Toward the Detection of Polyglot Files

Authors: Luke Koch, Sean Oesch, Mary Adkisson, Samantha Erwin, Brian Weber, Amul Chaulagain

UT-Battelle Business Sensitive

ORNL is managed by UT-Battelle LLC for the US Department of Energy
What is a Polyglot?

• A single file that is fully valid in two different formats

• Example: JPG+JAR polyglot can execute Java code or yield an image depending on the interpreting program.

• Danger: Interpreting programs simply ignore content from second format

• Large-scale analysis of leading commercial malware detectors revealed failure to detect 100% of polyglot malware in data set

---

1Robert A Bridges, Sean Oesch, Miki E Verma, Michael D Iannacone, Kelly MT Huffer, Brian Jewell, Jeff A Nichols, Brian Weber, Justin M Beaver, Jared M Smith, et al. 2020. Beyond the Hype: A Real-World Evaluation of the Impact and Cost of Machine Learning-Based Malware Detection. arXiv preprint arXiv:2012.09214 (2020).
Generating Polyglots

• Open-source *Mitra* tool combines compatible files into polyglots
  – Created by Ange Albertini
  – [https://github.com/corkami/mitra](https://github.com/corkami/mitra)

• Creates 4 types of polyglots from pairs of donor files
  – Stacks: file 2 appended to end of file 1
  – Parasites: file 2 placed inside comment markers of file 1
  – Zippers: both files placed within each other's comment markers
  – Cavities: file 2 placed inside padding space of file 1

• *Mitra* attempts all 4 types of combination on each pair of files
Hexdump of JPEG+JAR Stack-type Polyglot

JPEG magic number

File Start

00000000: ffd8 ffe0 0010 4a46 4946 0001 0101 0060 .......JFIF......`

End of Image Marker for JPEG

JAR Start

0000ff40: dd4e ad27 73c8 7c37 fd23 118e 5123 ffd9 .N.'s.l7.#..Q#...
0000ff50: 504b 0304 1400 0800 0800 6a5b 4341 0000 PK............j[CA...
0000ff60: 0000 0000 0000 0000 1400 0400 4d45 ............ME
0000ff70: 5449 2d49 4e46 2f4d 414e 4946 4553 542e TA-INF/MANIFEST.
0000ff80: 4d45 feca 0000 ad57 5d53 ea3a 147d 77c6 MF...[w]S...}w.

First Local File in JAR

Presence of Manifest file differentiates JAR from Zip

UT-Battelle Business Sensitive
Polyglot Data Set for ML/DL Training

| Data Set   | Train | Test  |
|------------|-------|-------|
| Monoglot   | 31,199| 7,799 |
| Polyglot   | 25,210| 6,303 |

Polyglot

| Type       | Train | Test  |
|------------|-------|-------|
| Stack      | 10,267| 2,574 |
| Parasite   | 10,301| 2,542 |
| Zipper     | 1,795 | 463   |
| Cavity     | 2,847 | 724   |

UT-Battelle Business Sensitive
Existing Tools for Polyglot Detection

- *File* is the ubiquitous tool for identifying a file’s format
- *TrID* is a similar tool used by Virustotal
- *Binwalk*, a file carving utility, has been used to detect polyglots
- *Polyfile*, a DARPA-funded tool, is a utility designed to analyze the structure of abnormal files, including polyglots
Existing Tool Performance

Recall, precision, F1 score for all 4 tools

| Tool   | F1 Score | Recall | Precision |
|--------|----------|--------|-----------|
| Polydet| 44.86    | 33.3   | 68.68     |
| TriID  | 18.02    | 11.19  | 46.38     |
| Binwalk| 36.16    | 24.08  | 72.56     |
| File   | 28.95    | 42.58  | 80.43     |

Breakdown of file utility’s performance

| Tool   | F1 Score | Recall | Precision |
|--------|----------|--------|-----------|
| Cavity | 97.26    | 95.58  | 98.99     |
| Zipper | 99.25    | 100    | 98.51     |
| Parasite| 98.75   |        |           |
| Stack  | 90       |        |           |
Machine Learning – Input Features

Round 1 Feature Vector

Round 2 Feature Vector

Linux File Utility

Mime Type

Hexadecimal Byte Values

Counts

0x0
0x255
121
305

Model

UT-Battelle Business Sensitive
Deep Learning – Input Features

Round 1 Feature Vector → Raw bytes → MalConv2

UT-Battelle Business Sensitive
ML/DL Performance: Round 1

Results of several ML models and one deep learning model

- MalConv2 trains on the raw bytes of the file
- The ML models, on the other hand, train on a byte occurrence vector
- This 255-character vector summarizes the occurrence of all possible hexadecimal values per file
ML Performance: Round 2

- New feature vector for ML models consists of byte occurrence vector concatenated with mime-type output from file utility
- MalConv2 did not improve from additional feature
- Tuned CatBoost performed best across all metrics
Conclusion and Next Steps

• Polyglot files present a serious challenge for existing utilities
• ML/DL show promise for detection
• Need to perform tuning and feature selection to maximize performance
• Need to demonstrate high throughput for practical use in industry
Questions?

• Contact: kochlr@ornl.gov or oeschts@ornl.gov
Supplemental: Detection of Novel Polyglots

• Tested our models against small dataset of novel polyglot malware
  – Cannot release data due to NDA
• Random Forest: 75% accuracy
• MalConv2: 83.80% accuracy
• CatBoost: 99.4% accuracy
Supplemental: Deep Learning Model Selection

• We chose Malconv2
  – Binary classifier designed for malware detection
  – 1-D convolutional neural network

• Reasons for selection
  – File type agnostic
  – Feature vector = raw bytes
  – Larger input capacity (16MB) than competing models