

Internet measurement data management challenges

Marina Fomenkov and kc claffy
{marina,kc}@caida.org
CAIDA, University of California, San Diego

1. Motivation

Effective Internet measurement raises daunting issues for the research community and funding agencies. Improved understanding of the structure and dynamics of Internet topology, routing, workload, performance, and vulnerabilities remains disturbingly elusive, in part for lack of realistic and representative datasets available to scientific researchers. The dearth is understandable; measurement of operational Internet infrastructure involves managing more complex and interconnected dimensions than measurement in most scientific disciplines: logistical, financial, methodological, technical, legal, and ethical. CAIDA has been navigating these challenges with modest success for fifteen years, collecting, coordinating, curating, and sharing data sets for the Internet research and operational community in support of Internet science. Our three current biggest challenges which we hope to explore at the workshop are: sustainable collection, curation, and storage of large volumes of data; privacy-respecting sharing; and long-term archiving for reproducibility.

2. Overview of CAIDA data

A complete list of the data that we currently collect and offer to the research community is available at [1]. It includes both ongoing data collections, data sets covering a single event, and archived collections. As of 2011, CAIDA captures and curate datasets from three primary sources: (i) macroscopic topology data (both IPv4 and IPv6) with the Archipelago infrastructure; (ii) traffic traces at Internet core backbone links; and (iii) traffic traces from the UCSD Network Telescope. Table 1 shows the amount of data resulting from our ongoing data collection operations in 2010.

2.1 Active measurement data

Funded by NSF and DHS, Archipelago (Ark) is CAIDA's newest active measurement infrastructure. As of June 2010 it

2010 Data Sets	Size	Compressed
IPv4 Routed /24 Topology	1.6 TB	509.2 GB
DNS Names for IPv4 Routed /24 Topology	24.2 GB	6.3 GB
AS Links for IPv4 Routed /24 Topology	500.7 MB	124.2 MB
Macroscopic Internet Topology Data Kit (ITDK)	13.5 GB	2.6 GB
IPv6 Topology	1.8 GB	519.2 MB
Internet backbone traces	6.9 TB	4.1 TB
Network Telescope Data	61 TB	33 TB
DNS root/gTLD RTT Dataset	762.6 MB	762.6 MB

Table 1: Data CAIDA Regularly Collected in 2010

consisted of 54 PCs deployed around the world, operating as a coordinated secure platform capable of performing various active Internet measurements. We continue to extend Ark in geographic scope as well as function. With Ark we generate and share the following data sets.

IPv4 raw traceroute data. Since September 2007, we have been using the Ark platform to support ongoing global Internet topology measurement and mapping. To our knowledge, Ark gathers the largest set of IP topology data for use by academic researchers [1]. Ark monitors continuously measure IP-level paths to a dynamically generated list of IP addresses covering all /24 prefixes (about 7.4 million) in routed IPv4 address space. Measurement parallelization allows us to cycle through probing each routed /24 prefix in about two days. Over the lifetime of Ark, we have collected more than 4 billion traceroutes (1.6 TB of data).

DNS annotations. We execute DNS lookups of all IP addresses seen in the Ark IPv4 traceroutes. Using a customized bulk DNS lookup service that is capable of millions of DNS lookups per day, we attempt DNS lookups as soon as possible after we collect topology data (within 1-2 days) since host names may change. This collection system yields two datasets: (i) a simple IP-to-hostname map; and (ii) raw DNS query/response traffic generated by the lookup service. DNS annotations are valuable for many analyses as a host name often indicates machine type, organizational affiliation and geographical information.

Derived data sets: router-level and AS-level IPv4 Internet graphs. CAIDA's topology measurement project discovers IP interfaces as components of traceroute-inferred forward paths. To make this data more useful, we estimate which of these interfaces belong to the same router, a process called *alias resolution*. We developed a tool to perform alias resolution with unprecedented accuracy and completeness at Internet-scale (millions of addresses), allowing us to derive router-level Internet graphs from traceroute and publicly available BGP data.

Our AS-level topology map is another data set derived from raw traceroute measurements. Using publicly available BGP data, we map the IP addresses in gathered IP paths to the AS numbers that advertise the longest IP prefixes matching each IP address. If two consecutive IP hops in a trace resolve to different ASes, we interpret it as a link between these ASes. The set of these links constitutes an AS-level topology graph. We post the adjacency matrix of the Internet AS-level graph on a daily basis.

We then annotate the links and nodes in our AS-level graphs to facilitate further research and analysis. AS node annotations may label different types of ASes, e.g., large or small Internet Service Providers (ISPs), exchange points, universities, customer enterprises, etc. AS link annotations represent business relationships between AS nodes, e.g., customer-

to-provider, peer-to-peer. To infer these AS relationships, we use multi-objective optimization heuristics developed at CAIDA and elsewhere. Our most recent topology data set offering is the Internet Topology Data Kit (ITDK) [1], which includes a stand-alone set of raw IPv4 topology data traces collected over a certain period of time (usually, a two-week window) and all corresponding derived and annotated graphs.

IPv6 raw traceroute data. We started measurements of IPv6 topology in December 2008, using six IPv6-capable Ark monitors. In June 2011, 26 of our monitors were IPv6-capable and conduct continuous probing of BGP-announced IPv6 prefixes (/48 or shorter, nearly 4,000 prefixes as of December 2010). Each Ark monitor probes a single random destination in each prefix; a full probing cycle takes 48 hours. We plan to expand measurements of IPv6 topology and performance, including enabling DNS mapping for IPv6 topology data.

2.2 Passive (traffic) measurement data.

Core link traffic data. The logistical, financial, and technical obstacles to collecting and managing Internet data are most acute for traffic data. Security, privacy, and legal concerns, as well as the cost of monitoring equipment which must be upgraded every few years to keep up with changes in the underlying infrastructure, severely limit our options for collecting and sharing traffic data. Currently, CAIDA hosts data collected by two passive monitors on U.S. Tier 1 backbone links at Internet exchange points in San Jose, CA and Chicago, IL. We continuously post near-real-time graphs of application breakdown and geographic information about sources and destinations of the observed traffic [1]. Additionally, every month we attempt to capture a one hour trace on each link (typically, a few hundred GB of data), clean and anonymize the data, and make these traces available to researchers, albeit with significant policy restrictions on use.

Unsolicited traffic data. Network telescopes have been used to observe Internet *background radiation*, i.e. unsolicited traffic sent to unassigned address space. The routing system carries the traffic to this *darkspace* because the address space is being announced globally, but there is no response back to the traffic sources since there are no hosts in the darkspace to respond. In June 2011 the UCSD Network Telescope captured about a hundred gigabytes of compressed trace data per day. Until this year we were packaging and releasing anonymized traces from the UCSD Network Telescope. To increase the data's utility to cybersecurity researchers, we are now trying to extend our real-time reporting methods to support early detection and visualization of changes in the character of unsolicited traffic, and experimenting with providing vetted research analysts with near real-time access to the most recent 30 days of raw packet traces from the telescope.

3. Sharing CAIDA data

3.1 Access policies

CAIDA's approach to data sharing is guided by the goals of improving the integrity of Internet science, respecting the sensitivities in different types of data, and navigating the associated technology, legal, and ethical challenges. Efficient,

appropriate, and flexible disclosure control techniques facilitate the development and validation of scientific models. Some of our datasets require that users agree to an Acceptable Use Policy (AUP), but are otherwise freely available. In 2010, the most popular datasets in these category, the Code-Red Worm Dataset and AS Relationships, were accessed by 1259 and 1189 unique visitors (machines), correspondingly.

Access to other datasets is restricted to academic researchers (except in a few countries with U.S. export restrictions), U.S. government agencies, and CAIDA members. The access is also subject to AUPs designed to protect the privacy of monitored communications, ensure security of network infrastructure, and comply with agreements with our data providers, when applicable. The two most popular 2010 datasets in this category were: Anonymized Internet Backbone Traces (185 requests) and Active Topology (including IPv4, IPv6, and ITDK) datasets (163 requests). We received about 14% more requests in 2010 than in 2009, and approved 20% more. Almost 80% of the vetted users that were granted access in 2010 actually download data from our web servers.

3.2 User support infrastructure

We support user requests and inquiries by maintaining extensive web-based data distribution services. When users request data via a web form they receive immediate automated acknowledgment of the request and are subscribed to a corresponding mailing list that provides a direct communication channel for user feedback. CAIDA maintains several mailing lists of researchers (organized by the type of data they requested) as well as a public list for general announcements regarding CAIDA data. CAIDA staff also regularly answers questions sent to data-info@caida.org.

Our data administrator generally responds to data requests within three business days, with more than 50% of queries being answered within one day. Per our usage agreements for each protected dataset, we conduct periodic (at least annual) surveys of our data users to request a summary of research results and pointers to any resulting publications. We also solicit feedback on the usability of our datasets, any difficulties users had with the data, and other datasets researchers would like to analyze. Resources allowing, CAIDA makes custom datasets available to researchers with special requests.

3.3 Research enabled by CAIDA data

A broad research community has benefited from our active topology measurements for over a decade. Our web site lists known publications by non-CAIDA authors using CAIDA data [2]. Researchers have requested and downloaded topology data to support research in the areas of: routing on overlay networks; routing policy; modeling IPv4 and IPv6 AS-level topology and BGP behavior; alias resolution and router-level topology discovery; improving anycast implementations; metrics for describing scale-free networks; evaluating router responsiveness to probes; peer-to-peer system scalability; improving visualization of complex systems; geolocation; modeling of delay; improved traceback for network attacks; and improved packet marking/filtering. Our Internet topology data also supports non-CS-related fields that study complex networks, such as physics, biology, and finance.

The backbone traffic data we host has been used for studies in Internet traffic classification and modeling, performance

modeling, monitoring and filtering techniques, intrusion detection, and traffic generation. Our UCSD Network Telescope data has supported studies of DOS attacks and various Internet worms and their victims. Data from our telescope was also used to parameterize round a model of the top speed of flash worms, estimate the “worst-case scenario” economic damages from such a worm, and analyze the pathways of their spread and potential means of defense.

4. Challenges and Open Issues

We are dealing with three interrelated challenges in data-intensive Internet research: sustainable collection, curation, and storage of large volumes of data; privacy-respecting sharing; and long-term archiving for reproducibility.

4.1 Sustainable collection, curation & storage

The volume of data accumulated by CAIDA are becoming prohibitively expensive to store, limiting the number of researchers who can make use of the data. The situation is worse during malicious activity outbreaks when data volumes can increase sharply, yet rapid analysis and response are necessary. The speed, scope, and strength of today’s automated malicious software demand real-time sources of data that can match the dynamics of the threat. Although the technical and policy obstacles are intimidating, we are now experimenting with near real-time sharing of live traffic data.

To support even a few researchers with real-time access to traffic data, the compute and storage systems must have reliability and performance characteristics more often associated with high-transaction commercial enterprises than academic institutions. These systems require large file systems built with redundancy, reusable parts and hot spares. Associated compute servers must handle multi-terabyte analysis, with reliable uptime and timely job completion. Administration of such resources requires dedicated system administrators with experience managing data processing pipelines. As an example, this year we had to move away from standard operating system tools such as *fsck* to check the consistency and health of file systems because the sizes of the underlying disk partitions exceed the available memory needed by the tool.

The economic implications also demand continual consideration, although we still have no formal methods to perform cost-benefit analyses of storage and archiving demands for a given set of data or research. Experience has shown us that user interest in any particular dataset decreases with time, yet longitudinal studies require long-term historical data. Balancing usability versus costs to ascertain which data is “worth keeping” past a given research project’s lifetime – and who pays for the archiving costs – is an open problem in our field.

4.2 Privacy-sensitive sharing

We have developed a privacy-sensitive data-sharing framework that integrates the best available (anonymization) techniques to protect privacy without obliterating all utility in the data, with a policy approach that applies standard privacy principles and obligations of researchers data providers. We use this framework to evaluate our own and proposed data-sharing techniques along two primary criteria: (1) how they address privacy risks; and, (2) how they achieve utility objectives [3]. Recognizing that privacy risk management is a

collective action problem, our framework contains this risk by transitively replicating the collection, use, disclosure and disposition controls to any user of the data.

External advisory groups such as Institutional Review Boards (IRB’s) may serve a key function in the framework (CAIDA’s approved IRB application is available on our web site), although many IRBs are not yet equipped to evaluate Internet research that does not involve direct human interaction but involves human activity, e.g. observation of web transactions, email, or generic traffic. Open questions include: What are users’ perceptions of privacy and confidentiality in network traffic, and how are they changing? What are the legal constraints on collecting and disclosing network data for research purposes? How does one identify a potentially at-risk population in a network trace?

4.3 Archiving for reproducibility

Per NSF’s new Data management policy [4], NSF scientists are now required to provide a 2-page supplementary document to describe plans for managing and sharing the data products of proposed research. We view this policy as a fine starting point, but NSF must seek and embrace more concrete metrics for evaluation of a given plan, e.g., how reproducible is the work? (and how do we define reproducibility?)

The reproducibility expectation raises many related issues. Who and how should determine a data set’s lifetime? If a particular restricted-use dataset was analyzed and published in a paper, and the Researcher is not allowed to archive (or perhaps even see) the data, whose responsibility is it to maintain this dataset to allow reproducibility of the scientific result? As data volumes grow, who should pay storage and archiving costs for scientific work more than 1, 5, 10 years old? When is purging the raw data underlying our discoveries justified? Is it time to develop policy guidelines that uphold a reasonable standard of scientific reproducibility in our field, and ease decision-making for researchers, data curators, and system administrator?

ACKNOWLEDGMENTS. CAIDA data collections are supported by NSF CNS-0958547 and OCI-0963073 and DHS (S&T) N66001-08-C-2029 and NBCHC070133. The views expressed here represent only the authors.

5. References

- [1] Cooperative Association for Internet Data Analysis. CAIDA Data - Overview, April 2011. <http://www.caida.org/data/overview/>.
- [2] Cooperative Association for Internet Data Analysis. Non-CAIDA Publications using CAIDA Data, June 2011. <http://www.caida.org/data/publications/bydate/index.xml>.
- [3] Erin Kenneally and Kimberly Claffy. Dialing Privacy and Utility: A Proposed Data-sharing Framework to Advance Internet Research. *IEEE Security and Privacy (S&P)*, July 2010. http://www.caida.org/publications/papers/2009/dialing_privacy_utility/.
- [4] National Science Foundation. NSF Data Management Plan Requirements, 2010. http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.