Can we leverage data science to improve retail data security?

Presenters:
Ashwin Swamy, Omega ATC
Thomas Duncan, Omega ATC
Agenda

• Housekeeping
• Presenters
• About Conexxus
• Presentation
• Q & A
Housekeeping

This webinar is being recorded and will be made available in approximately 30 days.

- YouTube (youtube.com/conexxusonline)
- Website Link (conexxus.org)

Slide Deck
- Survey Link – Presentation provided at end

Participants
- Ask questions via webinar interface
- Please, no vendor specific questions

Email: info@conexxus.org
Conexxus Host
Allie Russell
Conexxus
arussell@conexxus.org

Moderator
Kara Gunderson
Chair, Data Security Committee
POS Manager, CITGO Petroleum
kgunder@citgo.com

Speakers
Ashwin Swamy
Director, Omega ATC
ashwin.swamy@omegaatc.com

Thomas Duncan
Security Strategist, Omega ATC
thomas.duncan@omegaatc.com
About Conexxus

• We are an independent, non-profit, member driven technology organization

• We set standards…
  – Data exchange
  – Security
  – Mobile commerce

• We provide vision
  – Identify emerging tech/trends

• We advocate for our industry
  – Technology is policy
<table>
<thead>
<tr>
<th>Month/Date</th>
<th>Webinar Title</th>
<th>Speaker</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 24, 2018</td>
<td>QIR Program Update</td>
<td>Chris Bucolo Todd Rosen</td>
<td>ControlScan</td>
</tr>
<tr>
<td>June 21, 2018</td>
<td>Leveraging data science to improve retail data security</td>
<td>Thomas Duncan Ashwin Swamy</td>
<td>Omega ATC Omega ATC</td>
</tr>
<tr>
<td>July 26, 2018</td>
<td>Roadmap to a Vulnerability Disclosure Program</td>
<td>Joe Basirico</td>
<td>Security Innovation</td>
</tr>
<tr>
<td>August 23, 2018</td>
<td>Moving Toward Outdoor EMV</td>
<td>Linda Toth</td>
<td>Conexxus</td>
</tr>
<tr>
<td>September 2018</td>
<td>TBD</td>
<td>George Sconyers</td>
<td>Omega ATC</td>
</tr>
<tr>
<td>November 2018</td>
<td>Building a Scalable Security Engineering Team</td>
<td>Joe Basirico</td>
<td>Security Innovation</td>
</tr>
</tbody>
</table>
NACS Show
October 7-10, 2018
Las Vegas, NV

Conexxus thanks our 2018 Annual Diamond Sponsors!
Can we leverage data science to improve retail data security?

- Greater intelligence and accuracy
- Aligning security with overall analytics strategy
- Easing the burden of security management
- Grounding data science in reality

CONEXXUS
solve forward
But first, let’s understand the problem.
I was dismayed to learn this weekend about a Tesla employee who had conducted quite extensive and damaging sabotage to our operations. This included making direct code changes to the Tesla Manufacturing Operating System under false usernames and exporting large amounts of highly sensitive Tesla data to unknown third parties.

The full extent of his actions are not yet clear, but what he has admitted to so far is pretty bad. His stated motivation is that he wanted a promotion that he did not receive. In light of these actions, not promoting him was definitely the right move.

However, there may be considerably more to this situation than meets the eye, so the investigation will continue in depth this week. We need to figure out if he was acting alone or with others at Tesla and if he was working with any outside organizations.
I was dismayed to learn this weekend about a Tesla employee who had conducted quite extensive and damaging sabotage to our operations. This included making direct code changes to the Tesla Manufacturing Operating System under false usernames and exporting large amounts of highly sensitive Tesla data to unknown third parties.

The full extent of his actions are not yet clear, but what he has admitted to so far is pretty bad. His stated motivation is that he wanted a promotion that he did not receive. In light of these actions, not promoting him was definitely the right move.

However, there may be considerably more to this situation than meets the eye, so the investigation will continue in depth this week. We need to figure out if he was acting alone or with others at Tesla and if he was working with any outside organizations.
An Analogous C-Store Attack

A bad actor (internal or external) attempts to log into the POS system, site controller, fuel controller, or Electronic Payment Controller.

Social engineering attacks are frequent at c-stores. Callers or visitors can frequently pose as vendors with minimum resistance.

If the bad actor succeeds in accessing a system, he or she may try to access particular files – either to exfiltrate sensitive data or simply to disrupt operations.

In the case of stealing sensitive data, the bad actor will attempt to export or send the data using a variety of methods, such as simply sending an email, covertly changing the web address where payments are sent, using a USB drive, etc.
Human behavior leaves a trace

In both the Tesla and c-store example, multiple actions were taken by the bad actor – logging into one or more systems, accessing a file or program, and removing the data.

Hence, in these scenarios, any event logs related to user access, file behavior, and data transmission (USB drive, email, DNS settings) are related and collectively form a “fingerprint.”

The question is how to understand and detect these relationships.
So how does this relate to data science?

We must be able to filter, detect, and alert on specific events AND patterns of events.
It’s crucial to see log data from multiple perspectives...
Common problems managing a SIEM program

Endpoints and “vectors” yield a discrete set of alerts:

• No understanding of relationships between machine activity by time.
• No understanding of relationships between different processes.
• Events are not understood in historical context.
Centrally collecting logs is not enough:
• SIEM tools must be able to correlate events in a time series format.
• SIEM tools must work alongside other analysis tools that provide methods for regression and clustering.
• SIEM tools must allow the security teams the ability to make models smarter.
Common problems managing a SIEM program – events are not easily relatable by time

<table>
<thead>
<tr>
<th>Time</th>
<th>Firewall</th>
<th>POS</th>
<th>Back Office</th>
<th>EPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:00:00</td>
<td>FW Event A</td>
<td>POS Event A</td>
<td>BO Event A</td>
<td>EPC Event A</td>
</tr>
<tr>
<td>01:01:00</td>
<td>FW Event B</td>
<td>POS Event B</td>
<td>BO Event A</td>
<td>EPC Event B</td>
</tr>
<tr>
<td>01:02:00</td>
<td>FW Event C</td>
<td>POS Event A</td>
<td>BO Event B</td>
<td>EPC Event C</td>
</tr>
<tr>
<td>01:03:00</td>
<td>FW Event D</td>
<td>POS Event C</td>
<td>BO Event C</td>
<td>EPC Event D</td>
</tr>
</tbody>
</table>
SIEM – the modern approach to dumping logs
False Positives. Positively Negative.

1. A retail chain can produce thousands of alerts per year in a SIEM platform.
2. Standard application updates and processes can inadvertently trigger alerts.
3. Security teams do not always have an easy method for prioritizing alerts – no queuing methodology.
Security teams need to review alerts that truly matter, or else it might be too late.

**ATTACK TIMELINE**

- **Research & preparation**
  - Attackers find any weakness & target information on any device or service
  - 24–48 hours

- **Attacker undetected (data exfiltration)**
  - Mean dwell time 150+ days (varies by industry)
  - You may be under attack (or already compromised) and unaware

- **First host compromised**
- **Domain admin compromised**
- **Attack discovered**
  - Attackers often target Active Directory and admins to gain access to business assets
The 24-48 hour compromise window is all that is needed for initial exfiltration of data.
Can we leverage data science to improve retail data security?

- Greater intelligence and accuracy
- Aligning security with overall analytics strategy
- Easing the burden of security management
- Grounding data science in reality
Data Science is a methodology.

1. Identifying What Data Needs to be Collected and What You Want to Predict
2. Collecting the Data
3. Standardizing and Cleaning the Data
4. Exploring the Data
5. Developing Machine Learning Algorithm
6. Developing and Testing the Model with Data
7. Monitoring Performance of the Model
8. Release/Production
Identify what data needs to be collected.

- Identifying What Data Needs to be Collected and What You Want to Predict
- Collecting the Data
- Standardizing and Cleaning the Data
- Developing Machine Learning Algorithm
- Exploring the data
- Developing and Testing the Model with Data
- Monitoring Performance of the Model
- Release/Production
Identify what data needs to be collected.

Standard: PCI Data Security Standard (PCI DSS)
Version: 1.0
Date: May 2016
Author: Effective Daily Log Monitoring Special Interest Group
        PCI Security Standards Council
Identify what data needs to be collected.
We are trying to **classify** and **predict** malicious behavior
We are also trying to classify and predict what is NOT malicious behavior
Data must be sent, received, and stored.

Identifying What Data Needs to be Collected and What You Want to Predict

Collecting the Data

Standardizing and Cleaning the Data

Developing Machine Learning Algorithm

Exploring the data

Developing and Testing the Model with Data

Monitoring Performance of the Model

Release/Production
Data must be in a format that is usable.

- Identifying What Data Needs to be Collected and What You Want to Predict
- Collecting the Data
- Standardizing and Cleaning the Data
- Exploring the data
- Developing Machine Learning Algorithm
- Developing and Testing the Model with Data
- Monitoring Performance of the Model
- Release/Production
Data must be **explored** before it can be understood and applied.
Let’s explore!
ELK Stack by Elastic – an open source analytics toolkit
Methodology for demonstration

- Visualize live data to form connections and patterns
- Automated Scripts
- Real-life workflow

Elasticsearch Service

Deploy Hosted Elasticsearch and Kibana on AWS and GCP

Spin up a fully-loaded deployment on the cloud provider you choose. As the company behind Elasticsearch, we bring our real-world support to your Elastic clusters in the cloud.

FREE TRIAL!!

Start Free Trial

NO CREDIT CARD!
Objective: Attempt to gain access to a user account / file share by guessing the password for a known username.

Tips -

- Research the environment
- DON’T GET CAUGHT
  - Consider common thresholds for security alerting

**Attack #1 – brute force password attack**
Google = the hacker’s primary analytics tool

...need to find username, let's Google it!
PCI Jiu Jitsu – using the rules against you

PCI Standards for accounts?
- Lock account after 6 failed attempts
- Disable account for 30 minutes

Windows defaults?
- NONE! (Please make sure this isn't you!)
  - We will assume PCI standards

<table>
<thead>
<tr>
<th>Policy</th>
<th>Security Setting</th>
<th>Security Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account lockout duration</td>
<td>Not Applicable</td>
<td>0 passwords remembered</td>
</tr>
<tr>
<td>Account lockout threshold</td>
<td></td>
<td>42 days</td>
</tr>
<tr>
<td>Reset account lockout counter after</td>
<td></td>
<td>0 days</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 characters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disabled</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Disabled</td>
</tr>
</tbody>
</table>
Correlating events between endpoints helps identify human behavior.
Comparing login patterns can give insight into behavior.

Correlation between failed and successful login from legitimate user.
Comparing login patterns can give insight into behavior.

- **Multiple machines being targeted**
- **User Access Threshold**
  - Amount of failed logins before an alert is triggered
- **Identifiable time sequence**
Comparing login patterns can give insight into behavior.

No correlation between failed and successful logins
Log correlation helps visualize malicious behavior.

**BEFORE**
- Common thresholds
- Isolated alerts
- No visualization

**AFTER**
- Behavioral insights
- Aggregated alerts
- Visualization
Demo - Correlating events from multiple data sets helps identify and confirm anomalies in file integrity monitoring.
Change in pattern correlates with compromise of system.

Visual of successful compromise
Not all FIM is created equal.
Supporting data for anomaly detection from other sources.

File change anomaly
Other FIM considerations...

Logging the start of an application

VS.

Changes in folder path of application that has not started

Changes in folder path of the recently started application.
Without proper correlation to filter false positives or accurately assign priority, FIM will remain “noisy.”

BEFORE
• No context to changes
• Frequent alerts
• First come first serve alerting

AFTER
• Better understanding of changes
• Prioritization
• Fewer false positives
PCI rules for file integrity monitoring are the absolute baseline. Weekly review of alerts might be too late!

- Fewer false positives
- Faster response time
- Greater insights into machine behavior

11.5 Deploy a change-detection mechanism (for example, file-integrity monitoring tools) to alert personnel to unauthorized modification (including changes, additions, and deletions) of critical system files, configuration files, or content files; and configure the software to perform critical file comparisons at least weekly.
Can we leverage data science to improve retail data security?

Greater intelligence and accuracy

Aligning security with overall analytics strategy

Easing the burden of security management

Grounding data science in reality
Can we leverage data science to improve retail data security?

Greater intelligence and accuracy

Easing the burden of security management

Aligning security with overall analytics strategy

Grounding data science in reality
Apply insights gained during exploration.

- Identifying What Data Needs to be Collected and What You Want to Predict
- Collecting the Data
- Standardizing and Cleaning the Data
- Developing Machine Learning Algorithm
- Visualizing the Data (exploratory data analysis)
- Developing and Testing the Model with Data
- Monitoring Performance of the Model
- Release/Production
You must train before you can learn.

- Identifying What Data Needs to be Collected and What You Want to Predict
- Collecting the Data
- Standardizing and Cleaning the Data
- Developing Machine Learning Algorithm
- Visualizing the Data (exploratory data analysis)
- Developing and Testing the Model with Data
- Monitoring Performance of the Model
- Release/Production
Applying insights from our exploration.

\[ X = \text{multiple login attempts over regular time intervals in 2 or more machines with a sudden stoppage in login attempts after 1 successful login} \]

\[ Y = \text{specific pattern of file change activity} \]

- A pair of \( X \) and \( Y \), \((X_1,Y_1)\), should be assigned a higher probability and fall into the queue as a higher priority alert. This ensures that more likely signs of malicious behavior are being addressed first.
- Whereas the specific pattern of file change activity may represent only a 5% chance of being a sign of malicious behavior, the conditions of \( X \) AND \( Y \) being met could represent a 33% chance of malicious activity.
Applying insights from our exploration.

\( X = \text{multiple login attempts over regular time intervals in 2 or more machines with a sudden stoppage in login attempts after 1 successful login} \)

\( Y = \text{specific pattern of file change activity} \)

- When the alert is examined, the information collected in the SIEM has already established a potential relationship between X and Y. Hence, the investigation and forensic process is dramatically shortened.
- The fingerprint is readily available to see; no searching required.
Penetration testing that truly helps you “learn”...

• With data science, penetration testing can be even more valuable to security operations by feeding real data that truly represents malicious behavior.

• A SIEM program that leverages data science does not simply close tickets. Every alert is assigned a TRUE or FALSE value, which helps to train and refine models produced by the algorithm.
Analytics tools give you access to expert communities

- Hackers share knowledge and collaborate. Security teams too often operate in isolation. Knowledge exists in silos.
- Communities formed through analytics tools provide access to libraries of pre-built models, forums, data sets, open source code, and expert advice.
- Users of log management tools often share machine learning plugins.
Can we leverage data science to improve retail data security?

- Greater intelligence and accuracy
- Aligning security with overall analytics strategy
- Easing the burden of security management
- Grounding data science in reality
Can we leverage data science to improve retail data security?

- Greater intelligence and accuracy
- Aligning security with overall analytics strategy
- Easing the burden of security management
- Grounding data science in reality
Example: door counters – how marketing innovation can also support security efforts.

- Modern people traffic counters are able to measure inbound and outbound foot traffic, thereby providing an accurate count of overall store traffic at a given point in time.
- Many people counter models are able to export data in .csv formats for further analysis. Log analysis platforms can serve as an ideal tool for collecting and analyzing this data – with the potential to implement alerting.
- While data from people/traffic counters is typically used to measure the success of marketing efforts or the conversion rate of foot traffic, it can also be used for security information and event management.
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Transaction ID</th>
<th>Transaction Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:05:00 PM</td>
<td>120</td>
<td>$10.50</td>
</tr>
<tr>
<td>15:06:30 PM</td>
<td>125</td>
<td>$6.85</td>
</tr>
<tr>
<td>16:07:55 PM</td>
<td>279</td>
<td>$22.10</td>
</tr>
<tr>
<td>17:03:00 PM</td>
<td>400</td>
<td>$31.10</td>
</tr>
</tbody>
</table>
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Total Foot Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:00 – 3:59 PM</td>
<td>123</td>
</tr>
<tr>
<td>4:00 – 4:59 PM</td>
<td>59</td>
</tr>
<tr>
<td>5:00 – 5:59 PM</td>
<td>28</td>
</tr>
<tr>
<td>6:00 – 6:59 PM</td>
<td>93</td>
</tr>
</tbody>
</table>
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Total Foot Traffic</th>
<th>Transaction Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:00 – 3:59 PM</td>
<td>123</td>
<td>61</td>
</tr>
<tr>
<td>4:00 – 4:59 PM</td>
<td>59</td>
<td>28</td>
</tr>
<tr>
<td>5:00 – 5:59 PM</td>
<td>73</td>
<td>37</td>
</tr>
<tr>
<td>6:00 – 6:59 PM</td>
<td>81</td>
<td>80</td>
</tr>
</tbody>
</table>
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Firewall</th>
<th>POS</th>
<th>Back Office</th>
<th>Foot Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:00:00</td>
<td>FW Event A</td>
<td>POS Event A</td>
<td>BO Event A</td>
<td>18</td>
</tr>
<tr>
<td>01:01:00</td>
<td>FW Event B</td>
<td>POS Event B</td>
<td>BO Event A</td>
<td>20</td>
</tr>
<tr>
<td>01:02:00</td>
<td>FW Event C</td>
<td>POS Event A</td>
<td>BO Event B</td>
<td>20</td>
</tr>
<tr>
<td>01:03:00</td>
<td>FW Event D</td>
<td>POS Event C</td>
<td>BO Event C</td>
<td>18</td>
</tr>
</tbody>
</table>

“Hey John, we detected an unauthorized USB drive with a malicious file at 3AM. Our data shows only six people in the store at that time. Please contact the store manager and see if you can learn anything more.”
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Firewall</th>
<th>POS</th>
<th>Back Office</th>
<th>Conversion Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:00:00</td>
<td>FW Event A</td>
<td>POS Event A</td>
<td>BO Event A</td>
<td>49%</td>
</tr>
<tr>
<td>01:01:00</td>
<td>FW Event B</td>
<td>POS Event B</td>
<td>BO Event A</td>
<td>63%</td>
</tr>
<tr>
<td>01:02:00</td>
<td>FW Event C</td>
<td>POS Event A</td>
<td>BO Event B</td>
<td>20%</td>
</tr>
<tr>
<td>01:03:00</td>
<td>FW Event D</td>
<td>POS Event C</td>
<td>BO Event C</td>
<td>18%</td>
</tr>
</tbody>
</table>
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Firewall</th>
<th>POS</th>
<th>Back Office</th>
<th>Conversion Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:00:00</td>
<td>FW Event A</td>
<td>POS Event A</td>
<td>BO Event A</td>
<td>49%</td>
</tr>
<tr>
<td>02:00:00</td>
<td>FW Event B</td>
<td>POS Event B</td>
<td>BO Event A</td>
<td>63%</td>
</tr>
<tr>
<td>03:00:00</td>
<td>FW Event C</td>
<td>POS Event A</td>
<td>BO Event B</td>
<td>20%</td>
</tr>
<tr>
<td>04:00:00</td>
<td>FW Event D</td>
<td>POS Event C</td>
<td>BO Event C</td>
<td>18%</td>
</tr>
</tbody>
</table>
Example: door counters – how marketing innovation can also support security efforts.

<table>
<thead>
<tr>
<th>Time</th>
<th>Firewall</th>
<th>POS</th>
<th>Back Office</th>
<th>Conversion Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:00:00</td>
<td>FW Event A</td>
<td>POS Event A</td>
<td>BO Event A</td>
<td>49%</td>
</tr>
<tr>
<td>02:00:00</td>
<td>FW Event B</td>
<td>POS Event B</td>
<td>BO Event A</td>
<td>63%</td>
</tr>
<tr>
<td>03:00:00</td>
<td>FW Event C</td>
<td>POS Event A</td>
<td>BO Event B</td>
<td>20%</td>
</tr>
<tr>
<td>04:00:00</td>
<td>FW Event D</td>
<td>POS Event C</td>
<td>BO Event C</td>
<td>18%</td>
</tr>
</tbody>
</table>

Outage, maintenance, security incident, etc.  

Business impact.

CONEXXUS solve forward
Can we leverage data science to improve retail data security?

- Greater intelligence and accuracy
- Easing the burden of security management
- Aligning security with overall analytics strategy
- Grounding data science in reality
New concepts are fun to learn about, but is this really feasible?

**EXPLORE** Security teams can get started by simply downloading and exploring tools. Widely shared libraries and data sets can help you get started. Take small steps forward.

**PARTNER** Managed security service providers should increasingly bring a rigorous data science approach to their operations; always test them for their knowledge.

**UNDERSTAND** As the Tesla example shows, business operators need to understand high-level concepts related to attack patterns and methods of detection.
Can we leverage data science to improve retail data security?

- Greater intelligence and accuracy
- Aligning security with overall analytics strategy
- Easing the burden of security management
- Grounding data science in reality
Speakers

Ashwin Swamy
Director, Omega ATC
ashwin.swamy@omegaatc.com

Thomas Duncan
Security Strategist, Omega ATC
thomas.duncan@omegaatc.com
• Website: www.conexxus.org
• Email: info@conexxus.org
• LinkedIn Profile: Conexxus.org
• Follow us on Twitter: @Conexxxusonline