UMBC

AN HONORS UNIVERSITY IN MARYLAND

Towards a Malicious Software Search Engine

John Seymour

Information Retrieval

5/13/2014

Motivation

Why do we care about having the google equivalent for malware?

- Learning Tool
- Post-Triage Analysis
- Debugging
- Feature Selection

```
"RPCLTC1.DLL": 50,
"RPCLTC3.DLL": 51, #NOTE: This is misspelled in the literature as RPCTLC3.DLL
"RPCLTC5.DLL": 52, #NOTE: This is misspelled in the literature as RPCTLC5.DLL
"RPCLTC6.DLL": 53, #NOTE: This is misspelled in the literature as RPCTLC6.DLL
"RPCLTS3.DLL": 54, #NOTE: This is misspelled in the literature as RPCTLS3.DLL
"RPCLTS5.DLL": 55, #NOTE: This is misspelled in the literature as RPCTLS5.DLL
"RPCLTS6.DLL": 56, #NOTE: This is misspelled in the literature as RPCTLS6.DLL
"RPCNS4.DLL": 57,
```

"Attempt" 1

- MySQL
  - Need MySQL in environment, which is a pain to set up and a pain to use
  - Also is an exploitation surface itself
  - In general, not very friendly

Attempt 2

- Use Linux strings command
  - Large amount of useful output
  - Large amount of useless output, too
  - Probably useful, but requires more domain knowledge and time

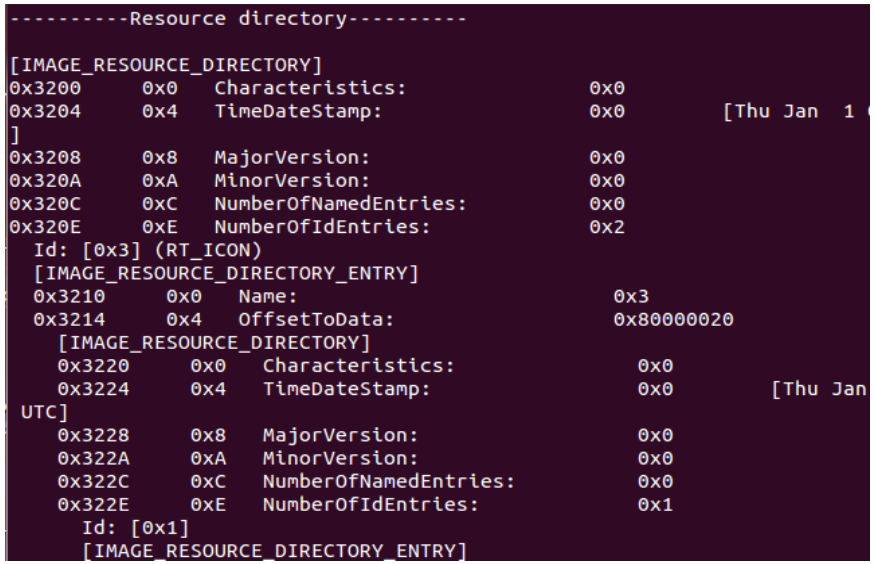
```
sandbox@sandbox-VirtualBox:~/Desktop$ strings viruses/Trojan-Downloader.Win32.Sm
all.afvx
BFD: viruses/Trojan-Downloader.Win32.Small.afvx: Warning: Ignoring section flag
IMAGE_SCN_MEM_NOT_PAGED in section .text
BFD: viruses/Trojan-Downloader.Win32.Small.afvx: Warning: Ignoring section flag
IMAGE_SCN_MEM_NOT_PAGED in section .rdata
BFD: viruses/Trojan-Downloader.Win32.Small.afvx: Warning: Ignoring section flag
IMAGE_SCN_MEM_NOT_PAGED in section .data
VPVVh
0x%X
whh2
hDdk
FindPspTerminateProcessAddr ret=0x%X
SVW`3
j"PSV
Y_^[
TSVW
```

Extracting Features

- So we know “strings” won't work too well
- How about capturing metadata of the file?
  - pefile can extract a lot of info from an executable
  - But, standard

frequency analysis

doesn't look useful...



My own research: Attempt 3

- Idea: Only index the important stuff

1)Use pefile to extract DLL's of executables to files

2)Use grep

```
sandbox@sandbox-VirtualBox:~/Desktop/viruses$ grep user32.dll *
bash: /bin/grep: Argument list too long
```

So that doesn't really work either...

My own research: Attempt 4

Well, how about applying IR concepts?

Previous Work

- Lots of work on Bag of Words model
  - Mostly classification, not user queries
- Vector space model has been used before
  - Defining the index vocabulary is important
  - Also, mostly for classification, not as much IR
- 2006: Websense/Malware Search
- In my own experience, most people just keep a malware database

My Malware Search Engine

- Size of corpus: 61.5G of executables

1)Extract DLLs using pefile python utility

  - Decreases size to about 1G

2)Build the dictionary and index

3)Query

Won't waste time re-explaining the project that you all have done already

Adapting the class project:

- Starting from executable, so have to actually extract the tokens to index
- Different format of information; no need to strip HTML tags, punctuation, stopwords, or remove terms that occur once in the corpus
- Normalizing the term weights based on the document length turns out to be incredibly useful

Some Challenges

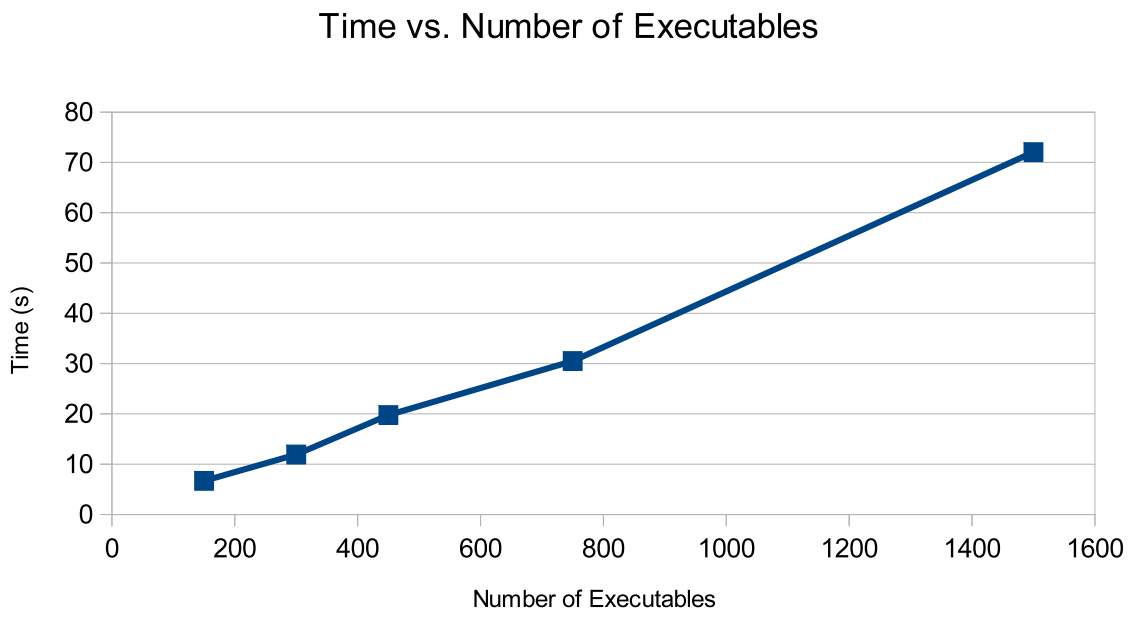
- Executables can be packed, which means pefile won't be able to extract features
- Executables can obfuscate their import tables

End Result

```
sandbox@sandbox-VirtualBox:~/Desktop/ExtractingFeatures/CMSC676$ time ./retrieve.sh user
32.dll
Backdoor.IRC.Agent.e, 0.537302039971307
Ares.exe, 0.41534148401312626
CompMgmtLauncher.exe, 0.41486208325219043
calc.exe, 0.38938264126230593
asr_ldm.exe, 0.3595071704989409
BitLockerWizard.exe, 0.33188966660175234
Backdoor.BAT.Comlabat.04, 0.33188966660175234
cidaemon.exe, 0.33188966660175234
BitLockerWizardElev.exe, 0.33188966660175234
chatclient.exe, 0.3264767090848666

real    0m1.515s
user    0m1.416s
sys     0m0.072s
```

Indexing time



Final notes

- Able to index entire 61.5G corpus in about 5.5 hours
- A lot of the time is dependent on parsing the executable
- Querying time is actually fairly static- querying entire corpus only takes a couple seconds

Questions?


Malware Search
Project Information
Total result count: 4
Worm.SCO.A
Aliases: Win32.Novarg.A@mm, Email-Worm.Win32.Mydoom.a
Win32.Mydoom.AQ@mm
Aliases: Email-Worm.Win32.Mydoom.am, Worm.Mydoom.AV
Win32.Mydoom.AT@mm
Aliases: Email-Worm.Win32.Mydoom.gen, Worm.Mydoom.AM
Win32.Mydoom.M@mm
Aliases: Worm.Mydoom.M, Email-Worm.Win32.Mydoom.m