FloCon 2020

Using Data to Defend

JANUARY 6–9, 2020 | SAVANNAH, GEORGIA

Data Driven Security Challenges

Timothy Shimeall, Ph.D.

CERT Situational Awareness Group

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

FloCon 2020

Using Data to Defend

JANUARY 6–9, 2020 | SAVANNAH, GEORGIA

Data Driven Security Challenges

Carnegie Mellon University
Software Engineering Institute

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

Document Markings

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM19-0618

Overview

Introduction

Challenges

Data to drive research

Summary

Introduction

Definitions

- Data-driven: supported by live capture of network configuration, usage, attacks, and defenses in a measurable manner
- Security: Protection of the authenticity, confidentiality, integrity, or availability of a network and its data.
- Challenges: Issues and conflicts in the conduct of research

Motivation:

- Experience in evaluating methods for network security, particularly methods that scale to and above Internet security providers
- Desire to provide methods of use in realistic security defense of networks
- Inclination for security methods that can generalize across networks and attacks

Challenges

Reproducibility of methods

Scalability of analytics

Amenability to unclean data

Applicability to evolving threats

Adaptability to encrypted traffic

Reproducibility

Methods that can be reconstructed in an evaluation or usage environment duplicating results from the development environment

- Described in sufficient detail
  - What functionally is the method, step by step?
  - What related work is essential to reproduction?
- Identified all necessary parameters for constructing or adapting to different environment
  - What constitutes a similar environment?
  - What assumed knowledge of environment, usage, attacks, and defenses?
- Presented results of sufficient clarity that comparison is feasible
  - Precision
  - Transformations

Scalability

Providing sufficient detail for utility vs. confusing results

Showing data displays too crowded to display data (plot goes grey)

Representing multiple axes of variation effectively

Generating results in reasonable amounts of time

Back-hauling data to central point for processing vs. federating distributed data sources processed locally

Ensuring known provenance of results (data goes brown)

Unclean Data

Network attackers (and defenders) don’t produce clean data

- Traffic artifacts (exponential back off, repeated termination, scanning, distraction)
- Deception and concealment engineered in (protocols, ports, endpoints)

Data is almost never normally distributed

- Network behaviors driven by work cycle and network stacks, not individuals
- Attack behaviors: noisy or invisible

Power-law distributions are often not useful

Often need to transform data before it can be effectively used or displayed

- Clean and regularize
- Scale and measure

Evolving threats

Security is largely unique in Computer Science: a motivated, intelligent, and resourced set of actors is actively engaged in defeating our efforts.

- Motivated: our “win” isn’t necessarily their “lose”
- Intelligent: our assumptions are their opportunities
- Resourced: they can afford the expensive options

Attackers have been shown to statistically shift activity in small number of days

- New attack options
- New vulnerabilities
- Defeating countermeasures

Encrypted data

On production networks, more than 50% of all traffic is now encrypted:

- HTTPS is the vast bulk of this
- SSH and VPN technologies form much of the remainder

Fraction is increasing at a linear rate of about 3-7% per year

Attack surface is now mainly in the encrypted data

- Web exploits / code insertions
- Email (webmail) with attachments

Responses:

- Roll it back (unworkable)
- Proxy it (break security to enhance security; shifting targets)
- Use content-agnostic or content-inferential methods (blind spots)

Data to Drive Research

Realism vs. sensitivity

Attack presence and frequency

Benign presence and frequency

Abnormality other than attack

Realism

Obtaining data representative of production networks while avoiding data too sensitive

- Personally-identifying information / academic records
- Protected health information
- Financial access information
- Containing incriminating / embarrassing content

Non-disclosure and publication review requirements

Provably transforming data to remove protected information

- Substitution
- Conflation
- Fuzzing
- Randomization

Traffic Frequency

	Known	Unknown
Malicious	Expected	Evolving
Ambiguous	Noise	??
Benign	Needed	Allowed

Finding proper mix of content

Malicious too frequent:

- Compromised or unclean

Ambiguous too frequent:

- Over-chaotic

Benign too frequent

- Over-controlled

Known too frequent

- Generated data

Unknown too frequent

- Can’t differentiate

Summary

Finding advanced methods that actually work

Providing development, training, and evaluation data sets that realistically represent real networks of scale

Finding the threat where it is, not where we’d like to look