Cyber Attack Attribution using Malware Artifacts

University of Houston

Team Members

Michael Tran
mhtran4@uh.edu

Alec Davila
andavil2l@uh.edu

Kevin Chen
ckchen2@uh.edu

Tu Van
tnnguyen66@uh.edu

Attribution

Definition: The action of regarding something as being caused by a person or thing.

Applications:

- Authorship attribution in disputed works
- Plagiarism detection
- Author profiling
- Stylistic inconsistency detection in collaborative work
- Malware attribution

Project Overview

- Attribute malicious PE executables to their respective malware families
- Assumptions:
  - Non-generic malware families are written by a single author or group of authors
  - Infeasible to derive “real” ground truth without claim of ownership
  - Family attribution is best proxy for ground truth
- Commonly cited features of malware for attribution analysis:
  - Static features - PE header information, printable strings, opcodes, etc.
  - Dynamic features - network traffic, file system changes, API calls, etc.
  - Hybrid analysis

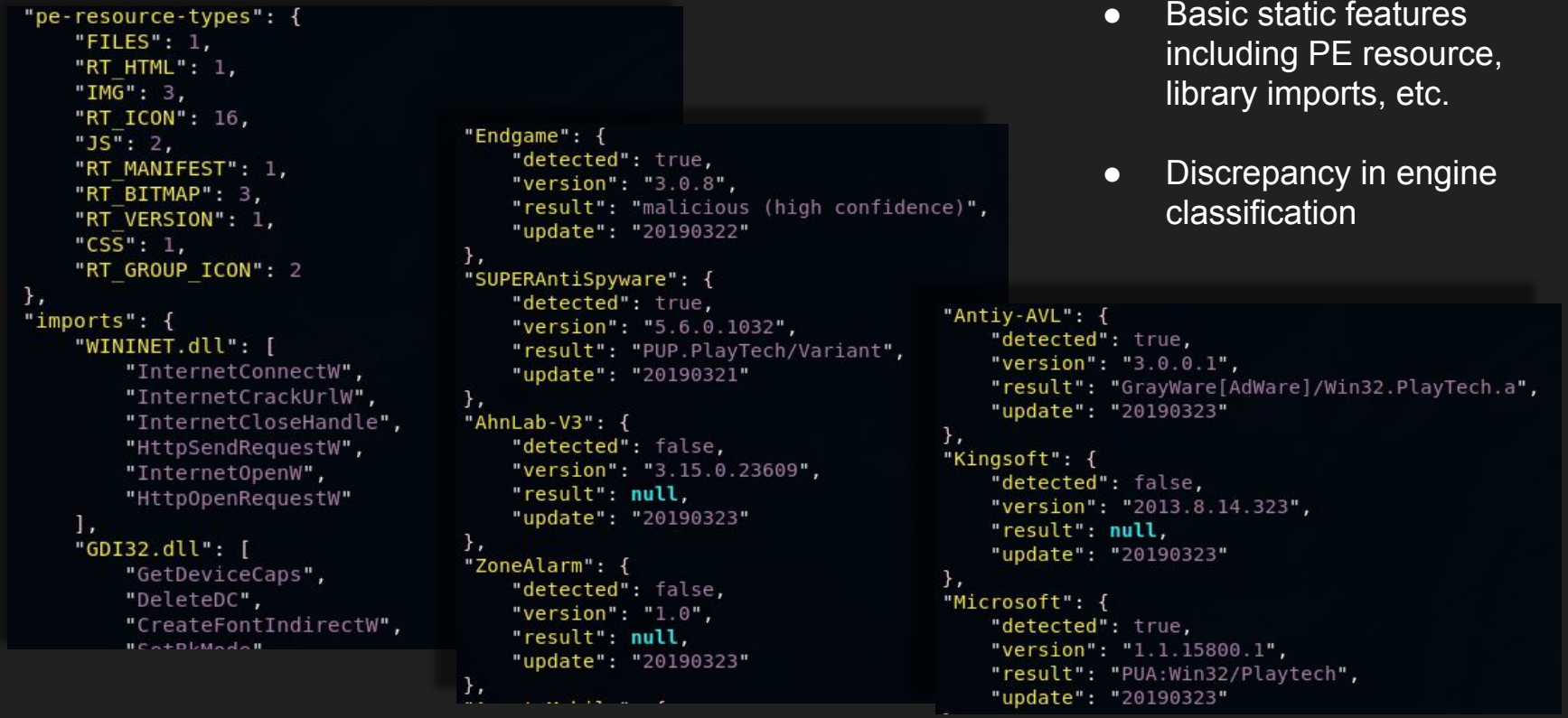
Static Features - VirusTotal

- One year's worth of samples spanning Feb 2018 - Feb 2019 collected using VirusTotal API
- 80,000 reports generated so far
- Projected to be around 100,000 total reports
- VirusTotal reports returned on Windows executables uses a combination of different tools such as ExifTool, sigcheck, TrID, etc.

Dynamic Features - Cuckoo

- Trial run on ~1,500 samples gathered from Feb 2019
- 2 minutes per sample
- Many more features than VirusTotal including API calls, network traffic information, manipulated registries, behavioral descriptions, etc.

Sample VirusTotal Report



- Basic static features including PE resource, library imports, etc.
- Discrepancy in engine classification

Sample VirusTotal Report


```
"magic": "PE32 executable for MS Windows (GUI) Intel 80386 32-bit",
"sigcheck": {
    "product": "IESettings",
    "verified": "Signed",
    "description": "IESettings",
    "file version": " 4, 2, 0, 19",
    "signing date": "10:50 PM 2/23/2019",
    "x509": [
        {
            "name": "GlobalSign CodeSigning CA - SHA256 - G3",
            "algorithm": "sha256RSA",
            "valid from": "12:00 AM 6/15/2016",
            "valid to": "12:00 AM 6/15/2024",
            "serial number": "48 1B 6A 07 26 D2 E8 3F 26 02 D4 82 5A CD",
            "cert issuer": "GlobalSign",
            "thumbprint": "090D03435EB2A8364F79B78CB173D35E8EB63558",
            "valid_usage": "Code Signing, 0.5.5.7.3.9"
        },
        {
            "name": "Cloud Installer",
            "algorithm": "sha256RSA",
            "valid from": "11:36 PM 7/26/2017",
            "valid to": "11:36 PM 8/26/2019",
            "serial number": "1A A0 48 89 F3 75 02 18 6E 69 3B BF",
            "cert issuer": "GlobalSign CodeSigning CA - SHA256 - G3",
            "thumbprint": "F04D0D19560828B694DD69F7EDB7CD498BABEDC7",
            "valid_usage": "Code Signing"
```

Sample Cuckoo Report


```
{
    "markcount": 3,
    "families": [],
    "description": "Creates a slightly modified copy of itself",
    "severity": 3,
    "marks": [
        {
            "type": "generic",
            "description": "Possibly a polymorphic version of itself",
            "file": {
                "yara": [],
                "sha1": "ab967d05608b41b83e88bd6cbfacc52f5ad9f638",
                "name": "49d2989a485d112a_mrsys.exe",
                "filepath": "C:\\Users\\sandy\\AppData\\Roaming\\mrsys.exe",
                "type": "PE32 executable (GUI) Intel 80386, for MS Windows",
                "sha256": "49d2989a485d112a84cb5f57f4023c618356dead1847cfd889a13c1a25ec129e",
                "urls": [
                    "http://don.service-master.eu/gate.php",
                    "http://www.ibsensoftware.com/",
```

```
    "process_path": "C:\\Users\\sandy\\AppData\\Local\\Temp\\3254156\\RunBoosterSetup64_3231.exe",
    "process_name": "RunBoosterSetup64_3231.exe",
    "pid": 2564,
    "summary": {
        "file_created": [
            "C:\\Program Files\\RunBooster\\RCX2974.tmp",
            "C:\\Program Files\\RunBooster\\WinDivert.dll",
            "C:\\Windows\\System32\\drivers\\WinDivert64.sys",
            "C:\\Program Files\\RunBooster\\RCX2859.tmp",
            "C:\\Program Files\\RunBooster\\RunBoosterService64.exe",
            "C:\\Program Files\\RunBooster\\msvcr110.dll",
            "C:\\Program Files\\RunBooster\\Uninstall.exe",
            "C:\\Program Files\\RunBooster\\RCX276E.tmp",
            "C:\\Program Files\\RunBooster\\RunBoosterUpdateTask64.exe"
```

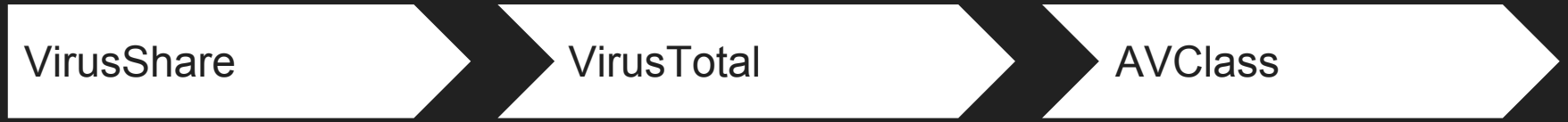

- Example of Cuckoo event descriptions and process information
- JSON formatting convenient for data extraction
- ~ 40MB of data per file with a much larger variation in size across samples

Sample Cuckoo Report


```
{
    "status": 403,
    "src": "192.168.56.101",
    "resp": {
        "path": "/opt/cuckoo/storage/analyses/78/network/f12266072605c17a027b5d5b4748649999749395",
        "sha1": "f12266072605c17a027b5d5b4748649999749395",
        "md5": "fb5d6c52840f3c4609217ea428de4c90"
    },
    "sha1": "f12266072605c17a027b5d5b4748649999749395",
    "protocol": "http",
    "dst": "13.32.168.182",
    "req": {
        "path": "/opt/cuckoo/storage/analyses/78/network/d3618aae46a0423c9b13d4386268dfb3ec7f51fc",
        "sha1": "d3618aae46a0423c9b13d4386268dfb3ec7f51fc",
        "md5": "7c0d3cd296ff27d299c291e9671193f6"
    },
    "request": "POST http://one.mountaincanvas.pw/installer.php?affId=1462&instId=803&ho_trackingid=HO3
7d5d2c9ca4102300e73c3e6a49f&v=3&kid=hqmrb21ag2gqnrmqav6 HTTP/1.1\r\nHost: one.mountaincanvas.pw\r\nConne
encoded\r\nContent-Length: 146",
    "uri": "http://one.mountaincanvas.pw/installer.php?affId=1462&instId=803&ho_trackingid=HO3556423205
4102300e73c3e6a49f&v=3&kid=hqmrb21ag2gqnrmqav6",
    "response": "HTTP/1.1 403 Forbidden\r\nServer: CloudFront\r\nDate: Sun, 24 Mar 2019 23:09:06 GMT\r\
t\r\nVia: 1.1 da94bfa4529bf05a5b62b3f058727c6c.cloudfront.net (CloudFront)\r\nX-Amz-Cf-Id: qWQvrT8e0t8j-
    "host": "one.mountaincanvas.pw",
    "dport": 80,
    "path": "/opt/cuckoo/storage/analyses/78/network/f12266072605c17a027b5d5b4748649999749395",
    "sport": 49189,
    "method": "POST",
    "md5": "fb5d6c52840f3c4609217ea428de4c90"
},
```


- Example of network traffic information
- Some interesting features worth exploring include dport, sport, dip, host, uri

Gathering and Labeling Samples



Free and **unrestricted** “repository of malware samples to provide security researchers, incident responders, forensic analysts, and the morbidly curious access to samples of live malicious code.”

~ 70 AntiVirus engines deployed to classify and label malware based on family

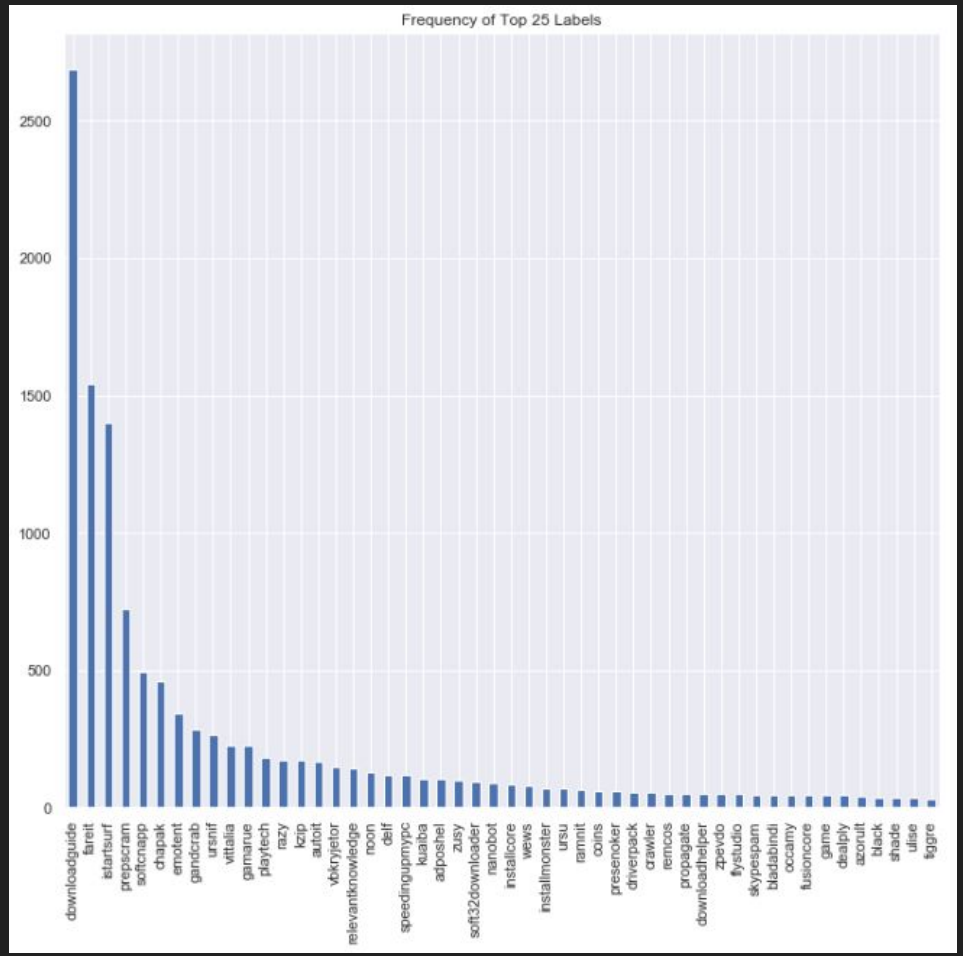
Extracts family classification from each VirusTotal engine and assigns to each sample a malware family based on plurality voting

Data Extraction

- From ~15,000 samples of our dataset, we prune out all samples where there are less than 3 samples per class label
    - Left with 12,364 Samples
- Data is split out for training/testing/validation by 80/10/10
- We ensure that each split has at least one of each class in it
- Plans to increase pruning threshold when we finish gathering samples

Top Extracted Classes

- Found dominant classes in our samples: downloadguide, fareit, istartsurf
- Many sparse classes with 3-4 samples



Proof of Concept - Results

- Our current POC application extracts features from Static VT and runs it through a standard classifier
  - Features: trid, pe-resource-list
  - Classifier: NB
- 332 Distinct Classes/Families
- 60% Accuracy

Proof of Concept - Features Extracted

classification	hash	trid	pe_resource_list
emotent	0023397c9133	Win32 Dynamic Link Library (generic)	6b1e4ec9d89888370de6c023cf6c
zamg	00234b2ed9b	Win32 Executable MS Visual C++ (generic)	af536bca0a3facd368746430337
uniblue	002656a01ee	Win32 Executable MS Visual C++ (generic)	f4724f68448073e0465a65c8b204
downloadguide	0036f96361d8	Win32 Executable MS Visual C++ (generic)	abcb0193ed76d190556c3748136
speedingupmypc	006ee1a621a	Win32 Executable Delphi generic	91d89ac0b1bbaefe1506d0225bf
coins	009984f522ec	Win32 Dynamic Link Library (generic)	539dc26a14b6277e87348594ab7
softcnapp	00d344db47fa	Win32 Executable MS Visual C++ (generic)	f59f62e7843b3ff992cf769a3c608
crypmod	0199f54551e3	Win32 Executable MS Visual C++ (generic)	b9470bdd2fdda587cfff0253565b
istartsurf	022a6339f2a8	Win32 Executable (generic)	4c3dbfd9423c428190bede1659a
zenpak	02c4184f38fc2	Win32 Executable MS Visual C++ (generic)	da23a6f22ad564ad14e6b127694
chapak	0308db8bed4	Win64 Executable (generic)	cea464faec611ad26c8a6fe32645
prepscram	0342c4e817b4	Win64 Executable (generic)	96a296d224f285c67bee93c30f8a
autoit	05463f06155d	Win32 Executable MS Visual C++ (generic)	9803deebb424e82f73c26dc00c0
vbkryjetor	05b45ebea3f2	Win32 Executable Microsoft Visual Basic 6	a3ed1d9f3fe6092528f5b385649a
vittalia	0610729d92ea	Win32 Executable MS Visual C++ (generic)	7b99f0e5e7a3db2de9f02622f1ac
ursu	062bdd8a42c3	Win32 Executable (generic)	2313003b1ba00596efd2dd9d5e5

Issues Encountered

- Gathering binaries
- Obtaining ground truth
- Hardware
  - Limited ability to perform dynamic analysis on representative sample
- Feature selection
  - Obfuscation
  - Variability in reports

Next Steps

- Focus analysis efforts on VirusTotal reports
- Extract more features from JSON data
- Feature selection
  - Determine feature frequency across all samples
  - Choose deterministic, high frequency features using measures such as mutual information, chi-squared, information gain, etc.
- Assess different classification models
  - Naive Bayes, SVM, Random Forest
- Explore options for large-scale, distributed dynamic analysis
  - Stratified random sampling

Sponsors

Matthew Elder

William La Cholter

Johns Hopkins University
Applied Physics Laboratory

References

AVClass
https://github.com/malicialab/avclass

Saed Alrabaee, Paria Shirani, Mourad Debbabi, Lingyu Wang. (2017) On the Feasibility of Malware Authorship Attribution. arXiv:1701.02711 [cs.CR].

Shalaginov Andrii, Banin Sergii, Dehghantanha Ali, Franke Katrin. Machine Learning Aided Static Malware Analysis: A Survey and Tutorial. (2018) arXiv:1808.01201v1 [cs.CR].

Sihwail Rami, Omar Khairuddin and Ariffin K.A.Z. (2018) A Survey on Malware Analysis Techniques: Static, Dynamic, Hybrid and Memory Analysis. In International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 4-2. (pp. 1662-1671)

VirusShare.com
https://virusshare.com/

VirusTotal.com
https://virustotal.com/