Measuring and Modeling the Label Dynamics of Online Anti-Malware Engines

Shuofei Zhu\textsuperscript{1}, Jianjun Shi\textsuperscript{1,2}, Limin Yang\textsuperscript{3}
Boqin Qin\textsuperscript{1,4}, Ziyi Zhang\textsuperscript{1,5}, Linhai Song\textsuperscript{1}, Gang Wang\textsuperscript{3}

\textsuperscript{1}The Pennsylvania State University
\textsuperscript{2}Beijing Institute of Technology
\textsuperscript{3}University of Illinois at Urbana-Champaign
\textsuperscript{4}Beijing University of Posts and Telecommunications
\textsuperscript{5}University of Science and Technology of China
VirusTotal

• The largest online anti-malware scanning service
  – Applies 70+ anti-malware engines
  – Provides analysis reports and rich metadata

• Widely used by researchers in the security community
Challenges of Using VirusTotal

• Q1: When VirusTotal labels are trustworthy?
Challenges of Using VirusTotal

• Q1: When VirusTotal labels are trustworthy?
• Q2: How to aggregate labels from different engines?
• Q3: Are different engines equally trustworthy?

<table>
<thead>
<tr>
<th>Engines</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>McAfee</td>
<td>⚠️</td>
</tr>
<tr>
<td>Microsoft</td>
<td>✔️</td>
</tr>
<tr>
<td>Kaspersky</td>
<td>✔️</td>
</tr>
<tr>
<td>Avast</td>
<td>⚠️</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Challenges of Using VirusTotal

• Q1: When VirusTotal labels are trustworthy?
• Q2: How to aggregate labels from different engines?
• Q3: Are different engines equally trustworthy?
• Surveyed 115 top-tier conference papers that use VirusTotal
• Our findings:
  – Q1: rarely consider label changes
  – Q2: commonly use threshold-based aggregation methods
  – Q3: often treat different VirusTotal engines equally
Overview

• Q1: the impact of label changes (label flips)

• Q2: threshold-based label aggregation methods

• Q3: the correlation between VirusTotal engines

Equally trustworthy?
• Q1: the impact of label changes (label flips)
• Q2: threshold-based label aggregation methods
• Q3: the correlation between VirusTotal engines
Data Collection of the Main Dataset

• We chose “fresh” files without prior VirusTotal history
  – Sampled 14,423 files submitted for the first-time on 08/31/2018
  – Roughly half were labeled as “benign” by all engines on day-1
  – The rest were labeled as "malicious" by at least 1 engine on day-1

• We collected “daily” VirusTotal labels over one year
  – Use rescan API to force VirusTotal to scan the samples everyday
  – Data collection window: 08/31/2018 – 09/30/2019

• Data Preprocessing
  – 341+ million data points from 65 engines
Outline

• Q1: the impact of label changes (label flips)

• Q2: threshold-based label aggregation methods

• Q3: the correlation between VirusTotal engines

Equally trustworthy?
We model the label dynamics by sequences of “0” and “1”
- ✔️ (benign): 0 ➖ (malicious): 1

A Flip: 0 → 1 or 1 → 0
- hazard flip: temporary, lasts only one day
- non-hazard flip: long term, lasts at least two days
### Characteristics of Flips

<table>
<thead>
<tr>
<th>No. of flips (10K)</th>
<th>% of engines</th>
<th>% of files</th>
</tr>
</thead>
<tbody>
<tr>
<td>weeks</td>
<td>normalized flips per engine</td>
<td>normalized flips per file</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>hazard flips</td>
<td></td>
</tr>
<tr>
<td></td>
<td>all flips</td>
<td>9%</td>
</tr>
</tbody>
</table>

Both flips and hazard flips widely exist across scan dates, engines and files.
• How long to wait for a file’s labels to become stable?
• Stable file: all engines' labels on the file do not change any more

Waiting for longer time does not guarantee to have more stable files.
• Q1: the impact of label changes (label flips)

• Q2: threshold-based label aggregation methods

• Q3: the correlation between VirusTotal engines

Equally trustworthy?
Aggregated Label Stabilization

- Many researchers use a threshold \( t \) to aggregate engines' labels
  - A file is considered as malicious, when \( \geq t \) engines detect the file

- How flips impact this aggregation policy?
  - Influenced files: files with both benign and malicious aggregated labels
  - Measure % of influenced files for different \( t \)

\[
\begin{array}{c|ccccc}
\text{Day} & 1 & 2 & 3 & 4 & 5 & \ldots \\
\hline
\text{McAfee} & 1 & 1 & 1 & 1 & 0 & \ldots \\
\text{Microsoft} & 1 & 1 & 1 & 0 & 0 & \ldots \\
\text{Kaspersky} & 0 & 1 & 0 & 0 & 0 & \ldots \\
\text{\ldots (62 engines)} & \text{\ldots (all 0)} & & & & & \\
\text{Aggregated labels} & 1 & 1 & 1 & 0 & 0 & \ldots \\
\end{array}
\]

\( (t = 2) \)
Many researchers use a threshold \((t)\) to aggregate engines' labels

- A file is considered as malicious, when \(\geq t\) engines detect the file

How flips impact this aggregation policy?

- Influenced files: files with both benign and malicious aggregated labels
- Measure % of influenced files for different \(t\)

Flips can heavily influence labeling aggregation results when threshold \(t\) is too small or too large.
• Q1: the impact of label changes (label flips)

• Q2: threshold-based label aggregation methods

• Q3: the correlation between VirusTotal engines

Equally trustworthy?
Temporary Labeling Similarity

• How to compute the similarity between engines A and B?
  – Compute the similarity between the two labeling sequences for each file
  – Compute the average sequence-level similarity over all the files

• An example for sequence-level similarity

  engine A on file X: \( (0, 1, 0, 0, 0, 0, 0, 0, 0, 0)01000... \)
  \( (1, 1, 1, 5)0(0, 0, 0, 7, 0, ...) \) cosine

  engine B on file X: \( 0000000001000000100... \)
  \( (0, 0, 0, 7, 1, 1, 1, 4, ...) \) cosine

0.87
Label Correlations Between Engines

- Avast
- AVG
- K7GW
- K7AntiVirus
- Gdata
- ESET-NOD32
- BitDefender
- Ad-Aware
- Emsisoft
- MicroWorld-eScan
Label Correlations Between Engines

- MicroWorld-eScan
- Emsisoft
- GData
- Ad-Aware
- BitDefender

The diagrams show the label trends over 300 days.
There are groups of engines with strong correlations in labeling decisions.
Outline

• Q1: the impact of label changes (label flips)

• Q2: threshold-based label aggregation methods

• Q3: the correlation between VirusTotal engines

PE

APK

Ground-Truth Dataset

no ground-truth

has ground-truth

Equally trustworthy?
How we create “fresh” ground-truth samples?
- Obfuscating ransomware to create malware
- Obfuscation + compiling open-source software to create goodware

Findings:
- Obfuscation brings many false positives
  • Even for high-reputation engines
- $3 \leq t \leq 15$ can produce good aggregation results
  • As long as the benign files are not obfuscated
- Inconsistency exists between the desktop and the VirusTotal versions

More results in our paper...
Conclusion and Takeaways

• A paper survey on how researchers use VirusTotal
• Data-driven methods to validate labeling methodologies

• Takeaways and suggestions
  – Data preprocessing
    • Submit the same files in 3 consecutive days to detect hazards
    • No need to wait over long time
  – Threshold-based label aggregation
    • Stable: when t is within a reasonable range (2-20)
    • Correctness: t = 3 to 15 when benign files are not obfuscated
  – Correlation and causality exists between engines
  – High-reputation Engines are not always accurate
Thank you!

• Also thanks to my collaborators

• Contact
  – sfzhu@psu.edu

• Artifact