“ X's economic espionage has reached an intolerable level, and I believe the US and our allies in Europe and Asia have an obligation to confront X and demand they put a stop to this piracy.”

- U.S. Mike Rogers, October, 2011

“ It is unprofessional and groundless to accuse the X’s Military of launching cyber attacks without any conclusive evidence.”

- X Defense Ministry, January, 2013

Overview

- Introduction
- Final Progress
- Dataset
- Data Processing
- Machine Learning Model
- Results of ML
- Conclusions/Lessons Learned

Introduction

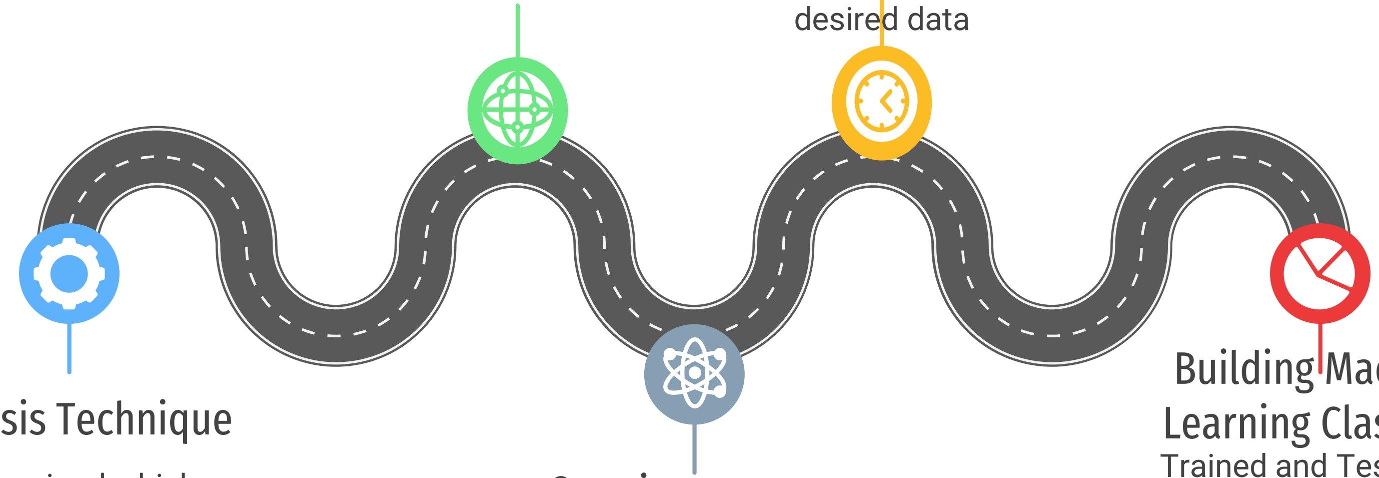
What is Cyber-Attack Attribution using Malware Artifacts?

Cyber-attack attribution using malware artifacts is the process of attempting to trace back a piece of code or malware to a perpetrator of a cyberattack

What is our focus?

Use **Machine Learning** to attribute a nation-state sponsored **APT** malwares to the source country.

Final Progress



Analysis Technique

Determined which analysis technique to use

Data Collection

Extracted data from Github and VirusTotal

Creating our own dataset

Fetched attributes from VirusTotal's reports

Data Preprocessing

Extensive data pre processing was adopted to customise into desired data

Building Machine Learning Classifier

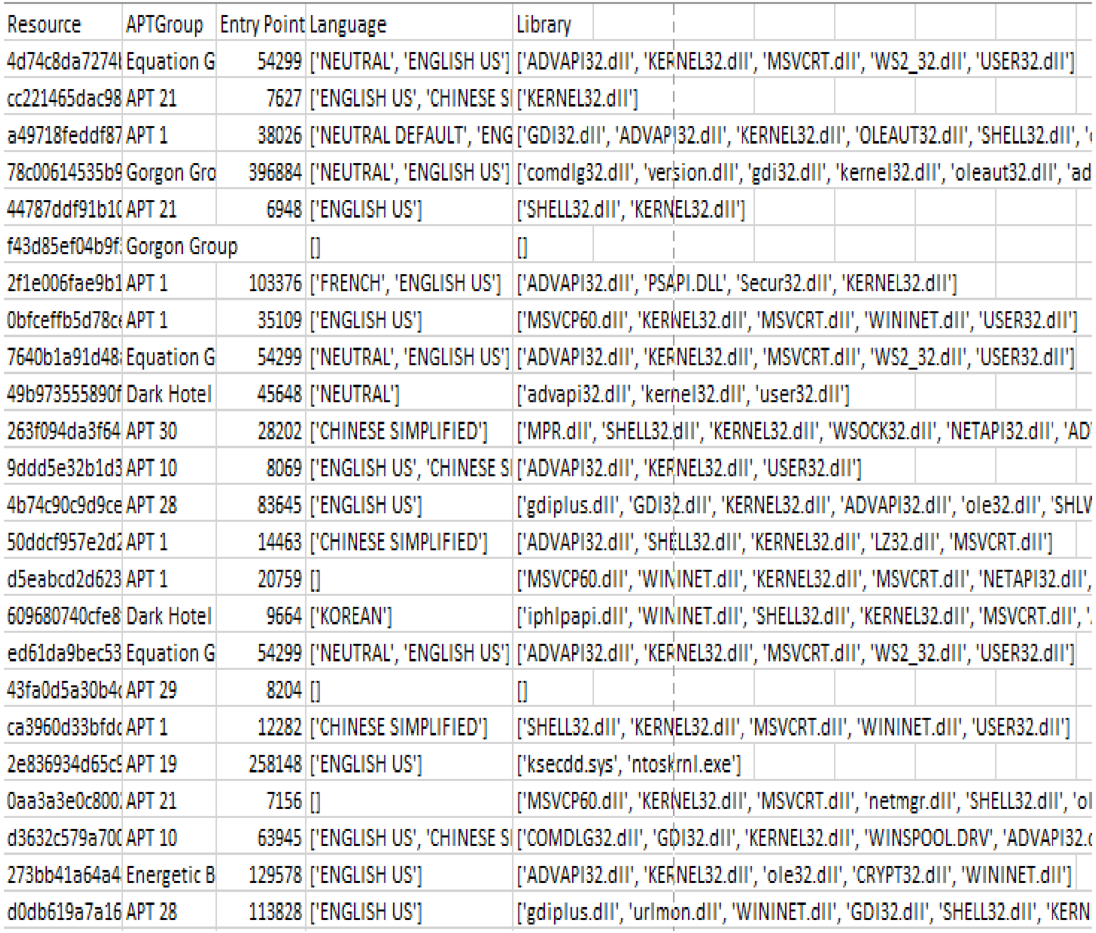
Trained and Tested on Random Forest Classifier

Dataset

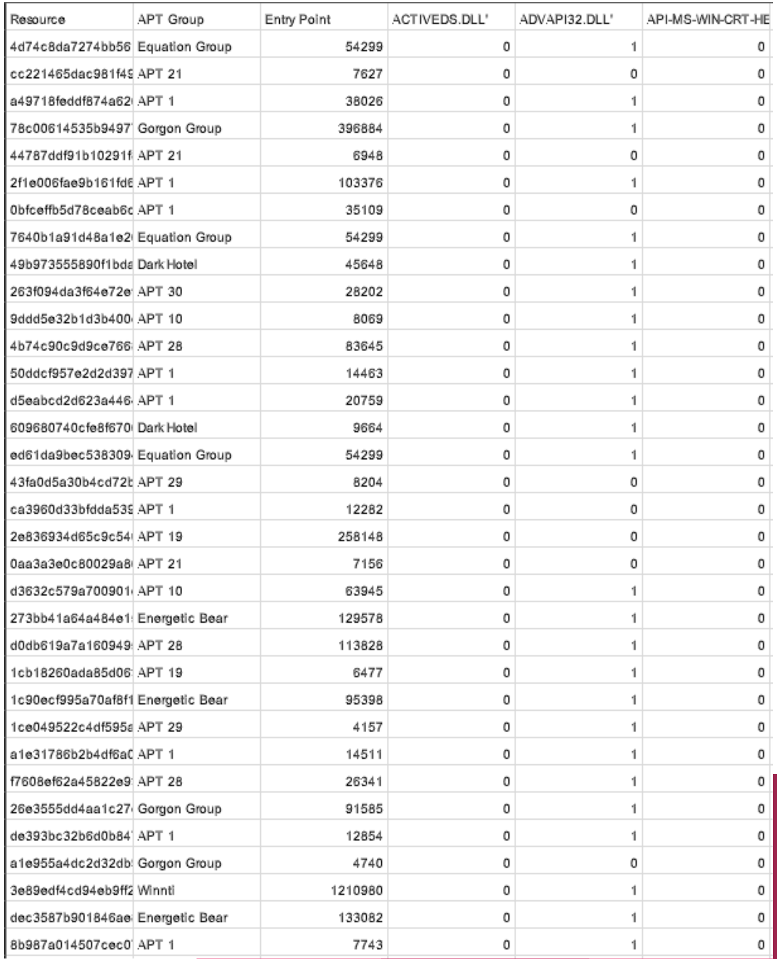
- ***GitHub Repo***, containing ~4,500 country-sponsored malware samples.
- ***PEiD***, detecting packed and unpacked malwares.
- ***Packed Malwares*** were eliminated.
- ***Unpacked Malwares***, with 3591 samples used for analysis.
- ***VirusTotal Developer API***, used in fetching reports for the malware samples.

Data Extraction and Preprocessing

- Attributes extracted from VirusTotal and converted as raw dataset:
    - *Resource*
    - *APT group*
    - *pe-entry-point*
    - *pe-resource-langs*
    - *imports*
- **3591** malware samples corresponding to **12** APT groups
- Attributes with numerical values (pe-entry-points) were kept unchanged
- imports and pe-resource-langs encoded using one hot encoder
- APT group encoded using label encoding
- **729 rows** with null values for pe-entry-point removed
- **148 features** and **2862 rows** in the preprocessed data.



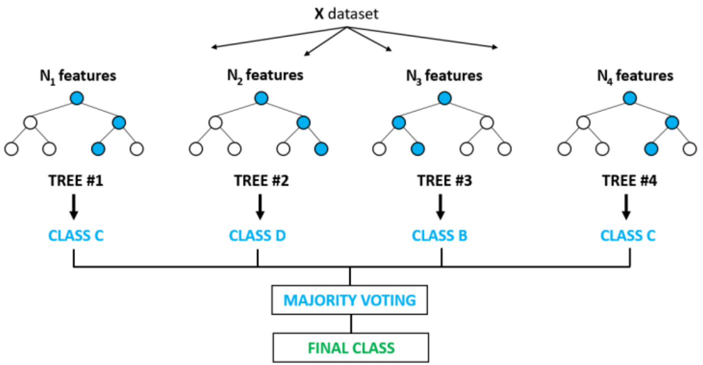
Resource	APTGroup	Entry Point	Language	Library
4d74c8da7274l	Equation G	54299	['NEUTRAL', 'ENGLISH US']	['ADVAPI32.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'WS2_32.dll', 'USER32.dll']
cc221465dac98	APT 21	7627	['ENGLISH US', 'CHINESE SI	['KERNEL32.dll']
a49718feddf87	APT 1	38026	['NEUTRAL DEFAULT', 'ENG	['GDI32.dll', 'ADVAPI32.dll', 'KERNEL32.dll', 'OLEAUT32.dll', 'SHELL32.dll', '
78c00614535b9	Gorgon Gro	396884	['NEUTRAL', 'ENGLISH US']	['comdlg32.dll', 'version.dll', 'gdi32.dll', 'kernel32.dll', 'oleaut32.dll', 'ad
44787ddf91b10	APT 21	6948	['ENGLISH US']	['SHELL32.dll', 'KERNEL32.dll']
f43d85ef04b9f:	Gorgon Group		[]	[]
2f1e006fae9b1	APT 1	103376	['FRENCH', 'ENGLISH US']	['ADVAPI32.dll', 'PSAPI.DLL', 'Secur32.dll', 'KERNEL32.dll']
0bfceffb5d78c	APT 1	35109	['ENGLISH US']	['MSVCP60.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'WININET.dll', 'USER32.dll']
7640b1a91d48	Equation G	54299	['NEUTRAL', 'ENGLISH US']	['ADVAPI32.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'WS2_32.dll', 'USER32.dll']
49b973555890f	Dark Hotel	45648	['NEUTRAL']	['advapi32.dll', 'kernel32.dll', 'user32.dll']
263f094da3f64	APT 30	28202	['CHINESE SIMPLIFIED']	['MPR.dll', 'SHELL32.dll', 'KERNEL32.dll', 'WSOCK32.dll', 'NETAPI32.dll', 'AD
9ddd5e32b1d3	APT 10	8069	['ENGLISH US', 'CHINESE SI	['ADVAPI32.dll', 'KERNEL32.dll', 'USER32.dll']
4b74c90c9d9ce	APT 28	83645	['ENGLISH US']	['gdiplus.dll', 'GDI32.dll', 'KERNEL32.dll', 'ADVAPI32.dll', 'ole32.dll', 'SHLV
50ddcf957e2d2	APT 1	14463	['CHINESE SIMPLIFIED']	['ADVAPI32.dll', 'SHELL32.dll', 'KERNEL32.dll', 'LZ32.dll', 'MSVCRT.dll']
d5eabcd2d623	APT 1	20759	[]	['MSVCP60.dll', 'WININET.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'NETAPI32.dll',
609680740cfe8	Dark Hotel	9664	['KOREAN']	['iphlpapi.dll', 'WININET.dll', 'SHELL32.dll', 'KERNEL32.dll', 'MSVCRT.dll', '
ed61da9bec53	Equation G	54299	['NEUTRAL', 'ENGLISH US']	['ADVAPI32.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'WS2_32.dll', 'USER32.dll']
43fa0d5a30b4	APT 29	8204	[]	[]
ca3960d33bfd	APT 1	12282	['CHINESE SIMPLIFIED']	['SHELL32.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'WININET.dll', 'USER32.dll']
2e836934d65c	APT 19	258148	['ENGLISH US']	['ksecdd.sys', 'ntoskrnl.exe']
0aa3a3e0c800	APT 21	7156	[]	['MSVCP60.dll', 'KERNEL32.dll', 'MSVCRT.dll', 'netmgr.dll', 'SHELL32.dll', 'ol
d3632c579a70	APT 10	63945	['ENGLISH US', 'CHINESE SI	['COMDLG32.dll', 'GDI32.dll', 'KERNEL32.dll', 'WINSPOOL.DRV', 'ADVAPI32.(
273bb41a64a4	Energetic B	129578	['ENGLISH US']	['ADVAPI32.dll', 'KERNEL32.dll', 'ole32.dll', 'CRYPT32.dll', 'WININET.dll']
d0db619a7a16	APT 28	113828	['ENGLISH US']	['gdiplus.dll', 'urlmon.dll', 'WININET.dll', 'GDI32.dll', 'SHELL32.dll', 'KERN

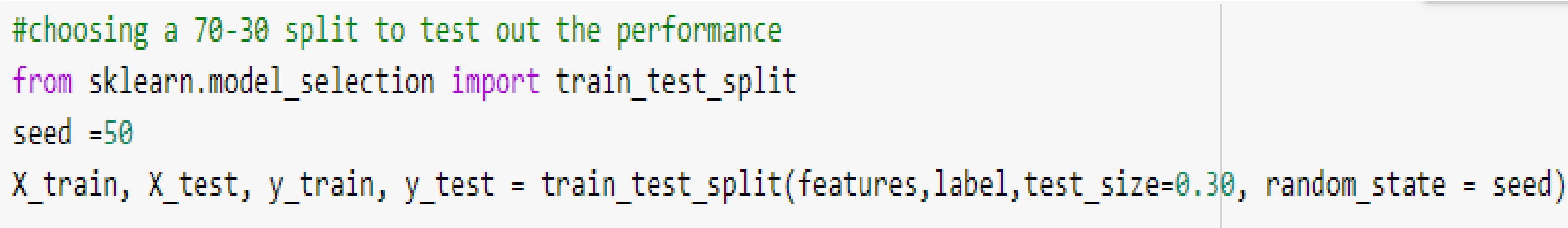


Resource	APT Group	Entry Point	ACTIVEDS.DLL'	ADVAPI32.DLL'	API-MS-WIN-CRT-HE
4d74c8da7274bb56	Equation Group	54299	0	1	0
cc221465dac981f49	APT 21	7627	0	0	0
a49718feddf874a62	APT 1	38026	0	1	0
78c00614535b9497	Gorgon Group	396884	0	1	0
44787ddf91b10291f	APT 21	6948	0	0	0
2f1e006fae9b161fd6	APT 1	103376	0	1	0
0bfceffb5d78ceab6c	APT 1	35109	0	0	0
7640b1a91d48a1e2	Equation Group	54299	0	1	0
49b973555890f1bda	Dark Hotel	45648	0	1	0
263f094da3f64e72e	APT 30	28202	0	1	0
9ddd5e32b1d3b400	APT 10	8069	0	1	0
4b74c90c9d9ce766	APT 28	83645	0	1	0
50ddcf957e2d2d397	APT 1	14463	0	1	0
d5eabcd2d623a446	APT 1	20759	0	1	0
609680740cfe8f670	Dark Hotel	9664	0	1	0
ed61da9bec538309	Equation Group	54299	0	1	0
43fa0d5a30b4cd72b	APT 29	8204	0	0	0
ca3960d33bfdda539	APT 1	12282	0	0	0
2e836934d65c9c54	APT 19	258148	0	0	0
0aa3a3e0c80029a8	APT 21	7156	0	0	0
d3632c579a700901	APT 10	63945	0	1	0
273bb41a64a484e1	Energetic Bear	129578	0	1	0
d0db619a7a160949	APT 28	113828	0	1	0
1cb18260ada85d06	APT 19	6477	0	1	0
1c90ecf995a70af8f1	Energetic Bear	95398	0	1	0
1ce049522c4df595a	APT 29	4157	0	1	0
a1e31786b2b4df6a0	APT 1	14511	0	1	0
f7608ef62a45822e9	APT 28	26341	0	1	0
26e3555dd4aa1c27	Gorgon Group	91585	0	1	0
de393bc32b6d0b84	APT 1	12854	0	1	0
a1e955a4dc2d32db	Gorgon Group	4740	0	0	0
3e89edf4cd94eb9ff2	Winnti	1210980	0	1	0
dec3587b901846ae	Energetic Bear	133082	0	1	0
8b987a014507cec0	APT 1	7743	0	1	0

Machine Learning

- Random forest classifier model
- 70% data was assigned to training and 30% to testing
- Used scikit-learn RandomForestClassifier

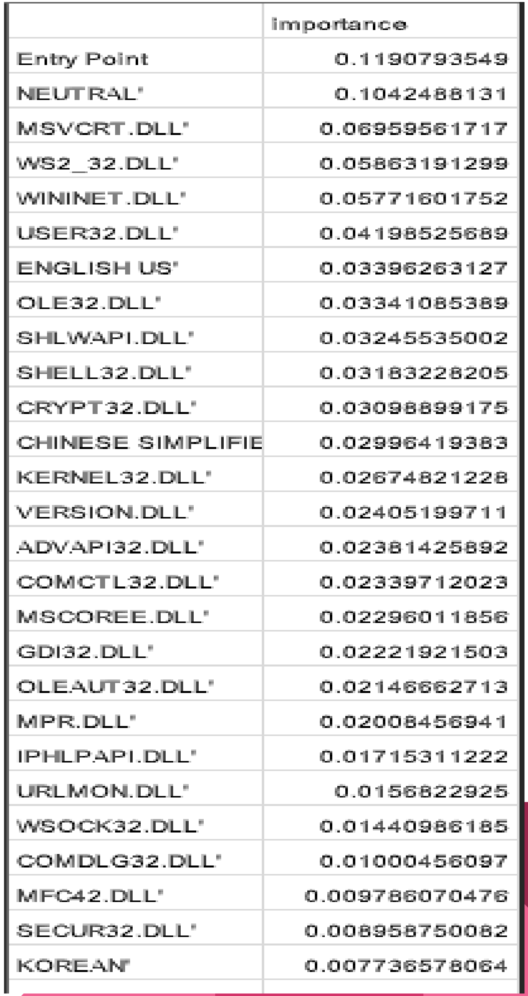




```
#choosing a 70-30 split to test out the performance
from sklearn.model_selection import train_test_split
seed =50
X_train, X_test, y_train, y_test = train_test_split(features,label,test_size=0.30, random_state = seed)
```

Results of ML

- ***58%*** initial accuracy
- ***83%*** accuracy after hyperparameter tuning
- Cross validation to validate the results of model
- ***86%*** accuracy for 20 fold random cross validation
- ***"Entry point"*** was the most valuable artifact

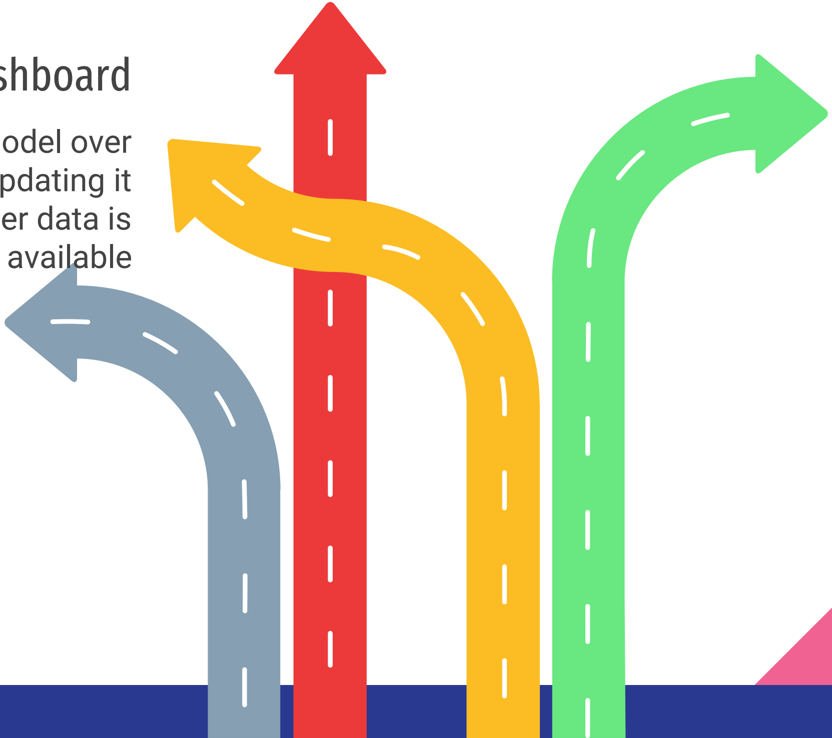


	importance
Entry Point	0.1190793549
NEUTRAL'	0.1042488131
MSVCRT.DLL'	0.06959561717
WS2_32.DLL'	0.05863191299
WININET.DLL'	0.05771601752
USER32.DLL'	0.04198525689
ENGLISH US'	0.03396263127
OLE32.DLL'	0.03341085389
SHLWAPI.DLL'	0.03245535002
SHELL32.DLL'	0.03183228205
CRYPT32.DLL'	0.03098899175
CHINESE SIMPLIFIE	0.02996419383
KERNEL32.DLL'	0.02674821228
VERSION.DLL'	0.02405199711
ADVAPI32.DLL'	0.02381425892
COMCTL32.DLL'	0.02339712023
MSCOREE.DLL'	0.02296011856
GDI32.DLL'	0.02221921503
OLEAUT32.DLL'	0.02146662713
MPR.DLL'	0.02008456941
IPHLPAPI.DLL'	0.01715311222
URLMON.DLL'	0.0156822925
WSOCK32.DLL'	0.01440986185
COMDLG32.DLL'	0.01000456097
MFC42.DLL'	0.009786070476
SECUR32.DLL'	0.008958750082
KOREAN'	0.007736578064

Conclusion

- Model was quite effective even with less features and less data.
- Accuracy may increase with more features and more training data.
- Less features due to time constraint
- Lack of availability of APT Malware datasets
- Involved rigorous data pre-processing
- Generation of Reports from VirusTotal takes 14 hours.

Future Scope



Dynamic analysis

Live Dashboard

Hosting the model over Website and updating it as newer data is available

Extract more features

entropy, no of sections, and lots more

Neural network classification

Deep Learning Techniques may help in better cyber attribution

References

[1] Boot, Coen. Applying Supervised Learning on Malware Authorship Attribution. Diss. Radboud University Nijmegen, 2019.

[2] Aylin Caliskan, Fabian Yamaguchi, Edwin Dauber, Richard Harang, Konrad Rieck, Rachel Greenstadt, Arvind Narayanan "When Coding Style Survives Compilation: De-Anonymizing Programmers from Executable Binaries." N.p., 2015. Web. https://arxiv.org/abs/1512.08546

[3] Ferhat Ozgur Catak, Ahmet Faruk Yazı,Ogerta Elezaj, Javed Ahmed "Deep Learning Based Sequential Model for Malware Analysis Using Windows Exe API Calls." PeerJ. Computer science 6 (2020): e285–. Web. https://peerj.com/articles/cs-285/

[4] Haddadpajouh, Hamed ; Azmoodeh, Amin ; Dehghantanha, Ali ; Parizi, Reza M. "MVFCC: A Multi-View Fuzzy Consensus Clustering Model for Malware Threat Attribution." IEEE access 8 (2020): 139188–139198. Web.

[5] Kaspersky, Eugene, et al. "The Power of Threat Attribution-Kaspersky Threat Attribution Engine" https://media.kaspersky.com/en/business-security/enterprise/threat-attribution-engine-whitepaper.pdf

[6] Rosenberg, Ishai ; Sicard, Guillaume ; David, Eli "DeepAPT: Nation-State APT Attribution Using End-to-End Deep Neural Networks." Artificial Neural Networks and Machine Learning – ICANN 2017 Lecture Notes in Computer Science, 2017, pp. 91–99., doi:10.1007/978-3-319-68612-7_11.

[7] VirusTotal Developer's API: https://developers.virustotal.com