Machine Learning Toolkit for System Log File Reduction and Detection of Malicious Behavior

RALPH P. RITCHEY

DR. RICHARD PERRY

Outline

- Background
- Research Overview
- Research Goals
- Data Sources
- Toolkit Design
- Results

Background

- Cybersecurity historically relied heavily on network-based detection.
- Signature-based detection uses deep packet inspection to examine payload content.
- Network traffic is increasingly using more encryption as time passes.
- Encryption blinds signature-based detection.

Research Overview

- Alternative data sources are needed to offset the decrease in signature-based network detection.
- System logs generated on devices record a wide range of information from activity taking place on the device.
- These logs can be voluminous in size depending on the purpose of the device or the activity logged.
- Apply a machine learning based approach using truncated singular value decomposition chained with k-means.

Research Goals

- Reduce log files for cybersecurity purposes:
  - Remove routine entries
  - Retain indicators of malicious activity
- Why reduce the size?
  - Transport off device
  - Centralize the reduced data for cybersecurity use
- Limit resource utilization to ensure real-world applicability

Data Sources

- Synthetically generated
  - Publicly available
  - Australian Institute of Technology (AIT)
  - Labels provided
- Real-world
  - Not publicly available
  - Originated from publicly accessible servers
  - No labels

Data Sources – Synthetic Data (AIT)

HTTPD Log file format:

```
<srcip> - - [dd/mon/yyyy:hh:mm:ss] "<request>" <statuscode> <bytes> "<referrer>" "<useragent>"
```

Basic log file statistics:

Server	Log Lines	Malicious Lines	Malicious/200 Status	% Malicious	File Size
mail.cup.com	148,534	6,789	475	5%	36MB
mail.insect.com	169,340	6,973	665	4%	43MB
mail.onion.com	81,963	6,429	129	8%	22MB
mail.spiral.com	100,445	7,370	1,047	7%	24MB

Server	Unique Src IPs
mail.cup.com	4
mail.insect.com	3
mail.onion.com	3
mail.spiral.com	4

Data Sources – Real-World

HTTPD Log file format:

```
<srcip> - - [dd/mon/yyyy:hh:mm:ss] "<request>" <statuscode> <bytes>
```

Basic file statistics:

Server	Log Lines	Malicious Lines	% Malicious	File Size
Alpha	180,782	18,990	11%	18MB
Beta	72,488	15,671	22%	15MB
Gamma	68,442	18,476	27%	13MB
Delta	438,208	57,505	13%	36MB

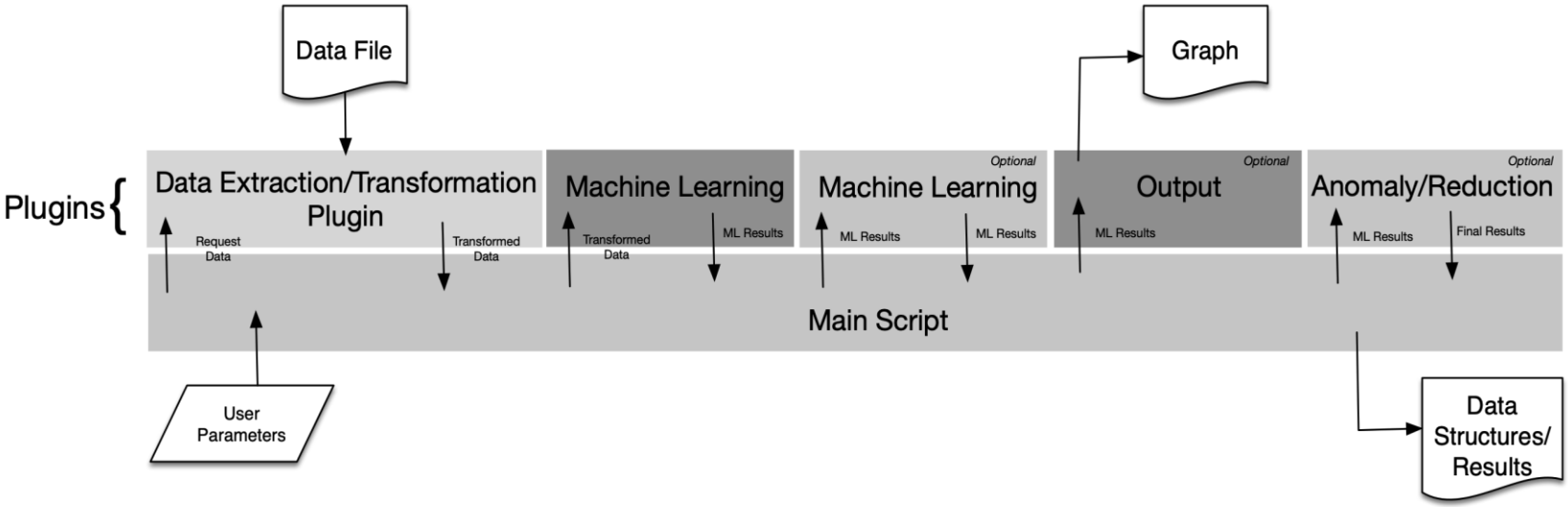
Server	Unique Src IPs
Alpha	3,817
Beta	5,725
Gamma	6,494
Delta	36,647

Data Sources – Extracted Features

- Source IP (client)
- Bytes Sent
- HTTP Status Code
- Referrer*
- Command
- Request (URL + Parameters)
- Useragent*

*Not included in real-world data set

Toolkit Design



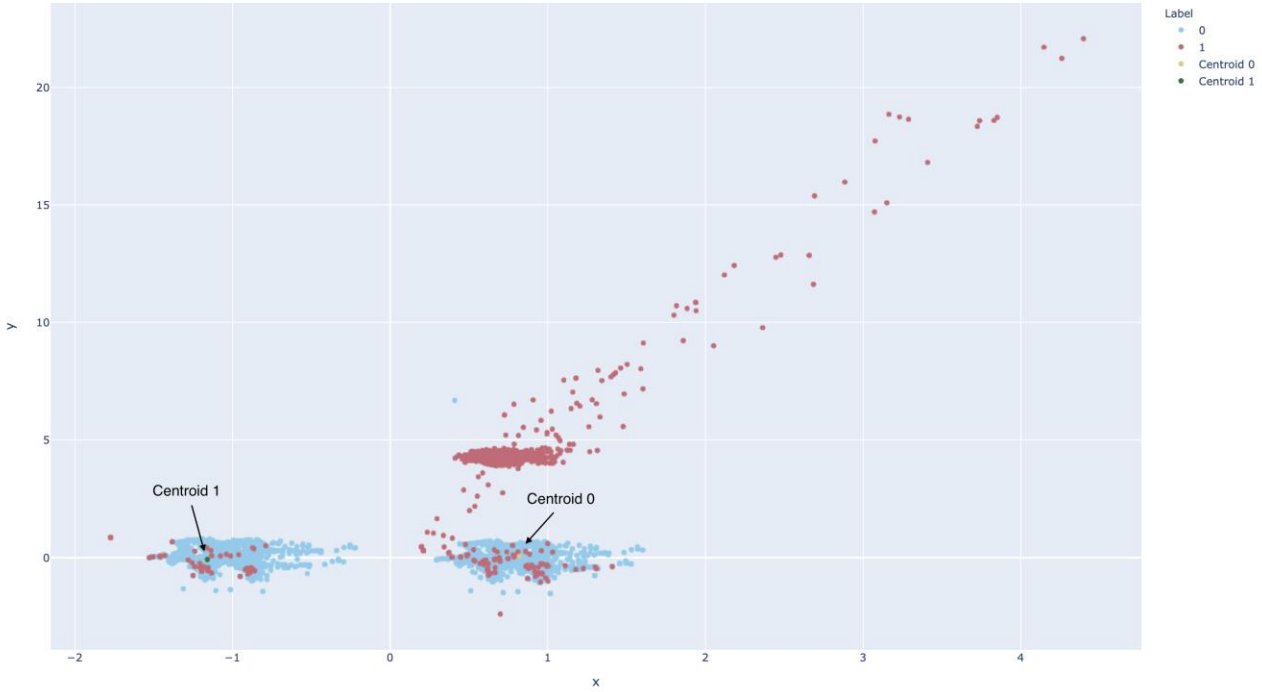
Results – Synthetic Data Experiment

URL and User-agent Splitting

Server	Detection Methodology	TP	FP	TN	FN	# 200 Status
mail.cup.com	Kmeans + Cosine Similarity	6,691	80,296	61,449	98	0
mail.cup.com	Kmeans	6,691	80,296	61,449	98	383

Server	Detection Methodology	Possible File Reduction %
mail.cup.com	Kmeans + Cosine Similarity	41.44%
mail.cup.com	Kmeans	41.44%

Results – Synthetic Data Experiment



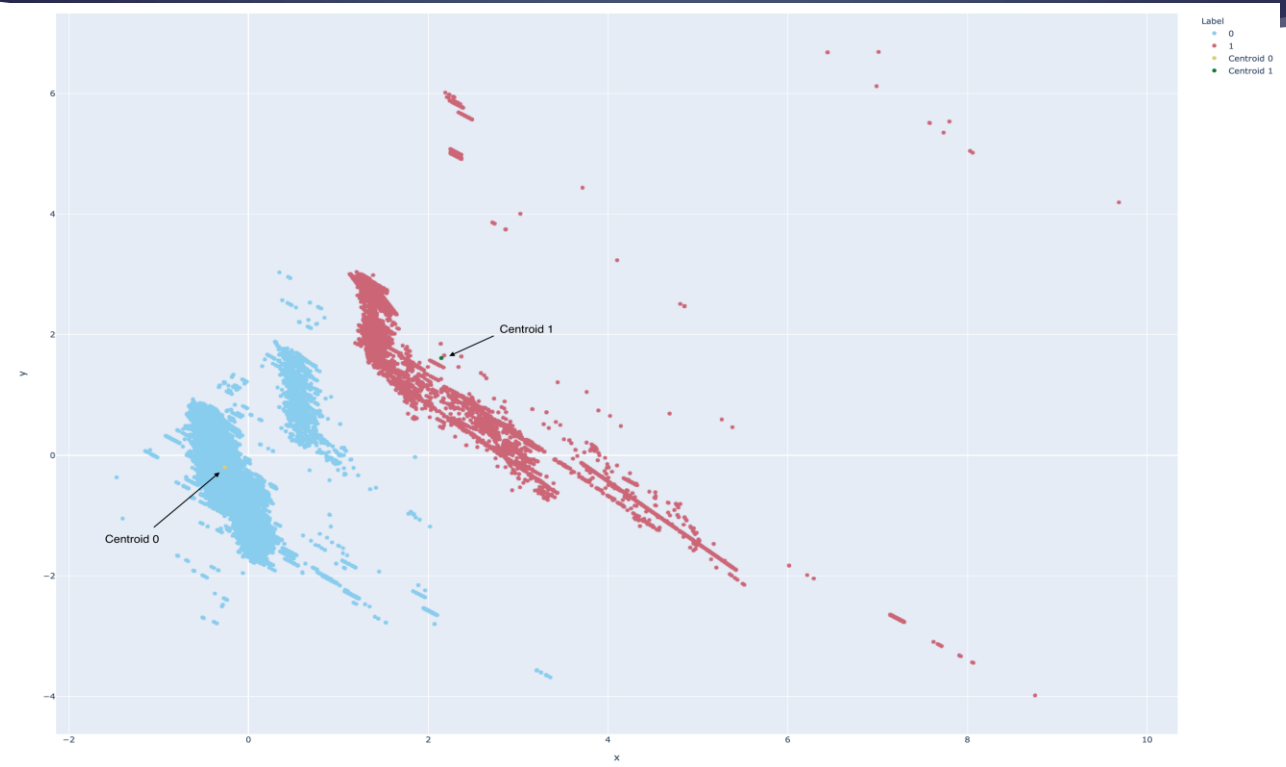
Results – Real-World Data Experiment

URL Splitting

Server	Detection Methodology	TP	FP	TN	FN
alpha	Kmeans + Cosine Similarity	18,990	8,346	153,446	0
	Kmeans	18,990	978	160,814	0

Server	Detection Methodology	Possible File Reduction %
alpha	Kmeans + Cosine Similarity	84.88%
	Kmeans	88.95%

Results – Real-World Data Experiment



Results – Resource Utilization (Time)

Execution Run	File Processing	TSVD	K-Means	Cosine Similarity	Total
1	49.504	1.359	0.796	0.718	52.377
2	47.169	1.227	0.782	0.690	49.869
3	46.418	1.238	0.785	0.684	49.125
Average Time:	47.697	1.275	0.788	0.697	50.457

8685 Lines Per Second

Execution Run	File Processing	TSVD	K-Means	Cosine Similarity	Total
1	49.125	0.742	0.989	0.656	51.511
2	49.370	0.733	0.993	0.682	51.778
3	49.164	0.803	0.996	0.730	51.693
Average Time:	49.219	0.759	0.993	0.689	51.661

8482 Lines Per Second

Results – Resource Utilization (Memory)

Execution Run	Raw Data	DataFrame	Reduced DataFrame
1	173	598	538
2	173	598	538
3	173	598	538
Average Memory:	173	598	538

Execution Run	Raw Data	DataFrame	Reduced DataFrame
1	200	380	320
2	200	380	320
3	200	380	320
Average Memory:	200	380	320

Values are in MB

Conclusions

- The approach achieved:
  - High accuracy identifying log lines triggered by malicious activity
  - Significant log file size reduction
- Limited resource utilization during application of the approach
- Quick execution even on older, modest desktop hardware
- The approach transfers to real-world application