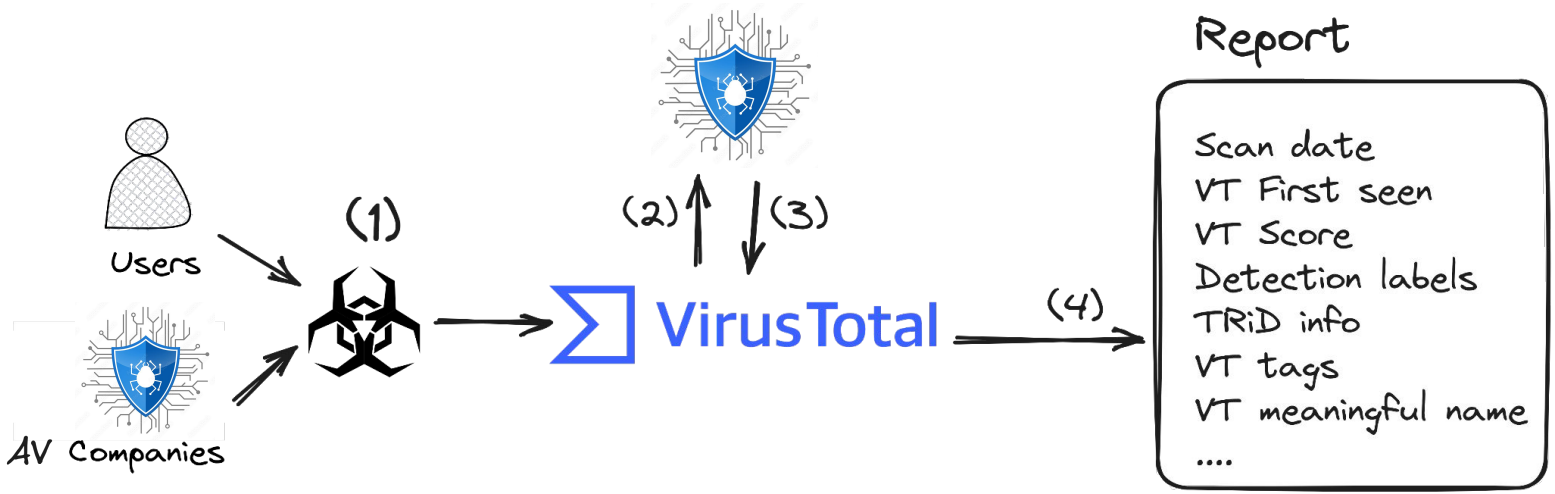
A Deep Dive into the VirusTotal File Feed

Kevin van Liebergen, Juan Caballero,
Platon Kotzias, Chris Gates

VirusTotal (VT)

- First characterization VT file feed
- Telemetry comparison
- Used for **malware detection** and **labeling**
- Source for **collecting malware** and for **identifying** new **threats**



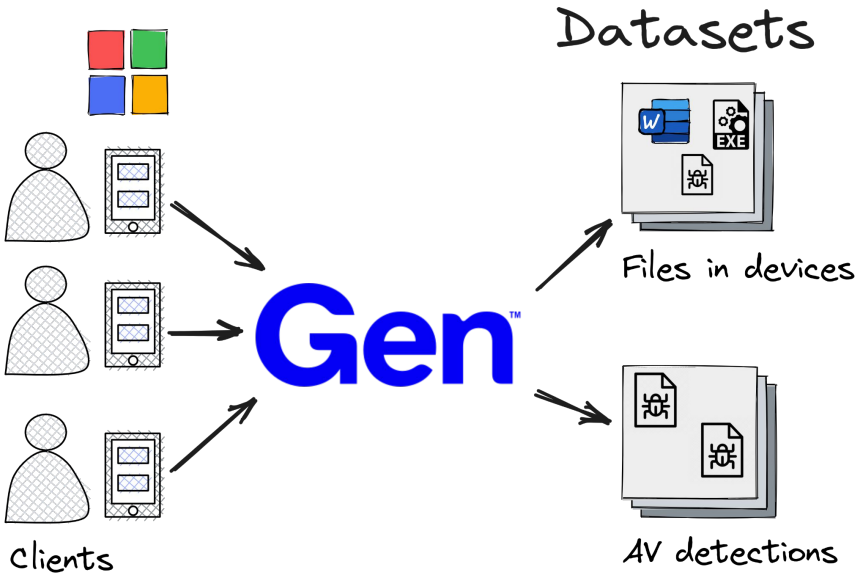
VT File Feed

- Stream of **reports** for submitted files
- We do **not collect** the binaries
- One year collection from Dec/2020 - Nov/2021

Data	All
Reports	328M
Samples	235M

- Research questions
  - How diverse is the file feed?
  - Does it allow building malware datasets for different filetypes?
  - How fresh are the samples it provides?
  - What is the distribution of malware families it sees?

Security Vendor Datasets



- Metadata of clients files
- 2 Windows datasets
- Millions of devices

- Research questions
  - How different are the views from telemetry and VT file feed?
  - Who observes samples faster?

Most Related Work

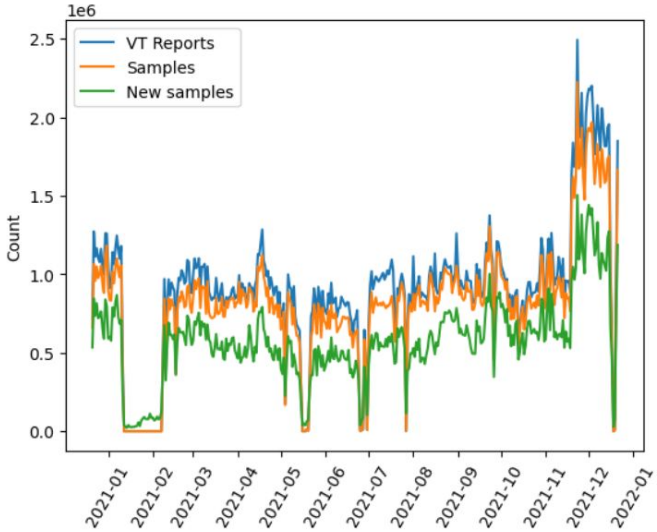
- **VT URL feed** characterization [Pen et al. ’19]
  - We characterize the **VT file feed**

- Characterization of AV malware feed for **one day** [Ugarte-Pedrero et al. ’18]
  - **One year comparison**

- Malware ecosystem measurements
  - **Windows** (e.g., [Lever et al. ’17])
  - **Android** (e.g., [Suarez-Tangil et al. ‘20])
  - **Linux** (e.g., [Cozzi et al. ‘18])
  - VT file feed contains **many filetypes**

Feed Analysis

Daily Volume and Freshness



Daily	Median
Reports	1.8M
Samples	1.6M
New samples	1.1M

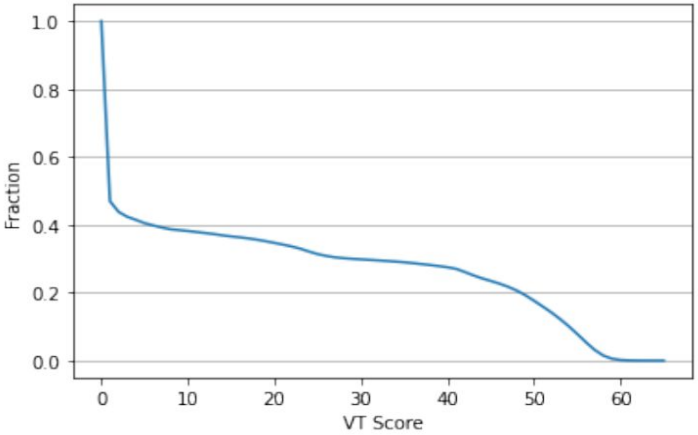
- VT file feed is **not** a **malware feed** (≥ 50% undetected samples)
- **69%** of the samples are **new**

Filetypes

- The feed is a good **source** of samples to create **datasets** for multiple filetypes

Filetype	Samples	Perc
peexe	155M	65.9%
javascript	21M	8.9%
html	12M	5.3%
pdf	11M	4.8%
apk	8M	3,4%
text	5M	2.1%
NULL	4M	1.7%
zip	4M	1.6%
Other	14M	5.9%
ALL	235M	100.0%

AV Detections



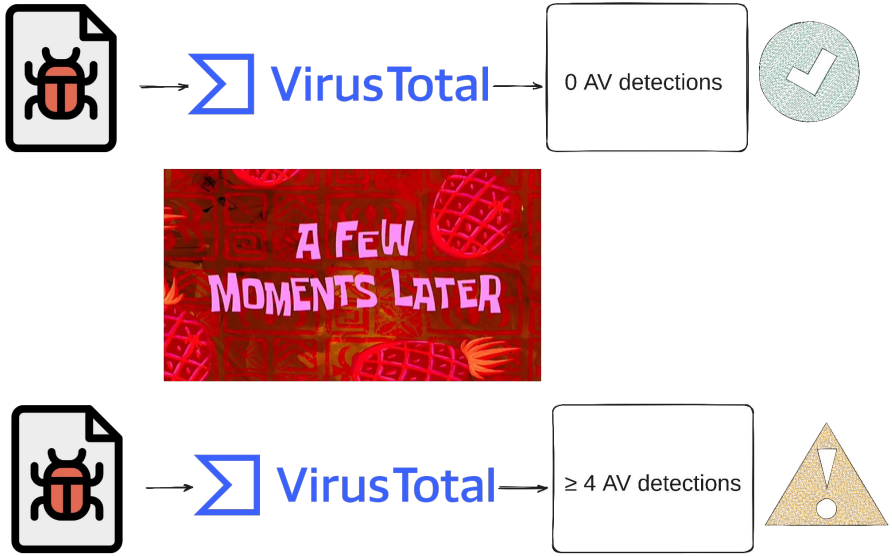
Reverse ECDF for first report

VT Score	Samples (%)
= 0	53
≥ 1	47
≥ 4	41
≥ 10	35

- We consider a file **malicious** if **≥ 4** detections
- Increasing the threshold decreases number of malware

Originally Fully UnDetected (FUD) Malware

- Zero detections on first VT observation
- Later considered malicious ≥ 4 engines
- **600K samples** originally FUD



Family Labeling

Family	Class	Samples
berbew	backdoor	19M
dinwod	downloader	9M
virlock	virus	7M
pajetbin	worm	7M
sivis	virus	6M
lamer	virus	4M
salgorea	downloader	3M
vobfus	worm	3M
drolnux	worm	2M
griptolo	worm	2M

Peexe top 10 families

Family	Class	Samples
smsreg	pup	616K
ewind	pup:adware	430K
hiddad	pup:adware	219K
fakeadblocker	pup:adware	82K
airpush	pup:adware	80K
revmob	pup:adware	78K
dowgin	pup:adware	68K
dnotua	pup	65K
kuguo	pup:adware	63K
mobidash	pup:adware	40K

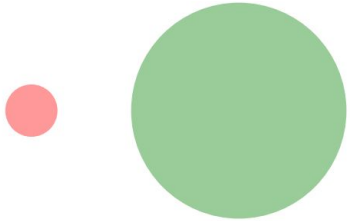
APK top 10 families

- Released updated **AvClass** taxonomy (v2.8.0)
- 62% of samples labeled on first sight
- Feed is
  - Diverse, 4.9K families with at least 100 samples
  - Good **source** to create **datasets** for multiple **malware families**

Comparison with Telemetry

Comparison: Volume, Intersection, Delay

- **Volume**: **Telemetry 17x** more
- **Malware**: VT file **feed 16x** more
- Malware samples **largely disjoint** (1.2% overlap)
- Devices see malicious samples **4.4 hours earlier**
  - 61% samples first seen in user devices
  - But, 39% first seen by VT



1:17 relation

Families

Family	Class	Samp.
berbew	backdoor	**19M**
dinwod	downloader	**9M**
virlock	virus	**7M**
pajetbin	worm	**7M**
sivis	virus	**6M**
lamer	virus	**4M**
salgorea	downloader	**3M**
vobfus	worm	**3M**
drolnux	worm	**2M**
griptolo	worm	**2M**

Feed peexe top 10 families

Family	Class	Dev.	Samp.
winactivator	pup	**2.0M**	10.8K
utorrent	pup	**1.6M**	1.3K
installcore	pup	**1.5M**	46.7K
webcompanion	pup	**1.4M**	2.5K
dotsetupio	pup	**1.1M**	0.2K
iobit	pup	**0.9M**	4.3K
opensupdater	pup	**0.7M**	14.9K
opencandy	pup	**0.5M**	9.3K
offercore	pup	**0.5M**	0.3K
driverreviver	pup	**0.5M**	0.6K

Telemetry peexe top 10 families

- Family **distribution widely differs** between VT file feed and telemetry
- **Number** of family **samples** may **not capture** real **impact** on devices

Conclusions

- VT file feed is a great source for malicious and benign files
  - 1.6M daily samples, 50% benign
  - 69% daily samples are new
  - Diverse: 4.9K families at least 100 samples

- Detected 600K originally FUD samples

- Comparison with security vendor datasets
  - Security vendor datasets much larger but less malware
  - Largely disjoint samples (1.2% overlap)
  - Malware first seen 4.4 hours earlier in devices
  - Widely different family distribution by infected devices

Questions?

- Paper

https://kevinliebergen.github.io

kevin.liebergen@imdea.org

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

A Deep Dive into the VirusTotal File Feed

Kevin van Liebergen, Juan Caballero,
Platon Kotzias, Chris Gates

Backup Slides

Feature Extraction

- 21 features, 17 from VT reports, 4 derived from those. VT reports split into:
  - Sample: Have the same values across all scans
  - Scan: May differ across scans
  - Derived

Feature	Scope	Type	peexe	apk
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	✓	✓
md5	sample	cryptohash	✓	✓
package_name	sample	string	✗	✓
sha1	sample	cryptohash	✓	✓
sha256	sample	cryptohash	✓	✓
trid_filetype	sample	string	✓	✓
detection_labels	scan	string list	✓	✓
scan_date	scan	timestamp	✓	✓
sig_verification_res	scan	string	✓	✗
vt_meaningful_name	scan	string	✓	✓
vt_score	scan	integer	✓	✓
avc_family	derived	string	✓	✓
avc_tags	derived	string list	✓	✓
avc_is_pup	derived	bool	✓	✓
filetype	derived	string	✓	✓

Code Signing

- VT supports code signatures extraction
- **5.6%** samples have **code signing** signature
  - 56% are APKs
  - 43% are Windows PE files
- 91.3% of APKs are signed
- 3.7% of peexe are signed

Discussion

- Most popular Windows families differ between the feed and the telemetry
  - Telemetry highly dominated by PUP, VT file feed by virus and worms
- Data collection issues on 39 days
- VT file feed has little overlap with the telemetry (1.2%-1.8%)
- Family labeling limited by the AV labels
- VT reports lacks a unified filetype field
  - This depends on our granularity

Future Work

- Comparative analysis over **different AV engines**
- Replace threshold-based detection approach with **machine-learning models**
- **Threat Hunting**