CAIDA Security Datasets to Support AI Research

Elena Yulaeva, Bradley Huffaker, Ricky K. P. Mok, kc claffy,
CAIDA/University of California San Diego

CAIDA has been actively involved in generating, curating, and disseminating labeled Internet measurement datasets tailored for ML and AI applications, particularly within the domain of Internet transport layer security research. CAIDA has designed and prototyped an open-source catalog (https://catalog.caida.org/) which facilitates the discovery of appropriate datasets and provides links to associated publications, shared code by authors, and relevant metadata. Our talk will highlight several annotated datasets that are often used in ML/AI applications.

Dataset name	Fields/Labels	AI use
Internet Topology Data Kit (ITDK) dataset https://catalog.caida.org/dataset/ark_itdk	Inferred: Router ASN and hostnames, router geolocation, IP interfaces on router.	Infrastructure property classification; Infer semantic meanings of hostnames; Improve geolocation accuracy
Anonymized Two-Way Internet Passive Traces dataset https://catalog.caida.org/dataset/passive_merged_pcap and statistics metadata https://catalog.caida.org/dataset/passive_metadata	Protocol, ports, packet counts, packet size distribution https://www.caida.org/catalog/datasets/trace_stats/	Optimize networks, replicate real traffic, anomaly detection and prediction in network traffic, graph models based on available network parameters.
UCSD Network Telescope raw pcap data https://catalog.caida.org/dataset/telescope_live	Protocol, size, source/destination: IPs, ports, timestamp, TTL	Create/compare AI/ML model/algorithms that detect (DoS and other) attacks, Malware detection/modeling
UCSD NT Telescope Aggregated Flow dataset https://catalog.caida.org/dataset/corsaro_flowtuple	Timestamp. Dest IP network, geolocation of source IP, Source ASN, inferred spoofing classifications	
UCSD NT Telescope Randomly-Spoofed Distributed Denial-of-Service (RSDoS) Attacks dataset https://catalog.caida.org/dataset/telescope_daily_rsdos	Inferred attack metadata: Timestamp. Target IP, protocol, number of /16 subnets that received packets from victim, # of packets attributed to attack, geolocation of target IPs	
UCSD NT Telescope InfluxDB Time Series https://www.caida.org/projects/stardust/docs/data/timeseries/	Geolocation, network protocol, ASN, spoofing classifications	