2007-2459
Revised: October 7, 2009

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

Executive Summary: Research and development targeted at identifying and mitigating Internet security threats requires current network data. To fulfill this need, the Cooperative Association for Internet Data Analysis (CAIDA), a program at the University of California's San Diego Supercomputer Center which is based at the University of California, San Diego (UCSD), will collect packet header data from large backbone ISPs (so long as we have access to links, which is not guaranteed) and the UCSD Network Telescope, IPv4 and IPv6 topology data, and real-time monitors to view traffic on monitored links. We will curate this data, in some cases anonymize, and distribute to the network and security research community. In light of progress and pitfalls encountered in the first two years of this process, and in the face of increased concerns over policy obstacles to cybersecurity research, in September 2009 we re-aligned our statement of work to better support what PREDICT needs to accomplish in the next two years -- community building and demonstrated responsiveness to current public and private sector needs in Cybersecurity S&T research. We have replaced fixed data set collection intervals with a more flexible approach designed to better meet the current needs of researchers, including eventual access to real-time traffic data from the telescope for vetted security researchers. Many of our deliverables have changed in support of these new objectives.

Technical Approach:

We now critically depend on the Internet for our professional, personal, and political lives. This dependence has rapidly grown much stronger than our comprehension of its underlying structure, performance limits, dynamics, and evolution. Fundamental characteristics of the Internet are perpetually challenging to research and analyze, and we must admit we know little about what keeps the system stable. As a result, researchers and policymakers currently analyze what is literally a trillion-dollar ecosystem essentially in the dark, and agencies charged with infrastructure protection have little situational awareness regarding global dynamics and operational threats. To make matters worse, the few data points suggest a dire picture, shedding doubt on the Internet's ability to maintain and strengthen its role as the world's communications substrate.

The current lack of data documenting both malicious and benign Internet traffic impedes security threat mitigation efforts because there are;

- no realtime datasets available to allow those responsible for high-security sites to differentiate between general attacks and those targeting their installations,
- no easily available traces containing traffic from current high-speed networks to use in development, testing, and comparison of mitigation technologies,
- limited availability and evaluation of tools for anonymizing data sets for protected sharing with researchers.

The state-of-the-art in the development of security technologies could be improved through coordinated data collection and distribution efforts. We propose the collection, curation,

anonymization[1], and distribution of Internet data to support research and development activities, and to participate as a Data Provider and Data Host in the Protected REpository for the Defense of Infrastructure against Cyber Threats (PREDICT) program.

This basic fundamental research is being performed on a reasonable efforts basis.

CAIDA's network data collection capabilities include:

Passive network monitors

Each monitor consists of a pair of 2-unit servers instrumented with either an off the shelf NIC or an Endace DAG high-performance data collection card. The servers are time-synchronized with stratum-1 time servers to allow comparison of trace data collected at disparate locations.

The Archipelago (Ark) active measurement platform:

Ark is a new platform designed, developed, and deployed by CAIDA for optimized, coordinated active network measurements. In September 2009 we had 35 Ark monitors, 10 of them IPv6-capable. Ark supports a variety of macroscopic Internet active measurement projects, including the `scamper` IPv4/v6 topology discovery tools. Existing skitter monitors were upgraded to Ark monitors during the first year of this project.

The UCSD Network Telescope:

The UCSD Network Telescope consists of a large piece of globally announced IPv4 address space. The telescope contains almost no legitimate hosts, so inbound traffic to nonexistent machines is anomalous in some way. Because the network telescope contains approximately 1/256th of all public IPv4 addresses, we receive roughly one out of every 256 packets sent by an Internet worm with an unbiased random number generator. The telescope has enabled us to provide a unique global view of the spread of some Internet worms. The advent of the Conficker worm and its associated traffic load on the telescope has changed our approach to sharing telescope data. In 2009 we are transitioning from a model of static trace sharing (which is of extremely limited utility to researchers, especially when anonymized as it must be for Phase 1 of PREDICT) and indefinite storage of data on CAIDA servers, to a model of real-time data sharing with vetted researchers, but only storing a 30-day window of history. Many of the new tasks and deliverables below are intended to overcome current obstacles to achieving this objective, or are otherwise responsive to current public and private sector needs in Cybersecurity S&T research.

Adaptive Netflow:

Adaptive NetFlow, deployable through an update to router software addresses shortcomings of NetFlow by dynamically adapting the sampling rate to achieve robustness without sacrificing accuracy. Thus collection infrastructure remains intact during flooding attacks, sampling rates are automatically tuned to data volume, and flow data reporting interacts well with applications that operate on time-binned data. Adaptive NetFlow has been incorporated into CoralReef, CAIDA's passive measurement software suite, and is available for data collection on high-speed links.

1 Or other appropriate measures of privacy protection.

We will provide access to data for researchers in several ways in accordance with UCSD policy. We will maintain one or more data servers to allow researchers to download data via secure login and encrypted transfer protocols. We will receive, format, transfer data to, and return USB hard drives to researchers who wish to access datasets whose volume prohibits timely data download via the Internet. We will provide a near-realtime, interactive graphical interface to passive monitors and the Network Telescope to allow researchers a continual view of statistics of Internet traffic on these links and to allow them to identify time periods containing traffic characteristics of interest for further investigation using captured traces. Finally, we will experiment with a model for real-time sharing of the telescope data/monitor, using a new framework we will develop for balancing privacy and utility in Internet research.

Statement of Work:

We propose to pursue the installation of existing equipment to monitor OC48 and OC192 links, including running the CoralReef report generator and collecting packet traces, as allowed by link owners and Data Providers.

We propose to collect, process, and distribute data from the following sources:

- Internet backbone and peering links (when links and monitors are available). Will include raw traces and statistical summaries.
- The UCSD Network Telescope, with real-time data sharing supported under a new manuscript entitled, "Internet Data Sharing Framework For Balancing Privacy and Utility", which we will also publish in 2009.
- scamper running on the Ark infrastructure, collecting IPv4 and IPv6 network topology as discovered via continuous, active traceroute-like probing (including all /24 networks of the IPv4 address space). In conjunction with BGP routing tables from RouteViews or RIPE, this data allows us to create and serve Autonomous System (AS)-level topology graphs updated weekly (for use in virus, worm, botnet spread propagation research, routing security database support, infrastructure stability and vulnerability analysis).
- Realtime (or close to real-time) detailed traffic reports (from CoralReef report generator software) from any available OC48/GigE, OC192/10GigE links (subject to the approval of the Data Provider), and from the UCSD Network Telescope, to provide data on current threats and help researchers identify periods of interest in collected trace data.

We will also continue to distribute previously collected data of observable interest to researchers, based on user requests. Historical data will allow previously impossible longitudinal analysis of threat evolution over the last several years.

We will help evolve the PREDICT program with updated Memoranda of Agreement to allow sharing of telescope data in real-time in accordance with UC policy. We will help develop appropriate PREDICT infrastructure to serve the evolving needs of the research and development communities.

Sharing of sensitive network data with researchers is almost always blocked on the need to protect personally identifying information, but there has been little attention thus far by the

research community in analyzing and comparing existing anonymization schemes for data leakage and other performance characteristics. We will investigate current and proposed anonymization schemes that support PREDICT's goal to protect privacy while supporting cybersecurity research. In the first year (2008) we will make available via the web an initial taxonomy of known tools, techniques, related publications and known issues. We will also provide suggestions to PREDICT data providers on the use of current and future data anonymization schemes to increase security and privacy. We will update this web page and set of suggestions as technology develops in future years of the project.