Big Data Sharing for Security: Is “Big Data” Worth It?

Moderator:
Sven Dietrich, City University of New York

Panelists:
Jean Camp, Indiana University Bloomington
Michael Collins, RedJack & CyberGreen
David Dittrich, University of Washington

December 7, 2017
Motivation

- State-of-the-art big data collection efforts bring together information from multiple security and non-security data sources
  - network, the desktops and servers, embedded devices in our public spaces and private homes
  - we derive newfound insights for normal and malicious behavior using machine-learning and data mining techniques

The pile gets soaked with data and starts to get mushy over time, so it's technically recurrent.  

Source: xkcd
Problem

• To better understand the Internet and their users:
  • we need to ensure reproducibility of these results
  • we may (must?) share the raw or processed data sets for independent verification and validation

• Dilemma: how do we balance the technical ability to share vs. the intended values, and the realized values of sharing?

• Frameworks in existence to describe sharing of data sets
  • Ethical guidelines (Menlo Report)
  • Challenges ahead with technical implementation and the recognition of the social impact
Panel overview

• Ongoing challenges arising from several contexts:
  – ethical data sharing with human subjects
  – network data sharing from controversial/criminal cases, and large-scale (Internet-scale) data collection
  – the related anonymity and privacy concerns
Michael Collins
RedJack & CyberGreen
The Corpus Dream

- Corpus: a large structured set of reference data
- AI has a long history of corpora for testing training and learning
  - USC-SIPI archive, Calgary Corpus
- For security, we’ve tried to construct similar
  - LL LARIAT dataset
- Proposition: We’ve inherited this tendency from AI

Image from USC-SIPI image archive
The Problem with Corpora

• Skeptical that general anonymity is feasible, or productive
  – Risks of hidden dependencies
  – We are looking for structure

• Data has a half-life
  – Have you looked at QUIC lately?

• Two experiments on the same dataset is not reproducibility
  – Maybe!
Building from Experiment Back

• Risk is an experiment-by-experiment issue
  – Consider EDA to be a private task done on internal and constrained datasets
  – Once the phenomenon is identified, then move onto an experiment with a shareable and constrained dataset – limit the fields!
  – Consider the timing and goals
    • Do you need yesterday’s Internet?
Are You Doing Security Research?

- The history of computer security, *especially intrusion detection*, is tightly intertwined with AI
- Are you solving a computer security problem or an AI problem?
- Are you *discovering* something about attacker behavior, the Internet, malware?
What Are The Next Steps?

• How do you take your research to an operational floor?
  – How does it integrate into analyst workflow?
  – How does it combine with analyst tools?
  – What are the performance constraints? Real performance constraints.

• Answers I don’t like:
  – It’s threat intelligence!
  – It’s a new detector!
  – You have to install my completely incompatible tool that is unrelated to the rest of your ops environment and involves this one attack I may have made up!
David Dittrich
University of Washington / Independent Consultant
Research with “Big Data”

• Use existing data?
  – Most common open public data sets not adequate
    • KDD Cup1999, DARPA1999, DARPA2000, LBL-PKT-4, Semifinals of FIFA World-Cup98
    – Some restricted data sets (e.g., DHS IMPACT)
  • Make your own data?
    – Very hard, not representative
  • Use “production” data
    – “Live fire” is most accurate
    – Your institution’s data
      • Good luck
    – Bring production data to you?
      • IRB issues, problems with anonymization, reputational risk
    – Go to production data?
      • IRB only applies to federal grant funded studies
      • Protection through NDA, contract
Applause from you and 7 others

Ivan Campos  Follow
@SlalomBoston consultant - currently serving as a canary in virtual coal mines.
Nov 4 · 6 min read

Private Data on Public Clouds

Low Hanging Cloud Leaks Reign Over Data Privacy Fears

Wrestling fans. Voters in Chicago. Military veterans. Over the past four months, members of these groups have all had their personal information exposed. Characterized as “cloud leaks”, this post will look to better understand what can be done to secure private data on public clouds.

“53% of organizations using cloud storage services such as Amazon Simple Storage Service (Amazon S3) had inadvertently exposed one or more such services to the public.”

— RedLock Cloud Infrastructure Security Trends (October 2017 Edition)
By default, all Amazon S3 resources are private as only the AWS account that created the resource can access them; however, this hasn't prevented the following twelve incidents from occurring over the past four months:

- **July**—**WWE customer data leaked**
  https://mackeepersecurity.com/post/world-wrestling-entertainment-leaks-3-million-emails

- **July**—**Dow Jones customer records leaked**
  https://www.upguard.com/breaches/cloud-leak-dow-jones

- **July**—**Republican National Committee records leaked via Deep Root Analytics**
  https://www.upguard.com/breaches/the-rnc-files

- **July**—**Verizon customer records leaked via third-party vendor**
  https://www.upguard.com/breaches/verizon-cloud-leak

- **August**—**Highly sensitive military data leaked via Booz Allen Hamilton**
  https://www.upguard.com/breaches/spy-games

- **August**—**Chicago voting records leaked**
  https://www.upguard.com/breaches/cloud-leak-chicago-voters

- **September**—**Resumes of military veterans with “Top Secret” US government security clearance leaked**
  https://www.upguard.com/breaches/cloud-leak-tigerswan

- **September**—**Verizon’s Distributed Vision Services data leaked**
The RNC Files: Inside the Largest US Voter Data Leak

Updated on November 28, 2017 by Dan O'Sullivan

Filed under: data breaches
DEVELOPERS: we don't need ethics
ALSO DEVELOPERS: here's how to root everyone's computer

11/28/17, 16:29

54 Retweets  173 Likes
Menlo Report

• Builds on the “Common Rule” principles

<table>
<thead>
<tr>
<th>Principle</th>
<th>Philosophy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respect for Persons</td>
<td>Deontological</td>
</tr>
<tr>
<td>Beneficence</td>
<td>Consequentialist</td>
</tr>
<tr>
<td>Justice</td>
<td>Deontological</td>
</tr>
<tr>
<td>Respect for Law and Public Interest</td>
<td>Deontological</td>
</tr>
</tbody>
</table>

• Published in Federal Register (vol. 76, no. 249, 28 Dec. 2011, p. 81517.)
• Consideration of “human harming” vs. strict “human subjects”
• Start with comprehensive stakeholder analysis
Simon Adler takes us down a technological rabbit hole of strangely contorted faces and words made out of thin air. And a wonderland full of computer scientists, journalists, and digital detectives forces us to rethink even the things we see with our very own eyes.

Oh, and by the way, we decided to put the dark secrets we learned into action, and unleash this on the internet.
Jean Camp
Indiana University Bloomington
A Tool That Looks Different to Different People
Where Does it Fit?
What is ML Big Data?

- How many routes updates does it take?


- How may data points?

What can be shared

- Certificate data
  - Phishing ✔
  - Normal ✔
  - Rogue
- Routing data
  - CAIDA
- Insider threat
Where Does it Fit?

A Modest Suggestion

- Identify a clear goal
- A structured problem
- Attackers constrained
- Identify failures
- Sample, audit, review, repeat
Examples Domains

- Certificates
  - Highly constrained structure
  - Relationships between components
  - Must present it to connect

- Routing
  - Can diffuse fake data
  - But must diffuse it to leverage attacks

- Insider threat
  - Defeat the goal not the individual
Certificates

- Relationships between fields
  - IndiaCCA
  - Cannot change just one field
    - Usually
    - Cryptography applies
Certificates

- Goals
  - Flags incidents
  - Delay trust
  - Early identification of the suspicious
Not Goals

- Remove human intelligences
- Perfect identification
- One time investigation
Time is a Necessary Variable

- Remove human intelligences
- Perfect identification
- One time investigation
- Intelligent adversaries respond
Human Intelligence is Critical

• Identify failures
• LinkedIn identifies success!!
  – Jobs identified!
  – Member retained!
Was This Success?

- Looking for
  - Consulting
  - Advisory Board positions
  - Board positions
- Found
  - Not that.
Problematic Solutions

- What is the defense against inadvertent correlated foolishness.
  - Transparency
  - Reproducibility
Transparency

• Limits the power of the analytical tools
• Allows us to understand semantics
• Attackers have it too
Reproducibility

- Reproducibility
  - Time is a critical variable
  - Fails with an intelligence adversary
    - Constant vigilance!
  - When and how do we detect this
Big Data Is Not Enough
Big Data vs, and, or Small Data

• Big data augments human intelligence
  – Constant vigilance is inhuman
  – Part of human-centered solutions

• Small, artisanal data augments big data