Opening the Blackbox of VirusTotal: Analyzing Online Phishing Scan Engines

Peng Peng¹, Limin Yang², Linhai Song³, Gang Wang²

¹Virginia Tech
²University of Illinois at Urbana-Champaign
³The Pennsylvania State University
Widely-used Online Scanning Engine: VirusTotal

- **Research usage**: data labeling, system evaluation
  - Malware scanning: IMC’18, CCS’18, USENIX Security’18, etc.
  - IP and URL scanning: IMC’14, IMC’18, WWW’18, etc.

- **Industry and government usage**: threat alerts

U.S. Cyber Command (Department of Defense)
The Challenge of the “Blackbox”

New VirusTotal hash causes drop in antivirus detection rates

Questions were raised about how antivirus vendors use the VirusTotal database after a researcher highlighted a significant drop in malware detection rates following an upload of a new VirusTotal hash.

By Michael Heller, Senior Reporter

Better Malware Ground Truth: Techniques for Weighting Anti-Virus Vendor Labels

Need a better understanding of how VirusTotal works
Research Questions

• Preliminary exploration
  • Investigating the workflow of VirusTotal and its 60+ security vendors
  • Using phishing and URL scanning engine for the case study

• Aspects interested
  • Detection accuracy
  • Label update delay
  • Result inconsistency
  • Reaction to phishing take-down
Experiment Settings
Detailed ethical considerations discussed in the paper.

- VirusTotal Database
  - Scan API
  - Submit URLs
  - Retrieve Reports
- VirusTotal Service
- Phishing Sites Under Our Control
- Accuracy and Delay
- Result Consistency
- Monitor Traffic
- 68 third-party Vendors
  - Registered domains, no prior history
- 18 out of 68 vendors have their own APIs
Using “Google Safe Browsing” as an example third-party vendor

Phishing page URL submitted to Phishing page URL submitted to 3rd-party scanner

“Submit URL to VirusTotal Scanning twice a week”
VirusTotal Labels Over Time

Result of 1st scan gets updated after 2nd scan

VirusTotal failed to detect IRS page during the first week

Avg. malicious labels per site

Time (weeks)
Detection Accuracy

- 18 PayPal sites + 18 IRS sites (36 sites in total)
- Only **15 (out of 68)** VirusTotal vendors have flagged one or more sites

<table>
<thead>
<tr>
<th>Rank</th>
<th>Vendor</th>
<th>Detected Total</th>
<th>Detected PayPal</th>
<th>Detected IRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Netcraft</td>
<td>26</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>Emsisoft</td>
<td>26</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>Fortinet</td>
<td>26</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Sophos</td>
<td>23</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>CRDF</td>
<td>17</td>
<td>14</td>
<td>3</td>
</tr>
</tbody>
</table>

Highly synchronized actions: need further research
Label Inconsistency
Between VirusTotal scan and the Vendors’ own API scan

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Brand</th>
<th>VirusTotal Before Week 2</th>
<th>Vendor Snapshot Results (week 2)</th>
<th>VirusTotal After Week 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forcepoint</td>
<td>PayPal</td>
<td>benign</td>
<td>malicious</td>
<td>benign</td>
</tr>
<tr>
<td>Fortinet</td>
<td>IRS</td>
<td>benign</td>
<td>malicious</td>
<td>benign</td>
</tr>
<tr>
<td>Netcraft</td>
<td>IRS</td>
<td>benign</td>
<td>malicious</td>
<td>malicious</td>
</tr>
<tr>
<td>5 more vendors</td>
<td>PayPal</td>
<td>benign</td>
<td>malicious</td>
<td>benign</td>
</tr>
</tbody>
</table>

Vendors scan results different from VirusTotal results (of the same vendor)
Reaction to Phishing Take-down

- 15 vendors detected at least phishing sites
- Only 4 vendors have some “reaction” to phishing site take-down

Change all malicious labels to benign
Additional Experiments

- Robots.txt
  - Most scanners would ignore robots.txt

- Impact of Obfuscation
  - Redirection: URL shortener
  - Image-based: overlay the login form on top of the webpage screenshot
  - PHP obfuscation: replace variable names and output ASCII encoding

<table>
<thead>
<tr>
<th>Obfuscation Method</th>
<th>Redirection</th>
<th>Image</th>
<th>PHP Code</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. # Vendors Detected the Site</td>
<td>12</td>
<td>4.5</td>
<td>2</td>
<td>12.1</td>
</tr>
</tbody>
</table>
Discussion

• Takeaways of using VirusTotal for phishing scanning
  • Accuracy is not high on “fresh” phishing sites
  • Scan the URL twice to force the VirusTotal database to update the scan result
  • Cross-check the labels with the vendors’ own APIs

• Future Work
  • Correlation of labels from different scanners
  • Aggregation strategies to improve label quality
  • Scanner performance on different phishing brands

Experiment data available: https://github.com/whyisyoung/virustotal
Thank You

Please kindly forward your questions to the author

pengp17@vt.edu
Backup Slides
Screenshots for Phishing Sites