**Malware Categorisation Using Machine Learning**

A dissertation submitted in partial fulfilment of

the requirements for the degree of

BACHELOR OF ENGINEERING in Computer Science

in

The Queen's University of Belfast

By

Jon Kyle

09/05/2017

Table of Contents

[Acknowledgements 3](#_Toc479964391)

[Abstract 3](#_Toc479964392)

[Introduction 3](#_Toc479964393)

[Literature Review and Problem Specification 4](#_Toc479964394)

[Dynamic Analysis 5](#_Toc479964395)

[Static Analysis 5](#_Toc479964396)

[PE File Structure 6](#_Toc479964397)

[Proposed Machine Learning Features 7](#_Toc479964398)

[File Entropy 7](#_Toc479964399)

[Ratio of Code to Initialised Data 7](#_Toc479964400)

[Major Image Version 7](#_Toc479964401)

[Number Of Sections 7](#_Toc479964402)

[Common DLL Imports 8](#_Toc479964403)

[Dimensionality Reduction Techniques 9](#_Toc479964404)

[Principle Component Analysis (PCA) 9](#_Toc479964405)

[Clustering Algorithms 10](#_Toc479964406)

[System Requirements Specification 11](#_Toc479964407)

[Assumptions made during development 11](#_Toc479964408)

[System constraints 11](#_Toc479964409)

[Functional Requirements 11](#_Toc479964410)

[Non-functional requirements 11](#_Toc479964411)

[User Interface requirements 12](#_Toc479964412)

[User Characteristics 12](#_Toc479964413)

[Design 13](#_Toc479964414)

[System Architecture 13](#_Toc479964415)

[User Interface Design 13](#_Toc479964416)

[Software Design 13](#_Toc479964417)

[Implementation 15](#_Toc479964418)

[Choice of languages/libraries 15](#_Toc479964419)

[Development Environment 16](#_Toc479964420)

[Key Implementation Decisions 16](#_Toc479964421)

[Implementation of important functions/Algorithms 16](#_Toc479964422)

[Implementation of individual components 16](#_Toc479964423)

[Testing 17](#_Toc479964424)

[Testing approach 17](#_Toc479964425)

[System Evaluation and Experimental Results 18](#_Toc479964426)

[Conclusion 19](#_Toc479964427)

[References 20](#_Toc479964428)

[Appendices 22](#_Toc479964429)

[Meeting Minutes 22](#_Toc479964430)

[PE32 Class Full UML 29](#_Toc479964431)

# **Acknowledgements**

# **Abstract**

Malware categorization techniques play a critical role in safeguarding existing technical infrastructure. Malware comes in many forms and uses many different attack vectors to gain access to computers, to cause undesired and potentially harmful behaviors to occur.

This project presents a heuristic-based static analysis tool which extracts features from the binary representation of WIN32 Portable Executable files. These features are then analysed using a number of dimensionality reduction techniques including Principal Component Analysis and Non-Negative Matrix Factorisation, to maximize the variance within the data. This processed data is then used as an input to the KMeans clustering algorithm to categorise the files into clusters.

# **Introduction**

Malware can often cause harm to companies and in some cases individuals through the interception of network data, encryption or deletion of files, and in extreme cases can cause catastrophic failure on a hardware level through infecting or replacing existing firmware.

The symptoms of malware differ, depending on the malware type or family; As the symptoms differ, so does the potential risk to systems infected by the malware.

Categorization of malware allows us to make an approximation of the potential risk a file poses. Though there are many existing anti-virus scanners available, the results of these scanners can disagree when it comes to the categorization of a piece of malware, and in some rare cases an anti-virus may classify a file as safe, when in fact it is not.

Malware comes in many forms and employs many different propagation techniques to deliver payloads to vulnerable machines. Common propagation techniques include Virus, and Worm which self-replicate locally or across networks. Payloads can range from relatively benign adware, to ransomware and rootkits.

Since 2012 the number of network-capable devices has increased from 8.7 Billion to 28.4 Billion and is expected to increase to 50.1 Billion by 2020 [1], this poses a problem for Security Operation Centres as this increasing number of devices gives attackers an increasing number of targets and opportunities to exploit the networked nature of these devices for malicious purposes [2]. Regardless of platform, there is likely no perfect way of detecting or categorizing malicious files as the evasion and propagation techniques used by attackers increase in complexity, but as the techniques used by attackers evolve we need a similar evolution in detection techniques in order for Security Operations Centres to quickly and accurately assess the potential risk associated with a file and put the necessary mitigations and remediation in place.

For this project, we will be focusing on heuristic based static analysis of the malware file structure to determine its relative risk. Our goal is to investigate this method of malware categorization, which will involve data-mining a set of Windows PE files and make use of existing machine learning libraries and clustering algorithms to categorise these files.

# **Literature Review and Problem Specification**

**Exploration of Existing Malware Detection Methods**

Most anti-virus software looks at syntactic signatures of a file to determine whether it is malicious, these signatures can be made up of combinations, or patterns, of instructions which have been previously flagged as malicious. At a higher level, anti-virus software can look at the hash of a file and compare to a database of hash values from known malware samples. The disadvantage of the signature based approach most anti-virus software takes is that small changes to malware source code can result in a signature which may not have been seen before and therefore not in the most up-to-date database of signature definitions. Similarly, these signature based methods provide little protection from zero-day attacks. Metamorphic and polymorphic malware also causes issues for signature based detection systems as one piece of malware can have multiple signatures.

There has been some exploration into semantic-based detection systems as a way to combat the previously mentioned methods of evading syntactic signature-based systems. These systems seek to identify *what* an executable is doing rather than *how* it is doing it through use of abstractions and templates. [3]

## **Dynamic Analysis**

Dynamic analysis involves executing the malware in a sandboxed environment, often on a virtual machine. The run-time behaviors of the malware are then observed through the tracking of system events and network traffic. This approach to malware analysis excels as it removes a large amount of the guess work which would be present using static analysis because the malware is being run in a relatively realistic environment. [4]

Dynamic Analysis has its drawbacks, there is a non-zero chance of infecting other machines on the same network. Researchers have also observed malware which behaves differently when run in a virtualized environment[5], [6], which prevents accurate categorization.

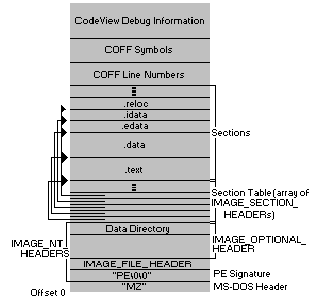
## **Static Analysis**

Static analysis is a technique used to inspect properties of a piece of malware without loading the executable into memory. It involves looking at the heuristics of the malware file for features which reveal behaviors the malware may exhibit if it were executed. The obvious advantage of this approach is that, since the malware is not being executed on the machine, then there is no chance of the malware spreading to other machines on the network and this also prevents damage to the host operating system. Static analysis also has limitations, such as its inability to fully analyze the behavior of a binary file which uses self-modifying code techniques or relies on data which is not static e.g. The current date/time. [7]. Accurate results can be computationally expensive to obtain. The processing time poses a problem for systems which rely on real-time detection of threats, although this is not the case for all static analysis systems as near real-time detection was observed with the PE Miner tool created by Shafiq et al. [8]

Static analysis has also been shown to be susceptible to reduction in effectiveness when used with obfuscated binaries [9],[10].

## **PE File Structure**

The Windows Portable Executable File Format is an extension of the Common Object File Format. There is a large amount of information available from an executable file including its target operating system, the date and time it was compiled and the DLLs it imports. A typical PE32 file consists of several sub-structures shown below.



**Fig1.** The PE File Format [11]

Each of the components of a PE File have a clearly designed specification which can be found at [11].

This specification forms the basis for our static analysis system, allowing us to create a parser which is able to extract information and features of the file quickly and conveniently. There are structural differences between 32 bit binaries and 64 bit binaries. For our solution we are targeting 32 bit binaries only.

## **Proposed Machine Learning Features**

### File Entropy

This is a number between 0 and 8 and represents the randomness of the data within the executable. According to the results proposed by Lyda and Hamrock [12], packed files and encrypted executables usually have a higher value for entropy when compared to non-packed and non-encrypted executables. Since a large proportion of malware use these techniques while benign files rarely use them we can use this entropy value as an indicator of whether a file has been packed or encrypted and therefore if it is benign or malicious. This indicator alone would obviously not be good enough to distinguish between a benign or malicious with 100% accuracy but it is a useful metric nonetheless.

### Ratio of Code to Initialised Data

This is the ratio of executable code to initialised data within the PE File. During my initial research, mining data from my dataset of malicious and benign files, I observed that benign files often have a large amount of data but a small amount of executable code in comparison. Conversely I observed that malicious files tend to have little to no data but a larger proportion of executable code. To normalise these observations and make them more meaningful I have chosen to express this ratio as a single machine learning feature as opposed to two distinct features consisting of the size of code and size of initialised data.

### Major Image Version

This number is the version of the executable. According to Raman [13] malicious files often have a value of zero and benign files often have a higher value. I observed the same pattern during my initial research and as such this feature has been the best indicator of whether a file is benign or malicious, that I have come across.

### Number Of Sections

This is the number of section headers in the section header table. As Yonts[14] observed, the number of sections in an executable can be a very good indicator of whether the file is benign or malicious. In general, benign files range from 0-10 sections while malicious files almost always have 3 or 4 sections. I was also able to observe this pattern when data mining my dataset.

### Common DLL Imports

The DLLs imported by an executable paint a very good picture of the functionality and behaviour of the executable. For example, if an executable is importing Wsock32.dll then you can surmise that the executable makes use of networking. Although there will be overlap between the imports of benign files and malicious files, we should be able to classify the malware by family using this information as behaviours differ between families. We will be using the same technique Shafiq et. Al. [8] proposed where we look for a particular import and set a flag to true or false depending on the whether it was present or not.

|  |  |
| --- | --- |
| **DLL Name** | **Function** |
| WS2\_32.dll | Networking |
| Wsock32.dll | Networking |
| Kernel32.dll | Memory Management/ IO operations |
| WININET.DLL | Networking |

## **Dimensionality Reduction Techniques**

Dimensionality reduction techniques seek to find the subset of data which best represents the whole collection of data for a particular object. That is, dimensionality reduction involves mapping a high-dimensionality dataset to a lower-dimension sub-space by filtering out data which is deemed non-essential. This has the benefit of providing an easier visualization of the whole data-set as it is not practical to represent anything with more than 3 dimensions on a graph. Low-dimensional datasets also improve performance when processing data with machine learning algorithms.[15]

### Principle Component Analysis (PCA)

Principle component Analysis is the most common form of dimensionality reduction and it achieves this though factorization of a matrix (our dataset) into two sub matrices, one of which is the Principle Component Model which best represents the data in the original matrix, this is the data which will be plotted on a graph, and the other is considered noise. [16]

### Non-Negative Matrix Factorisation (NMF)

Much like Principle Component Analysis, Non-Negative Matrix Factorisation involves factorizing a high-dimensionality matrix into two smaller matrices. Where this approach differs is that it operates under the constraint that the numerical data which makes up the data-set or matrix is non-negative. As a result of this constraint NMF can result in a different representation of the original matrix.[17]

## **Clustering Algorithms**

### KMeans

The K-Means algorithm divides provided data-points into K clusters such that the sum of squares within each cluster is minimized.[18][19] A crucial component of the KMeans algorithm is the number of clusters which the algorithm is instructed to create. Unlike other clustering algorithms such as MeanShift the KMeans algorithm does not decide the number of clusters internally based on the data.

There are several techniques which can be used to determine the optimal number of clusters, one of which is silhouette metric. This number is essentially a score for how well each data-point fits into its cluster. This metric is expressed as a score between -1 and 1. A score tending towards -1 would indicate that there may be to many or too few clusters and that the data does not belong in this cluster. A score tending towards 1 would indicate that the data has been clustered correctly. A score around 0 indicates that the data may fit more than one cluster [18].

The Elbow Method is an alternative technique for determining the optimal number of clusters for a given dataset. This involves calculating the cost function for a given value of K starting with k = 2 and increasing K by 1 until the cost function of K drops drastically [18]. The cost function can be defined as the sum of squared errors, that is, the sum of the squared linear distance of each point to its nearest centroid. The reason the we see the cost of K decrease as K increases is that, as the number of centroids increases, the distances from each point to its nearest centroid decreases.

### Mean Shift

The Mean Shift algorithm is a form of hierarchical clustering algorithm. It works by iteratively shifting each data-point towards its nearest peak. These peaks are identified using a probability density function calculated from the given dataset. To start, each point is identified as a peak. We need to provide a kernel bandwidth value which will indirectly determine how many clusters will end up with. The kernel bandwidth is a radial distance from each point. If other points fall within this radial distance then the mean value for all of these points is used as a peak. This process is then repeated iteratively until all points shift towards their nearest peak. [20] This forms clusters of data-points.

# **System Requirements Specification**

## **Assumptions made during development**

????

## **System constraints**

As previously mentioned, this project is only concerned with WIN32 binary files and not with their 64 bit equivalents. The system also does not make any attempt to decrypt or unpack malware files. This is due to the large number of implementations of packing algorithms and encryption algorithms used by attackers. It is unrealistic to have perfect conditions for each of these implementations. The extracted feature of file entropy will help combat this as this value gives a good indication of whether a file is malicious or not. Another possible approach to this problem is to attempt to unpack or decrypt the malware using existing tools before analysis with our solution.

## **Functional Requirements**

1. Provide methods for the extraction of all features outlined in the COFF specification [11]
2. Extract chosen features from a directory of WIN32 binary files
3. Provide a percentage measure for how accurate the clustering was.
4. Display the output of clustering algorithms in 2d and 3d graphs

## **Non-functional requirements**

If this project is a success it should:

1. Be able to distinguish between benign and malicious files with greater than 95% accuracy.
2. Be able to distinguish between different malware families with greater than 80% accuracy.
3. Provide an easy to use tool for static analysis of a PE32 file. i.e small setup time and cost with minimal configuration required to run.
4. Be platform independent for increased usability.
5. Handle errors gracefully and not crash
6. Be scalable to large datasets (1000+ files)
7. Data-mine a file in no more than 5 seconds

## **User Interface requirements**

The user interface for this tool should:

1. Allow the user to browse to a chosen directory
2. Display this directory for reference
3. Have a button for performing analysis on the specified directory.
4. Have tabs which switch between 2d and 3d plots
5. Allow the user to specify which clustering algorithms and dimensionality reduction algorithms to use
6. Show output from each of the selected algorithms
7. Show a percentage accuracy value for each of the selected algorithms

## **User Characteristics**

This tool is aimed at Security Operation Centres as well as InfoSec researchers in a professional environment.

# **Design**

## **System Architecture**

## **User Interface Design**

## **Software Design**

1. Browse to a directory of PE32 files. For each file create an instance of the custom PE32 class. This will expose properties of the original PE File which have been parsed from the file on disk.
2. For each PE32 object, extract our chosen machine learning features to a row in a CSV file.

|  |  |
| --- | --- |
| Feature | Source |
| File Entropy | PEFILE.GetFileEntropy() |
| Code/ Data Ratio | PEFILE.GetCodeDataRatio() |
| Major Image Version | PEFILE.GetMajorImageVersion() |
| Number Of Sections | PEFILE.NumberOfSections() |
| CommonDll Imports | PEFILE.GetImports() - map this list of imports to an array of true or false values for the presence of our chosen dlls. |

1. Perform Principle Component Analysis on the high dimensionality data in the CSV file and project to a lower dimensional subspace suitable for plotting. Store this in a Numpy array.
2. Cluster data from array using KMeans algorithm and plot data points on a 2d Matplotlib scatter graph.



Example K=2 KMeans clustering for malicious (red) and benign(green) files.

1. Use labels assigned by KMeans Algorithm in conjunction with filenames to determine the accuracy of the clustering. Note: for the above clustering malicious files were prefixed with MAL and benign files prefixed with BEN. This allows us to quickly compare the clustering of files.

**Clustering accuracy = Total number of correctly clustered files X 100**

**Total number of files analysed**

# **Implementation**

## **Choice of languages/libraries**

This project was implemented in Python 3.5. The reasoning for this is that Python is ideal for rapid prototyping; often requiring significantly less code than other object-orientated languages such as Java and C++ due to its dynamic typing. There are of course trade-offs with this choice, namely performance in comparison to C++ and other low level languages.

Due to the specialised requirements of this project, I made use of several libraries to aide in the rapid development of a working piece of software.

For the machine learning component of the project, I opted to use scikit-learn. This is an open source machine learning library for python which has implementations of common clustering algorithms such as KMeans and MeanShift. There are also several implementations for dimensionality reduction/decomposition available, namely Principle Component Analysis (PCA), Non-Negative Matrix Factorisation (NMF) and Fast Independent Component Analysis (FastICA). In my solution I make use of both PCA and NMF. To cluster the decomposed data,I made use of the KMeans and MeanShift algorithms.

For the purposes of this project I will be using both previously discussed methods for determining the most appropriate value for K.

I have chosen to use Matplotlib for data visualisation as it provides the ability to plot graphs in 2d and 3d with relatively little code. The data-mined features initially be too high-dimensional to plot on a graph, so we use PCA and NMF to project the data to a lower dimensional subspace which retains most of the variance from the initial dataset [15]. We then use the output of this as our input for the KMeans algorithm and then subsequently plot the clustered data using Matplotlib.

To implement a rudimentary GUI which displays the plotted graphs and evaluation of the clustering labels, I used PyQt which is a cross platform python binding used for rapid GUI development.

## **Development Environment**

Due to the malicious nature of the Windows PE files being analysed, my host operating system for my development environment was Mac OS X. This ensured that the malware files did not execute, as OS X uses the ELF format for executables, similar to a Linux environment.

I used Spyder as my Python IDE which is packaged with the Anaconda Python data-science platform. Anaconda comes with the Pylab package which includes Numpy and Matplotlib. My reasons for choosing the Anaconda Python platform were its ease of set-up and its ability to create multiple python environments with different versions of referenced libraries. It also comes bundled with scikit-learn and PyQt which minimises the time required to set-up a full development environment.

To create my GUI, I used the tools contained within QtCreator IDE. QtCreator is a cross platform C++ IDE which comes bundled with QtDesigner. QtDesigner is a WYSIWYG GUI design tool, which outputs a .ui file that can be used to generate a PyQt GUI class, using the pyuic5 command line tool. This solution to designing and creating a GUI minimises time spent creating complex layouts which could be better spent elsewhere in the development of the solution. It also allows for significantly more complex layouts than those possible with built in Python GUI libraries such as TKinter.

## **Key Implementation Decisions**

Use of Existing libraries – scope of project does not include re-implementing machine learning algorithms / dimensionality reduction algorithms

## **Implementation of important functions/Algorithms**

Calculation of entropy

Reading BigEndian Bytes

Reading LittleEndian Bytes

Conversion of Bytes to Strings

## **Implementation of individual components**

PEFILE/PE32 class

Peparse command line tool

GUI Windows and Tabs

# **Testing**

## **Testing approach**

# **System Evaluation and Experimental Results**

# **Conclusion**

# **References**

[1] Statista, “IoT: number of connected devices worldwide 2012-2020 | Statista,” 2015. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. [Accessed: 14-Apr-2017].

[2] D. Peraković, M. Periša, and I. Cvitić, “Analysis of the IoT impact on volume of DDoS attacks,” 2015.

[3] M. Christodorescu, S. Jha, S. A. Seshia, D. Song, and R. E. Bryant, “Semantics-aware malware detection,” in *2005 IEEE SYMPOSIUM ON SECURITY AND PRIVACY, PROCEEDINGS*, 2005, pp. 32–46.

[4] M. Egele, T. Scholte, E. Kirda, and S. Barbara, “A survey on automated dynamic malware analysis techniques and tools,” *ACM Comput. Surv.*, vol. V, no. 2, pp. 1–49, 2011.

[5] B. Lau and V. Svajcer, “Measuring virtual machine detection in malware using DSD tracer,” *J Comput Virol*, vol. 6, pp. 181–195, 2010.

[6] T. Petsas, G. Voyatzis, E. Athanasopoulos, M. Polychronakis, and S. Ioannidis, “Rage Against the Virtual Machine: Hindering Dynamic Analysis of Android Malware,” in *Proceedings of the Seventh European Workshop on System Security*, 2014, p. 5:1--5:6.

[7] S. Gadhiya, K. Bhavsar, and P. D. Student, “Techniques for Malware Analysis,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 4, pp. 2277–128, 2013.

[8] M. Zubair Shafiq, S. Momina Tabish, F. Mirza, and M. Farooq, “‘PE-Miner: Mining Structural Information to Detect Malicious Executables in Realtime’ in Recent Advances in Intrusion Detection,” Springer Science + Business Media, 2009, pp. 121–141.

[9] M. Christodorescu and S. Jha, “Static analysis of executables to detect malicious patterns,” *SSYM’03 Proc. 12th Conf. USENIX Secur. Symp.*, vol. 12, pp. 12–12, 2003.

[10] A. Moser, C. Kruegel, and E. Kirda, “Limits of Static Analysis for Malware Detections,” *Acsac*, pp. 421–430, 2007.

[11] “Microsoft Portable Executable and Common Object File Format Specification,” 2015. [Online]. Available: http://download.microsoft.com/download/9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184a/pecoff\_v83.docx.

[12] R. Lyda and J. Hamrock, “Using Entropy Analysis to Find Encrypted and Packed Malware,” *IEEE Secur. Priv.*, vol. 5, pp. 40–45, 2007.

[13] K. Raman, “Selecting Features to Classify Malware,” *InfoSec Southwest 2012*, pp. 1–5, 2012.

[14] J. Yonts, “Attributes of Malicious Files,” 2012. [Online]. Available: https://uk.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979. [Accessed: 08-Dec-2016].

[15] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[16] S. Wold, K. H. Esbensen, and P. Geladi, “Principal Component Analysis,” *Chemom. Intell. Lab. Syst.*, vol. 7439, no. August, pp. 37–52, 1987.

[17] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inf. Process. Syst.*, no. 1, pp. 556–562, 2001.

[18] T. M. Kodinariya and P. R. Makwana, “Review on determining number of Cluster in K-Means Clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 2321–7782, 2013.

[19] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” *Source J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

[20] Y. Cheng, “Mean Shift, Mode Seeking, and Clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, 1995.

# **Appendices**

## Meeting Minutes





## PE32 Class Full UML

