Transparent Computing Engagement 5 Data Release

This collection of files comes from the DARPA Transparent Computing (TC) program, generated during Engagement #5. This README provides a basic description of the program as well as an overall manifest and description of each file.

DARPA is releasing these files in the public domain to stimulate further research. Their release implies no obligation or desire to support additional work in this space. The data is released as-is. DARPA makes no warranties as to the correctness, accuracy, or usefulness of the released data. In fact, since the data was produced by research prototypes, it is practically guaranteed to be imperfect. Nonetheless, as this data represents a very large repository of semantically rich and structured data, DARPA believes that it is in the best interests of the Department of Defense and the research community to make them freely available.

Transparent Computing Program

The Transparent Computing program is a DARPA effort to develop technologies and an experimental prototype system to provide both forensic and real-time detection of Advanced Persistent Threats (APTs) as well as proactive enforcement of desirable policies.

Background

Modern computing systems act as black boxes in that they accept inputs and generate outputs but provide little to no visibility of their internal workings. This greatly limits the potential to understand cyber behaviors at the level of detail necessary to detect and counter some of the most important types of cyber threats, particularly APTs. APT adversaries act slowly and deliberately over a long period of time to expand their presence in an enterprise network and achieve their mission goals (e.g., information exfiltration, interference with decision making and denial of capability). Because modern computing systems are opaque, APTs can remain undetected for years if their individual activities can blend with the background “noise” inherent in any large, complex environment. Beyond the APT problem, a lack of understanding of complex system interactions interferes with (and sometimes completely inhibits) efforts to diagnose and troubleshoot less sophisticated attacks or non-malicious faulty behavior that spans multiple applications and systems.

The TC program aims to make currently opaque computing systems transparent by providing high-fidelity visibility into component interactions during system operation across all layers of software abstraction, while imposing minimal performance overhead. The program will develop technologies to record and preserve the provenance of all system elements/components (inputs, software modules, processes, etc.); dynamically track the interactions and causal dependencies among cyber system components; assemble these dependencies into end-to-end system behaviors; and reason over these behaviors, both forensically and in real-time. By automatically or semi-automatically “connecting the dots” across multiple activities that are individually legitimate but collectively indicate malice or abnormal behavior. TC has the potential to enable the prompt detection of APTs and other cyber threats, and allow complete root cause analysis and damage assessment once adversary activity is identified. In addition, the TC program will integrate its basic cyber reasoning functions in an enterprise-scale cyber monitoring and control construct that enforces security policies at key ingress/exit points, e.g., the firewall.

The TC program web page can be found at https://www.darpa.mil/program/transparent-computing

Due to its size, the material is hosted by Five Directions Inc. on their Google Drive at https://drive.google.com/open?id=1QIbUFWAGq3Hpl8wVdzOdIoZLFxkII4EK

A description of the files contained in that directory is provided later on in this document.

TC Program Technical Areas (TAs)

The TC program is divided into multiple TAs, each with one or more participants. The TAs and their roles are:

TA1: Tagging and Tracking. TA1 performers are developing approaches for tagging and tracking interactions between components on a computing platform in order to allow events to be linked to each other (e.g., parent and child processes, file accesses by processes, etc.) TA1 performers represent different technical approaches, platforms of focus, and level of fidelity. TA1 performers have to balance system overhead against tracking detail (and hence volume of generated metadata). Other than contextual information (e.g., papers) and ground truth, the data files contained in this release were generated by TA1 performers. The TA1 performers are identified by the project names "cadets", "clearscope", "fivedirections", "marple", "theia", and "trace".

TA2: Detection and Policy Enforcement. TA2 performers are developing techniques and systems needed to construct causal graphs that link TA1-generated events, and then reason over them to detect APT activity in real time and forensically. This release contains no data or information relating to TA2 performers.

TA3: Architecture. The TA3 performer is developing the overall TC architecture to enable the TA1 and TA2 technologies to work together as a system, as well as the data interchange and storage requirements needed by all performers. To assist users of this data set, the TA3 performer, "starc", has provided annotations for each of the data streams identifying attack events performed by TA5.1 on TA1 systems.

TA5.1: Adversarial Challenge Team (ACT). The TA5.1 performer is developing tools and techniques that draw on actual APTs to instantiate realistic attacks and behaviors in the engagement exercises to allow an evaluation of the performance of the TC system and components. This data release contains a detailed description of the actions taken by the ACT during Engagement #5, along with specific Indicators of Compromise that should be (but are not always) present in the data.

Engagement #5

In May 2019, the TC program conducted the last of five planned engagement exercises. In this exercise, it was planned that three instantiations of each TA1 performer would be set up on separate host platforms. The exercise would start with a period of benign data generation, wherein a scripted set of activities was run on each host and all performers knew that these activities were being executed. After the benign data generation, the TA5.1 team was given control of the test range and began a series of activities to reflect the activities of new and existing APTs across the test range. Benign background traffic was run continuously during this time and malicious activities were only conducted during the period from approximately 9am to 5pm on weekdays to allow TA2 performers to staff their interfaces and provide real-time detection alerts with a reasonable level of manpower.

After the live engagement period, each TA2 performer was given additional time to conduct forensic-type examinations of their data and provide detailed information about initial APT detections as well as any they might have missed during the live period.

File manifest and descriptions

operational_event_log.md: This file contains a log of activity on the TC E5 test range.

ground truth/ - tc_ground_truth_report_e5_update.pdf: This file contains the ground truth, as constructed by TA5.1, to help guide TA1/TA2 evaluation. As the engagement took place in an isolated network environment, all the IP addresses and domain names referenced are fictional and bear no relationship to routable and valid addresses and hostnames respectively.

tools/ - ta5-java-consumer.tar.gz: This archive includes the following: - java code that can be used to parse the avro binary files and execute methods when a record type is read that can perform further analysis. Executing the java consumer requires Java version 1.8. Building the software from source requires maven. - A script that reads the avro binary file and writes out json - A script that reads some data from an avro binary file and executes some semantic checks.

schema/ - TCCDMDatum.avsc: This is the machine-readable CDM Avro schema file.

- CDM20.avdl: This is a human readable version of the CDM schema which includes notes. It is used by the code in ta3-java-consumer to create the avsc file.
- cdm.pdf: This document describes the concepts and details of the CDM schema in human readable form.

data/ - cadets/ - clearscope/ - fivedirections/ - marple/ - theia/ - trace/

Each of the data files in this directory tree is the output from a particular TA1 performer during the Engagement #5 activity period.

Jacob Torrey
Program Manager
DARPA/I2O

Original release: Sept xx, 2019