Malware Characterization using Compiler-based Graphs

Tristan Vanderbruggen

Dept of Computer & Information Sciences
University of Delaware
Malware can:
- Send spam
- Steal private information
- Ransom data
- Be used for warfare

Current technologies inadequate
- Automation => millions of variants
- Zero day exploits

Discovering new malware
- Manual analyses
Malware Characterization

Bytes
- Shannon Entropy
- Bytes N-grams
- Strings

Codes
- Instruction N-grams
- Statistics
  - Function, Blocks
  - Calls, Branches
State-of-the-art: Bytes Analysis

State-of-the-art: Bytes Analysis

ASCII

Shannon Entropy

Shannon Entropy

CISC 850: Cyber Analytics
Our Approach: Code Analysis

Reverse engineering: Radare2
Extract code from: x86, JAR, JS, etc
Compiler Internal Representation

Instruction Flow Graph
Control Flow Graph
Call Graph
Statistics
Instructions 1-grams
Instructions 2-grams
Compiler-based Graphs
Malware Characterization

Bytes-Entropy Histogram

- Signature of the binary
- 256 x 256 positive integers
- Fast to extract (~ms)

Disassembled Code

- Multiple graphs
- No fixed size
- Long reverse engineering (~s)
Malware Classification

- Millions to classified
  - Reversing Labs
  - Virus Total

Models trained often
Support Vector Machine

- Linear Separator
- Kernel Trick
- Graph Kernels

Problem: **Poor Scaling**
Deep Learning

Old concept

- Large model expensive
- Scale with dataset size
- Convolutional neural networks
Deep Learning on Byte-Entropy Histogram
Deep Learning on Graphs

Adjacency Matrix

Node Features

M
Eigenvectors of the graph-Laplacian

[Bronstein et al., 2016] *Geometric deep learning: going beyond euclidean data*
Graph Spectral Features

**Adjacency Matrix**

**Node Features**

**Projection Matrix**

\[ W \text{ eigenvectors of graph-Laplacian (largest eigenvalues)} \]

\[ W \times M \text{ features} \]
Malware Classification

CISC 850: Cyber Analytics
Malware Stream

Reversing Labs
- Billions of malware
- Curated streams

Financial stream
- 1.2 millions
- 40+ families

Selected
- families with more than 1,000 malware
## Malware Features for Deep Learning

<table>
<thead>
<tr>
<th>Characterization</th>
<th>Format</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>bytes-entropy histogram</td>
<td>matrix</td>
<td>256 x 256</td>
</tr>
<tr>
<td>Global</td>
<td></td>
<td></td>
</tr>
<tr>
<td>statistics</td>
<td>vector</td>
<td>43</td>
</tr>
<tr>
<td>1-grams</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>2-grams</td>
<td>2809</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>statistics</td>
<td>matrix</td>
<td>20 x 23</td>
</tr>
<tr>
<td>1-grams</td>
<td>20 x 53</td>
<td></td>
</tr>
<tr>
<td>2-grams</td>
<td>20 x 2809</td>
<td></td>
</tr>
<tr>
<td>Block</td>
<td></td>
<td></td>
</tr>
<tr>
<td>statistics</td>
<td>matrix</td>
<td>20 x 10</td>
</tr>
<tr>
<td>1-grams</td>
<td>20 x 53</td>
<td></td>
</tr>
<tr>
<td>2-grams</td>
<td>20 x 2809</td>
<td></td>
</tr>
<tr>
<td>Operations</td>
<td>statistic</td>
<td>matrix</td>
</tr>
<tr>
<td>Characterization</td>
<td>Format</td>
<td>Size</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>bytes-entropy histogram</td>
<td>matrix</td>
<td>256 x 256</td>
</tr>
<tr>
<td>Global</td>
<td>statistics</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>1-grams</td>
<td>vector</td>
</tr>
<tr>
<td></td>
<td>2-grams</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>statistics</td>
<td>20 x 23</td>
</tr>
<tr>
<td></td>
<td>1-grams</td>
<td>matrix</td>
</tr>
<tr>
<td></td>
<td>2-grams</td>
<td>20 x 2809</td>
</tr>
<tr>
<td>Block</td>
<td>statistics</td>
<td>20 x 10</td>
</tr>
<tr>
<td></td>
<td>1-grams</td>
<td>matrix</td>
</tr>
<tr>
<td></td>
<td>2-grams</td>
<td>20 x 2809</td>
</tr>
<tr>
<td>Operations</td>
<td>statistic</td>
<td>matrix</td>
</tr>
<tr>
<td>Characterization</td>
<td>Format</td>
<td>Size</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>Global Statistics</td>
<td>matrix</td>
<td>256 x 256</td>
</tr>
<tr>
<td>Global 1-grams</td>
<td>vector</td>
<td>53</td>
</tr>
<tr>
<td>Global 2-grams</td>
<td>vector</td>
<td>2809</td>
</tr>
<tr>
<td>Function Statistics</td>
<td>matrix</td>
<td>20 x 23</td>
</tr>
<tr>
<td>Function 1-grams</td>
<td>matrix</td>
<td>20 x 53</td>
</tr>
<tr>
<td>Function 2-grams</td>
<td>matrix</td>
<td>20 x 2809</td>
</tr>
<tr>
<td>Block Statistics</td>
<td>matrix</td>
<td>20 x 10</td>
</tr>
<tr>
<td>Block 1-grams</td>
<td>matrix</td>
<td>20 x 53</td>
</tr>
<tr>
<td>Block 2-grams</td>
<td>matrix</td>
<td>20 x 2809</td>
</tr>
<tr>
<td>Operations Statistic</td>
<td>matrix</td>
<td>20 x 2</td>
</tr>
</tbody>
</table>
5-fold cross-validation

- **Testing pools**
- **Cross-validation pool**
  - Many DNNs
- **Selection**
  - K best DNNs
- **DNN ensemble**
  - Consensus

5 testing pools * 4 CV pools * 10 DNNs = **200 DNNs**
Too large

CISC 850: Cyber Analytics
<table>
<thead>
<tr>
<th>Models</th>
<th>1 best</th>
<th>2 best</th>
<th>5 best</th>
</tr>
</thead>
<tbody>
<tr>
<td>bytes-entropy histogram</td>
<td>39.2%</td>
<td>32.8%</td>
<td>27.3%</td>
</tr>
<tr>
<td>statistics</td>
<td>60.5%</td>
<td>53.2%</td>
<td>44.4%</td>
</tr>
<tr>
<td>Global</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-grams</td>
<td>22.9%</td>
<td>20.8%</td>
<td>19.1%</td>
</tr>
<tr>
<td>2-grams</td>
<td>20.7%</td>
<td>19.2%</td>
<td>18.5%</td>
</tr>
<tr>
<td>statistics</td>
<td>26.5%</td>
<td>23.8%</td>
<td>20.3%</td>
</tr>
<tr>
<td>Function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-grams</td>
<td>12.3%</td>
<td>12.0%</td>
<td>10.8%</td>
</tr>
<tr>
<td>2-grams</td>
<td>30.5%</td>
<td>27.2%</td>
<td>25.7%</td>
</tr>
<tr>
<td>statistics</td>
<td>15.8%</td>
<td>14.8%</td>
<td>13.6%</td>
</tr>
<tr>
<td>Block</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-grams</td>
<td>10.3%</td>
<td>9.9%</td>
<td>9.9%</td>
</tr>
<tr>
<td>2-grams</td>
<td>39.2%</td>
<td>35.1%</td>
<td>31.4%</td>
</tr>
<tr>
<td>Operations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>statistic</td>
<td>13.6%</td>
<td>11.3%</td>
<td>10.4%</td>
</tr>
</tbody>
</table>
## Best Results: DNN Ensemble

<table>
<thead>
<tr>
<th>Models</th>
<th>1 best</th>
<th>2 best</th>
<th>5 best</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEH &amp; Global level</td>
<td>16.0%</td>
<td>15.3%</td>
<td>13.8%</td>
</tr>
<tr>
<td>Compiler Graphs</td>
<td>8.4%</td>
<td>8.0%</td>
<td>8.0%</td>
</tr>
<tr>
<td>All Features</td>
<td>6.9%</td>
<td>6.5%</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

More than twice as good when adding Compiler-based Graphs
Improving Deep Learning Model

Projection Matrix

\[ W \text{ eigenvectors of graph-Laplacian (largest eigenvalues)} \]

Node Features

\[ W \times M_1 \text{ features} \]

\[ W \times M_2 \text{ features} \]
New DNN model

- Less parameters to learn
  - Train on full malware dataset
Dataset for the class

- Files
  - `financial-1000`
  - `All`?

- New features
  - ASCII strings
  - More classifications sources (VirusTotal)

- Lessons learned
  - Bytes-Entropy Histograms
    - 16 entropy bins (16x256)
    - Different sliding windows (1024 by 256 and 2048 by 128)
  - Assembly Graphs
    - Removed 2-grams
    - 1-grams at operation level
QUESTIONS?

Characterization
Deep-learning
Datasets
Results
CISC850: DevOps

Tristan Vanderbruggen
Dept of Computer & Information Sciences
University of Delaware
Tools

- Communication
  - Slack
- Version Control
  - GitHub
  - git
- Scrum
  - Waffle I/O
- Cloud
  - EC2
  - S3
Teams

- Analysis 1
  - Leonardo De La Rosa
  - Yang Yang
  - Yuhao Peng

- Analysis 2
  - Sean Kilgallon
  - Ashwag Altayyar
  - Peng Su

- Chatbot
  - Aman Sawhney
  - Abhijeet Srivastava
  - Anupam Basu

- Graphs
  - Ian Lantzy
  - Fan Li
  - Paul Soper

- Visual Analytics 1
  - Wanxin Li
  - Zicheng Liu
  - Ezeanaka Kingsley
  - Abdulrahaman Alshammari

- Visual Analytics 2
  - Yujun Zeng
  - Ruikai Zheng
  - Hancheng Zhao
  - Ruijie Xi

- Machine Learning
  - Vinit Singh
  - Abhilash Parthasarathy
  - Mingxing Gong
● Slack
  ○ [https://ud-cisc850.slack.com/signup](https://ud-cisc850.slack.com/signup)
  ○ Team lead create channel
    ■ Invite other members
    ■ Invite Dr Cavazos
  ○ Send me **private message** with:
    ■ name
    ■ project
    ■ GitHub username
Version Control & Scrum

- **Version Control: GitHub**
  - [link](https://github.com/cavazos-lab/)
  - Find repo for your group
    - `spring-2017-CISC850-XXX`

- **Scrum: Waffle IO**
  - [link](https://waffle.io/cavazos-lab/spring-2017-CISC850-XXX)
  - Create backlogs:
    - Documentations
    - Project design
- Start instance
- SSH in the instance
- git CLI
  - `sudo yum install git vim emacs`
  - `git clone https://github.com/cavazos-lab/XXX XXX`
  - `cd XXX`
  - `vim README.md`

  **Make changes: each member fill a section**
  - `git status`
  - `git add README.md`
  - `git commit -m "update README with ..."`
  - `git push origin master`
  - `???

- More on GIT:
QUESTIONS?

AWS

GIT

Scrum