VAMO: Towards a Fully Automated Malware Clustering Validity Analysis

Roberto Perdisci and ManChon U
Malware Clustering

• Clustering malware into *families* is useful

Malware Triage

- **Red**
  - Respond Immediately
  - 1 hr SLA

- **Orange**
  - Research Immediately
  - 4 hr SLA

- **Yellow**
  - Research within 12 hrs
  - 24 hr SLA

- **Green**
  - Daily Event Report
  - Review within 2 days

Use to generate better signatures

New family
Malware Clustering Research

• Bailey et al. *Automated classification and analysis of internet malware* (RAID'07)

• Bayer et al. *Scalable, behavior-based malware clustering* (NDSS'09)

• Hu et al. *Large-scale malware indexing using function-call graphs* (CCS'09)

• Perdisci et al. *Behavioral clustering of http-based malware and signature generation using malicious network traces* (NSDI'10)

• Jang et al. *Bitshred: feature hashing malware for scalable triage and semantic analysis* (CCS'11)
Validating Clustering Results

• How do we know clustering output is good?
  – Need a reference clustering to compare
  – Challenge: unsupervised learning
  – Limited or no ground truth

• Reference clustering (previous work)
  – Use multiple AV scanners
  – Extract family names from AV labels
  – Samples that are assigned the same label by majority of AVs are considered in same family
Drawbacks of Majority Voting

- Different AV vendors use different notation
  - Different family names
  - One-to-many mapping
  - Missing and inconsistent labels

- Difficult to find majority consensus
  - Samples with no consensus are excluded
Drawbacks of Majority Voting

Majority consensus found only for a fraction of dataset!

\[
\frac{2,658}{14,212} = 18.7 \%
\]

Bayer et al. (NDSS’09)  
Our malware dataset  
5.6% of 1.1M samples

Reference clustering built using majority voting not representative of dataset

Li et al. *On challenges in evaluating malware clustering* (RAID'10)

“existing approaches to obtaining ground-truth data for malware clustering evaluation biases results by isolating those instances that are simple to cluster”
VAMO

Validity Analysis of Malware-clustering Outputs

• Research Goals
  – Consider entire malware dataset for validation
  – No manual mappings between AV labels
  – Deal with AV naming inconsistencies
  – Fully automated
Third-party malware clustering system

- Historic Malware Archive
- Malware Dataset
- Malware Clustering Process
- Clustering Results

AV Label Graph

Build Reference Clustering

Validity Analysis

VAMO

- Enables tuning clustering parameters
- Allows comparison of different systems
AV Label Graph

- Learns mappings between AV labels
- Labels that often appear “together” are considered similar

Node = <AV>_<Family>
Edge Weight = Label distance

**Sample X**
- **McAfee**
  - W32/Virut
- **Avira**
  - TR/Drop
- **Trend**
  - PE_VIRUT

**Historic Malware Archive**

<table>
<thead>
<tr>
<th>AV Label Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>McAfee</td>
</tr>
<tr>
<td>Avira</td>
</tr>
<tr>
<td>Trend</td>
</tr>
<tr>
<td>Sample X</td>
</tr>
<tr>
<td>W32/Virut</td>
</tr>
<tr>
<td>TR/Drop</td>
</tr>
<tr>
<td>PE_VIRUT</td>
</tr>
</tbody>
</table>
Building Reference Clustering

- Measure distance between each pair of malware samples in dataset

---

**Sample X:**

\[ d = \text{med}(d_1, d_2, ..., d_n) \]

**Sample Y:**

\[ d_1 \]

---

AV1_Lx

AV2_Lx

... AVn_Lx

AV1_Ly

AV2_Ly

... AVn_Ly
Building Reference Clustering

• Apply average-linkage hierarchical clustering on distance matrix
Computing Validity Indices

Third-Party Clustering Results

Validity Analysis

Reference Clustering

3rd-party Clustering Results

External Validity Index (e.g., Jaccard)

Quality Index = max(validity index)

AV label-based Reference Clustering

University of Georgia
Dept. of Computer Science
VAMO v.s. Majority Voting

• Which one can better tolerate AV label inconsistencies?

• Experimental Setup
  – Synthetic Dataset
    • complete ground truth
  – 3k samples in historic archive
  – 15 families, 200 samples each
  – 3 AVs (assume identical notation)
  – 300 samples in 3rd-party dataset

• Simulating AV Label Inconsistencies
  – Missing Labels
  – Label “Flips”
VAMO v.s. Majority Voting

- VAMO’s reference clustering *agrees more closely with ground truth*

- External validity indices
  - Rand Index
  - Jaccard Coefficient
  - Folkes-Mallows
  - F1 Index
    - Precision-Recall
VAMO in Practice

• Real-world malware dataset: 2,026 samples
• 3rd-party clustering algorithm: Bayer et al. (NDSS 2009)
  – Distance matrix based on system events
  – Hierarchical clustering
    • $L = \text{cut height}$

• VAMO’s configuration
  – ~1M samples AV labels
  – 4 validity indices

<table>
<thead>
<tr>
<th>$l$</th>
<th>clusters</th>
<th>Rand</th>
<th>Jaccard</th>
<th>Folkes-Mallows</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>674</td>
<td>0.8767</td>
<td>0.2086</td>
<td>0.4494</td>
<td>0.7100</td>
</tr>
<tr>
<td>0.20</td>
<td>451</td>
<td>0.9172</td>
<td>0.5438</td>
<td>0.7308</td>
<td>0.7918</td>
</tr>
<tr>
<td>0.30</td>
<td>313</td>
<td>0.9205</td>
<td>0.5777</td>
<td>0.7482</td>
<td>0.7948</td>
</tr>
<tr>
<td>0.31</td>
<td>301</td>
<td>0.9792</td>
<td>0.8924</td>
<td>0.9434</td>
<td>0.8436</td>
</tr>
<tr>
<td>0.32</td>
<td>291</td>
<td>0.9790</td>
<td>0.8916</td>
<td>0.9430</td>
<td>0.8431</td>
</tr>
<tr>
<td>0.33</td>
<td>288</td>
<td>0.9759</td>
<td>0.8782</td>
<td>0.9357</td>
<td>0.8496</td>
</tr>
<tr>
<td>0.34</td>
<td>286</td>
<td>0.9759</td>
<td>0.8782</td>
<td>0.9357</td>
<td>0.8467</td>
</tr>
<tr>
<td>0.35</td>
<td>280</td>
<td>0.9758</td>
<td>0.8775</td>
<td>0.9353</td>
<td>0.8479</td>
</tr>
<tr>
<td>0.36</td>
<td>274</td>
<td>0.9757</td>
<td>0.8772</td>
<td>0.9352</td>
<td>0.8467</td>
</tr>
<tr>
<td>0.37</td>
<td>261</td>
<td>0.9721</td>
<td>0.8614</td>
<td>0.9265</td>
<td>0.8433</td>
</tr>
<tr>
<td>0.38</td>
<td>255</td>
<td>0.9721</td>
<td>0.8613</td>
<td>0.9265</td>
<td>0.8424</td>
</tr>
<tr>
<td>0.39</td>
<td>248</td>
<td>0.9722</td>
<td>0.8623</td>
<td>0.9270</td>
<td>0.8421</td>
</tr>
<tr>
<td>0.40</td>
<td>241</td>
<td>0.9721</td>
<td>0.8617</td>
<td>0.9268</td>
<td>0.8401</td>
</tr>
<tr>
<td>0.50</td>
<td>187</td>
<td>0.9585</td>
<td>0.8081</td>
<td>0.8971</td>
<td>0.7937</td>
</tr>
<tr>
<td>0.60</td>
<td>142</td>
<td>0.9260</td>
<td>0.7070</td>
<td>0.8366</td>
<td>0.7489</td>
</tr>
<tr>
<td>0.70</td>
<td>113</td>
<td>0.8527</td>
<td>0.5614</td>
<td>0.7354</td>
<td>0.7260</td>
</tr>
<tr>
<td>0.80</td>
<td>85</td>
<td>0.7789</td>
<td>0.4659</td>
<td>0.6656</td>
<td>0.7124</td>
</tr>
</tbody>
</table>
Conclusion

• VAMO: automated malware clustering validity analysis
• Discussed drawbacks of validation approaches based on majority voting
• Compared VAMO vs. majority voting in a controlled setting
• Demonstrated a practical application of VAMO over a real-world malware dataset
Limitations

• Beware of feature mismatch
  – AVs categorize malware based on reversing
  – Malware clustering systems use different features
    (e.g., behavioral)
• AV labels “evolve” in time
  – Samples detected using heuristics labeled as *generic*
  – Later, AVs may re-assign samples to a more specific family
• Heuristics-based detection more and more common
  – Will most samples be labeled as *generic* in the future?
  – Do AV customers care about reliable malware naming?
### One-to-many mapping of family names

<table>
<thead>
<tr>
<th>MD5</th>
<th>McAfee</th>
<th>Avira</th>
<th>TrendMicro</th>
</tr>
</thead>
<tbody>
<tr>
<td>ec34ca31</td>
<td>M=W32/Virut.gen</td>
<td>A=W32/Virut.AX</td>
<td>T=PE_VIRUT.D-1</td>
</tr>
<tr>
<td>c2276216</td>
<td>M=W32/Virut.gen</td>
<td>A=TR/Drop.VB.DU.1</td>
<td>T=PE_VIRUT.XO-1</td>
</tr>
<tr>
<td>089ae4f5</td>
<td>M=W32/Virut.gen</td>
<td>A=WORM/Korgo.U</td>
<td>T=PE_VIRUT.D-4</td>
</tr>
<tr>
<td>8ba552c9</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8cb0ab6c</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b0b75f70</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a306b4e7</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>337a2cf4</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>62d18c7e</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8dbca633</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ac433383</td>
<td>M=W32/Virut.n</td>
<td></td>
<td>T=PE_VIRUX.A-3</td>
</tr>
<tr>
<td>cae61d9e</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td>T=PE_VIRUT.XO-2</td>
</tr>
<tr>
<td>7cc795f1</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td>T=PE_VIRUT.D-1</td>
</tr>
<tr>
<td>8de5214b</td>
<td>M=W32/Virut.gen.a</td>
<td></td>
<td>T=PE_VIRUT.XY</td>
</tr>
<tr>
<td>4d26cb0a</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td>T=PE_VIRUT.D-1</td>
</tr>
<tr>
<td>9fb75631</td>
<td>M=W32/Virut.n</td>
<td></td>
<td>T=PE_VIRUX.A-3</td>
</tr>
<tr>
<td>229004b9</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td>T=PE_VIRUT.XO-1</td>
</tr>
<tr>
<td>28a85d8a</td>
<td></td>
<td></td>
<td>T=PE_VIRUT.XO-1</td>
</tr>
<tr>
<td>663c5f6c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de6f1e00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ff43bca</td>
<td>M=</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ea580f6d</td>
<td>M=W32/Virut.n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a844eeff</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4f8613fd</td>
<td>M=W32/Virut.gen</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Missing Labels**

- M=
- M=

**Inconsistent Labels**

- T=PE_VIRUT.XO-4
- T=PE_VIRUX.A-3
- T=PE_VIRUT.XO-1