Empirical Analysis of Factors Affecting Malware URL Detection

Marie Vasek & Tyler Moore
SMU, Dallas, TX
mvasek@smu.edu

eCrime Researchers Summit

September 17, 2013
So Much Malware
So Much Malware

System is infected!

WARNING!
Your system is infected

Report

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Alert level</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-case.win32</td>
<td>Spyware</td>
<td>Average</td>
</tr>
<tr>
<td>CoolWebSearch(CWS).win32</td>
<td>Spyware</td>
<td>Danger</td>
</tr>
<tr>
<td>I-Worm.Sobig</td>
<td>Virus</td>
<td>High</td>
</tr>
<tr>
<td>Backdoor.SdBot.gen</td>
<td>Virus</td>
<td>Critical</td>
</tr>
<tr>
<td>TrojanDropper.JS.Mimail</td>
<td>Virus</td>
<td>Critical</td>
</tr>
</tbody>
</table>

Total infected files detected: 45

Recommendation: click the "Repair infected" button to remove all spyware and viruses from your computer.
So Many AntiVirus Products
Research Questions

Our hypothesis: At least part of the difference in performance across blacklists can be explained by characteristics of the malware URL or the blacklisting service.

1. Impact of type of malware URL on blacklisting?
   ○ Payload type (exploit kit, Fake AV, etc.)
   ○ URL features (IP address, path)

2. Impact of type of blacklisting service?
   ○ Blocking or advisory
   ○ Cost
# Table of Contents

- Intro
- Data Sources
- Methodology
- Data
- Factors
- Results
- Conclusion
# Malware Domain List

- Publicly Available
- Low Volume
  - 722 URLs in our collection
  - *BUT* it’s free, open, & no data-sharing issues

### Malware Domain List

<table>
<thead>
<tr>
<th>Date (UTC)</th>
<th>Domain</th>
<th>IP</th>
<th>Reverse Lookup</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013/07/11 12:54</td>
<td><a href="http://www.keeperstributes.com/calc/images/ac0094cbbe/?wMw1mLulWYt9VbzxXO5IDMxIzN3QDOSUD02w3LJmYjRT0wAzYh9ycldWYtl2LjxWYj9SbvNmlZVGd1JWayRXZrF2cwVWZr5yd3d3LvoDc0RHa8NnZ">www.keeperstributes.com/calc/images/ac0094cbbe/?wMw1mLulWYt9VbzxXO5IDMxIzN3QDOSUD02w3LJmYjRT0wAzYh9ycldWYtl2LjxWYj9SbvNmlZVGd1JWayRXZrF2cwVWZr5yd3d3LvoDc0RHa8NnZ</a></td>
<td>69.16.206.133</td>
<td>host2.jbatkins.com.</td>
<td>exploit kit</td>
</tr>
<tr>
<td>2013/07/11 12:54</td>
<td><a href="http://www.lowes-pianos-and-organs.com/images/d521c2a038/?zAXbu4Wah12XtNHf0YTNwTNw1JiM0ATMzIDvgzMwEmMjFjM1">www.lowes-pianos-and-organs.com/images/d521c2a038/?zAXbu4Wah12XtNHf0YTNwTNw1JiM0ATMzIDvgzMwEmMjFjM1</a></td>
<td>67.222.109.112</td>
<td>d15.altserver.com.</td>
<td></td>
</tr>
</tbody>
</table>
VirusTotal

- Publicly Available
- API accessible
- Google-Owned
- Checks each URL against 38 blacklists
Table of Contents

Intro

Data Sources

Methodology

Data

Factors

Results

Conclusion
What We Do

1. Check MDL for new URLs.
2. Scan new URLs through VT.
3. Scan new URL every other hour for 48 hours after.
4. Afterwards, scan every day for 14 days.
Table of Contents

Intro
Data Sources
Methodology
Data
Factors
Results
Conclusion
How Blacklisting Detection Rates Vary

![Histogram showing how blacklist detection rates vary across different blacklist services. The x-axis represents the blacklist service (rank order), and the y-axis represents the percent of URLs detected. The distribution shows a decreasing trend in detection rates as the blacklist service increases in rank order.]
How Blacklisting Detection Rates Vary
Percentage of URLs Detected by at least $n$ Services
Blacklisting over Time

![Graph showing the mean number of positives over hours since URL appeared in MDL. The x-axis represents hours since URL appeared in MDL, ranging from 0 to 400 hours. The y-axis represents the mean number of positives, ranging from 2 to 6.]
Table of Contents

Intro
Data Sources
Methodology
Data
Factors
Results
Conclusion
Research Questions

Our hypothesis: At least part of the difference in performance across blacklists can be explained by characteristics of the malware URL or the blacklisting service.

1. Impact of type of malware URL on blacklisting?
   - Payload type (exploit kit, Fake AV, etc.)
   - URL features (IP address, path)

2. Impact of type of blacklisting service?
   - Blocking or advisory
   - Cost
Malware Type
IP/Domain

127.0.0.1/bad.html

icanhazcats.com/milk.php
Path?

fluffybunnies.org/
goldphish.net/bubbles.php
## Type of Malware

<table>
<thead>
<tr>
<th>Malware Type</th>
<th>#</th>
<th>%</th>
<th>IP/Domain</th>
<th>#</th>
<th>%</th>
<th>Path?</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executable</td>
<td>175</td>
<td>24%</td>
<td>IP Address</td>
<td>124</td>
<td>17%</td>
<td>Has Path</td>
<td>675</td>
<td>93%</td>
</tr>
<tr>
<td>Fake AV</td>
<td>65</td>
<td>9%</td>
<td>Domain</td>
<td>598</td>
<td>83%</td>
<td>No Path</td>
<td>47</td>
<td>7%</td>
</tr>
<tr>
<td>Styx</td>
<td>51</td>
<td>7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blackhole Lnd.</td>
<td>149</td>
<td>21%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>282</td>
<td>39%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Blocks?

Web Page Blocked

Access to the web page you were trying to visit has been blocked in accordance with company policy. Please contact your system administrator if you believe this is in error.

**User:** 129.119.9.166

**URL:** www.malwareconference.org/index.php?option=com_docman_task=cat_view_gid=79_Itemid=51

**Category:** malware-sites
Costs Money
## Blacklist Types

<table>
<thead>
<tr>
<th>Blocks?</th>
<th>Costs?</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>Paid AV</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>Free AV &amp; Web Browser Blacklists</td>
</tr>
<tr>
<td>✗</td>
<td>✓</td>
<td>Malware Site Checkers</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>Malware URL lists</td>
</tr>
</tbody>
</table>
# Blacklist Types

<table>
<thead>
<tr>
<th>Blocks?</th>
<th>#</th>
<th>%</th>
<th>Costs?</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks Users</td>
<td>22</td>
<td>58%</td>
<td>Costs Money</td>
<td>17</td>
<td>45%</td>
</tr>
<tr>
<td>Doesn’t Block</td>
<td>16</td>
<td>42%</td>
<td>Free</td>
<td>21</td>
<td>55%</td>
</tr>
</tbody>
</table>
Table of Contents

Intro

Data Sources

Methodology

Data

Factors

Results

Conclusion
Logistic Regression: URL ever Blacklisted by Service?

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Odds Ratio</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td>-2.304</td>
<td>0.100</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>URL features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP address</td>
<td>-0.004</td>
<td>0.996</td>
<td>0.945</td>
</tr>
<tr>
<td><strong>Has a Path?</strong></td>
<td>-0.196</td>
<td>0.822</td>
<td>0.031</td>
</tr>
<tr>
<td>Executable</td>
<td>1.017</td>
<td>2.765</td>
<td>0.000</td>
</tr>
<tr>
<td>Fake AV</td>
<td>-0.838</td>
<td>0.433</td>
<td>0.000</td>
</tr>
<tr>
<td>Styx</td>
<td>0.746</td>
<td>2.109</td>
<td>0.000</td>
</tr>
<tr>
<td>Blackhole Landing Page</td>
<td>0.196</td>
<td>1.217</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Malware Blacklist Features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks Users?</td>
<td>0.611</td>
<td>1.843</td>
<td>0.000</td>
</tr>
<tr>
<td>Costs Money</td>
<td>0.298</td>
<td>1.347</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 1144.501, \ p \text{ value} = 0.000 \]
Survival Regression: Blacklist Timing

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Odds Ratio</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP address</td>
<td>0.056</td>
<td>1.058</td>
<td>0.210</td>
</tr>
<tr>
<td>Has a Path?</td>
<td>-0.207</td>
<td><strong>0.811</strong></td>
<td>0.012</td>
</tr>
<tr>
<td>Executable</td>
<td>0.896</td>
<td><strong>2.449</strong></td>
<td>0.000</td>
</tr>
<tr>
<td>Fake AV</td>
<td>-0.814</td>
<td><strong>0.443</strong></td>
<td>0.000</td>
</tr>
<tr>
<td>Styx</td>
<td>0.750</td>
<td><strong>2.118</strong></td>
<td>0.000</td>
</tr>
<tr>
<td>Blackhole Landing Page</td>
<td>0.179</td>
<td><strong>1.196</strong></td>
<td>0.000</td>
</tr>
<tr>
<td>Malware Blacklist Features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks Users?</td>
<td>0.538</td>
<td><strong>1.713</strong></td>
<td>0.000</td>
</tr>
<tr>
<td>Costs Money</td>
<td>0.300</td>
<td><strong>1.351</strong></td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ R^2 = 0.055 \]
Survival Regression

Fraction of websites unblacklisted after x hours

Hours since URL appeared in MDL

IP address
domain
Survival Regression
Survival Regression

- Hours since URL appeared in MDL
- Fraction of websites unblacklisted after x hours
- Exe
- Styx
- Blackhole
- Fake AV
- other
Survival Regression

- Blocks & Costs Money
- Doesn't Block & Costs Money
- Blocks & Free
- Doesn't Block & Free

Fraction of websites unblacklisted after x hours vs. Hours since URL appeared in MDL.
Table of Contents

Intro
Data Sources
Methodology
Data
Factors
Results
Conclusion
Summary

- Known Exploit kits more likely to be blacklisted
  - ... but Fake AV is not ..?
- Executables more likely to be blacklisted
- Blocking users → more likely to blacklist
- Costing money → more likely to blacklist
Limitations

- **MDL → low volume**
  - other exploit kits
  - AS / DNS features
  - Malicious vs. Hacked?

- **MDL → used as input to blacklists**
  - Blacklisted b/c found on VT or found organically?
  - Are things on MDL still bad on day $n$?

- **VT → not representative sample of malware blacklists**
  - Blacklists not on VT?
  - Blacklists on VT that didn’t blacklist anything?
Questions?