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

Cooperative Agreement FA8750-12-2-0326, CAIDA, UCSD

PROJECT MANAGEMENT PLAN

1. Project Description.

1.1. Research Objectives.

The Department of Homeland Security (DHS) has developed the Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) project to provide vetted researchers with current network operational data in a secure and controlled manner that respects the security, privacy, legal, and economic concerns of Internet users and network operators. The objectives of our project, "Supporting Research and Development of Security Technologies through Network and Security Data Collection", is to participate in PREDICT as Data Provider and Data Host, and also provide PREDICT project support for PI meetings, metrics development, outreach, and Application Review Board (ARB). We will perform this basic fundamental research on a reasonable efforts basis.

1.2. Public Problem Description.

The Internet has become critical infrastructure for almost every aspect of American life. Commerce, business, government, education, military, and society as a whole rely on networked computers for data communication, distribution, and dissemination. Yet the emergence of new security threats continues to outpace the development of new technologies aimed to ensure the security, integrity, and privacy of digital information. Researchers require current data on Internet security threats, including samples of normal and malicious Internet traffic, malicious software samples, logs from machines compromised in targeted attacks, and other data to develop hardware and software that would protect against and mitigate the effects of hacking attempts and malicious software.

1.2.1. Public Research Goals/Contribution

As **Data Provider**, we will collect, curate, anonymize (if necessary), and archive Internet measurement data of the following types:

a) Internet Topology Data including Internet Topology Measured from Ark Platform (IPv4 Routed /24 Topology, IPv4 Routed /24 DNS Names, and IPv6 Topology) and Internet Topology Data Kits (containing router-level topology data, router- to-AS assignments, geographic location of each router, and DNS lookups of all observed IP addresses);

b) Blackhole Address Space data including (near) real-time and archived samples of the UCSD Network Telescope Data.

We will also continue to provide the following legacy data:

a) Active Internet Topology Measurements with skitter (collected in 1998-2008);

b) OC48 Peering Point IP Packet Headers (collected in 2002-2003).

As **Data Host**, we will manage, curate, and share these data with vetted security researchers. We will maintain, upgrade, and expand as necessary our data storing and serving capabilities.

To **support the PREDICT project** we will give feedback to PREDICT Coordinating Center (PCC) regarding PREDICT portal utility and user experience, and provide statistics on data collection and usage by request. We will also participate in ARB activities and will review data requests submitted to PREDICT.

1.2.2. Expected Impact

For almost 15 years, CAIDA has been known in the Internet research community as one of the leading experts collecting and sharing Internet measurement data. As of December 2012, external researchers not affiliated with CAIDA published nearly 400 studies using CAIDA data.

Datasets provided by CAIDA via PREDICT framework potentially will have the following impact in the field of Internet security research:

a) Internet Topology data can be used for modeling and simulation of malware propagation and containment measures, infrastructure stability vulnerability assessments, longitudinal studies of Internet topology evolution, Internet address mapping and inferences.

b) Near-real-time and archived Blackhole Address Space data can enable study of the origin and characteristics of Internet pollution, analyzing malware activity, developing efficient mitigation strategies, and monitoring Internet censorship or outage events on a global scale.

c) Topology legacy data can be used to study the historical evolution of macroscopic connectivity and performance of the Internet.

d) OC48 traffic legacy data supports research on Internet traffic and classification, including analysis of security-related events.

Examples of past impact to the external community can be found on the CAIDA web site at http://www.caida.org/data/publications/bydate/.

2. Technical Approach.

2.1. Detailed Description of Technical Approach.

In order to accomplish the proposed Data Provider tasks, we will make use of our existing unique data collection capabilities including:

- The Archipelago (Ark) active measurement platform.

Ark is an infrastructure designed, developed, and deployed by CAIDA for optimized, coordinated active network measurements. We currently have more than a hundred Ark monitors (37 of them IPv6 capable) deployed on 6 continents in 40 countries and controlled by a central server at CAIDA. We plan to continue to grow Ark by adding approximately one or two monitors per month. In the last year, Ark growth accelerated as we began to use inexpensive miniature (credit-card size) Raspberry Pi devices as Ark monitors. This unique measurement platform supports a variety of macroscopic Internet active measurement projects, including ongoing IPv4/v6 topology discovery with the *scamper* tool, the Spoofer project, topology-on-demand probing, and various ad-hoc measurements and assessments benefiting empirical Internet research.

- The UCSD Network Telescope.

The UCSD Network Telescope consists of a large piece of globally announced IPv4 address space. This address space contains almost no legitimate hosts, so inbound traffic to non-existent machines is unsolicited, and anomalous in some way. Our network telescope contains approximately 1/256th of all public IPv4 addresses, so it receives roughly one out of every 256 packets sent by malicious software with an unbiased random number generator. The network telescope gives researchers the opportunity to observe and analyze anomalous traffic which comprises a significant portion of Internet activity. This traffic results from a variety of events, including scanning of address space by attackers or malware looking for vulnerable targets, backscatter from randomly spoofed source denial-of-service attacks, the automated spread of Internet worms and viruses, and misconfiguration (e.g. mistyping an IP address).

In order to accomplish the proposed Data Host tasks, we deploy and maintain expansive data hosting infrastructure (described in a separate document). Over the course of the project, we will make additional hardware purchases and upgrades. In addition to serving archived data, we will support near-real-time access to the most recent two months of the UCSD Network Telescope data by providing high-level compute and storage systems with adequate reliability and performance characteristics including redundancy, reusable parts and hot spares. Associated compute servers will handle multi-terabyte analysis, with reliable uptime and timely job completion.

In order to accomplish the proposed project support tasks, we will use conference rooms available on the UCSD campus and teleconferencing facilities provided by the San Diego Supercomputer Center.

2.2. Comparison with Current Technology.

2.2.1. Archipelago active measurement infrastructure.

Underpinning many Internet topology studies are data sets collected by *traceroute*-based measurements. By sending probe packets to the destination, *traceroute* methods capture a sequence of interfaces along the forward path from the source to a given destination.

Several research infrastructures have conducted *traceroute*-based active measurements in the past but are no longer funded (Surveyor, AMP NLANR, skitter, DIMES). Currently, WIDE's (Japan) Gulliver project is deploying active measurement hosts in developing countries while RIPE NCC's (Netherlands) Atlas project supports thousands of active probes primarily in the RIPE NCC service region to measure Internet connectivity and performance from those regions. PlanetLab, a primarily academic network testbed for distributed computer systems research, also supports limited but useful active network measurements. For example, iPlane used the PlanetLab infrastructure in addition to public traceroute servers to gather Internet topology information. Other active measurement projects have conducted traceroute and performance measurements utilizing fully decentralized software, such as plug-ins to BitTorrent applications intended to improve performance of that application.

With its focus on coordination, security, and long-term continuity, Ark has distinct but complementary features to these platforms, and fills a recognized gap in the research community. It is currently one of the most advanced and well-known infrastructure platforms available for active *traceroute*-based measurements of the Internet topology.

2.2.2. UCSD Network Telescope.

In the last decade, *network telescopes* have been used to observe Internet "background radiation" (IBR), i.e. unsolicited traffic sent to unassigned address space (“darkspace”). The routing system carries the traffic to darkspace because its address is being announced globally, but there is no response back to the traffic sources since there are no hosts in darkspace. Observing such one-way traffic allows visibility into a wide range of security-related events: scanning of address space by hackers looking for vulnerable targets, backscatter from denial-of-service attacks using random spoofed source addresses, the automated spread of worms or viruses, and various misconfigurations (e.g., mistyping an IP address). The observed packets represent mostly failed attempts to open connections, or other malware-related behavior. In the last two years this type of traffic has significantly increased due to botnet-related activities such as Conficker's scanning and p2p signaling.

University of Michigan and Team Cymru also conduct IBR measurements using their own portions of unassigned address space. Comparative analysis of IBR data collected at different network locations should provide a more comprehensive, multi-dimensional view of malicious Internet activities.

3. Schedule and Milestones.

3.1. Schedule Graphic.

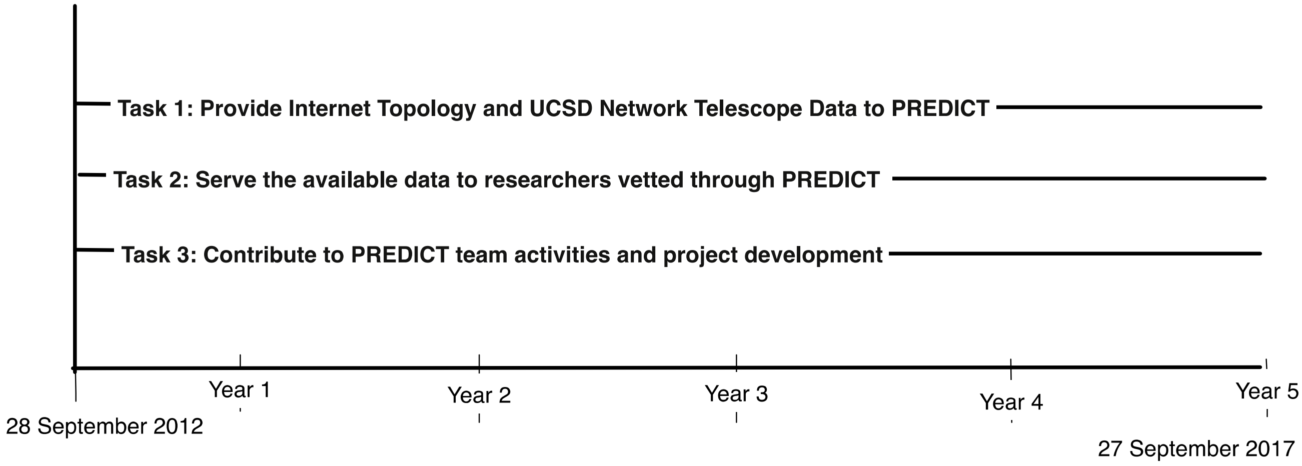
Task 1: Provide Internet Topology and UCSD Network Telescope Data to PREDICT
Task 2: Serve the available data to researchers vetted through PREDICT
Task 3: Contribute to PREDICT team activities and project development
Year 1
Year 2
Year 3
Year 4
Year 5
28 September 2012
27 September 2017

Figure 1 CAIDA Task Schedule

3.2. Detailed Individual Task Decriptions.

a) **Task 1**: provide Internet measurement data via PREDICT (ongoing).

We will curate and archive the collected IPv4 and IPv6 Topology data during the whole project period. We will also periodically (period is TBD) derive "Internet Topology Data Kits" which include annotations of inferences relevant to analyzing the Internet as critical infrastructure: router-level topology inferences, router-to-AS assignments, and geographic locations of each router.

We will also collect and curate Network Telescope data maintaining a sliding two-month window of the most recent data during the whole project period. We will maintain and upgrade as necessary the UCSD Network Telescope monitoring infrastructure and operate a real-time interactive monitor displaying Telescope traffic statistics of interest. Periodically (period is TBD), we will archive selected samples of the UCSD Network Telescope data and catalog them in PREDICT.

b) **Task 2**: serve the available data to researchers vetted through PREDICT (ongoing).

For each researcher whose data request is approved by ARB and PCC and who signs the appropriate Data Use Agreement with CAIDA, we establish an account giving access to CAIDA data servers. Our system administrator monitors the usage of accounts keeping track of data downloads and other activities. Our data administrator regularly contacts data users via specialized mailing lists used for data announcements, updates, and reporting requests. He also collects statistics of data requests and resulting publications, and provides user support regarding various data issues. Our webmaster maintains online information on available datasets, data descriptions, access policies, publication lists, and other project relevant matters.

We did not plan any major hardware purchases in Year 1 of the project.

b) **Task 3**: support PREDICT project activities (ongoing).

CAIDA researchers will work closely with PCC to optimize PREDICT portal utility, convenience, and overall user experience. We will synchronize CAIDA's internal data stewardship policies with PREDICT's framework and will revise and update CAIDA PREDICT MOAs as and if needed to support new types of data. We will contribute to PREDICT marketing and outreach efforts. CAIDA

PIs will participate in monthly PREDICT status teleconferences and take turns with other PREDICT PIs to host project PI meetings.

4. Deliverables Description.

a) Project Management Plan (this document) – annually
b) Hosting Infrastructure Description – annually
c) Technical Status Report – quarterly
d) Financial Status Report – monthly
e) Presentations and Research Papers – as available
f) Final Report – at the end of the project

5. Technology Transition and Technology Transfer Targets and Plans.

Not applicable