Learning from Big Malwares

Linhai Song, Heqing Huang, Wu Zhou, Wenfei Wu, and Yiying Zhang
Combating Malwares is Critical

• Definition of malwares
  – A variety of hostile or intrusive software

• Malwares are common and severe
  – 140 million new malwares appeared in 2015
  – 2 millions attempts to steal money via online bank

• Fighting malwares is increasing important
How to Fight Malwares?

Vulnerability Avoidance

Threat Prevention

Understanding Malwares

End User Education
Why Studying Big Malwares?

• Previous works on studying malwares
  – Provide invaluable insights
  – Only on a limited amount of malwares

• Studying big malwares
  – “Big”: in large scale and with high diversity
  – Exposes new insights
VirusTotal (VT)

• An online service to analyze suspicious files
  – Containing a huge amount of real-world files
  • 43 million suspicious files submitted last Nov.
  – Applying a host of latest anti-virus engines
  – Providing rich metadata
Existing Usage of VirusTotal

- Anti-virus vendors in industry
  - Identify FPs and FNs in their products
  - Fail to consider correlations

- Researchers in academia
  - Identifying users using VT as a test platform
  - Very few other works
Research Opportunities

- Which types of vulnerabilities are more likely to be exploited?

Threat Prevention

- How effective responses are to new security threats?
- Could we apply machine learning techniques on the VirusTotal data?
- How malwares spread?

Vulnerability Avoidance

End User Education

VirusTotal
Contributions

- An early-stage empirical study on VT data
  - Temporal analysis
    - Submission frequency and family generation rate
    - Burstiness of malwares
  - Distribution study
    - Skewness of malware families
    - Identifying hot malware families
- Identifying key research opportunities from VT
Outline

• Introduction
• Empirical Study on VirusTotal Data
• Research Opportunities
• Conclusion
Outline

• Introduction
• Empirical Study on VirusTotal Data
• Research Opportunities
• Conclusion
Empirical Study

Step 1

Downloads

Data Collection

Temporal Properties

Distribution Properties
Data Collection

• What to collect for each submission?
  – Metadata
    • File information: size, type
    • Submission information: timestamp, ID, country
    • Different hashes: ssdeep, sha256, md5
  – Analysis results
    • Roughly 50 engines used for each file

• All 43 million submissions in 2015/11
Preprocessing

- Focusing on PE files
- Merging redundant submission reports
- Leveraging Microsoft engine
  - Identifying malwares from benign files
  - Deciding malwares’ families

![Pie chart showing file types and largest number of submissions.]

![Graph showing submissions, PE submissions, and PE malwares over time.]

- PE 40%
- Audio+Video 3%
- Office 3%
- Text 4%
- Image 4%
- PDF 6%
- ZIP 7%
- Web page 7%
- Android 7%
- Other 5%
- Unknown 11%
Basic Properties

- Most malwares submitted once in 2015/11
  - Average submission number is 1.17
- Most malwares > 16 KB && < 2MB
- Most malwares are 32-bit

[Graph showing file size distribution with peaks at 16 KB and 2 MB]
Empirical Study

Step 2

Temporal Properties

Downloads

Data Collection

Distribution Properties
**Observation 1**: 100-400 new malware families appear each day.
Temporal Locality (I)

• Definition
  – How bursty malwares in the same family appear

• Cache mechanism
  – Cache design
    • Address: malware family
    • Time: submission timestamp
    • Cache hit: new submission’s family in the cache
  – Cache setting
    • Setting block size to be 1, no prefetching, LRU
**Observation 2**: The occurrence of malwares in each family has strong temporal locality.
Temporal Locality (III)

- Online malware occurrence prediction
  - Updating cache content once a day
  - Fixing cache size to be 200
Empirical Study

Data Collection

Temporal Properties

Distribution Properties

Step 3
• Submitted from 164 countries
• Top 5 countries include
  – Canada, USA, China, France, and Germany
**Observation 3:** Distributions of malwares are highly skewed in countries and malware families.
Outline

• Introduction
• Empirical Study on VirusTotal Data
• Research Opportunities
• Conclusion
• Information on VirusTotal
  – Metadata fields
  – Static features from executable
  – Dynamic behaviors

• Correlation mining
  – Which features/behaviors are more suspicious?
  – Which features/behaviors are ignored?
Evaluating Vendors’ Reports

• 50+ different engines used for each submission
  – Detailed detection results
  – How detection results change

• Questions to answer?
  – Are there influences between different vendors?
  – How to combine results from different vendors?
Studying Other File Types

- We only study PE files
- Question to answer?
  - How other malicious files distribute?
  - How other malicious files behave?
Machine Learning

- A huge set of labeled malwares on VirusTotal
- How about applying machine learning?
  - Training models using VT data
  - Using trained models to detect/classify malwares
- Questions to answer?
  - Which features on VT are useful?
  - Whether extracting features not on VT scalable?
Conclusion

• An early-stage empirical study on VT data
  – Temporal properties
  – Distribution properties

• Research Opportunities
  – Leveraging more information
  – Mining correlations
  – Applying machine learning
Thanks a lot!