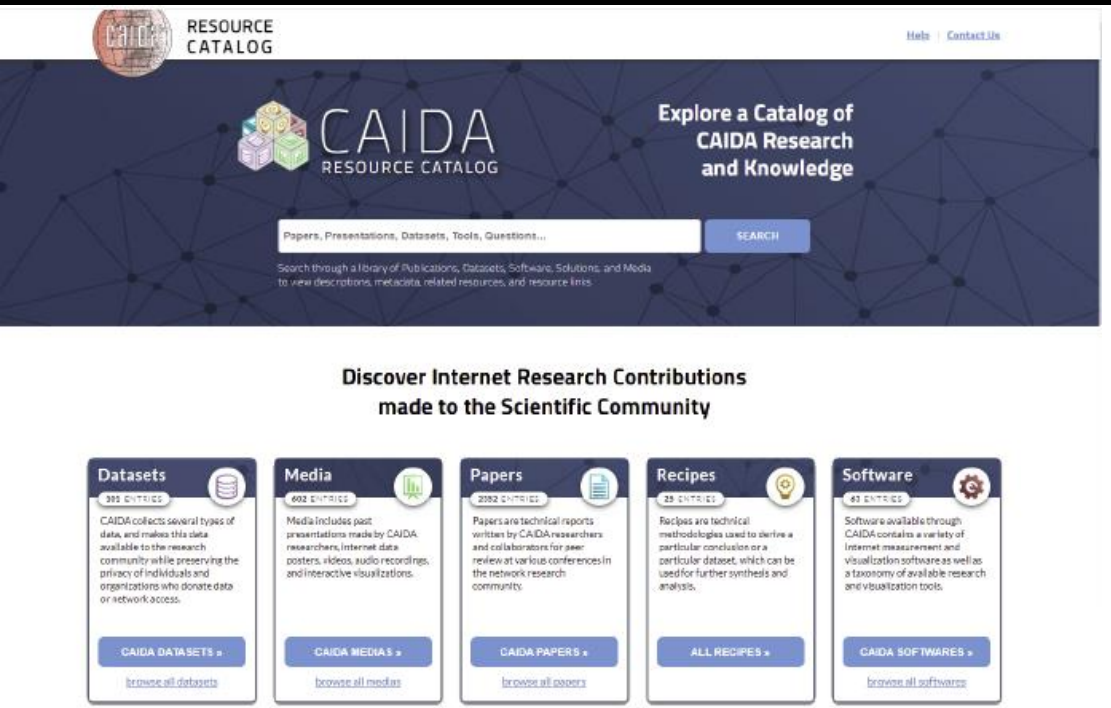
Elena Yulaeva, CAIDA/UCSD

CAIDA DATA ACQUISITION



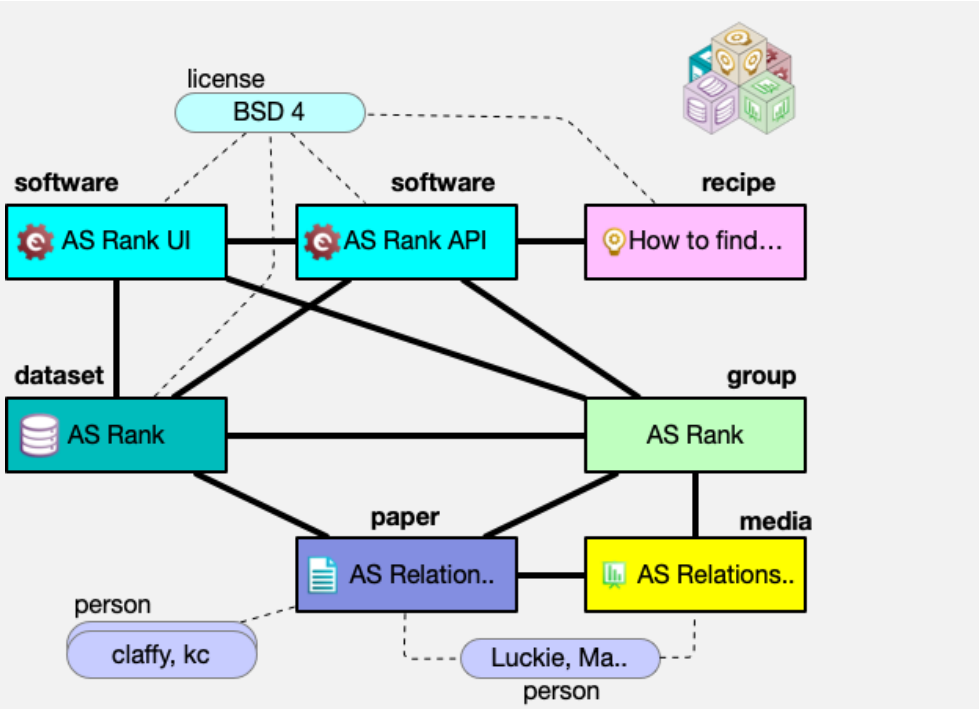
DATA MANAGEMENT -- CAIDA RESOURCE CATALOG

catalog.caida.org



Search through a library of Publications, Datasets, Software, Solutions, and Media to view descriptions, metadata, related resources, and resource links

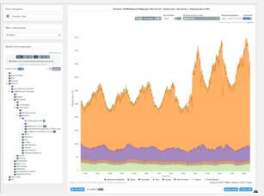
Currently contains 102 Datasets, 607 Media, 2367 Papers, 25 Recipes and 71 Software entries



DATA SHARING

Downloadable files: Public & Restricted (vetted researchers, Acceptable Use Agreements)

Interactive Web Services (e.g.ioda, manic, asrank)



Programmatic access to data streams

BGPSTREAM

APIs

Access to measurement infrastructure

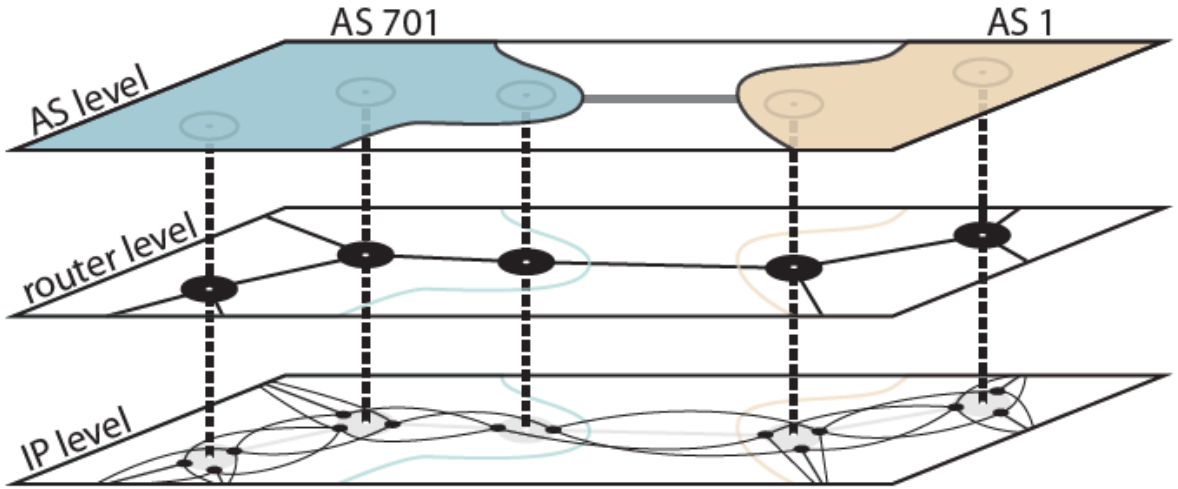
Archipelago

Send code to data

LAYERED INTERNET

Different Internet layers use different resource Identifiers.

CAIDA provides measurements of all these layers.

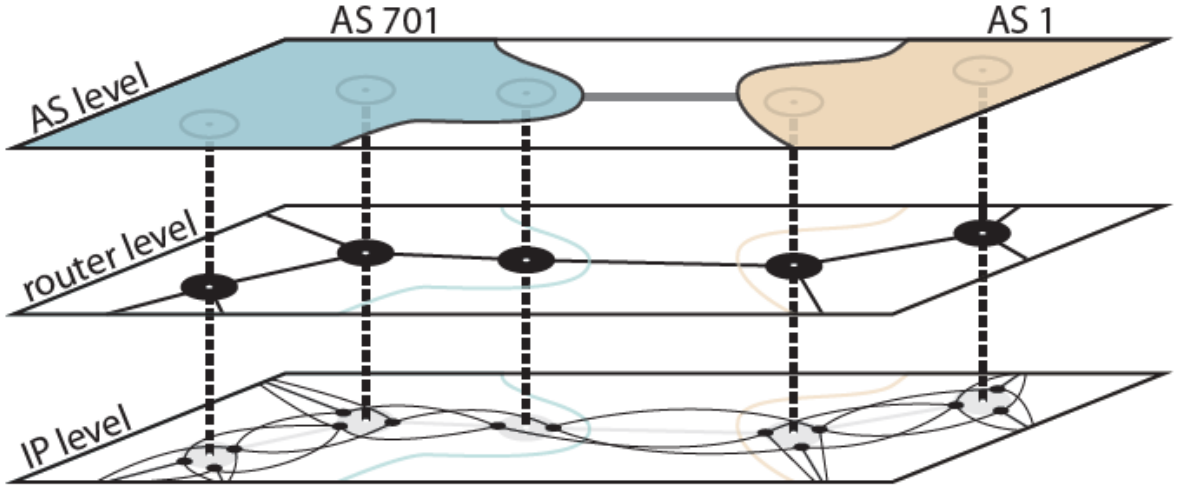


3 layers of Internet topology

IPV4/IPV6 ADDRESSES

Internet Protocol (IP) address is the numerical id used to connect and route machines connected to the Internet.

	IPv4	IPv6
prefix:	4.35.235.0/24	2800:5f0:801::/48
address:	23.8.3.45	231:18f:f53::1

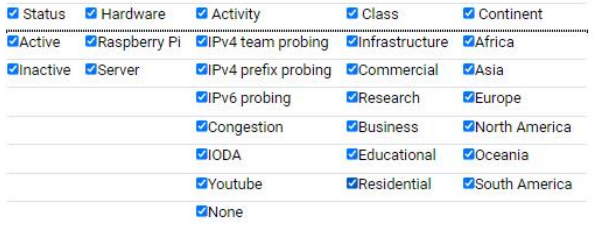


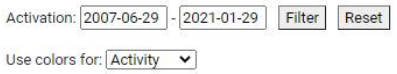
3 layers of Internet topology

ACTIVE MEASUREMENT INFRASTRUCTURE

Archipelago (Ark) Topology Measurement Infrastructure

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





- **Multiple teams of geographically distributed Ark monitors (Raspberry Pi)**
- **219 nodes; 164 AS; 167 cities; 63 countries**
- **Ongoing topology measurements (since 2007):**

IPv4: *traceroute measurement to all the routed /24 networks in the IPv4 address space probing over 10 million /24's*

IPv6: *Each Ark monitor probes all announced IPv6 prefixes (/48 or shorter) once every 48 hours.*

- **On demand measurements:**

3 measurement systems; MANIC, VELA, PERISCOPE

IPV4 & IPV6 PREFIX-PROBING TRACEROUTE

IPv4/IPv6, ark, ongoing, IP topology

https://www.caida.org/catalog/datasets/ipv4_prefix_probing_dataset

Daily traceroutes to one address of each announced BGP prefix.

Each day, we derive a new set of announced prefixes using a sliding 7-day window of BGP data. We then produce daily traceroutes to each announced prefix.

Availability:

- *IPv4 & IPv6*
- *Starts: 2015-12*
- *Update frequency: Daily*
- *Data older than 1 year is publicly available*
- *Access to the most recent data should be requested*
- *Format: warts*
- *RTTs; initial TTL; response IPID, TOS, and TTL, and size; ICMP type/code*

Use cases:

- *Internet topology discovery*
- *Latency, lose rates*

IPV4 /24 PREFIX-PROBING TRACEROUTE

https://www.caida.org/catalog/datasets/ipv4_routed_24_topology_dataset/

Finer grain measurements, every /24 rather than prefix, at slower rate.

Roughly every 3 days, we derive a new set of announced prefixes using a sliding 7-day window of BGP data. We then produce daily traceroutes to each /24 in each announced prefix from a randomly selected monitor.

Availability:

- *IPv4*
- *Starts: 2007-07*
- *Update frequency: roughly 3 days*
- *Data older than 1 year is publicly available*
- *Access to the most recent data should be requested*

Use cases:

- *Topological differences not captured by prefix*

IPV6 /48 PREFIX-PROBING TRACEROUTE

IPv6, ark, ongoing, IP topology

https://www.caida.org/catalog/datasets/ipv6_routed_48_topology_dataset/

Finer grain measurements, every /48 rather than every routed prefix

Traceroute paths to the ::1 address within each /48 in each routed prefix.

Availability:

- *IPv6*
- *Starts: 2008-12*
- *Update frequency: Daily*
- *Publicly available*

Use cases:

- *Topological differences not captured by prefix*

IPV4 ROUTED/24 DNS NAMES

IPv4/IPv6, ark, DNS, hostnames, ongoing, IP topology

https://www.caida.org/catalog/datasets/ipv4_dnsnames_dataset/

https://www.caida.org/catalog/datasets/ipv6_dnsnames_dataset/

Fully-qualified domain names for IP addresses seen in the traces of the IPv4 Routed /24

DNS query and response traffic resulting from the DNS lookups

Availability:

- *IPv4*
- *Starts: 2008-03 (IPv4), 2014-06 (IPv6)*
- *Update frequency: Daily*
- *Data older than 1 year is publicly available*
- *Access to the most recent data should be requested*
- *Format: ascii*
- *Variables: one DNS mapping per line – timestamp IP_address hostname*

Use cases:

- *Router and host hostnames*
- *DNS traffic*
- *DNS servers*

MANIC MEASUREMENT AND ANALYSIS OF INTERNET CONGESTION

IPv4/IPv6, ark, congestion, on demand, IP topology, congestion

https://manic.caida.org/ (web) https://api.manic.caida.org/v1/ (API)

A prototype system to monitor interdomain links and their congestion state.

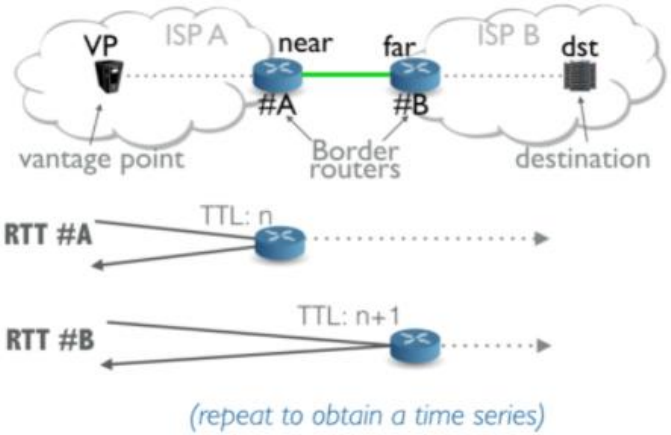
Its dataset is collected using the Time-Sequence Latency Probing (TSLP) method from Ark Vantage points, after running the bdrmap algorithm to automatically infer the IP links (the near IP and far IP addresses) interconnecting neighboring Autonomous Systems (ASNs).

Availability:

- *Grafana front end interface https:/viz.manic.caida.org*
- *--contains dashboards depicting latency time series to the endpoints of interdomain links*
- *Request access manic-info@caida.org*

Use cases:

- *localize and quantify interdomain congestion*



INTERFACE TO PUBLIC MEASUREMENT SERVERS

IPv4/IPv6, on-demand, IP topology

https://www.caida.org/catalog/software/looking-glass-api/

Middleware interface to looking glass server nodes with options to retrieve measurements status, search by type of measurement, and create new measurements.

Publicly-accessible overlay that unifies Looking Glasses (LGs) into a single platform and automates the discovery and use of LG capabilities. The system architecture combines crowd-sourced and cloud hosted querying mechanisms to automate and scale the available querying resources.

Availability:

- *HTTP[S] GET and POST requests to the Periscope API*
- *API responses in json format*
- *Each authorized user is issued a public key and a private key*
- *Request access https://www.caida.org/catalog/software/accounts/periscope_request/*

Use cases:

- *localize and quantify inter-domain congestion*

ON-DEMAND ACTIVE MEASUREMENTS

IPv4/IPv6, on demand, ark, IP topology

https://catalog.caida.org/details/software/vela

On-demand topology measurements on Ark. Users can conduct ping and traceroute measurements in IPv4 and IPv6 using ICMP, UDP, or TCP from any Ark monitors.

Availability:

- *Two interfaces: (1)command-line and (2)web-based*
- *API supports mons,trace,ping,get operations*
- *API responses in json format*
- *Python client code examples*
- *Request account https://www.caida.org/catalog/software/accounts/vela_request/*

Use cases:

- *discovery of the full potential value of massive raw Internet end-to-end path measurement data sets*

MIDAR: MONOTONIC ID-BASED ALIAS RESOLUTION

IPv4, on demand, router

https://www.caida.org/catalog/software/midar/ (web)

https://www.caida.org/projects/ark/vela/midar-api/ (API)

IP alias resolution -- identifies IPv4 addresses belonging to the same router (aliases) using shared monotonic IP ID counters.

The MIDAR web API delivers access to MIDAR's functionality. Relies on backend applications and databases (including ITDK) that implement the job queue and handle execution of MIDAR.

Availability:

- *Downloadable software*
- *Web VELA MIDAR API (a run of 10k takes about 1-2 hours)*
- *Request API access: ark-info@caida.org*

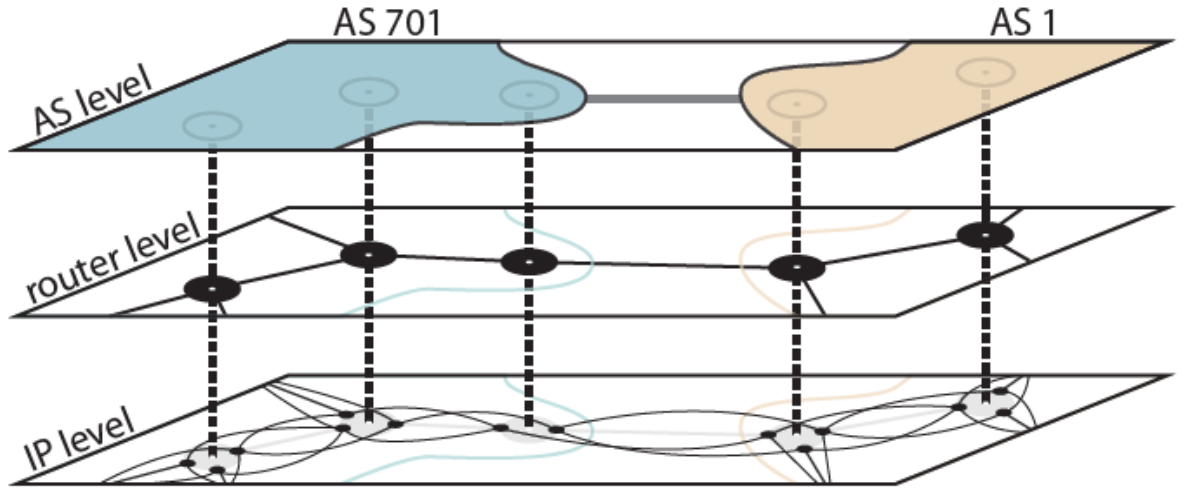
Use cases:

- *Mapping IP addresses to routers*

AUTONOMOUS SYSTEM NUMBER (ASN)

Autonomous System Numbers (ASN) are the macroscopic routing ID. Can be used to map IP addresses to organizations and organization level policies.

ASN	Organization
3356	Level 3



3 layers of Internet maps.

AS RANK

ASN, ongoing, geolocation, bgp, ark, ASN topology

https://asrank.caida.org/ (web) https://api.asrank.caida.org/v2 (API)

Inferred business (routing) relationships between ASes. ASes and organizations are ranked by their customer cone size, which is the number of their direct and indirect customers.

AS Rank is CAIDA's ranking of Autonomous Systems (AS) (which approximately map to Internet Service Providers) and organizations (Orgs) (which are a collection of one or more ASes). This ranking is derived from topological data collected by CAIDA's Archipelago Measurement Infrastructure and Border Gateway Protocol (BGP) routing data collected by the Route Views Project and RIPE NCC.

Availability:

- *GraphQL API – supports mixed record queries*
- *Restful API*

Use cases:

- *AS classification (Transit/Access/Content)*
- *Realistic models of Internet topology, routing, workload, and performance*

AS RELATIONSHIPS

ASN, ongoing, ark, ASN topology

https://www.caida.org/catalog/datasets/as-relationships/

Contains AS links annotated with inferred relationships. Each file contains a full AS graph derived from a set of RouteViews BGP table snapshots, 5-days of Ark traceroutes (serial-2), and multilateral peering.

Availability:

- *Two datasets: (1)* ***serial-1*** *– AS relationships inferred from BGP; and (2)* ***serial-2*** *that contains additional links inferred from the BGP communities and ark traceroutes*
- *IPv4 & IPv6 BGP*
- *Ongoing datasets*
- *Serial-1 available since 2004, serial-2 since 2015*
- *Publicly available for downloads*

Use cases:

- *AS business relationships*
- *Internet topology*
- *Traffic routing*

IPV6 AS LINKS

ASN, ongoing, ASN topology

https://www.caida.org/catalog/datasets/ipv6_aslinks_dataset/

Regular snapshots of AS links derived from the ongoing traceroute-like IP-level ark topology measurements

Data from the IPv6 Topology Dataset are processed by using RouteViews BGP data to identify the Autonomous System (AS) associated with each responding IP address and collapsing the original probed IP paths into a set of links between ASes.

Availability:

- *Ongoing datasets*
- *Available since 2008s*
- *Publicly available for downloads*

Use cases:

- *Studying topology at the AS level*

[OLDER] PREFIX-TO-AS MAPPING

ASN, IPv4/IPv6, ongoing, IP prefixes

https://www.caida.org/catalog/datasets/routeviews-prefix2as/

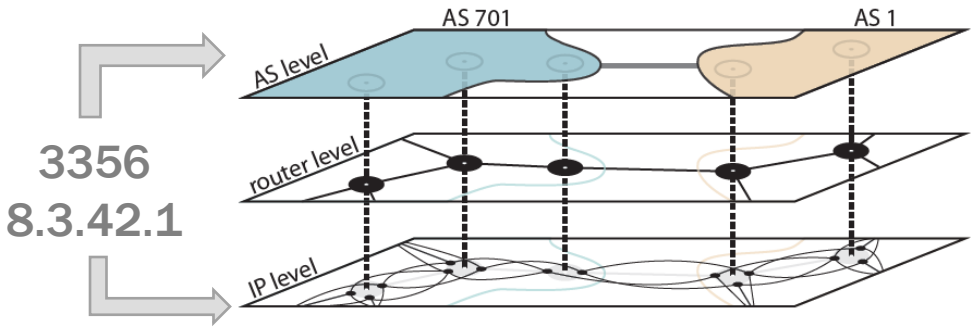
Contains IPv4/IPv6 Prefix-to-Autonomous System (AS) mappings derived from RouteViews data.

Availability:

- *Ongoing daily datasets*
- *Available since 2005 for IPv4 and 2007 for IPv6*
- *Publicly available for downloads*

Use cases:

- *Used to map IP addresses to prefixes and ASNs*

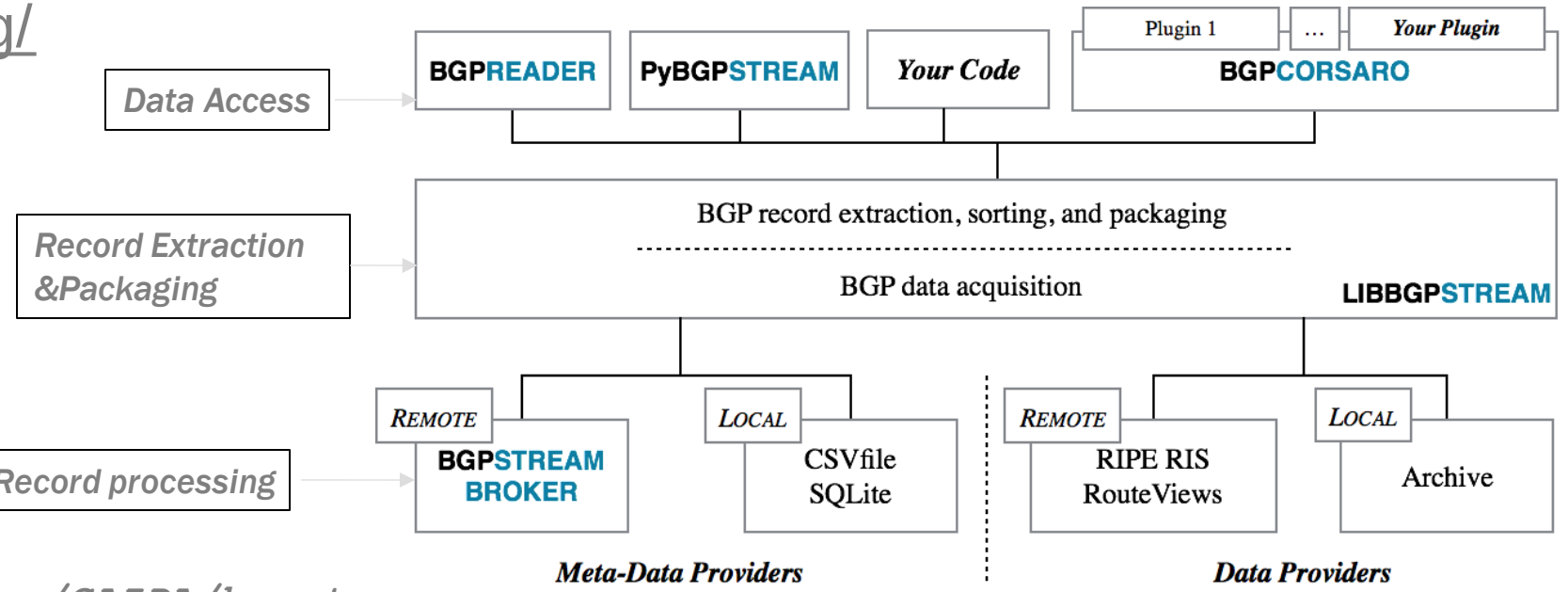


BGP DATA PROCESSING FRAMEWORK

ASN, IPv4/IPv6, IP prefixes, ASN topology

https://bgpstream.caida.org/

An open-source software framework for live and historical BGP data analysis



Availability:

- *Download and install*
- *Clone from* ***https://github.com/CAIDA/bgpstream***
- *2 programming APIs for accessing a stream of BGP data*
  https://bgpstream.caida.org/docs/api

Use cases:

- *Studying Internet topology*
- *Studying Internet security*

BGPSTREAM

BGP DATA PROCESSING FRAMEWORK

ASN, IPv4/IPv6, IP prefixes, ASN topology

caida

https://github.com/CAIDA/bgpview

A set of highly optimized data structures and libraries to facilitate the inference of "Global" routing tables at a finer granularity than the RIB dumps provided by the RouteViews and RIPE RIS projects.

Designed to make it simple to write analysis code by periodically walking through these routing tables and performing analysis on the entire "global" routing table ("inferred" RIB). All of the details of obtaining the raw BGP data, processing it, and inferring the routing table for each peer are abstracted away from the user.

Availability:

- *Realtime and offline operation modes*
- *Install BGPView and its dependencies from CAIDA's apt package mirror https://pkg.caida.org/os/ubuntu/bootstrap.sh*
- *Build from source*
- *BGPView live feed publicly accessible – intall BGPView software on the local server and configure Kafka interface to consume data from CAIDA Kafka cluster http://bgpview.bgpstream.caida.org:9192/*
- *Offline analysis mode – pass views in memory*

https://bgpstream.caida.org/docs/api

Use cases:

- *Topology (e.g. Prefix-to-as mapping)*
- *Detect/analyze snomalous event (e.g. hijacking)*

INTERNET TOPOLOGY DATA KIT (ITDK)

ASN, IPv4/IPv6, router, ongoing, IP/router/ASN topology

https://www.caida.org/catalog/datasets/internet-topology-data-kit

Each ITDK consists of 2 related IPv4 router-level topologies that differ in accuracy and completeness; router-to-AS assignments; geo location ; DNS lookups and observed IP addresses.

We use a subset of the IPv4 Routed/24 Topology dataset containing traceroutes to randomly-chosen destinations in each routed /24 BGP prefix.

Availability:

- *Ongoing: starting 2010-01*
- *Update frequency: 1-2 per year*
- *Data older than 1 year is publicly available*
- *Access to the most recent data should be requested*
- *Format: ascii*

Use cases:

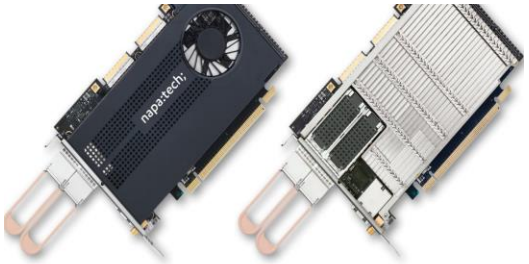
- *Analyze topology at the router-level*
- *Identify prevalence of routers by vendor, region, AS*
- *Labeled data set for use with ML/AI techniques*

ANONYMIZED PASSIVE TRACES

Two-way traffic, passive, anonymized

https://www.caida.org/catalog/datasets/passive_dataset/

Anonymized packet headers from commercial 10GB backbone link. Working to resurrect on 100 GB links (2022)



Availability:

- *April 2008 – January 2019*
- *Monthly one hour traffic capture (pcap format)*
- *Academic researchers upon request*

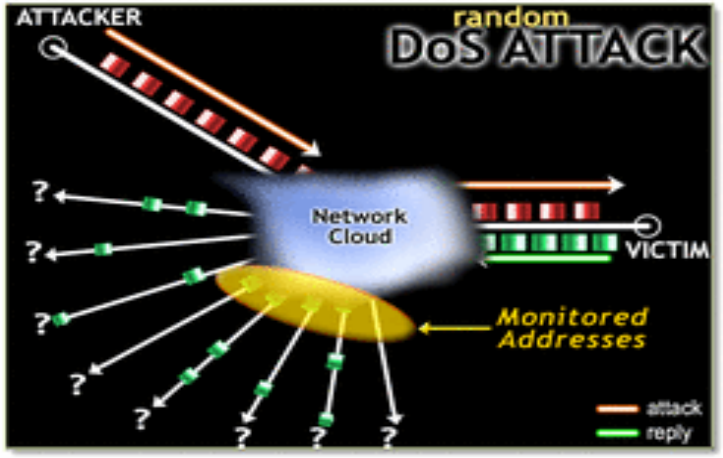
Use cases:

- *Internet traffic classification, especially for p2p application identification*
- *Use neural networks approaches to develop a classifier of peer-to-peer traffic*
- *Flow-based machine learning solution to classification of encrypted traffic*
- *Produce background benign traffic for verifying our ML algorithms on detecting specific flow patterns among production traffic*
- *Construct prototype ML analytics using Kubeflow against pcap data*
- *Use ML to introduce proactive flows in Software-Defined-Networking*
- *Evaluating new network telemetry systems based on programmable switch hardware*
- *Evaluate existing ML approaches for web proxy cache replacement*

UCSD NETWORK TELESCOPE

darknet, passive, unanonymized

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



Globally routed /9 and /10 network with non-allocated IP addresses -- continuous view of anomalous unsolicited traffic (Internet Background Radiation, IBR)

Three datasets available to academic researchers by request. Must be analyzed on CAIDA/UCSD machines. (Working to expand)
Also live stream (nDAG) + time series database with Grafana

Use cases:

- *Detection of DDoS attacks*
- *Test ML model to detect and forecast anomalies in computer networks, particularly software-defined networks*
- *Building a course around big data analysis and machine learning of network traffic to identify security behavior within the network*
- *Identify and categorize behavior of various network intrusions*
- *Detection of address space scans*
- *Botnets*
- *Worms*

RAW TELESCOPE PCAP DATA

darknet, passive, unanonymized, PCAP

https://www.caida.org/catalog/datasets/telescope-near-real-time_dataset/

Raw traffic traces available in near-real time as one-hour long compressed pcap files.

Availability:

- *Most recent 30-day data available on disk*
- *Historical data (starting 2013-07) archived at NERSC*
- *Each pcap contains 1 hour of data, size ~100 GB*
- *Packets are unanonymized and not truncated*
- *Must be analyzed at CAIDA machines (VM, can bring your code)*
- *Analysis libraries installed -- Libtrace, wandiocat*

TELESCOPE AGGREGATED FLOW DATA

darknet, passive, non-anonymized, FlowTuple

https://stardust.caida.org/docs/data/flowtuple/

Aggregated representation of traffic traces - enables a more efficient processing and analysis for many research use cases. Includes meta-data associated with the corresponding source IP address of each FlowTuple

Availability:

- *Ongoing since 2008*
- *Stored in CAIDA Openstack Swift storage*
- *Apache Avro format*
- *Analysis libraries available -- PyAvro-STARDUST or the Pyspark STARDUST API*
- *Must be analyzed on CAIDA/UCSD machines (VM)*

RANDOMLY-SPOOFED DISTRIBUTED DENIAL OF-SERVICE (RSDOS) DATA

darknet, passive, non-anonymized, denial of service attack

https://stardust.caida.org/docs/data/dos

Meta-data of randomly spoofed DoS attacks.
Data generated by processing 5-minute intervals of raw telescope data and extracting response packets sent by victims of RSDoS attacks.

Availability:

- *Daily updates*
- *2008 – August 2021: ascii format*
- *2020-07 – ongoing avro format*
- *Stored in CAIDA Openstack Swift*
- *Must be analyzed at CAIDA machines (VM)*

TELESCOPE LIVE TRAFFIC (NDAG)

darknet, passive, non-anonymized, DAG

https://stardust.caida.org/docs/data/ndag/

Two streams of nDAG (network DAG capture card) packets: raw packets (same as raw pcap) and tagged packets (processed by corsarotagger, additional metadata tags added)

The tags are:

- *Source port (or ICMP type for ICMP packets)*
- *Destination port (or ICMP code for ICMP packets)*
- *Transport protocol*
- *Flow hash value: a 32 bit hash of the fields in the packet that define which flowtuple the packet belongs to.*
- *A bitmask showing which built-in filters were matched by the packet*

Availability:

- *Access by request (via Limbo VM)*
- *libtrace analysis programs pre-configured to run eight processing threads*

TELESCOPE TIME SERIES (GRAFANA)

darknet, passive, non-anonymized,

https://stardust.caida.org/docs/data/timeseries

Includes the least amount of information but is the most efficient and easy-to-use data type.

Vars include:

- ***packets per second***
- ***bits per second***
- ***unique source IPs per minute***
- ***unique source ASNs per minute***
- ***unique destination IPs per minute***

Availability:

- *Access by request (via Grafana Dashboard)*
- *Either create an account or sign in using Github account*

SPOOFER

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

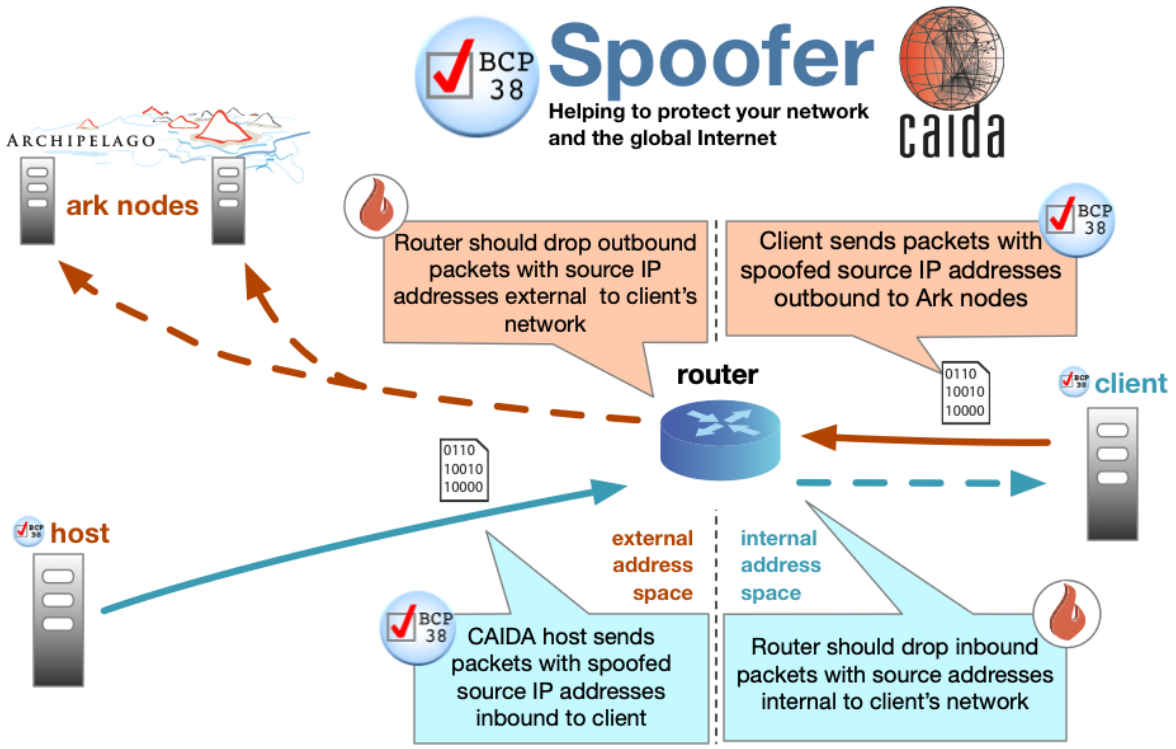
Client-server system that periodically tests a network's ability to both send and receive spoofed packets with forged source IP addresses.

Availability:

- *Downloadable client software*
- *Results public (to /24) by default (opt-out)*
- *Crowd-sourced*
- *Users can request notifications of positive spoofing tests from their networks*

Usage:

- Understanding security hygiene properties
- Assess effectiveness of interventions



AS TO ORGANIZATION MAPPING

Meta-data, mapping, AS topology, WHOIS

https://www.caida.org/catalog/datasets/as-organizations/

Results of applying CAIDAs method – to map Autonomous Systems (AS) to organizations to quarterly bulk dumps of WHOIS databases from the five Regional Internet Registries, ARIN, LACNIC, RIPE AFRINIC and APNIC, and two National Internet Registries, KRNIC and JPNIC

Availability:

- *Quarterly updates (since 2004)*
- *Publicly available downloadable dataset*
- *Publicly available restful API https://api.data.caida.org/as2org/v1/*

Usage:

- *Mapping ASN to organizations*
- *Congestion studies*

INTERNET EXCHANGE POINTS (IXPS) DATASET

Meta-data, mapping, IP topology

https://www.caida.org/catalog/datasets/ixps/

Information about IXPs and their geo locations, facilities, prefixes, and member ASes --derived by combining information from PeeringDB, Hurricane Electric, Packet Clearning House (PCH), and GeoNames

IXP is a physical infrastructure used by Internet service providers (ISPs) and content delivery networks (CDNs) to exchange Internet traffic between their networks (Autonomous Systems - ASes). An IXP can be distributed and located in numerous data centers (aka facilities), and a single facility can contain mutliple IXPs. Each IXP has a prefix, or collection of prefixes, which are used by companies/ASes to address machines within the IXP infrastructure. An AS connected to a given IXP is known as a member of that IXP. Internet traffic exchange through an IXP makes use of Border Gateway Protocol (BGP) that recognizes ISPs and CDNs by their Autonomous System Numbers (ASNs).

Availability:

- *Quarterly updates (since 2018)*
- *Json format*
- *Publicly available downloadable dataset*

Usage:

- *Topology, Geolocation*

IP GEOLOCATION

IPv4/IPv6, IP Prefixes, ASN, geolocation

https://github.com/CAIDA/libipmeta

https://catalog.caida.org/details/software/pyipmeta

Library to support the execution of historical and realtime IP metadata lookups using Maxmind GeoIP, NetAcuity (Digital Element) geolocation and CAIDAPrefix-to-AS databases

Availability:

- *Install from tarball or git clone*

Usage:

- *IP geolocation*

Third party databases:

Maxmind: *https://catalog.caida.org/details/dataset/maxmind* -- Commercial geolocation database –archived at CAIDA (since 2008) for internal use only

Netacuity: Historic archive of Netacuity db files. Files on Netacuity (Digital Envoy) site are checked daily for change – Edge data recalculated and archived at CAIDA since 2011 for internal use only

PeeringDB: *https://www.caida.org/catalog/datasets/peeringdb/* daily snapshots of historic online PeeringDB database (online database of peering policies, traffic volumes and geographic presence of participating networks).

Publicly available for downloads

RESEARCH BASED ON CAIDA DATA

1. Malicious activities (spam, botnets, phishing, censorship)
2. Services (Video on Demand, IP TV, P2P, …)
3. Measurements Methodologies and Tools
4. Traffic Characterization (cloud, businesses, residential)
5. BGP Behavior / AS Topology
6. DNS Behavior and Performance
7. TCP Performance