Hunting for New Threats in a Feed of Malicious Samples

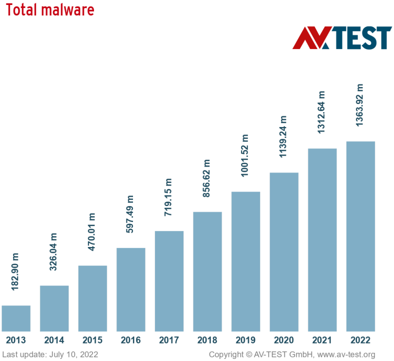
Kevin van Liebergen
Director: Juan Caballero (Instituto IMDEA Software)
Tutor: Javier Junquera (Universidad de Alcalá)

Universidad de Alcalá

July 22, 2022

Problem Space

- Increasing number of malware
  - Polymorphism = Malware variation to evade its detection
- Limited number of malware analysts
- Antivirus (AV) engines are not perfect and may not agree
- Online scanners analyze submitted samples with many AV engines regardless its filetype



Total malware

Last update: July 10, 2022

Copyright © AV-TEST GmbH, www.av-test.org

File Feeds

- File Feed = File dataset updated periodically
- Malware feed = File feed of malicious samples
- We compare four feeds on the same one-year period

Feed	Type	Free	Start	All Samples	New Samples in One Year
VT File Feed	File	✗	2004-06	>2,400,000K	209,600K
VirusShare	Malware	✓	2012-06-15	37,683K	1,400K
MalShare	Malware	✓	2017-09-14	4,721K	442K
MalwareBazaar	Malware	✓	2020-02-13	516K	178K

- VT File Feed collects 209M new samples over one-year
- We focus on the VT File Feed because of its massive volume

VT File (Report) Feed

- Report = Metadata of a submitted file
- Sample = Unique reports
- Hash = File compression function
  - Cryptographic hash = With similar inputs produces **different** outputs
  - Similarity hash = With similar inputs produces **similar** outputs (i.e., groups similar malware)
    - *tlsh*. Outperforms other similarity hashes
    - *vhash*. VirusTotal propietary hash

Feature Extraction

- 27 features, 23 directly from VT reports, 4 derived by tools such as AVCLASS. VT directly reports splitted into:
  - Sample: Should have the same values across all scans
  - Scan: May differ across scans
- Feed lacks a unified filetype
- AVCLASS: Malware labeling tool, extracts the malware family, a list of tags, and if the sample is a Potentially Unwanted Program (PUP)

Feature	Scope	Type	peexe	apk
authentihash	sample	cryptohash	✓	✗
cert_issuer	sample	string	✓	✓
cert_subject	sample	string	✓	✓
cert_thumbprint	sample	cryptohash	✓	✓
cert_valid_from	sample	timestamp	✓	✓
cert_valid_to	sample	timestamp	✓	✓
exiftool_filetype	sample	string	✓	✓
fseen_date	sample	timestamp	✓	✓
icon_hash	sample	cryptohash	✓	✓
imphash	sample	cryptohash	✓	✗
md5	sample	cryptohash	✓	✓
package_name	sample	string	✗	✓
richpe_hash	sample	cryptohash	✓	✗
sha1	sample	cryptohash	✓	✓
sha256	sample	cryptohash	✓	✓
tlsh	sample	fuzzyhash	✓	✓
trid_filetype	sample	string	✓	✓
vhash	sample	structhash	✓	✓
detection_labels	scan	string list	✓	✓
scan_date	scan	timestamp	✓	✓
sig_verification_res	scan	string	✓	✗
vt_meaningful_name	scan	string	✓	✓
vt_score	scan	integer	✓	✓
avc2_family	derived	string	✓	✓
avc2_tags	derived	string list	✓	✓
avc2_is_pup	derived	bool	✓	✓
filetype	derived	string	✓	✓

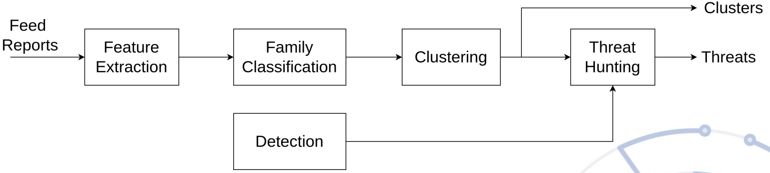
Threat Hunting

- Threat hunting = Finding interesting threats in a file feed
  - To send to a human analyst
- Threat = Malicious sample or cluster of similar malicious samples
- Interesting threat examples
  - Our goal
    - Undetected malicious samples
  - Other goals
    - Unclassified clusters, e.g., unknown family
    - New / quickly growing malicious clusters
- Threat hunting challenging due to huge volume, diversity

Our Threat Hunting Goal

- Find undetected malicious samples (with zero AV engine detections)
- Intuition:
  - Cluster all files, regardless if benign or malicious
  - Identify malicious clusters, i.e., with a majority of malicious samples that also contain samples with zero detections
    - Malicious samples = Samples with $\geq$ 4 detections
- We identify 190K potentially malicious samples in 29K clusters

Architecture Overview



State of the Art

Threat hunting works

- Graziano et al. developed an early detection approach while users submit first-stage samples to online scanners for **peexe** samples
- Huan et al. followed up the work but for **apk** samples
- Yuan et al. is a follow up of above works adding a scalability component
- Spotlight [Kaczmarczyck et al'20] threat hunting tool
  - The input is only **malicious** samples
  - Clusters ranking depends on the goal

Our work

- Our threat hunting approach may find samples regardless its **filetype**
- We include samples **regardless its AV detections**

State of the Art

VT Feed work

- Characterization of the VT **URL** Feed [Pen et al '19] measuring phishing websites
- Characterization the VT File Feed during **one day** [Ugarte-Pedrero et al'19]

Our work

- We characterize the VT **File** Feed during **one year**

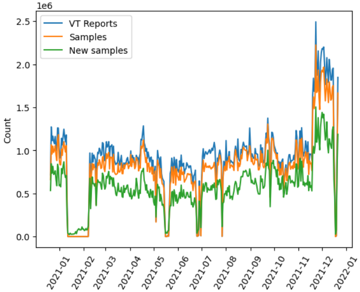
Contributions

- Threat Hunting
  - Evaluate two clustering approaches
  - Identify potentially malicious samples originally thought to be benign
- VT File Feed
  - One-year characterization of the VT File Feed

Table of Contents

1. Introduction
2. VT File Feed
3. Threat Hunting

Volume



	Mean	Median
Reports	1,681,470	1,879,952
Samples	1,493,410	1,680,520
New samples	1,028,370	1,120,242

Data	All	peexe	apk	other
Reports	328.3M	220.3M	15.9M	92.0M
Samples	235.7M	155.5M	8.2M	72.0M
New samples	209.6M	134.6M	5.6M	69.3M
Signed samples	13.3M	5.8M	7.5M	94.8K

- Collected 328M reports for 235M samples
- The 89% of the samples are new

VT File Feed Analysis: Daily statistics

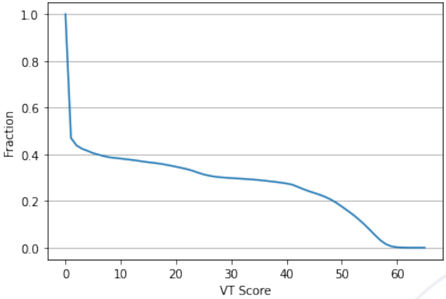


Figure: Reverse ECDF of each sample since 2021/11/19.

- VT File Feed is not a malware feed ($\geq$ 50% samples have zero detections)

Filetype distribution

Filetype	Samples	Perc
peexe	155,526,594	65.97%
javascript	21,048,404	8.93%
html	12,540,571	5.32%
pdf	11,346,815	4.81%
apk	7,992,206	3.40%
Other	24,843,745	11.56%
ALL	235,745,107	100.0%

Table: Top 5 filetypes of VT File Feed.

- The feed is a good source of samples to create malware datasets for especially *peexe* and *apk*

Family distribution

- The feed is diverse with 4.9K families with at least 100 samples. So, is a good source of samples to create malware datasets for a large variety of malware families

Filetype	Family	Class	Samples
peexe	FAM:berbew	backdoor	19,371,273
	FAM:dinwod	downloader	9,398,314
	FAM:virlock	virus	7,921,534
	FAM:pajetbin	worm	7,164,373
	FAM:sivis	virus	6,222,693
apk	FAM:smsreg	pup	616,406
	FAM:ewind	pup:adware	430,531
	FAM:hiddad	pup:adware	219,577
	FAM:fakeadblocker	pup:adware	82,715
	FAM:adlibrary:airpush	pup:adware	80,704
elf	FAM:xorddos	ddos	287,631
	FAM:mirai	backoor	163,525
	FAM:mirai:gafgyt	backoor	59,348
	FAM:tsunami	backoor	3,381
	FAM:mirai:hajime	downloader	2,499
macho	FAM:flashback	downloader	33,087
	FAM:mackontrol	backdoor	15,459
	FAM:mackeeper	pup	15,017
	FAM:evilquest	ransomware	7,070
	FAM:cimpli	pup:adware	5,444
doc	FAM:emotet	infosteal	24,643
	FAM:valyria	downloader	10,182
	FAM:thus	virus	4,917
	FAM:sagent	downloader	4,717
	FAM:donoff	downloader	2,437

Table: Top 5 families per top filetypes.

Table of Contents

1. Introduction

2. VT File Feed

3. Threat Hunting

Clustering

- The goal is to group similar samples, i.e., belongs to the same family
- Scalable approach to cluster 1.5M daily samples in less than 24h
- ↑ Precision = Same feature cluster is not split in many families

Clustering approaches:

- HAC-T [Oliver et al '20]
  - Cluster by *tlsh* feature
- Feature Value Grouping (FVG)
  - Equality comparison: Group samples with same feature, i.e., vhash, certificate thumbprint

Clustering Evaluation

- We evaluate the clustering on four popular malware ground truth datasets: Malicia, Malsign, AMD, and Drebin

Feature	Algor.	Clust.	Prec.	Recall	F1
authentihash	fvg	9,909	100%	0.5%	1.1%
avc2_family	fvg	284	97.0%	75.4%	84.8%
cert_thumb.	fvg	9,410	100%	1.8%	3.5%
icon_hash	fvg	9,766	99.9%	1.0%	1.9%
imphash	fvg	1,843	99.7%	5.7%	10.7%
richpe_hash	fvg	9,899	100%	0.6%	1.2%
vhash	fvg	900	98.8%	12.8%	22.7%
tlsh	hact-opt	3,772	99.9%	6.2%	11.8%
tlsh	hact	3,899	99.9%	3.8%	7.3%

Table: Malicia dataset (9.9K samples).

- FVG and HAC-T produces clusters with 97.0%-99.9% precision

Clustering Runtime

- FVG-vhash cluster 235M samples of the VT File Feed in 15 hours
- HAC-T does not finish to cluster one day with 2.2M samples

Threat Hunting

- Not-so-benign
  - Detect malicious clusters over FVG-vhash
  - Samples with zero detections in clusters
- We identify 190K potentially malicious samples in 29K clusters

Takeaways

VT File Feed

- Collected 328M reports for 235M samples
- The 89% of the samples are new
- VT File Feed is not a malware feed ($\geq$ 50% samples have zero detections)
- The feed is a good source of samples to create malware datasets for:
  - Especially *peexe* and *apk* filetypes
  - A large variety of malware families. The feed is diverse with 4.9K families with at least 100 samples

Threat Hunting

- FVG produce scalable clusters with 97.0%-99.9% precision
- We identify 190K potentially malicious samples in 29K clusters

Future Work

- Detect other threats
- Early detection
- Create alert rules while clustering
- Download, run, and analyze if the samples detected are really malicious
- Comparative analysis over different AV engines
- Investigate other scalable clustering approaches

Table of Contents

4 Backup slides

Limitations

- Validate results of potentially malicious samples
- Scalable clustering approach in terms of F1-score

Telemetry

- Telemetry file of an antivirus vendor with metadata of users
- Telemetry 17 times larger than VT File Feed
  - However, 8 times less malware
- New samples get detected a median of 4.4 hours early in the telemetry
- Can not make a systematic comparison with the telemetry file because its magnitude

Other Clustering Approaches

- Hierarchical Agglomerative Clustering (HAC)
- Hierarchical DBSCAN (HDBSCAN)
- Problem: $\mathcal{O}(n^2)$ complexity

Ground Truth Summary

Dataset	Plat.	Samples	Fam.	Collection
Malsign	Win	142,513	127	06/2012 - 02/2015
AMD	And	24,551	71	11/2010 - 03/2016
Malicia	Win	9,908	52	03/2012 - 02/2013
Drebin	And	5,560	179	08/2010 - 10/2012

Table: Ground truth datasets used to evaluate clustering.