

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 currently network data. To fulfill this need, 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 backbone/peering point data from large ISPs (depends on access to links which is not guaranteed), trace data from the UCSD Network Telescope, datasets on past (Code-Red, Witty) and future Internet worms, data on the IPv4 and IPv6 topologies, and realtime monitors to view traffic on monitored backbone/peering links and the UCSD Network Telescope. This data will be curated, anonymized, and distributed to the network security community.

Technical Approach:

Over the past two decades, the Internet has become critical infrastructure for almost every aspect of American life. Commerce, business, government, education, and even interpersonal relationships rely on networked computers for communication and data distribution. Yet the discovery of new security threats continues to outpace the development of new technologies to ensure the security, integrity, and privacy of digital information. The current lack of data documenting both malicious and benign traffic traversing the Internet impedes security threat mitigation efforts because there are:

- few or no ground truth examples of neoteric attacks in the wild, so focusing research and development to target current threats remains difficult
- no realtime datasets available to allow those responsible for high-security sites to differentiate between general attacks and those specifically targeting their installations
- no easily available traces containing traffic from current high-speed networks to use in development and testing mitigation technologies to minimize both false negatives and false positives in deployed infrastructure
- no canonical data sets with which to compare the efficacy of competing technologies that promise to detect or respond to a given threat
- 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, anonymization¹, and distribution of Internet data to support research and development activities, with the goal of eventual participation as a Data Provider and Data Host in the Protected REpository for the Defense of Infrastructure against Cyber Threats (PREDICT) program. PREDICT provides thoroughly vetted central

¹ Or other appropriate measures of privacy protection.

infrastructure designed to maximize ubiquitous data access while ensuring data security and privacy.

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 one time servers to allow interpolation of trace data collected at disparate locations. Currently one OC12 link at AMPATH is monitored, as well as several GigE links at UCSD. Data from the UCSD links cannot be redistributed.

The Archipelago (Ark) active measurement platform:

Ark is a new platform designed, developed, and deployed by CAIDA for optimized, coordinated active network measurements. We currently have 8 Ark monitoring locations but we expect to grow to 15 locations by July 08. skitter, a legacy active measurement project (to be replaced by Ark in 2008) collects traceroute data from 16 locations. Ark will support a variety of macroscopic Internet active measurement projects, including the scamper IPv4/v6 topology discovery tools. Existing skitter monitors will be upgraded to Ark monitors during the expected period of performance 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 always anomalous in some way. Because the network telescope contains approximately 1/256th of all IPv4 addresses, we receive roughly one out of every 256 packets sent by an Internet worm with an unbiased random number generator. Because we are uniquely situated to receive traffic from every worm-infected host, we provide a global view of the spread of Internet worms.

Adaptive Netflow:

Adaptive NetFlow, deployable through an update to router software addresses many 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. To enable counting of non-TCP flows, we also developed an optional Flow Counting Extension that can augment existing hardware at routers. Both our proposed solutions readily provide descriptions of the traffic of progressively smaller sizes. Transmitting these at progressively higher levels of reliability allows graceful degradation of the accuracy of traffic reports in response to network congestion on the reporting path. They also provide low, statistically provable error rates on sampled data. Adaptive NetFlow has been incorporated into CoralReef, CAIDA's passive measurement software suite, and is available for data collection on high-speed links.

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. Finally, 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 raw traces.

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 OC48/GigE, OC192/10GigE, and ISP peering point links (when links and monitors are available). Will include raw traces and statistical summaries.
- The UCSD Network Telescope, including data on random-spread Internet worms, distributed Denial-of-Service attacks, port and host scanning, and botnets. We will provide data on every worm or virus we deem important that is monitored by our measurement infrastructure. Quarterly traces will be coordinated with other collectors of Internet background radiation data to ensure broadly, time synchronized datasets for researchers.
- 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, including denial-of-service attack datasets, Code-Red and Witty worm datasets, UCSD Network telescope traces, and scamper topology data. This data will allow previously impossible longitudinal analysis of threat evolution over the last several years.

We will pursue participation in the PREDICT program via development of mutually acceptable Memoranda of Agreement and help to 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 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.