Machine Learning for Enterprise Security

Pratyusa K. Manadhata
Hewlett Packard Labs
manadhata@hpe.com
Outline

Evolution of enterprise security
Uses of machine learning
Strengths and weaknesses
Trends and research opportunities

Focus on problems, not on ML techniques
Enterprise security

Users

Information

Interactions

Infrastructure
A Security framework

Prevent

Detect

Recover
1st Generation: Point products
2nd Generation: Security information and event management systems (SIEM)

<table>
<thead>
<tr>
<th>Source IP</th>
<th>Destination IP</th>
<th>Event Type</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.16.110.234</td>
<td>172.16.110.234</td>
<td>GPI ATTACK RESPONSE to check returned root</td>
<td>7:41 PM</td>
</tr>
<tr>
<td>217.160.51.31</td>
<td>172.16.110.234</td>
<td>GPI ATTACK RESPONSE to check returned root</td>
<td>7:41 PM</td>
</tr>
<tr>
<td>172.16.110.234</td>
<td>172.16.110.234</td>
<td>GPI ATTACK RESPONSE to check returned root</td>
<td>7:41 PM</td>
</tr>
</tbody>
</table>

**Features:**

- **AntiVirus**
- **Unified UI**
- **Firewall**
- **False positive reduction**
3rd Generation: Security Operations Centers
1st Generation Point Products
Do one thing and do it “well”

- Anti-malware products
- Data leakage prevention products
- Firewalls
- Application firewalls
- ...

- Intrusion detection systems (IDS/IPS)
- Domain/IP Blacklists
- Sandboxes
- Web proxies
- ....
Product structure

**Platform**

- AV
- Firewall
- IDS
- ...

**Intelligence**

- AV Signatures
- Firewall Rules
- IDS Signatures
- ..
Anti-Malware products

Samples \rightarrow Label \rightarrow Benign \rightarrow Signature Generation

Samples \rightarrow Label \rightarrow Malware \rightarrow Signature Generation

Back End

Signature Matching

Front End

© Copyright 2015 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.
Sample labelling

Similarity

Machine Learning: Detect similarity at scale
Similarity is key in other detection mechanisms too

No “true” indicators of good/bad exist
Static analysis: looks similarity

- **Features**: PE Headers, instruction sequences,..
- Classification/Clustering
- False positives
- False negatives – easy to evade
Dynamic analysis: behavior similarity

- **Features**: system call sequences, API sequences, ..
- Classification/clustering
- Computational challenges
- Hard to run samples
Reputation analysis: behavior similarity

- Program behavior data from end points, no need for samples
- **Features:** popularity, source, destinations,…
- Lightweight data collection
- Lagging detection
Almost no machine learning on end devices
Complex and Fragile
Malicious domain detection

<table>
<thead>
<tr>
<th>Many Methods/Papers</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Classification/clustering/regression</td>
<td>• Syntactic properties of a domain name string</td>
</tr>
<tr>
<td>• Graph analysis</td>
<td>• HTTP/DNS protocol properties</td>
</tr>
<tr>
<td>• Pattern mining and matching</td>
<td>• Access patterns</td>
</tr>
<tr>
<td>• Statistical analysis</td>
<td>• Registration information</td>
</tr>
<tr>
<td>• ...</td>
<td>• Association with malware</td>
</tr>
<tr>
<td></td>
<td>• ...</td>
</tr>
</tbody>
</table>
A command & control domain used in the OPM breach

**opmsecurity.org**

source: https://www.threatconnect.com/opm-breach-analysis/
Observations and opportunities

ML helps with scalability, but not much with accuracy

Complex and fragile systems

ML on end devices instead of signature matching
2nd Generation
Security Information and Event Management
Correlation

Reduce alerts by grouping
Reduce false alarms
Security Information and Event Management

Billions of events / day

Filtering & Correlation

A few hundred events / day
How to generate correlation rules?
Market-basket Analysis
Observations

- Rules are heuristics driven/manually generated
- Correlation window order of minutes
3rd Generation Security Operations Centers
A large enterprise network

- HP IT supports 6 NGDC and 86 MCS
- 2.5B security events logged per day with ArcSight
- 140+ Windows Domain Controllers
- 1.2 million connected devices
- 450,000 end points protected with anti-virus
- 41K+ servers owned by HP IT
- 15K+ HPN switches
- 1,500+ enterprise HPN Routers
- Manage 150K+ mobile devices
- 11.5M+ Internet mails per day sent/received
- 597 IPS sensors deployed
- 2,000+ HP IT managed firewalls
- 450,000 mailboxes managed
- 970K+ scanned devices for vulnerabilities
- 450,000 IP Addresses including 2 contiguous Class A's
- 440K+ PCs deployed
- 39,000,000 IP Addresses including 2 contiguous Class A's
- 414K+ servers owned by HP IT
- 440K+ PCs deployed
Reducing false negatives

- AV
- DNS
- DLP
- LDAP
- DHCP
- IDS
- Firewall
- HTTP
Advanced persistent threat (APT) detection

- Infiltration
- Remote Control
- Discovery
- Exfiltration
Periodic DNS activity
Bursty DNS activity
Suspicious DNS traffic

Remote Control

Infiltration

DNS Exfiltration
HTTP/HTTPS
File formats

Discovery

Internal DNS/LDAP traffic
The data analysis approach holds promise

- Compromised account detection
- Lateral movement detection
- Anomalous user behavior
- Insider threat
- Preemptive detection – detecting early stages of an attack
Security Analytics
Data Lake

DNS          HTTP          IDS          ..          ..          AV

Magic
Scalable, Reliable, and Timely Detection
Challenges

- Infer human intent
- Adapt
- Reduce false alarms
- Preserve privacy
4th Generation Remediation & Recovery
Analyst sees an alert

192.168.0.23:43987 ➔ 203.45.65.201:1433 SQL Injection Attack 23Mar09 1930:003 user=Calvert
Analyst builds a context

192.168.0.23:43987 → 203.45.65.201:1433 SQL Injection Attack 23Mar09 1930:003 user=Calvert
Analyst follows a remediation plan

- Quarantine the infected machine
- Schedule/run clean up tools
- Schedule/run reimageing
A few minutes per alert
Target Ignored Data Breach Alarms

Target's security team reviewed -- and ignored -- urgent warnings from threat-detection tool about unknown malware spotted on the network.

SOCs don’t scale

Repetitive, Manual, and Error Prone
Machine learning to the rescue???
Context building as a learning problem

Predict the information needed for each alert

- alert characteristics
- source/destination
- protocol
- vulnerability
- ...

© Copyright 2015 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice.
Remediation plan as another learning problem

Predict remediation action for each alert

• context information
• device characteristics
• user characteristics
• location
• connectivity
• business needs
Incorporate analyst feedback
Summary
Current state

• ML has played a key role since early days of enterprise security

• ML helps achieve scale, but FPs/FNs remain

• One of the better tools at our disposal for enterprise security
Future: ML targeted toward enterprise security
Point products: Old problems require new solutions

Targeted new techniques

- False positives/False negatives

- Resource aware – power

- Advances in hardware and systems software
Security analytics and remediation: New problems

Targeted ML techniques for scalable and reliable detection

Holistic approach to detection and remediation

Incorporate user feedback
And finally, be judicious in using ML
Thank You

Pratyusa K. Manadhata
manadhata@hpe.com