APPLYING THE GUILT BY ASSOCIATION PRINCIPLE TO THREAT DETECTION WITH SPARSELY LABELED DATA

KEVIN ALEJANDRO ROUNDY
NORTONLIFELOCK RESEARCH GROUP (NRG)
“The Ass and His Purchaser” – by Aesop
Does Guilt By Association Apply in Computer Security?
Relationship Data

Relationship Data

Mobile Devices
PC’s / Laptops
IP Addresses

Apps
Software .exe .dll files
Alerts from AV, IDS, FW, ...

1
2
3
4
5
m
n
o
p
Challenges / Opportunities

1. Billions of nodes, trillions of edges
2. Miniscule amounts of labeled data of only one class
3. Massive imbalance between classes
4. Nodes representing rare entities for which we have very few observations
5. Cliques in the bipartite graph
Machine Hygiene

“Polonium: Tera-Scale Graph Mining and Inference for Malware Detection,” by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.
Machine Hygiene

“Polonium: Tera-Scale Graph Mining and Inference for Malware Detection,” by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.
“Polonium: Tera-Scale Graph Mining and Inference for Malware Detection,” by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.
Belief Propagation Parameterized for Polonium

Belief propagation is a message passing algorithm for inference on graphs using Bayesian principles.

File Prior
- .99 for GT good, .01 for GT bad, function of prevalence for other files

Machine Prior
- (aka node potential func) is a function of the machine reputation score

Edge Potential
- Machine reputation propagates along edges to files

Polonium: Tera-Scale Graph Mining and Inference for Malware Detection,” by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.
Smoke Detector

Managed Security Service Providers

MSSP running Security Operations Centers on behalf of 700 customers

100 products producing 2 trillion events per month, of 100,000 distinct event types

Normalized Event Table

<table>
<thead>
<tr>
<th>source</th>
<th>Target</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>172.18.1.6</td>
<td>172.18.11.4</td>
<td>Cross Site Scripting</td>
</tr>
<tr>
<td>172.18.1.6</td>
<td>172.18.11.4</td>
<td>SQL Injection Exploit II</td>
</tr>
<tr>
<td>172.18.1.6</td>
<td>172.18.11.4</td>
<td>Directory Traversal Attack</td>
</tr>
<tr>
<td>172.18.1.6</td>
<td>-</td>
<td>Sensitive Info. Disclosure</td>
</tr>
<tr>
<td>172.18.1.6</td>
<td>172.18.11.4</td>
<td>File Permission Bypass</td>
</tr>
<tr>
<td>172.18.1.6</td>
<td>172.18.11.4</td>
<td>Attempt to read PWD file</td>
</tr>
</tbody>
</table>
Challenges

Scale!

Security events are either neutral or bad, there are none that make you think a machine is not infected.

The noisier an event is, the more highly connected it is, and the more influential it is in the graph.

<table>
<thead>
<tr>
<th>Instance Count</th>
<th>Confidence</th>
<th>Signature Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>44,104,172,561</td>
<td>0.01</td>
<td>TCP Connection</td>
</tr>
<tr>
<td>21,396,843,738</td>
<td>0.00</td>
<td>Traffic</td>
</tr>
<tr>
<td>19,493,074,472</td>
<td>0.00</td>
<td>UPD Connection</td>
</tr>
<tr>
<td>15,161,094,870</td>
<td>0.00</td>
<td>Firewall</td>
</tr>
<tr>
<td>14,586,679,865</td>
<td>0.02</td>
<td>Teardown TCP Connection</td>
</tr>
<tr>
<td>6,905,911,034</td>
<td>0.00</td>
<td>Teardown UDP Connection</td>
</tr>
<tr>
<td>6,762,250,252</td>
<td>0.00</td>
<td>Flow Session Close</td>
</tr>
<tr>
<td>4,350,063,989</td>
<td>0.01</td>
<td>PIX-6-305012</td>
</tr>
<tr>
<td>3,790,388,695</td>
<td>0.01</td>
<td>Connection Discarded</td>
</tr>
<tr>
<td>2,626,258,118</td>
<td>0.02</td>
<td>TPC Cache Miss</td>
</tr>
<tr>
<td>2,381,224,704</td>
<td>0.00</td>
<td>HTTP Get</td>
</tr>
<tr>
<td>2,109,964,562</td>
<td>0.02</td>
<td>Packet Permitted</td>
</tr>
<tr>
<td>1,676,587,290</td>
<td>0.02</td>
<td>Connection Allowed</td>
</tr>
</tbody>
</table>
Approach: Learning from Semi-Supervised Data and Relationships

Events
- Cross Site Scripting
- SQL Injection Exploit II
- Directory Traversal Attack
- Sensitive Info. Disclosure
- File Permission Bypass
- Attempt to read PWD file
- System Infected CnC Traffic
- Traffic
- Trojan.Zbot

Machine, Time-Window pairs

Weight events based on the probability with which they occur on infected machines.

Symantec Research Labs
Part 2: Random Walk with Restart for Incident Ranking

A random particle starts traversing the graph from incident node $i$ (restart node)

Transmits to an out-neighbor with a probability $\propto$ to the edge weights

At each step, transports back to node $i$ with a small probability $c = 0.15$

Node $j$'s relevance to node $i$ is the probability that the particle will be found at node $j$

\[
\vec{r} = (1 - c) \tilde{W} \vec{r} + c \vec{e}
\]
Creepware

Intimate Partner Violence and Stalkerware

How To Spy Your Girlfriend's Android Phone
https://www.youtube.com/watch?v=ZL6zWUSaE6Y

With this simple app you can easily hear everything what your girlfriend is doing all day

May 9, 2016 - Uploaded by Nagy Sándor
Link: https://play.google.com/store/apps
Have you ever...
Stalkerware Discovery Through Co-Occurrence
Sequences of installed apps

1. Track a Phone by Number
2. Find My Friends
3. Live Mobile Location Tracker
4. SMS from PC / Tablet & MMS Text Messaging Sync
5. HelloSpy

1. GirlFriend Cell Tracker
2. System Services
3. Hidden Auto Call Recorder
4. Family Locator - GPS Tracking
5. SMS Forwarder
Searching for Guilt By Association In a Pure Form

Prior information limited to 18 covert stalkerware apps

No features

Parameterless algorithm
## Graph Mining Algorithm Walkthrough

<table>
<thead>
<tr>
<th>Round</th>
<th>Device scores</th>
<th>Normalized app scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.474</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.469</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.469</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.468</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.468</td>
</tr>
</tbody>
</table>

**Diagram:**
- **Mobile Devices:** 1, 2, 3
- **Apps:** m, n, o, p

The table shows device and app scores across different rounds, with normalized app scores calculated based on device scores.
False Positive Woes

- $\text{App}_q$ scores higher than $\text{App}_s$, which is intuitively wrong.

- The % of apps that are stalkerware-related is unknown but very low, many rare apps like $\text{App}_q$ co-occur with stalkerware by random chance, leading to high False Positive rates.
Controlling FP’s By Incorporating a Skeptical Prior

If we model $\theta$ as a binomial distribution with a Beta prior, we get a closed form solution for the posterior, whose mode is the Maximum a Posteriori estimate.

Mode of Beta (1.1, 185)

$$\text{Mode of Beta (1.1, 185)} = \frac{1.1 + \text{hits} - 1}{1.1 + 185 + \text{trials} - 2}$$

For a derivation see:
Sub-Category Counts in 2017 (similar distribution in 2018 and 2019)
Bombing

Message Bomber send 5000+ sms
1.8 for Android

Rajan Patel

Download APK (3.0 MB) Versions

Using APKPure App to upgrade Message Bomber, fast, free and save your internet data.
Cliques = Software Packages?

“Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs,” by Acar Tamersoy, Kevin A. Roundy, Duen Horng Chau. Knowledge and Data Mining Conference (KDD), Industry Track, 2011.
Software Files Are Not Distributed Individually
Files Get Distributed As Software Packages, Not Individually

<table>
<thead>
<tr>
<th>Device 1</th>
<th>Device 2</th>
<th>Device 3</th>
<th>Device 4</th>
<th>Device 5</th>
<th>Device 6</th>
<th>Device 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Aesop: Software Packages Form Cliques in the Install Graph

<table>
<thead>
<tr>
<th>File</th>
<th>Set of machines containing the file</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1$</td>
<td>$M_{f_1} = {m_2, m_4, m_5, m_8}$</td>
</tr>
<tr>
<td>$f_2$</td>
<td>$M_{f_2} = {m_3, m_5, m_7}$</td>
</tr>
<tr>
<td>$f_3$</td>
<td>$M_{f_3} = {m_1, m_3, m_5, m_6, m_7}$</td>
</tr>
</tbody>
</table>
Comparing File Install Lists at Scale

11 Billion files and 400 million machines

- Popular files occur up to 200,000 times
- Huge numbers of rare files (10.5 billion files occur only once)

MinHashing

- Map each machine to a random numeric ID with a hash function
- A file’s minhash is the smallest machine ID in its install list

Locality Sensitive Hashing

- We compute minhash $r$ times for each file, and a string composed of the $r$ minhashes is the file’s hash bucket
- Create $b$ bands of $r$ minhashes to get more coverage
Graph Transformation Within LSH Buckets

Files  Machines

Buckets exhibit homophily, we propagate labels within buckets
Aesop Detections on its First Day...

700 million benign files
1 million malware files
Guilt By Association Unlocks Power in Relationship Data

Mobile Devices
PC’s / Laptops
IP Addresses

Apps
Software .exe .dll files
Alerts from AV, IDS, FW, ...

1
2
3
4
5

m
n
o
p
Many Guilt-Based Security Systems @ NortonLifeLock / Symantec

Polonium

“Polonium: Tera-Scale Graph Mining and Inference for Malware Detection,” by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.

Aesop

“Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs,” by Acar Tamersoy, Kevin A. Roundy, Duen Horng Chau. Knowledge and Data Mining Conference (KDD), Industry Track, 2011.

Creepware


Smoke Detector


Adage

**Polonium**

“Polonium: Tera-Scale Graph Mining and Inference for Malware Detection,” by Duen Horng Chau, Carey Nachenberg, Jeffrey Wilhelm, Adam Wright, Christos Faloutsos. SIAM Data Mining Conference (SDM) 2011.

**Aesop**

“Guilt by Association: Large Scale Malware Detection by Mining File-relation Graphs,” by Acar Tamersoy, Kevin A. Roundy, Duen Horng Chau. Knowledge and Data Mining Conference (KDD), Industry Track, 2011.

**Creepware**


**Smoke Detector**


**Adage**