Malware Technical Exchange Meeting
26-28 July 2022

Malware Antivirus Scan Pattern Mining via Tensor Decomposition

Prajna Bhandary, Colin Vieson, Auguste Kiendrebeogo, Iman Adetunji, Robert J. Joyce, Maksim E. Eren, and Charles Nicholas
University of Maryland Baltimore County

UMBC DREAM LAB

Objectives

- Consistent and accurate malware naming conventions are important for malware detection and classification
- Anti-virus (AV) vendors use different labels in malware reports.
- Our work is the **first** to discover hidden patterns using tensor decomposition method with unsupervised Machine Learning represented with **3-dimensional tensors**.
- Identified **similar labeling patterns** across different AV vendors.

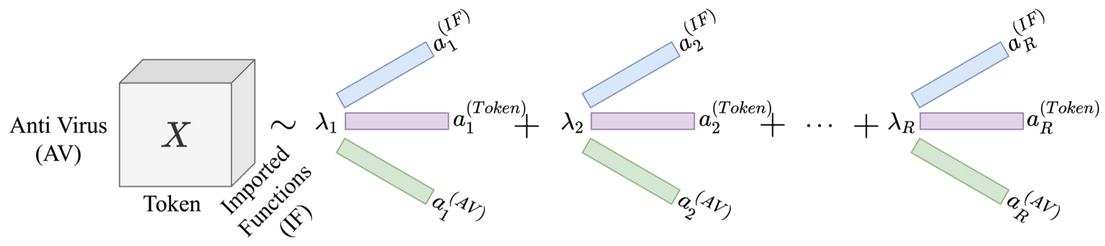
Methodology

- Multi-modal approach
- Combined **AV Scan labeling** dataset and **Motif** dataset[1] w.r.t MD5
- Data Cleaning: removing noisy tokens
- Build a 3-dimensional count tensor

$\mathbf{X} \in \mathbb{R}^{\text{AV x Tokens x Imported Function}}$

- **AV dimension** for each AV vendor, **Tokens dimension** for each individual token in alias, and **Imported Function dimension** for Windows API function imported by the sample.
- An entry $\mathbf{X}_{a,t,f}$ for number of times AV vendor '**a**' used the token '**t**' for each function '**f**' across each scanned malware in our dataset.
- We decompose **X** using pyCP-APR, Python implementation of CP-APR with GPU capability [2]

Canonical Polyadic Decomposition (CPD)



AV Scan Labels Example

1. Win32.Trojan-qqpass.Qqrob.Aljf
2. P2P-Worm.Win32.Darby.gen (v)
3. HEUR:Trojan.Win32.Generic
4. Malware-Cryptor.VB.gen.1
5. BehavesLike.Win32.Malware.tsc (mx-v)

Antivirus Vendors

NANO-Antivirus
TrendMicro-HouseCall
TrendMicro
Jiangmin
AegisLab

Tokens

wannacryptor cerber wannacry ransom maze sage spora locky packed petya

Imported Functions

advapi32.dll:cryptgenrandom
ole32.dll:coinitializeex
kernel32.dll:createdirectoryw
user32.dll:dispatchmessagea
shell32.dll:shellexecuteexw
kernel32.dll:expandenvironmentstringsw
kernel32.dll:movefileexw
mpr.dll:wnetopenenumw
shlwapi.dll:strstrw
kernel32.dll:lstrcatw
kernel32.dll:terminatethread
crypt32.dll:cryptstringtobinarya
kernel32.dll:createfilemappingw
kernel32.dll:lstrcmpiw

Results

- Most tensor components had generic tokens
- Component 7 clustered tokens " maze", " ransom", "locky", " wannacry", " spora", "petya", "sage" based on similarity
  - all of which suggests **ransomware**
- Clustered similar vendors like "TrendMicro-Housecall" and "TrendMicro", together
  - Use the same naming convention
- Imported functions: Top 30 Windows API functions used by malware -"cryptstringtobinarya", "cryptgenrandom" - indicative of ransomeware

Future Work

- Try the experiment on EMBER dataset
- Try other features on the tensor

References

1. Robert J. Joyce, Dev Amlani, Charles Nicholas and Edward Raff, "MOTIF: A Large Malware Reference Dataset with Ground Truth Family Labels. In The AAAI-22 Workshop on Artificial Intelligence for Cyber Security (AICS). Arxiv-2111.15031v1 https://github.com/boozallen/MOTIF
2. Maksim E. Eren, Juston S. Moore, Erik Skau, Elisabeth Moore, Manish Bhattarai, Gopinath Chennupati, and Boian S. Alexandrov. 2022. General-Purpose Unsupervised Cyber Anomaly Detection via Non-Negative Tensor Factorization. Digital Threats, (February 2022). https://doi.org/10.1145/3519602

Presented at the *Malware Technical Exchange Meeting* (MTEM), MIT Lincoln Laboratory, July 26-28, 2022.