
AFRL-RI-RS-TR-2021-058

# REAL-WORLD, LARGE SCALE NETWORK- AND HOST-LEVEL THREAT INTELLIGENCE

GEORGIA TECH RESEARCH CORPORATION

*MARCH 2021*

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-2021-058 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

**/ S /**
FRANCES A. ROSE
Work Unit Manager

**/ S /**
GREGORY J. HADYNSKI
Assistant Technical Advisor,
Computing & Communications 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)* |
|---|---|---|
| MARCH 2021 | FINAL TECHNICAL REPORT | NOV 2017 – SEP 2020 |

**4. TITLE AND SUBTITLE**

REAL-WORLD, LARGE SCALE NETWORK- AND HOST-LEVEL THREAT INTELLIGENCE

**5a. CONTRACT NUMBER**
FA8750-18-2-0038

**5b. GRANT NUMBER**
N/A

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Mustaque Ahamad
Paul Royal

**5d. PROJECT NUMBER**
DHSR

**5e. TASK NUMBER**
WL

**5f. WORK UNIT NUMBER**
SN

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Georgia Tech Research Corporation
505 10th ST NW
Atlanta GA 30318-5775

**8. PERFORMING ORGANIZATION REPORT NUMBER**

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

Air Force Research Laboratory/RITG
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-2021-058

**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**

Malicious software (malware) remains at the center of key threats on the Internet. Efforts to understand and defend against malware are thus critical to cyber security research and defense. However, these efforts often require large-scale malware data and due to various factors, many researchers and practitioners are left with an ongoing, unmet need for such data. Partnering with the DHS S&T the Georgia Tech College of Computing (GT CoC) leveraged the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT) program to make real-world, large-scale malware network- and host-level datasets available to hundreds of researchers and academic organizations.

**15. SUBJECT TERMS**

Malware data, cyber security, information security, Domain Name System (DNS), HyperText Transport Protocol (HTTP), Department of Homeland Security, Information Marketplace for Policy and Analysis of Cyber-risk and Trust (IMPACT)

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

## TABLE OF CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

## 1.0 SUMMARY

Malicious software (malware) remains at the center of key threats on the Internet. Efforts to understand and defend against malware are thus critical to cyber security research and defense. However, these efforts often require large-scale malware data and due to various factors, many researchers and practitioners are left with an ongoing, unmet need for such data.

By partnering with the Department of Homeland Security (**DHS**) Science and Technology Directorate (**S&T**) to participate in a program called the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (**IMPACT**), researchers at the Georgia Tech College of Computing (**GT CoC**) have leveraged their extensive malware collection and analysis experience to facilitate the availability of real-world, large-scale malware network- and host-level datasets. In providing such data to hundreds of **organizations, GT CoC** and **IMPACT** have filled a cyber security research and operational needs gap.

## 2.0 INTRODUCTION

Malicious software (malware) is used to create botnets that generate unsolicited email, to conduct Distributed Denial of Service attacks, and to steal sensitive information (e.g., financial information and intellectual property). Malware is widely leveraged by criminals and nation states alike. Efforts to understand and act upon the intentions of malicious programs are thus critical to cyber security research and defense. Such efforts, which include cyber threat discovery, compromise detection, and asset remediation, require the ability to observe and study current malware behavior.

Major information security organizations collect over 100,000 new malware samples every day. Each sample holds actionable network- and host-level information that, once derived via a dynamic analysis sandbox, has both research and operational utility. However, the sensitive nature of malware, the know-how and resources required to process the substantial volume of new samples that appear each day, and the commoditization of malware anti-analysis techniques have left many researchers and practitioners with an ongoing but unmet need for such data.

## 3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

To help empower the activities of both cyber security researchers and practitioners, the Georgia Tech College of Computing (**GT CoC**) makes large-scale malware network- and host-level information available through the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (**IMPACT**), a program within the Department of Homeland Security (**DHS**) Science and Technology Directorate's (**S&T's**) Cyber Security Division (CSD).

GT CoC's curation of large-scale malware network- and host-level data is a complex, multi-step process with prerequisites that include the ongoing collection of large volumes of new samples and the ability to process what is collected in a transparent, scalable manner. Through creation of large-scale collection infrastructure and the establishment of relationships with commercial and non-profit organizations, **GT CoC** has built a malware repository that, as of September 2020, contains over 260 million samples.

Through **IMPACT, GT CoC** provides safe representations of samples' network- and host-level actions, with supporting metadata. Descriptions of three datasets representative of **GT CoC's** efforts under the program are as follows:

**GT Malware Passive DNS Data Daily Feed** – This feed contains samples' use of the Domain Name System (**DNS**) in both raw (packet capture, or **PCAP**) and plaintext formats. The plaintext format comprises a set of four column entries that provide the executable's **MD5** hash, the date in which the executable was processed, the qname (domain name) of the **DNS** query, and (if the query was of type A) a resolution **IP** address for the domain name.

The wordle (or word histogram) shown in Figure 1 provides a simple, visual explanation of what this feed contains.

Figure 1: Malware Passive **DNS** Wordle (2020)

In Figure 1, higher counts of samples that queried for a domain name (or segment) are represented by its text having a larger font size. Benign domain names feature prominently because malicious software abuses the services hosted at them (e.g., malware accesses a search engine to perform clickfraud). Some malicious domain names also feature prominently because the malware instances that use them are both prolific and highly polymorphic, meaning that many programs distinct by hash are actually serial variants generated to make detection more difficult. Finally, some domains are named in an effort to masquerade them as legitimate infrastructure, but are indeed malicious.

Given malicious software's dependence on **DNS**, a common use of the **GT** Malware Passive **DNS** feed is the identification of domain names (via machine learning) that are used for adversary command and control (**C2**). As an example, in 2017 the WannaCry ransomware's kill switch domain began appearing in the feed and could, in turn, have been identified through various means. Once identified, action against **C2** domains can be taken to disrupt cyber threats and attacks.

**GT Malware HTTP Daily Feed** - This dataset contains a daily feed of HyperText Transport Protocol (**HTTP**) data. Supplemental metadata included with the feed associates each **URL** and **HTTP** object with a specific suspect Windows executable. Representations of this feed are available as **URL CSV** files, extracted **HTTP** object sets, and raw **PCAPs**.

**GT Malware API Call Daily Feed** - This dataset contains a daily feed of structured host-level **API** call information. Metadata included with the feed associates each **API** call log with a specific suspect artifact, which can include Windows executables, **PDF** files, and Microsoft Word documents. Each sample's interactions with the operating system is recorded, analyzed, and made available as structured plaintext. In addition, open source threat detection modules are used to make a determination about the provenance of the artifact. This determination is expressed as a maliciousness score; a score above 6.0 indicates the modules believe that the artifact is malicious.


```
...
        {
          "timestamp": "2020-12-29 00:27:38,047",
          "object": "file",
          "data": {
            "file": "C:\\Windows\\MSBLT.EXE"
          },
          "event": "write",
          "eid": 40
        }
      ],
      "anomaly": []
  },
  "malscore": 10.0,
...
```


Figure 2: Host-level Maliciousness Score

## 4.0 RESULTS AND DISCUSSION

As evidence of its usefulness in cyber research and defense, large-scale malware network- and host-level data offered by Georgia Tech's College of Computing (**GT CoC**) is among the most requested in **IMPACT's** ~15 year history. Per the **IMPACT** portal's statistics (of which a portion are represented in Table 1), the **GT** Malware Passive **DNS** Data Daily Feed is the most requested resource in **IMPACT** by more than a factor of two.

Table 1: IMPACT Top Three Most Requested Resources

| Resource Name | Data Host | Request Count | Popularity Rank |
|---|---|---|---|
| GT Malware Passive DNS Data Daily Feed | Georgia Tech | 242 | 1 |
| US Long-haul Infrastructure Topology | University of Wisconsin | 108 | 2 |
| CAIDA DDoS 2007 Attack Dataset | UCSD - Center for Applied Internet Data Analysis | 106 | 3 |

Table 2 provides a breakdown of approval counts for **GT CoC's** top three most requested resources, further highlighting their utility.

Table 2: **GT CoC** Most Popular Resource Counts

| Resource Name | Request Count |
|---|---|
| **GT** Malware Passive **DNS** Data Daily Feed | 242 |
| **GT** Malware Unsolicited Email Daily Feed | 92 |
| **GT** Malware **HTTP** Daily Feed | 58 |

**GT CoC's** resources comprise the ongoing publication of new datasets through feeds. Discrete datasets within each feed are produced by transforming the raw network- and host-level traffic of new samples processed each day into various activity-specific representations and then reviewing the resulting archive contents to ensure output safety and quality.

Table 3 summarizes counts of dataset distributions by year and reflects the popularity of the **GT** resources. In addition to their popularity, these feeds experienced a high rate of recurring retrievals from requestors who downloaded historical archives at the time of request approval and then returned to obtain new archives as they became available each day. The retrieval of daily feed archives by both new and previously approved requestors has resulted in over 175,000 dataset distributions over the period of performance.

Table 3: **GT CoC** Dataset Distributions by Year

| Year | Datasets Distributed |
|---|---|
| 2018 | 53,902 |
| 2019 | 63,610 |
| 2020 | 58,108 |

In providing large-scale malware network- and host-level data to approved requestors vetted by the **IMPACT** Coordinating Center (**ICC**), **GT CoC** and **DHS S&T CSD** have empowered the efforts of cyber security researchers and practitioners to whom this data would have otherwise been inaccessible. Large-scale malware network- and host-level datasets offered by **GT CoC** have enabled numerous cyber security activities – ranging from academic and commercial research that enhances understanding of cyber threats and cyber threat evolution, to improving the detection efficacy of commercial security tools, to enhancing the operational cyber defense of large commercial entities (e.g., major financials and organizations that operate critical infrastructure). Over 275 entities spanning academia, industry, and government have made use of **GT CoC** malware network- and host-level datasets; a list of organizations that received **GT** malware network- and host-level data through **IMPACT** is provided in Appendix B.

## 5.0 CONCLUSIONS

Efforts to understand and act upon the intentions of malicious software (malware) are critical to cyber security research and defense. Such efforts, which include cyber threat discovery, compromise detection, and asset remediation, require the ability to observe and study current malware behavior. However, the sensitive nature of malware, the know-how and resources required to process the substantial volume of new malware that appears each day, and the commoditization of malware anti-analysis techniques has traditionally left many researchers and practitioners with an ongoing but unmet need for data that enables such observation and study.

In partnering with the Department of Homeland Security (**DHS**) to participate in the Information Marketplace for Policy and Analysis of Cyber-risk and Trust (**IMPACT**), Georgia Tech's College of Computing (**GT CoC**) has leveraged its extensive malware collection and analysis experience to make real-world, large-scale malware network datasets widely available. As a result, hundreds of organizations from academia, industry, and government have requested and received over 175,000 **GT CoC** malware datasets through **IMPACT**, which has enabled a host of cyber security activities that include enhancement of cyber threat understanding, improvement in security tool detection efficacy, and better operational cyber defense. By producing and sharing hard-to-find, high value datasets curated in a manner that is mindful of the associated legal and ethical risks, **GT CoC** and **IMPACT** have filled an important research and operational needs gap.

## 6.0 REFERENCES

[1] Artem Dinaburg, Paul Royal, Monirul Sharif, and Wenke Lee. Ether: Malware Analysis via Hardware Virtualization Extensions. In Proceedings of the 15th ACM Conference on Computer and Communications Security (CCS 2008), Alexandria, VA, October 2008.

[2] Paul Royal, et al. PolyUnpack: Automating the Hidden-Code Extraction of Unpack-Executing Malware. In Proceedings of the 22nd Annual Computer Security Applications Conference (ACSAC 2006), Miami Beach, FL, December 2006.

[3] Paul Royal. Alternative Medicine: The Malware Analyst's Blue Pill. In Proceedings of Black Hat USA 2008, Las Vegas, NV, August 2008.

[4] Paul Royal. Automated Malware Analysis and Environment-Sensitive Malware. Presentation at the 2013 Malware Technical Exchange Meeting (MTEM 2013), Laurel, MD, July 2013.

[5] Paul Royal. Mirai and the Future of IoT Malware. Invited talk at the 32nd Annual Computer Security Applications Conference (ACSAC'16), Los Angeles, CA, December 2016.

## APPENDIX A – PUBLICATIONS AND PRESENTATIONS

[1] Paul Royal. Large-scale Malware Analysis and Data Sharing. Presentation at the 2018 Malware Technical Exchange Meeting (MTEM 2018), Pittsburgh, PA, July 2018.

[2] Paul Royal. Real-world, Large Scale Malware Network Data. Presentation at the 2019 DHS S&T Cyber Security Division R&D Showcase, Washington, D.C., March 2019.

## APPENDIX B – ORGANIZATIONS PROVIDED GT MALWARE DATA

9bplus
AFEX Inc
ANZ Bank
ASRC Federal Holding Company
Accenture
Akima LLC
Alchemy Data
AlienVault
Alliance Data Systems
Amgen
Arbor Networks
Area 1 Security
Argentine Federal Savings
Argonne National Laboratory
Asymmetric Security
BT Americas
Bank of America
Barclays
Baylor University
Ben Gurion University
Berkeley Research Group LLC
Binghamton University
Black Knight Financial Services
Blue Coat Systems
Bluescope Steel
Boeing
Booz Allen Hamilton
Branch Banking and Trust
Bridgewater Associates
Broadcom Inc
CERT Australia
CUBRC Inc
Cambridge Intelligence
Canadian Cyber Incident Response Centre
Capital Group Companies
Capitol Technology University
Carnegie Mellon Sotware Engineering Institute
Center for Army Analysis
CenturyLink
Check Point
Cisco Active Threat Analytics
Cisco Systems
CloudFlare
Coinbase
Comcast Corp
Communications Security Establishment
Concordia University of Edmonton
Core Security
Cox Communications
Critical Assets
Critical Path Security
CyberIQ Services Inc
CyberSoft Operating Corporation
Cylance Inc
Cynnovative
DARPA
DC3
DIR
Dalhousie University
DataRobot
Defence Science and Technology Organisation
Delft University of Tech
Dell SecureWorks
Deloitte
Department of Homeland Security
Deteque LLC
Dimension Data
Dissect Cyber Inc
DomainTools LLC
Drexel University
Duke University
Dynamic Semantics
EMC
ESET Canada Inc
Edinburgh Napier University
Emdeon
Emory University
Emporia State University
Endgame
Enlghten IT Consulting
Equifax
Experian
FNA
FireEye
Flashpoint Intel
Florida Atlantic University
Florida Institute of Technology
Forcepoint
GE Capital
GEICO
GMU
GTRI

Galois
George Mason University
George Washington University
Georgia Institute of Technology
Gladiator Technology
Government of Canada
Grover Group Inc
GuardiCore
HP
HYAS InfoSec Inc
Harvard University
Hewlett Packard Enterprise
Hiddenfigures Inc
Hoplite Industries Inc
ICANN
INCB
Icebrg
Indiana University
Infoblox
University of Illinois
Intel
Intelligent Automation Inc
Interset
Intrusion Inc
Iowa State University
JHU APL
Jacksonville State University
James Lay Consulting LLC
Johns Hopkins University
Kaspersky Labs Japan
Kennesaw State University
LBrands
Lancaster University
Lastline Inc
Level 3 Communications
LinkedIn
Lloyds Banking Group
Lockheed Martin
LogicHub
LookingGlass Cyber Solutions
MIT Lincoln Laboratory
MITRE Corporation
Magellan Midstream
MalShare
Malware Patrol
Malwarebytes
Marine Forces Reserve
Marymount University
MassMutual
MeasuredRisk
Merit Network Inc
Miami University of Ohio
Micro Focus
Mission Health
Mitre Corporation
Modus Operandi Inc
Morgan State University
NBCUniversal
NEUSTAR Inc
NPPD/CS&C/NCCIC/USCERT
NRI SecureTechnologies
NYUTandon
National Science Foundation  Phacil
National University Of Singapore
New York Life Insurance Company
New York University
Newport News Shipbuilding
Northwestern Mutual
OITC
Old Dominion University Norfolk VA
Open University
OpenDNS
Optiv Inc
PRS Technology
Pacific Northwest National Laboratory
Palantir
Palo Alto Networks
Patrick Desnoyers
Performanta Ltd
PerimeterX Inc
Perspecta Labs
Pfizer Inc
PhishLabs
PhishMe
Plixer
Plymouth University
PwC
Qintel
QuadMetrics Inc
RCMP
RSA
RSA Labs
Rackspace
Rapid7
Red Hat Inc
ReversingLabs
Riot Games
Rochester Institute of Technology

SRC Inc
SRI International
SSC
San Diego State University
San Jose State University
Sandia National Laboratories
School of Computing and Communications
Lancaster University
SecBI
SecureData EU
SecureInfo a Kratos company
Security Scorecard
SecurityTrails
Software Engineering Institute
Solutionary
Sony
Splunk
Support Intelligence Inc
Surfwatch Labs
Swansea University
Symantec
Syracuse University
T Rowe Price
TASKSTREAM
TSYS
Target Corporation
Team Cymru
Tel Aviv university
Telstra
Texas AM University
Textron Inc
The Home Depot
The MITRE Corporation
The Shadowserver Foundation
The Walt DIsney Company
ThreatQuotient
ThreatSTOP Inc
ThreatStream
ThreatTrack Security
Trend Micro Inc
UCSD Center for Applied Internet Data Analysis
UNC Charlotte
UNISYS Corp
UNIVERSITY OF NEVADA LAS VEGAS
US Naval Research Laboratory
USC Information Sciences Institute
USCERT
Under Armour
University at Albany State University of New York
University at Buffalo
University of California at Santa Cruz
University of Colorado
University of Illinois at Urbana Champaign
University of Maryland Baltimore County
University of Memphis
University of Michigan EECS Department
University of Nevada Reno
University of Oxford
University of South Florida
University of Southern California
University of Texas at San Antonio
University of Warwick
University of Washington
VITA Program
Verint Ltd
Verisign
Verizon Enterprise Solutions
Volexity
Walmart
Walmart Stores Inc
Wells Fargo Bank
Westfield Group
WinterWinds Robotics
Zingbox
Zions Bancorp
athenahealth
IBM
NCCIC
nominet
seclytics inc
topspin security
zvelo Inc

## LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

| | |
|---|---|
| C2 | Command and Control |
| CAIDA | Center for Applied Internet Data Analysis |
| CoC | College of Computing |
| CSD | Cyber Security Division |
| CSV | Comma-separated Values |
| DARPA | Defense Advanced Research Projects Agency |
| DHS | Department of Homeland Security |
| DNS | Domain Name System |
| GT | Georgia Institute of Technology |
| HTTP | HyperText Transport Protocol |
| ICC | IMPACT Coordinating Center |
| IMPACT | Information Marketplace for Policy and Analysis of Cyber-risk and Trust |
| IP | Internet Protocol |
| MD5 | Message Digest Algorithm, Version 5 |
| PCAP | Packet Capture |
| S&T | Science and Technology |
| UCSD | University of California San Diego |
| URL | Uniform Resource Locator |
| US | United States |