Understanding Real-World Malwares and Anti-Virus Engines

Anonymous Author(s)

ABSTRACT

Your abstract should go here. You will also need to upload a plain-text abstract into the web submission form.

CCS CONCEPTS

• **Security and privacy** → Use https://dl.acm.org/ccs.cfm to generate actual concepts section for your paper;

KEYWORDS

template; formatting; pickling

1 INTRODUCTION

Combating malware is important.

Combating malware needs efforts from the whole community. VirusTotal is a website, combining vendors? new detection techniques. VirusTotal is widely used in industry.

Beyond industry, academia also widely uses VirusTotal for different purposes.

Our measurement shows that research community is using VirusTotal in a wrong way.

VirusTotal is used in a wrong way.

The question we want to ask is whether academia uses VirusTotal in a correct way. If not, how should we use VirusTotal.

Contribution:

a. We survey more than 100 academic paper and summarize how researchers use VirusTotal. We find two usage patterns.

b. We collect big data from VirusTotal and use these data to show that the current usage of VirusTotal is wrong.

c. We build a prediction model to help better use VirusTotal

2 EMPIRICAL STUDY ON HOW VIRUSTOTAL IS USED IN ACADEMIC PAPERS

2.1 How we collect paper and characteristics of collected paper

a. Year distribution

b. Conference distribution

c. Topic distribution

2.2 Findings

a. Do not wait until results become stable

b. Treat vendors equally

2.3 Discussion

What if the current usage is not correct?

3 METHODOLOGY TO COLLECT VIRUSTOTAL DATA

3.1 The large data set

How the data set is built?

What information we can get? Basically, we need to explain the data format.

Basic properties of the data set

a. How many submissions every data?

b. Submission type distribution

c. The number of submissions for the same file

d. Engines used to scan a submission

Advantage:

Across categorization, such as file types

Covering a longer time.

3.2 The small data set

How the data set is built?

What information we can get? I mean the data format.

Basic properties of the data set

a. Detection results from the first scan

b. Vendor distribution

c. How VirusTotal update engines? Scanning time vs. update vs. version

3.3 Caveats

Discuss errors during our data collection.

4 FLIPPING AND STABLE STUDY

4.1 Hazard discussion

Todo: add hazard discussion in this part before discussing flipping.

Hazard caused by VT API? NO

Hazard caused by vendor misfunction? NO

conclusion: randomly appear, but quite frequently, affect lots of files

Conclusion: remove hazard before studying flipping?

Run experiments both with and without hazard

4.2 Flipping

a. For each file, we count of flipping vendors and of flipping times

b. For each vendor, we count of flipping, of flipping files, average flipping per file

4.3 Stable

a. We need some metrics to support our definition of ?stable? is reasonable.

b. Some metrics to show how long we need to wait until results become stable

4.4 Conclusions

5 INFLUENCE MODEL

5.1 Influence graph

How we model vendors? influence

5.2 Influence measurement

The four metrics

5.3 Results

a. On the small data set
  b. On the large data set**
  Justify file category
  Justify length of the data

5.4 Discussion

6 PREDICTION MODEL

6.1 Overview

What do we want to predict?
  a. Whether a file is stable now
  b. decision tree

6.2 Experiments and implications

7 RELATED WORKS

The AISec Paper

8 CONCLUSIONS

REFERENCES