AFRL-RI-RS-TR-2018-099

PROTECTED REPOSITORY FOR THE DEFENSE OF INFRASTRUCTURE AGAINST CYBER THREATS (PREDICT) DATASET DEVELOPMENT AND HOSTING

PACKET CLEARING HOUSE INC.

APRIL 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

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

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. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2018-099 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
ROBERT L. KAMINSKI
Work Unit Manager

/ S /
WARREN H. DEBANY JR
Technical Advisor, Information
Exploitation and Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

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

4. TITLE AND SUBTITLE

PROTECTED REPOSITORY FOR THE DEFENSE OF INFRASTRUCTURE AGAINST CYBER THREATS (PREDICT) DATASET DEVELOPMENT AND HOSTING

5a. CONTRACT NUMBER
FA8750-12-2-0329

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Ross Stapleton-Gary

5d. PROJECT NUMBER
HS53

5e. TASK NUMBER
PC

5f. WORK UNIT NUMBER
H1

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Packet Clearing House Inc.
572B Ruger St. The Presidio of San Francisco
San Francisco CA 94129-0920

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIG
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-2018-099

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
This effort focused on collection, curation and dissemination of datasets and data that could be collected by any researcher with internet access, but which benefited by either or both persistence and collection scope. Packet Clearing House (PCH) was a participant in the Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) and Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) programs from the former's inception through formal contract end in August 2017. PCH played a number of roles as dataset provider and host, and PCH personnel also developed or advised on technologies and projects including a trusted enclave for access to sensitive datasets, and the development and launch of the IMPACT portal site.

15. SUBJECT TERMS

Internet measurement, Internet physical layer topology, Internet maps, Internet risk analysis, Internet outage

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**ROBERT L. KAMINSKI**
U	U	U	UU	10	19b. TELEPHONE NUMBER (*Include area code*)

Table of Contents

1.0 Summary 1
2.0 Introduction 1
3.0 Methods, Assumptions, and Procedures 1
3.1 Creation, Curation and Hosting of Datasets 1
3.2 Serving as a “Host of Convenience” 2
3.3 PCH Secure Enclave 2
3.4 IMPACT Portal Site Development 3
3.5 Serving as a “Prime of Convenience” 3
4.0 Results and Discussion 4
5.0 Conclusion 5
6.0 Recommendations 5
8.0 References 6

1.0 Summary

Packet Clearing House (PCH) was a participant in the Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) and Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) programs from the former's inception through formal contract end in August 2017. PCH played a number of roles as dataset provider and host, and PCH personnel also developed or advised on technologies and projects including a trusted enclave for access to sensitive datasets, and the development and launch of the IMPACT portal site. PCH also provided administration for a number of subcontractors under both PREDICT and IMPACT, including Blackfire, the implementer for the IMPACT portal.

2.0 Introduction

Packet Clearing House (PCH) participation in the Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) and Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) programs from the former's inception through formal contract end in August 2017 covered the following primary activities:

- Creation, curation and hosting of datasets for cybersecurity researchers;
- Serving as a “host of convenience” for datasets from PREDICT and IMPACT non-performers;
- Creation of a “secure enclave” for researcher access to highly-sensitive datasets;
- Guidance in development of the IMPACT portal site, in support of IMPACT contractor Blackfire;
- Serving as a “prime of convenience” in management of subcontractors to the PREDICT and IMPACT programs.

3.0 Methods, Assumptions, and Procedures

3.1 Creation, Curation and Hosting of Datasets

Packet Clearing House, as a significant contributor to global Internet development and management, generates considerable data on Internet operations, e.g., in collection and curation of BGP routing announcements, infrastructure data (cable infrastructure and IXP siting), and Internet outage events. From PREDICT through the end of PCH's participation in IMPACT it has provided a number of datasets for research. Demand for PCH-provided datasets has been only modest, and is surpassed by PCH-hosted datasets not provided by PCH (see next section, “Serving as a 'Host of Convenience'”).

Over the course of the PREDICT/IMPACT programs PCH provided quarterly technical reports, and, with the advent of IMPACT, monthly dataset delivery metrics.

3.2 Serving as a "Host of Convenience"

PCH served as a "host of convenience" for non-PCH-provided datasets, including four groups that collectively received the bulk of researcher interest over the period of performance:

1. Scalable Network Monitoring program data from DARPA (aka "DARPA");
2. P2INGS program data from IARPA (aka "IARPA");
3. Annual logs of network traffic from the National Collegiate Cyber Defense Competition (aka "NCCDC"); and
4. Datasets on Malicious Insider activity created for DHS by MIT Lincoln Lab.

PCH hosted a number of other datasets on behalf of non-direct IMPACT participants, but those received few (if any) requests, e.g., the mirroring of an FCC-originated (and self-hosted) dataset on broadband statistics.

Accession of these datasets was very much *ad hoc*, and very little in the way of formal process was established. Complicating issues included one of authority over datasets, e.g., who could assert that a particular dataset could be shared, to what degree. In at least one case (DARPA), the parent agency provided a release, even though the program had ended and the administration (e.g., the original program manager) over the program no longer existed.

3.3 PCH Secure Enclave

While PCH created a secure enclave to provide "researcher to data" access to more sensitive data, this was never actually used in support of the IMPACT program. One factor that likely contributed to this outcome was that PCH itself was not a generator of highly-sensitive datasets, so had no first-hand need for the enclave; IMPACT's general lack of marketing of the availability of datasets (see below, Results and Discussion), and the chicken/egg problem of taking the initiative to seek out and acquire sensitive datasets in the absence of demonstrated demand may also have contributed.

3.4 IMPACT Portal Site Development

PCH was heavily involved in the transition from the PREDICT program portal site to the IMPACT, and in the site's development and evolution. Ross Stapleton-Gray served as a subject matter expert in the IMPACT portal design process, and was the lead researcher in support of integration of digital object identifiers (DOIs) into IMPACT dataset management.

Design and development of the portal was dependent on “performer as proxy,” where program principal investigators provided assumptions and preferences, as cybersecurity investigators and proxies for the broader current and potential user communities. No attempt was made to elicit input or feedback from non-performer users, other than informally, via the PIs.

DOIs in particular ought to be useful in facilitating user citation of IMPACT datasets, in providing a standard (and unbreaking) means to cite datasets used, that will point subsequent inquiries back to the IMPACT portal. (This is a “future-proof” solution—when the previous predict.org web site was turned off, and the domain lost when the contractor supporting the program failed to renew the registration, all previously published references to PREDICT on the web were effectively broken. If IMPACT should choose to transition to a new site, all of the DOIs can be readily updated to reflect such a move, and none of those references need break.)

3.5 Serving as a “Prime of Convenience”

Similar to its work as a “host of convenience” in taking on orphaned and otherwise unhosted datasets, Packet Clearing House served as a prime on behalf of the IMPACT program in managing Blackfire Technologies as a subcontractor, from the time when Blackfire was first funded under IMPACT, to its transition as a direct DHS awardee.

In addition to Blackfire, PCH managed four other subcontracting organizations or individuals: University of Washington (Dave Dittrich); RedJack, LLC; University of Illinois (Michael Bailey); and Erin Kenneally.

PCH's role in administration of subcontractors was limited to management of reporting and invoicing, and did not extend to substantive direction or management of the subcontracting activities; the work with Blackfire was the only subcontracting situation where PCH had an active role in the work of the subcontractor, through Dr. Stapleton-Gray's work as a subject matter expert and proxy for portal users in the development process, and in the implementation of digital object identifiers.

4.0 Results and Discussion

The PREDICT and IMPACT programs, over the period of Packet Clearing House's participation, were directed primarily to the collection, curation and dissemination of datasets focused on “open Internet” phenomena, i.e., data that could be collected by any researcher with Internet access, but which benefited by either or both persistence and collection scope. A range of other dataset types and sources, and roles for the program (e.g., in addressing challenges in reducing dataset sensitivity, to allow for a broader set of sources to made available) might have been pursued.

We also believe that marketing—both assessing and appreciating “customer” need, and raising awareness of PREDICT/IMPACT as a resource—was one of the weaker aspects of the programs. One could characterize the programs as “supply” driven, i.e., dataset collection and curation was in largest part determined by the collection activities of the direct program performers. While that is not to say that what was collected, curated and provided wasn't of use, there was only nominal means to understand what would have been of use, i.e., cybersecurity researcher needs were known more anecdotally than systematically.[1]

We believe that the demand for the datasets PCH hosted on behalf of non-participants (three entirely synthetic, and one, NCCDC, records of activity in an artificial environment) indicate a healthy and persistent demand for “reasonably lifelike network activity,” in the face of a near complete lack of real-world activity data that isn't open and readily observable. (As noted above, the largest part of the IMPACT collection is open and observable by many parties; the value added there by IMPACT performers is scope and persistence, e.g., collection from hundreds of sensor points, or over a period of years.) While it may be that such synthetic data were being sought out as a half-step, and would not be of such interest were more authentic “real” traffic available, that premise was never tested.

As of contract end, we were still delivering some of the current non-PCH datasets hosted by PCH to Colorado State University (contact: Christos Papadopoulos, christos@colostate.edu) for hosting there. As datasets are received and hosting effected, they can be reassigned on the IMPACT portal.

Processes for participation limited the growth of PREDICT/IMPACT, e.g., (and particularly in the earlier PREDICT phase) burdensome administrative procedures, few if any incentives for information sharing by other than direct participants, and the slow expansion of the program to non-U.S. participants.

5.0 Conclusion

PCH filled needed roles in PREDICT/IMPACT, in particular for hosting of datasets for which it was not the provider (orphan datasets from concluded federal programs, or for parties not enrolled as IMPACT hosts), and for subcontractor management.

Demand for synthetic data, albeit only anecdotally measured, suggests this as an area deserving of emphasis under IMPACT. Of the datasets hosted by Packet Clearing House, the largest number of requests for datasets were for the synthetic (DARPA, IARPA, Malicious Insider) and cybersecurity exercise (NCCDC) datasets, hosted by PCH but originated by other non-performer sources.

In general, only minimal testing of assumptions was performed—decisions on dataset targeting, collection and curation were “supply side” driven.

6.0 Recommendations

Based on our experience as a provider and host in the PREDICT/IMPACT programs, we would recommend that the program seek to address a number of factors that have and may in some cases still inhibit growth, given that the success or failure of a community resource like IMPACT depends very much on achieving a critical mass of membership and active use.

IMPACT faces a classic “crossing the chasm” challenge, needing to achieve network effects. The friction in current processes, even where enrollment/access have been accelerated, and the limited scope of international participation, make IMPACT a more rarified resource than it could be. While alternative approaches (e.g., allowing any party that chose to to post notification of datasets, a la the UCSD “DatCat” project) could be overly difficult to curate and to ensure quality, there may be a happier medium between the recent state of IMPACT (through our direct experience, which ended at contract end) and a much more inclusive approach.

Based on our experience with “host of convenience” offering, and the demand for those datasets we provided in that fashion, we would make three recommendations for IMPACT going forward:

1. IMPACT should charter new performers as “hosts of convenience,” to make it possible for “orphan” datasets and those by providers unwilling to assume an active role in IMPACT to be accessible to researchers. Our understanding is that this role may be in the process of being assigned to the University of Southern California Information Sciences Institute (USC/ISI) and Colorado State University

(CSU), and endorse those two organizations in that role;
2. Synthetic datasets ought to be embraced as a reasonable facsimile of authentic network traffic and activities, in the absence of actual data that can be as widely shareable as needed; and
3. More of an effort be made to explore making data of interest more shareable, which would include a focus on just what about various data are useful or needed, and the various means to identify, anonymize or otherwise “defuse” the risk of broader sharing of sensitive data.

7.0 Acronyms

BGP	Border Gateway Protocol
DARPA	Defense Advanced Research Projects Agency
DNS	Domain Name System servers
DHS	Department of Homeland Security
DOI	Digital Object Identifier
IARPA	Intelligence Advanced Research Projects Agency
IMPACT	Information Marketplace for Policy and Analysis of Cyber-risk & Trust
IP	Internet Protocol
ISP	Internet Service Provider
IXP	Internet Exchange Points
NCCDC	National Collegiate Cyber Defense Competition
PCH	Packet Clearing House
PREDICT	Protected Repository for the Defense of Infrastructure Against Cyber Threats
UCSD	University of California San Diego

8.0 References

1. Discussion of "Data for Cybersecurity Research: Process and 'Wish List'", on the IMPACT web site: https://www.impactcybertrust.org/forum/viewtopic.php?id=26