AFRL-RI-RS-TR-2018-146

SUPPORTING RESEARCH AND DEVELOPMENT OF SECURITY TECHNOLOGIES THROUGH NETWORK AND SECURITY DATA COLLECTION

UNIVERSITY OF CALIFORNIA SAN DIEGO -
CENTER FOR APPLIED INTERNET DATA ANALYSIS (CAIDA)

JUNE 2018

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2018-146 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /	**/ S /**
FRANCES A. ROSE	JOHN D. MATYJAS
Work Unit Manager	Technical Advisor, Computing & Communications Division Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* JUN 2018	**2. REPORT TYPE** FINAL TECHNICAL REPORT	**3. DATES COVERED** *(From - To)* SEP 2012 – DEC 2017

4. TITLE AND SUBTITLE SUPPORTING RESEARCH AND DEVELOPMENT OF SECURITY TECHNOLOGIES THROUGH NETWORK AND SECURITY DATA COLLECTION	**5a. CONTRACT NUMBER** N/A
	5b. GRANT NUMBER FA8750-12-2-0326
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Kimberly Claffy, Marina Fomenkov	**5d. PROJECT NUMBER** DHSP
	5e. TASK NUMBER UC
	5f. WORK UNIT NUMBER SD
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of California San Diego 9500 Gilman Drive, Dept 621 LaJolla CA 50854	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RITE 525 Brooks Road Rome NY 13441-4505	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL/RI
	11. SPONSOR/MONITOR'S REPORT NUMBER AFRL-RI-RS-TR-2018-146

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Research and development targeted at identifying and mitigating Internet security threats require current network data. To fulfill this need, researchers working for the Center for Applied Internet Data Analysis (CAIDA), a program at the San Diego Supercomputer Center (SDSC) which is based at the University of California, San Diego (UCSD), have been engaged in collecting packet-level data from the UCSD Network Telescope (which monitors a /8 IPv4 darknet), and IPv4 and IPv6 topology data from the Ark infrastructure. We curated and, as necessary, anonymized this data, and shared it with the vetted network and security researchers using the PREDICT/IMPACT portal and legal framework. We have also contributed to community building efforts that were responsive to public and private sector needs in Cybersecurity S&T research. To help further advance cybersecurity research, we provided access to this sensitive data – real-time traffic destined for blackhole address space – using a “bring-code-to- data” model on CAIDA machines. The major challenges in our approach were: sustainable collection, curation, and storage of large volumes of data, and enabling privacy-respecting sharing. To manage privacy risk without sacrificing research utility in our approach to data sharing, we collaborated with the PREDICT/ IMPACT legal team to develop, formalize, test, and use a privacy-sensitive data-sharing framework that integrated proven disclosure control techniques to protect privacy without obliterating all utility in the data, with a policy approach that relies upon standard privacy principles and obligations of researchers and data providers.

15. SUBJECT TERMS
Cybersecurity, Internet Measurements, Data Sharing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** FRANCES A. ROSE
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	20	**19b. TELEPHONE NUMBER (*Include area code)***

TABLE OF CONTENTS

1 SUMMARY

Research and development targeted at identifying and mitigating Internet security threats require current network data. To fulfill this need, researchers working for the Center for Applied Internet Data Analysis (CAIDA), a program at the San Diego Supercomputer Center (SDSC) which is based at the University of California, San Diego (UCSD), have been engaged in collecting packet-level data from the UCSD Network Telescope (which monitors a /8 IPv4 darknet), and IPv4 and IPv6 topology data from the Ark infrastructure. We curated and, as necessary, anonymized this data, and shared it with the vetted network and security researchers using the PREDICT/IMPACT portal and legal framework. We have also contributed to community building efforts that were responsive to public and private sector needs in Cybersecurity S&T research. To help further advance cybersecurity research, we provided access to this sensitive data – real-time traffic destined for blackhole address space – using the "bring-code-to-data" model on CAIDA machines.

The major challenges in our approach were: sustainable collection, curation, and storage of large volumes of data, and enabling privacy-respecting sharing. To manage privacy risk without sacrificing research utility in our approach to data sharing, we collaborated with the PREDICT/IMPACT legal team to develop, formalize, test, and use a privacy-sensitive data-sharing framework that integrated proven disclosure control techniques to protect privacy without obliterating all utility in the data, with a policy approach that relies upon standard privacy principles and obligations of researchers and data providers.

2 INTRODUCTION

Over the past two decades, 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 discovery of new security threats continues to outpace the development of new technologies aimed to ensure the security, integrity, and privacy of digital information.

The state-of-the-art in the development of security technologies can be improved through strategic coordinated data collection and distribution efforts. The Protected REpository for the Defense of Infrastructure against Cyber Threats (PREDICT) program and its successor the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program have responded to these needs by developing a centralized infrastructure designed to facilitate data access for cybersecurity research while ensuring data security and privacy. We participated in PREDICT/IMPACT as a **Data Provider** and **Data Host**. In the former role, we collected, curated, anonymized (if necessary), and archived Internet data to support cybersecurity research and development activities. In the latter role, we managed, maintained, and shared these data with vetted security researchers.

We performed this basic fundamental research on a reasonable efforts basis.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Data Provider Tasks

As **Data Provider**, we collected, curated, and archived various Internet measurement data sets of relevance for cybersecurity research and development activities. To procure these data sets, CAIDA researchers made use of unique Internet data producing capabilities: distributed Archipelago measurement infrastructure tailored specifically for active probing experiments and the UCSD Network Telescope that can observe the manifestations of malicious activities in the Internet. We also had access to network monitors collecting samples of backbone Internet traffic as long as conditions permitted and links were available – until the end of 2016.

• ***The Archipelago (Ark) active measurement platform***. Ark is a platform designed, developed, and deployed by CAIDA for optimized, coordinated active network measurements. It is used for a variety of macroscopic Internet active measurement projects in support of empirical Internet research, including ongoing and ad-hoc measurements. Ark monitors are deployed worldwide on 6 continents. Over the project period, we continued to grow Ark infrastructure by approximately 1-2 monitors per month.

We measured IPv4 and IPv6 Internet topology from Ark infrastructure, and mapped observed IP addresses to their DNS names in real-time. From this topology data we derived and heavily curated “Internet Topology Data Kits” which included annotations of inferences relevant to analyzing the Internet as critical infrastructure, namely: router-level topology inferences, router-to-AS assignments, and geographic locations of each router. In Year 4 of the project we also added an ongoing Prefix-Probing data set (described below).

• ***The UCSD Network Telescope***. The UCSD Network Telescope consists of a large piece of globally announced IPv4 address space (/8 segment). 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.

We collected pcap files (header and content) from the UCSD Network Telescope, instrumentation that monitors, strips the payload, and retains a sliding most recent two-month window of data on our machines, while archiving older data to an outside facility (NERSC).

• ***Passive network monitors***. In collaboration with a Tier 1 ISP we operated passive network monitors at two different locations on their backbone. Each bidirectional link monitor consisted of either a single server or a pair of 2-unit servers instrumented with an Endace DAG high-performance data collection card. The servers were time-synchronized with stratum-1 time servers to allow interpolation of trace data collected at disparate locations. Per our Research Agreement with the provider, after collecting a trace, we stripped the payload on the monitor before transferring files to CAIDA servers. Once on CAIDA servers, we compiled basic statistics, anonymized the raw traces using CryptoPAn prefix-preserving anonymization with the same key for all traces collected during a given calendar year, and packaged the data for release to vetted researchers.

The resulting datasets contain one-hour long samples of traffic collected quarterly during a calendar year.

Over the project period, CAIDA researchers also continually improved existing and developed new methodologies for data cleansing, curation, annotation, and creative integration with other relevant data independently available from 3rd parties (i.e., routing, geolocation, organizational data).

3.2 Data Host Tasks

As a **Data Host**, we stored (including backups), managed, and served the available data to vetted security researchers. CAIDA personnel maintained numerous data-serving hardware platforms hosted in the machine room at the San Diego Supercomputer Center (SDSC). Our system administrators have designed, configured and deployed these hosts to provide high availability for data collection, indexing, curation, and distribution. (Note that our data storage/processing infrastructure was not exclusively funded by the DHS PREDICT/IMPACT Project.)

As a general strategy, we have chosen to deploy several hosts with moderately large (20-40TB) locally attached disk systems that make use of the ZFS file systems. These configurations optimized cost of storage and availability for data consumers. We also ran several systems acting as web servers hosting project description pages and distributing data to vetted account holders. We used FreeBSD jails servers that mounted exported file systems from back end data servers.

We continued to use two repurposed nodes from the decommissioned SDSC Trestles Supercomputer to support our topology query system and to produce our flagship Internet Topology Data Kit (ITDK) datasets.

For UCSD Telescope data processing and visualization, we had access to 15 dedicated compute nodes and one I/O node on the SDSC Gordon supercomputer platform that stored and processed the indexed time-series data.

Finally, to archive older Telescope data, we made use of an Energy Research Computing Allocations Process (ERCAP) Allocation at the National Energy Research Scientific Computing Center (NERSC) facility, a division of the Lawrence Berkeley National Laboratory located in Berkeley, California. SDSC has high bandwidth connectivity (10 GB) with the NERSC.GOV domain allowing regular file transfers for archival of historical data.

One of the challenges inherent in Data Host activities was dealing with the huge data volumes. The data sets collected by CAIDA can grow to many tens of even hundreds of terabytes, limiting the number of researchers who can use the data. The problem is especially acute for the UCSD Telescope data where during malicious activity outbreaks, data volumes can increase sharply, yet rapid analysis and response are necessary. The speed, scope, and strength of today’s automated malicious software demand that relevant data are available in real time, matching the fast dynamics of the threat. As a Data Host, we supported such real-time access by providing high-level compute and storage systems with adequate reliability and performance characteristics including redundancy, reusable parts and hot spares.

4 RESULTS AND DISCUSSION

We successfully accomplished all the objectives of the project.

4.1 Technical Accomplishments

Our data offered via PREDICT/IMPACT included the following data sets:

a. Internet Topology Measured from Ark Platform

(i) *IPv4Routed/24 Topology* (forward IPv4 paths, reply Time-to-Live (TTL), Round-Trip-Time (RTT), and ICMP responses)

(ii) *IPv4Routed/24 DNS Names* (fully-qualified domain names for IP addresses in the IPv4 Routed /24 dataset)

(iii) *IPv6 Topology* (IP paths, RTT, TTL, and ICMP for IPv6)

(iv) As of the end of 2015, we also added the *IPv4 Prefix-Probing dataset* which consists of daily traceroutes to every announced BGP prefix from a subset of Ark monitors. Each monitor probes the entire set of targets independently, and completes exactly one pass of the target set every calendar day (aligned on UTC boundaries).

b. Internet Topology Data Kits (ITDK)

Over the project period, we produced 8 ITDKs. Each kit contains router-level topology data inferred from the measurement data using CAIDA alias resolution methodology, router-to-AS assignments, geographic location of each router, and DNS lookups of all observed IP addresses.

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, and Internet address mapping and inferences.

c. UCSD Real-time Network Telescope Data

The network telescope traffic results from a variety of security related 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).

To give researchers the opportunity to observe and analyze this anomalous traffic that comprises a significant portion of Internet activity, we continuously collected these packets and, after stripping the payload, stored them in one-hour long files in PCAP format. We made these files available in near-real-time (with 1-hour delay). To protect against risk of disclosure or mis-use of the data, we used a “bring-code-to-data” access mode to share a sliding most-recent approximately two-month window of data with vetted researchers. We deployed powerful compute servers to handle multi-terabyte data analysis, with reliable uptime and timely job completion. A dedicated system administrator with experience in managing data processing pipelines administered these facilities and helped users to navigate the process.

d. Archived Samples of UCSD Network Telescope Data

Using previously collected, older UCSD Network Telescope data, we created, archived, and cataloged in PREDICT/IMPACT portal a few sample data sets. In these data sets, we anonymized the destination IP addresses by zeroing the first octet of the IP address. We did not anonymize the source IP addresses, which mostly represent denial-of-service attack victims, and their anonymization would substantially diminish the research utility of this data.

Blackhole address space data supports study of the origin and characteristics of Internet pollution, evaluating various malware collection approaches, developing efficient mitigation strategies, and monitoring Internet censorship or outage events on a global scale.

e. Archived Samples of Internet Traffic

We collected and archived one-hour monthly samples of backbone Internet traffic. We removed the payload from all packets and anonymized IP addresses using CryptoPan prefix-preserving anonymization with anonymization keys changing annually.

This data is useful for research on Internet traffic characteristics, including application breakdown, security events, geographic and topological distribution, flow volume and duration, routing issues, and many others.

Over the project period, we continued to grow our collections of Topology and Telescope data and create new ITDKs, providing these timely data via the PREDICT/IMPACT portal. Our collection of backbone Internet traffic stopped in April 2016 when the monitored links were upgraded from 10 G to 100 G which our capture hardware could not handle.

Additionally, to support longitudinal analysis of infrastructure and threat evolution, we maintained and distributed previously collected data of interest to researchers. Our past collections available via PREDICT /IMPACT included:

f. Active Internet Topology Measurements with Skitter

This 4-TB archive represents the previous generation topology infrastructure measurements from CAIDA's Skitter infrastructure that preceded Ark. It contains forward IP paths and RTTs to hundreds of thousands of IPv4 destination addresses collected in 1998-2008 using the *skitter* probing tool from 24 monitors on 4 continents. These legacy measurements can be used to study the historical development of macroscopic connectivity and performance of the Internet.

g. OC48 Peering Point IP Packet Headers

This legacy data consists of three traces containing anonymized packet headers in PCAP format that were captured from an OC-48 link at a commercial peering point in California in 2002-2003. It supports research on Internet traffic and classification, including analysis of security-related events.

h. Other data

Finally, we obtained/created a few new data sets and started offering them through IMPACT. These novel data sets included:

- CAIDA DDoS 2007 Attack Dataset

This dataset contains approximately one hour of anonymized traffic from a distributed denial-of-service (DDoS) attack on August 4, 2007 (20:50:08 UTC to 21:56:16 UTC). Only attack traffic to the victim and responses to the attack from the victim are included in the traces. Non-attack traffic has as much as possible been removed.
http://www.caida.org/data/passive/ddos-20070804_dataset.xml

- IPv4 2013 Census Dataset

We inferred utilization of the IPv4 address space and taxonomized /24 address blocks as "ietf-reserved", "used", "routed unused", "unrouted assigned", "available". Each /24 address block in the dataset is also labeled with the Autonomous System Number that announced in BGP the longest network prefix containing that address.
http://www.caida.org/data/active/ipv4_2013_census_dataset.xml

- Geolocated Router Dataset

This dataset is a collection of router interface IP addresses geolocated to the city level. It includes 11,857 IP addressed geolocated based on their DNS names and 4,838 IP addresses geolocated based on RTT proximity to traceroute probes with known locations.
http://www.caida.org/publications/papers/2017/look_at_router_geolocation/

- UCSD Telescope Darknet Scanners dataset

This dataset contains IP addresses that were observed to conduct horizontal scans of the IPv4 address space monitored by the UCSD Network Telescope. It includes: IP addresses, scanned ports, scanned protocol, the timestamp when scan began, and scanning statistics for determining scanning strategy.
http://www.caida.org/data/passive/telescope-darknet-scanners_dataset.xml

- Border mapping dataset

This dataset consists of a set of border routers (between two ASes) inferred to be owned by the network hosting Ark Vantage Points (VPs) along with the set of neighbor routers connected to each border router.
http://www.caida.org/data/active/bdrmap_dataset.xml

4.2 Quantitative Metrics

Table 1 shows the number of files and the total volume of data collected during the report period (from October 1, 2012 until December 31, 2017) as well as cumulative size of the data at the end of this period. To report, we consolidate data into three principal categories: Ark Topology (including ITDKs and Prefix-Probing datasets), Telescope, and Traffic Traces. Note that the size of the Telescope data (measured in Petabytes – 10^{15} bytes, while the sizes of Ark Topology and

Traffic Traces data are measured in Terabytes – 10^{12} bytes) completely dominates the overall volume of data. Sizes of other data sets described in Section 4.1 are negligible in comparison with these major categories.

Collection	# of files	Size	On-disk size (compressed), 12/31/2017	Uncompressed size, 12/31/2017
Ark Topology	386097	26.0 TB	10.6 TB	34.1 TB
Telescope	129552	2.85 PB	1.30 PB	3.25 PB
Traffic Traces	12415	9.0 TB	.1 TB	30.6 TB

14

Table 1: Statistics of Data Collected During the Project Period

Table 2 below shows statistics of data requests and downloads during the report period (from October 1, 2012 until December 31, 2017). Note that in the course of the project we acquired considerable experience regarding what kind of Internet measurement data researchers seek and how they use and value available data. Our understanding of sensitivities intrinsic to our data also vastly improved. Therefore, we changed and adjusted modes of data access accordingly in order to make the whole sharing process more efficient and convenient and to better target the needs of our users. Some data were made public to ensure their availability and maximize their use, while access to other data remained restricted, requiring vetting of users and evaluating their proposed data usage. Statistics in Table 2 are integrated over the 5-year project period regardless of changes in the data access mode. The numbers of users appear much larger than the numbers of requests because they include users who downloaded our data publicly available in those categories; public data are extremely popular because they are easy to access and are sufficient for many research and developments purposes.

Dataset	Requests Received	Requests Granted	Users	Bytes Downloaded
Ark+Skitter Internet Topology	1700	1115	6616	78.9 TB
Telescope	1135	836	12852	19.0 TB
Traffic Traces	3840	2955	12271	294.5 TB
DDoS dataset	1342	862	531	2.84 TB

Table 2. Data requests and downloads during the report period

Finally, Table 3 shows the statistics of CAIDA data requests processed through IMPACT. Note that the previous PREDICT portal was practically unusable as it suffered from inefficient design, the absence of good Search functionalities, poor data choice and representation, and other numerous defects. An accounting system also was not implemented. Fortunately, many of those issues were solved or at least mitigated when the project changed from PREDICT to IMPACT at the end of 2015. The new version of portal was deployed in April 2016 and after the necessary testing and debugging, this new portal started accepting, processing, and counting accounts and data requests in July 2016. At that time, we began to increase our data offerings through the portal and promote its use among the users of our data. Therefore, request statistics in Table 3 are shown only for this time period, from July 2016 until (including) December 2017. The data downloads are included in Table 2 above.

Dataset	Requests Received
OC48 Passive Internet Traces	7
Archival Telescope Data	48
Near-real-time Telescope Data	5
Ark Internet Topology	13
DDoS	14

Table 3. IMPACT Data requests granted during July 2016 – December 2017 period

4.3 Other Achievements

CAIDA researchers have been active participants of the PREDICT/IMPACT project and over the years made significant contributions to team activities aimed at improving the portal usability and user experience. Selected highlights include:

- Edits and improvements to the data sharing legal framework developed for PREDICT/IMPACT. To document and codify the emerging structure, we went through 4 versions of Memoranda of Agreement (MOA) documents that defined and described rights and obligations of all parties involved in the project.
- Improvements on the front end of CAIDA's data sharing infrastructure. As our experience with data sharing mechanisms grew, we improved dataset representation and descriptions on CAIDA's web site, created a mailing list to keep the users of our data informed about latest changes and developments, and created a structure of responses to formalize handling of user requests.
- Improvements on the back end of CAIDA's data sharing infrastructure. We maintained and grew our hardware machine park, experimented with various approaches to backups, increased our storage to keep up with the growing data collection, and implemented internal databases to track data requests and distribution.
- Assisting developers with portal building and debugging. We made concerted efforts to work with RTI (PREDICT portal developer) and with its successor Blackfire (IMPACT portal developer): participated in design discussions, tested and helped debug new deployments, made numerous suggestions aimed at improving portal operations, look and feel, and the overall user’s experience.
- Co-organizing and hosting team meetings. We worked with PREDICT/IMPACT DHS Program Managers to organize project meetings and site visits at UCSD.
- Using portal data and procedures. CAIDA researchers not only contributed data to the portal, but also used data made available by other PREDICT/IMPACT team members to advance CAIDA's research programs. The benefits of these activities are two-fold: first, by applying for data as any outside user would do, we tested and evaluated usability factors, e.g., convenience and robustness, of portal operations. We were able to propose useful changes and modifications. Second, we demonstrated usefulness of PREDICT/IMPACT data offerings for cybersecurity research.

4.4 Deliverables

- Monthly Financial Reports, 10/2012 – 12/2017
- Quarterly Technical Reports, 2012 – 2017
- Project Management Plans, annually, 2012 – 2016
- Hosting Infrastructure Descriptions, annually, 2012 – 2016

4.5 Publications

The list of project related publications is in the reverse chronological order.

Gharaibeh, M., Shah, A., Huffaker, B., Zhang, H., Ensafi, R., and Papadopoulos, C.,"A Look at Router Geolocation in Public and Commercial Databases,” *Internet Measurement Conference (IMC)*, London, UK, 2017.

Jonker, M., King, A., Krupp, J., Rossow, C., Sperotto, A., and Dainotti, A., "Millions of Targets Under Attack: a Macroscopic Characterization of the DoS Ecosystem,*” Internet Measurement Conference (IMC),* London, UK, 2017.

Claffy, K., and Clark, D., "The 9th Workshop on Active Internet Measurements (AIMS-9) Report,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 47**, No. 5, pp. 4, 2017.

Orsini, C., King, A., Giordano, D., Giotsas, V., and Dainotti, A., "BGPStream: a software framework for live and historical BGP data analysis,*” Internet Measurement Conference (IMC),* Santa Monica, CA, 2016.

Claffy, K., "The 8th Workshop on Active Internet Measurements (AIMS8) Report,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 46**, No. 4, pp. 23-29, 2016.

Dainotti, A., Katz-Bassett, E., and Dimitropoulos, X., "The BGP Hackathon 2016 Report,” *ACM SIGCOMM Computer Communication Review Online (CCR-Online),* 2016.

Dainotti, A., Benson, K., King, A., Huffaker, B., Glatz, E., Dimitropoulos, X., Richer, P., Finamore, A., and Snoeren, A., "Lost in Space: Improving Inference of IPv4 Address Space Utilization,” *IEEE Journal on Selected Areas in Communications (JSAC),* **Vol. 34,** No. 6, pp. 1862-1876, 2016.

Czyz, J., Luckie, M., Allman, M., and Bailey, M., "Don't Forget to Lock the Back Door! A Characterization of IPv6 Network Security Policy,” *Network and Distributed Systems Security (NDSS),* San Diego, CA, 2016.

Zseby, T., Vázquez, F., King, A., and Claffy, K., "Teaching Network Security With IP Darkspace Data,” *IEEE Transactions on Education*, **Vol. 59**, No. 1, pp. 1-7, 2016.

Claffy, K., "The 7th Workshop on Active Internet Measurements (AIMS-7) Report,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 46**, No. 1, pp. 50-57, 2016.

Giotsas, V., Smaragdakis, G., Huffaker, B., Luckie, M., and Claffy, K., "Mapping Peering Interconnections to a Facility," *ACM SIGCOMM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, Heidelberg, Germany, 2015.

Orsini, C., King, A., and Dainotti, A.*, "BGPStream: a software framework for live and historical BGP data analysis,"* AFRL-PR-ED-TR-2015-10, Center for Applied Internet Data Analysis (CAIDA), 9500 Gilman Drive, La Jolla, CA, Oct 2015.

Benson, K., Dainotti, A., Claffy, K., Snoeren, A., and Kallitsis, M., "Leveraging Internet Background Radiation for Opportunistic Network Analysis*," Internet Measurement Conference (IMC)*, Tokyo, Japan, 2015.

Lehr, W., Kenneally, E., and Bauer, S., "The Road to an Open Internet is Paved with Pragmatic Disclosure & Transparency Policies," *Telecommunications Policy Research Conference (TPRC),* Washington, DC, 2015.

Kenneally, E., and Fomenkov, M., "Ethics Research & Development Summary: Cyber-security Research Ethics Decision Support (CREDS) Tool," *ACM SIGCOMM Workshop on Ethics in Networked Systems Research*, London, UK, 2015.

Clark, D., and Claffy, K., "Comments on Cybersecurity Research and Development Strategic Plan", *Networking and Information Technology Research and Development (NITRD) Program*, Jun 2015.

Dainotti, A., King, A., Claffy, K., Papale, F., and Pescapè, A.,"Analysis of a "/0" Stealth Scan from a Botnet," *IEEE/ACM Transactions on Networking*, **Vol. 23**, No. 2, pp. 341-354, 2015.

Raftopoulos, E., Glatz, E., Dimitropoulos, X., and Dainotti, A., "How Dangerous Is Internet Scanning? A Measurement Study of the Aftermath of an Internet-Wide Scan," *Traffic Monitoring and Analysis Workshop (TMA)*, **Vol. 9053**, No. 1, pp. 158—172, 2015.

Kenneally, E., "How to Throw the Race to the Bottom: Revisiting Signals for Ethical and Legal Research Using Online Data," *ACM SIGCAS Computers and Society*, 2015.

Dainotti, A., Squarcella, C., Aben, E., Claffy, K., Chiesa, M., Russo, M., and Pescapè, A., "Analysis of Country-wide Internet Outages Caused by Censorship," *IEEE/ACM Transactions on Networking,* **Vol. 22**, No. 6, pp. 1964-1977, 2014.

Alt, L., Beverly, R., and Dainotti, A., "Uncovering Network Tarpits with Degreaser," *Annual Computer Security Applications Conference (ACSAC)*, Los Angeles, CA, 2014.

Claffy, K., Clark, D., and Wittie, M., "The 6th Workshop on Active Internet Measurements (AIMS6) Report," *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 44**, No. 5, pp. 39-44, 2014.

Huffaker, B., Fomenkov, M., and Claffy, K., "DRoP:DNS-based Router Positioning,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 44**, No. 3, pp. 6-13, 2014.

Bailey, M., and Kenneally, E., "Cyber-security Research Ethics Dialogue & Strategy (CREDS) Workshop Report,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 44**, No. 2, pp. 76-79, 2014.

Zseby, T., King, A., Fomenkov, M., and Claffy, K., *"Analysis of Unidirectional IP Traffic to Darkspace with an Educational Data Kit"* AFRL-PR-ED-TR-2014-02, Cooperative Association for Internet Data Analysis (CAIDA), 9500 Gilman Drive, La Jolla, CA, Feb 2014.

King, A., Dainotti, A., Huffaker, B., and Claffy, K., "A Coordinated View of the Temporal Evolution of Large-scale Internet Events", *Computing,* **Vol. 96**, No. 1, pp. 53-65, 2014.

Dainotti, A., Benson, K., King, A., Claffy, K., Kallitsis, M., Glatz, E., and Dimitropoulos, X., "Estimating Internet address space usage through passive measurements,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 44**, no. 1, pp. 42-49, 2014.

Dittrich, D., Bailey, M., and Kenneally, E., *"Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Menlo Report",* AFRL-PR-ED-TR-2013-10, U.S. Department of Homeland Security, Washington, D.C. 20528, Oct 2013.

Claffy, K, "The 5th Workshop on Active Internet Measurements (AIMS-5) Report,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 43**, No. 3, pp. 101-107, 2013.

Benson, K., Dainotti, A., Claffy, K., and Aben, E., "Gaining Insight into AS-level Outages through Analysis of Internet Background Radiation,” *Traffic Monitoring and Analysis Workshop (TMA),* Turin, Italy, 2013.

Key, K., Hyun, Y., Luckie, M., and Claffy, K., "Internet-Scale IPv4 Alias Resolution with MIDAR,” *IEEE/ACM Transactions on Networking,* **Vol. 21**, No. 2, pp. 383-399, 2013.

Zseby, T., King, A., Brownlee, N., and Claffy, K., "The Day After Patch Tuesday: Effects Observable in IP Darkspace Traffic,” *Passive and Active Network Measurement Workshop (PAM),* **Vol. 7799**, pp. 273-275, 2013.

Dainotti, A., King, A., Claffy, K., Papale, F., and Pescapè, A., "Analysis of a "/0" Stealth Scan from a Botnet,” *Internet Measurement Conference (IMC),* pp. 1-14, 2012.

Dainotti, A., King, A., and Claffy, K., "Analysis of Internet-wide Probing using Darknets,” *Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS),* Raleigh, NC, 2012.

Zseby, T., and Claffy, K., "DUST 2012 Workshop Report,” *ACM SIGCOMM Computer Communication Review (CCR),* **Vol. 42**, No. 5, pp. 49-53, 2012.

4.6 Meetings and Presentations

Claffy, K., “DHS IMPACT Project: CAIDA update,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2017/impact_pi_dec/impact_pi_dec.pdf, Menlo Park, CA, December 2017.

Dainotti, A., “HI-Cube - Hub for Internet Incidents Investigation,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2017/hi_cube_hub/hi_cube_hub.pdf, Menlo Park, CA, December 2017.

Jonker, M., “Millions of Targets Under Attack: a Macroscopic Characterization of the DoS Ecosystem,” ACM Internet Measurement Conference (IMC), URL: http://www.caida.org/publications/presentations/2017/millions_targets_under_attack_imc/millions_targets_under_attack_imc.pdf, London, UK. November 2017.

Claffy, K., “Supporting Research and Development of Security Technologies Through Network and Security Data Collection,” DHS R&D Showcase and Technical Workshop, URL: http://www.caida.org/publications/presentations/2017/supporting_research_development_security_dhsrd/supporting_research_development_security_dhsrd.pdf, Washington, D.C, July 2017.

Dainotti, A., “IODA - Internet Outage Detection & Analysis,” Workshop on Active Internet Measurements (AIMS), URL: http://www.caida.org/publications/presentations/2017/ioda_aims/ioda_aims.pdf, La Jolla, CA, March 2017.

Claffy, K., “DHS IMPACT Project: CAIDA update,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2017/impact_pi_feb/impact_pi_feb.pdf, Marina del Rey, CA, February 2017.

Huffaker, B., “Interactive Access to Internet Topology Data,” Gateways Conference, URL: http://www.caida.org/publications/presentations/2016/interactive_access_internet_topology_gateways/interactive_access_internet_topology_gateways.pdf, La Jolla, CA, November 2016.

Claffy, K., “DHS IMPACT Project: CAIDA update,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2016/impact_pi_sep/impact_pi_sep.pdf, Madison, WI, September 2016.

Claffy, K., “DHS IMPACT Project: CAIDA update,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2016/impact_pi_jun/impact_pi_jun.pdf, Menlo Park, CA, June 2016.

Benson, K., “Leveraging Internet Background Radiation for Opportunistic Network Analysis,” UC San Diego Thesis Defense, URL: http://www.caida.org/publications/presentations/2016/leveraging_internet_background_radiation_kbenson/leveraging_internet_background_radiation_kbenson.pdf, San Diego, CA, June 2016.

Huffaker, B., “DHS IMPACT Project: CAIDA update,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2016/dhs_impact_project_impact/dhs_impact_project_impact.pdf, La Jolla, CA, February 2016.

Huffaker, B., “Autonomous Systems (AS) Introduction and Visualization,” Data Visualization (CSE199), URL: http://www.caida.org/publications/presentations/2016/as_intro_visualization_ucsd/as_intro_visualization_ucsd.pdf, La Jolla, CA, January 2016.

Benson, K., “Leveraging Internet Background Radiation for Opportunistic Network Analysis,” ACM Internet Measurement Conference (IMC), URL: http://www.caida.org/publications/presentations/2015/leveraging_internet_background_imc/leveraging_internet_background_imc.pdf, Tokyo, Japan, October 2015.

Dainotti, A., “Lost in Space: Improving Inference of IPv4 Address Space Utilization,” IRTF Workshop on Research and Applications of Internet Measurements (RAIM), URL: http://www.caida.org/publications/presentations/2015/lost_in_space_raim/lost_in_space_raim.pdf, Yokohama, Japan, October 2015.

Dainotti, A., “IODA- Internet Outages: Detection & Analysis,” DHS National Cybersecurity and Communications Integration Center (NCICC), URL: http://www.caida.org/publications/presentations/2015/ioda_fcc/ioda_fcc.pdf, Washington, D.C, April 2015.

Huffaker, B., “DHS PREDICT Project: CAIDA update,” DHS IMPACT PI Meeting, URL: http://www.caida.org/publications/presentations/2015/predict_pi_jan/predict_pi_jan.pdf, Washington, D.C., January 2015.

Kenneally, E., “Disclosure Control Update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2014/disclosure_control_predict_pi/disclosure_control_predict_pi.pdf, La Jolla, CA, June 2014.

Huffaker, B., “DRoP: DNS-based Router Positioning & DDec: DNS Decoding,” DHS Site Visit (CAIDA), URL: http://www.caida.org/publications/presentations/2014/drop_ddec_dhs/drop_ddec_dhs.pdf, La Jolla, CA, June 2014.

Kenneally, E., “How to Throw the Race to the Bottom: Harmonizing Ethical & Legal Issues with ICT Research Using Online Data,” Human Aspects of Security, Privacy & Trust, URL: http://www.caida.org/publications/presentations/2014/harmonizing_ethical_legal_issues_haspt/harmonizing_ethical_legal_issues_haspt.pdf, Crete, Greece, June 2014.
Kenneally, E., “ICT Research Ethics Update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2014/ict_research_ethics_predict_pi/ict_research_ethics_predict_pi.pdf, San Jose, CA, June 2014.

Fomenkov, M., “DHS PREDICT Project: CAIDA update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2014/predict_pi_may/predict_pi_may.pdf, Marina del Rey, CA, May 2014.

King, A., “Internet Garbage: Storage, Access, and Analysis,” Workshop on Network Data Storage, Access and Analysis (NDSAA), URL: http://www.caida.org/publications/presentations/2014/internet_garbage_ndsaa/internet_garbage_ndsaa.pdf, March 2014.

Fomenkov, M., “DHS PREDICT Project: CAIDA update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2014/predict_pi_jan/predict_pi_jan.pdf, Washington, D.C, Jan 2014.

Polterock, J., “DHS PREDICT Project: CAIDA update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2013/predict_pi_aug/predict_pi_aug.pdf, Ann Arbor, MI, August 2013.

Benson, K., “Gaining Insight into AS-level Outages through Analysis of Internet Background Radiation,” Traffic Monitoring and Analysis Workshop (TMA), URL: http://www.caida.org/publications/presentations/2013/gaining_insight_outages_tma/gaining_insight_outages_tma.pdf, Turin, Italy, April 2013.

Claffy, K., “DHS PREDICT Project: CAIDA update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2013/caida_update_predict_pi/caida_update_predict_pi.pdf, La Jolla, CA, March 2013.

Huffaker, B., “DATCAT Lesson Learned,” ISMA Workshop on Active Internet Measurement (AIMS), URL: http://www.caida.org/publications/presentations/2013/datcat_lessons_learned/, La Jolla, CA, February 2013.

Dainotti, A., “Lessons Learned by "Measuring" the Internet During/After the Sandy Storm,” FCC Workshop on Network Resiliency, URL: http://www.caida.org/publications/presentations/2013/lessons_learned_measuring_sandy_fcc/lessons_learned_measuring_sandy_fcc.pdf, Brooklyn, New York, February 2013.

King, A., “Toward Realtime Visualization of Garbage,” ISMA Workshop on Active Internet Measurement (AIMS), URL: http://www.caida.org/publications/presentations/2013/realtime_visualization_garbage/realtime_visualization_garbage.pdf, La Jolla, CA, February 2013.

Kenneally, E., “Why Are We Talking About Data Sharing,” ISMA Workshop on Active Internet Measurement (AIMS), URL: http://www.caida.org/publications/presentations/2013/talking_about_data_sharing/talking_about_data_sharing.pdf, La Jolla, CA, February 2013.

Dainotti, A., “Analysis of an Internet-wide Stealth Scan from a Botnet,” USENIX Large Installation System Administration (LISA), URL: http://www.caida.org/publications/presentations/2012/analysis_stealth_scan_lisa/analysis_stealth_scan_lisa.pdf, San Diego, CA, December 2012.

Kenneally, E., “Of Skunks and Canaries (and maybe rat holes),” Workshop on Internet Economics (WIE), URL: http://www.caida.org/publications/presentations/2012/skunks_and_canaries_wie/skunks_and_canaries_wie.pdf, La Jolla, CA, December 2012.

King, A., “A Coordinated View of Large-Scale Internet Events,” Workshop on Internet Visualization (WIV), URL: http://www.caida.org/publications/presentations/2012/coordinated_view_wiv/coordinated_view_wiv.pdf, Boston, MA, November 2012.

Dainotti, A., “Analysis of a "/0" Stealth Scan from a Botnet,” ACM Internet Measurement Conference (IMC), URL: http://www.caida.org/publications/presentations/2012/analysis_stealth_scan_imc/analysis_stealth_scan_imc.pdf, Boston, MA, November 2012.

Claffy, K., “DHS PREDICT Project: CAIDA update,” DHS PREDICT PI Meeting, URL: http://www.caida.org/publications/presentations/2012/caida_update_predict_pi/caida_update_predict_pi.pdf, Washington, D.C., November 2012.

Dainotti, A., “Analysis of Internet-wide Probing using Darknets,” BADGERS, URL: http://www.caida.org/publications/presentations/2012/analysis_darknets_badgers/analysis_darknets_badgers.pdf, Raleigh, NC, October 2012.

5 CONCLUSIONS

The PREDICT/IMPACT program is going on 18 years now, and has come a long way in terms of improving data architecture, legal support, and transparency. Much of the research and infrastructure enabled by the program would not exist otherwise.

Metrics for evaluation of the program have been a long-standing challenge, as the disincentives to share data in the first place still dwarf the effects of any other actions taken to expand use of the platform. Our current belief is that an auspicious direction for IMPACT's goals would be to try to position the program in the direct trajectory of emerging mandatory (or strongly incentivized) data-sharing initiatives, such as those required to acquire cybersecurity insurance, or to fulfill corporate, local, state, or federal disclosure requirements. We discussed these options at the final PI meeting of the program, and hope to undertake effort to support this direction in follow-up programs.

List of Symbols, Abbreviations and Acronyms

AS	Autonomous System
BGP	Border Gateway Protocol
CAIDA	Center for Applied Internet Data Analysis
DAG	Data Acquisition and Generation
DDoS	Distributed Denial of Service
DHS	Department of Homeland Security
DNS	Domain Name Service
ERCAP	Energy Research Computing Allocations Process
GB	Gigabytes
IMPACT	Information Marketplace for Policy and Analysis of Cyber-risk and Trust
I/O	Input/Output
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
ITDK	Internet Topology Data Kit
MOA	Memorandum of Agreement
NERSC	National Energy Research Scientific Computing Center
OC	Optical Carrier
PB	Petabytes
PCAP	Packet Capture
PREDICT	Protected Repository for the Defense of Infrastructure against Cyber Threats
RTI	Research Triangle Institute
RTT	Round-Trip-Time
S&T	Science and Technology
SDSC	San Diego Supercomputer Center
TB	Terabytes
TTL	Time-to-Live
UCSD	University of San Diego
UTC	Coordinated Universal Time
VP	Vantage Points