

An approach to fast malware classification with machine learning technique

Pham Van Hung

Faculty of Environment and Information Studies

Keio University

5322 Endo Fujisawa Kanagawa 252-0882 JAPAN

*Submitted in partial fulfillment of the requirements
for the degree of Bachelor*

Advisors:

Professor Hideyuki Tokuda

Professor Jun Murai

Associate Professor Hiroyuki Kusumoto

Professor Osamu Nakamura

Associate Professor Kazunori Takashio

Assistant Professor Rodney D. Van Meter III

Associate Professor Keisuke Uehara

Associate Professor Jin Mitsugi

Lecturer Jin Nakazawa

Professor Keiji Takeda

Copyright©2011 Pham Van Hung

Abstract of Bachelor Thesis

An approach to fast malware classification based on malware's meta-data using machine learning technique

With the rapid increase of malware, it is important for malware analysis to classify unknown malware files into malware families. By doing so, the behavior and characteristics of malware will be identified accurately. In this paper, an approach was introduced to perform fast malware classification based on meta-data of malware's file. A machine learning technique, called decision tree algorithm, is used to classify malware rapidly and correctly. Experimental results of the malware samples show that the system successfully determined some semantic malware similarities, especially showed their inner similarities in behavior and static malware characteristic.

Keywords:

Malware, static analysis, decision tree.

Pham Van Hung

Faculty of Environment and Information Studies

Keio University

Contents

List of Figures	5
List of Tables	6
1 Introduction	1
1.1 Background	1
1.2 Malware analysis problem	1
1.3 Approach	2
1.4 Thesis outline	2
2 Background	4
2.1 Growth of malware attack	4
2.2 Malware avoidance technique	5
2.3 Malware analysis technique	6
2.3.1 Dynamic malware analysis	6
2.3.2 Static malware analysis	7
2.4 Malware categories	9
2.4.1 Use virus total to detect the name of categories.	10
2.4.2 Using virus total to getting vendor name	10
2.5 Problems of malware name	10
2.6 Malware families are used in this paper	10
3 Related research	13
3.1 Flow graph	13
3.2 Optimizing decision tree in malware classification system by using Genetic Algorithm	14
3.3 Conclusion	15

4	Classification based on malware’s meta-data using decision tree approach	16
4.1	PE file format[25]	16
4.1.1	The PE File Format	16
4.1.2	The PE Header	19
4.2	Decision tree[27]	23
4.3	Classification based on malware’s meta-data using decision tree approach	25
5	Implementation	26
5.1	Environment	26
5.2	Overview	26
5.3	Classification based on machine learning technique	27
5.3.1	Meta-data	27
5.3.2	Create training data	28
5.3.3	Classification	28
6	Evaluation	31
6.1	Accuracy evaluation	31
6.2	Classification efficiency	32
7	Conclusion	34
7.1	Conclusion	34
7.2	Future work	34
	References	ii

List of Figures

2.1	Number of malicious program	5
2.2	SysAnalyser tool for dynamic analysis.	7
2.3	PEid identify the packer used by the malware author.	8
2.4	Ollydbg tool for static analysis.	8
2.5	IDA pro for static analysis.	9
2.6	Malware name is detected by antivirus engines.	12
4.1	The general layout of a PE file.	17
4.2	A section table for a typical PE file.	18
4.3	PE file format.	19
4.4	Layout a file in PE header format.	20
4.5	The data table.	24
4.6	Decision tree sample.	24
5.1	The system architecture.	27
5.2	Clustering method.	28
5.3	Worm autorun decision tree.	28
5.4	Malware classification system.	30
5.5	Worm autorun decision tree.	30
6.1	The best decision tree order.	32
6.2	Experimental result	32

List of Tables

2.1	Malware	11
-----	-------------------	----

Chapter 1

Introduction

1.1 Background

Malware is a general word for all types of malicious software. Malware includes Virus, Trojan house, Back door, Worm, and other malicious software which are characterized by malicious code. Because of the widespread use of the Internet, computer users face many dangerous propagations of malware. The modern malware's purpose is commonly to gain illegal profit. For example, a larger number of computers are infected by keylogger and 24.3 billion USD is leveraged by e-payment system losing [3]. In addition, According to 2010 Annual Security Report, on May 2010, tens of millions computer world wide were infected by email worm such as I LOVE YOU, LOVE LETTER, LOVEBUG[2]. As a result, preventing, detecting, and removing malware are very important for network security.

1.2 Malware analysis problem

This session will be a detail introduction of malware analysis of security professionals. Malware analysis is the process of analysing the purpose and functionally of a malware. Malware analysis is for understanding common areas that all viruses in a family share uniquely, and can thus create a set of signatures in order to detect malwares. In addition, the knowledge about the purpose and functionally of a malware is important for removal.

Commonly, when new malware was detected, dynamic and static malware analysis have been applied. Dynamic malware analysis technique executes malware in the Virtual Machine and uses ProcMon, RegShot, and other tools. These tools are used to identify the general behavioral analysis techniques such as network traffic analysis; file system; and other Window features: service, process,

the registry.

However, these dynamic techniques are susceptible to a variety of anti-monitoring defenses, as well as *time bombs* or *logic bombs* and can be slow and tedious to identify and disable code analysis techniques to unpack the code for examination [37]. Furthermore, it takes large amount of time to prepare environment to analyze malware such as virtual machine environment, but some malware can not be executed in the kind of environment.

With the static malware analysis technique, researchers perform reverse engineering using IDA Pro and Ollydbg tool to analyze malware based on its structure, to discover its purpose and functionality but it takes a lot of time to see the malware structure.

Malware analysis is necessary to understand the behavior of malware. As a result, malware signature is created to effectively detect malware. Nevertheless, it wastes much time to find out the behavior and characterization of malware.

With a vast amount of samples increased day by day, anti-virus industry and virus researchers is harder to analyze malware without information of new malware. In order to reduce time of malware analysis, it is necessary to have an automatic malware classification system.

1.3 Approach

Two following issues are focused on:

- Automatically perform fast malware classification based on malware file's meta-data using a machine learning technique, called decision tree algorithm to classify unknown malwares or subspecies rapidly and correctly.
- Help researcher to understand which family malware belongs to and detect some semantic information about malware.

For those reason, in this paper, an approach is proposed to perform fast malware classification based on malware's meta-data using machine learning technique, known as decision tree.

1.4 Thesis outline

The rest of this thesis is structured as follows:

- Chapter 2 describes malware meta-data, and the method in static classification of malware. Insight is provided to understand the purpose of method given in this research.

- Chapter 3 presents the other malware static classification approach.
- Chapter 4 gives approach, design and operation of malware classification system
- Chapter 5 mentions environment and implementation of static malware system.
- Chapter 6 shows the system evaluation method and results.
- Summary, the conclusion and future work are presented in chapter 7

Chapter 2

Background

At first, this chapter presents the growth of malware. Then, malware analysis technique is introduced. Furthermore, malware categories and the way to use Virus total service to identify malware names defined by anti-virus vendor are introduced. Finally, there are problems of using malware name to detect semantic meaning of malware.

2.1 Growth of malware attack

Malware can self-replicate recursively. For example, Mota is a kind of worm that proliferate itself by sending spam email to addresses harvested on local machine [21]. In addition, Downadup is a malware that receives file through a peer-to-peer function[?]. Malware infects system and spreads to the other systems by communication tools such as the Internet and related technologies.

Since the rise of widespread broadband Internet access, the number of malware samples has been rapidly increasing. According to the report issued by Kaspersky's research team, 205 million pieces of malware were detected and neutralized [28]. In addition, in 2010 the number of malware samples increased by 20 millions [7].Figure 2.1 shows the increase of malware from 2003 to 2009 by Kaspersky Lab. As a result of the fast growth malware, malware is still a huge problem in Internet security and network connectivity.

At this time, malware is easily created by malware creation tools, known the programs which are used by attacker to generate malware[4]. In addition, unlike early attack tools implementing one type of attack, such tools now can be changed quickly by upgrading or replacing their code. This causes rapidly evolving attacks and, at the extreme, results in polymorphic tools that self-evolve, change with each active instance of the attack. Therefore, a large amount of malware have challenged anti-virus vendor and researcher to avoid their analysis.



Figure 2.1: Number of malicious program

[7]

2.2 Malware avoidance technique

At the present time, malware is implemented with avoidance technique in order to invalidate static signature based method by anti-virus software and make analysis process more complicated. Avoidance technique can change malware signature and syntax without changing the behaviors of malware. The avoidance technique consists of code obfuscating and packing technique. Therefore, an avoidance technique causes more difficulty in malware analysis.

Code obfuscation changes the form of a malware instance to other form in order to invalidate signature based on detection technique; it consists of polymorphism and metamorphism [15]. Polymorphic technique modifies the representation of malware. Virus, worm, and other self-propagating software are often applied polymorphic technique such as encryption, data appending, and data prepending. Metamorphic malware automatically recodes itself when it distributed or executed[15]. Simple metamorphic techniques include: adding and varying lengths of NOP instructions, adding useless instruction, and looping the code segment. Advantage metamorphism techniques consist of function: function reordering, program flow modification, static structure modification.

Packing malware can compress the Win32 portable execution file by several tools such as UPX,

Winpack, and ExeCryptor. According to a study carried out by Panda Research, 78% of new malware used some kinds of file packing to evade detection. PE-packer is designed to reduce the size of malware through compression. The size of packed malware is small but it is bigger when it runs in system[17]. When uncompressed in memory, packed malware is normally executed. UPX and some PE-packer compress malware binary files and make malware analysis harder.

For the reason that modern malware is implemented with avoidance technique, detection method by the use of static signatures is criticized for being ineffective.

2.3 Malware analysis technique

This section describes two malware analysis techniques including: dynamic and static analysis. The detail of two techniques is described as follow:

2.3.1 Dynamic malware analysis

Dynamic malware analysis technique is technique to find out the purpose of malware by running it and making sure what will happen in system. In general, malware is executed in the Virtual Machine. After that, malware researchers use SysAnalyzer, Process Explorer, ProcMon, RegShot, and other tools to identify the general behavioral analysis techniques. For example, SysAnalyser is a great analysis tool to monitors in many aspects of system and process states such as running process, open ports, loaded drivers, injected libraries, key registry changes, APIs called by a target process, file Modifications, http, IRC, and DNS traffic. Figure 2.2 shows the example of using SysAnalyser tool to analyse malware.

The advantage of dynamic malware analysis technique is simple. The process of malware analysis is running malware and making sure what will happen in system.

However, there are some disadvantage of dynamic malware analysis technique. At first, these dynamic analysis techniques are vulnerable to a variety of anti-monitoring defenses such as *time bombs* and *logic bombs*. Additionally, these techniques can be slow and tedious to identify and disable. In addition, a dynamic analysis may fail to identify the malware if the behavior of malware is not logged during the analysis process. Furthermore, Much time is required to prepare environment for malware analysis, such as virtual machine environment or sandbox. Moreover, Some malware cannot be executed in virtual machine environment. In conclusion, dynamic technique is not enough for malware analysis.

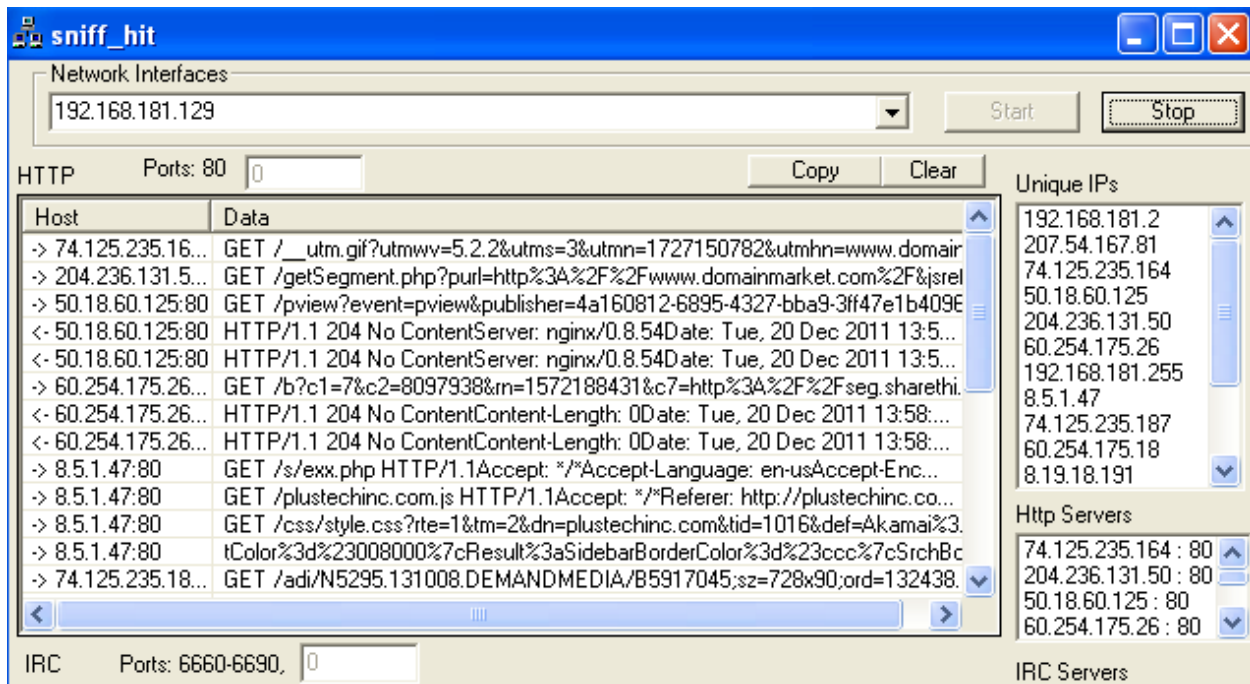


Figure 2.2: SysAnalyser tool for dynamic analysis.

2.3.2 Static malware analysis

Static malware analysis is a technique that identifies malware program without executing it. With the static malware analysis technique, researcher performs reverse engineering by using disassemble tools, decompile tools, source code analyzer tools such as IDA pro and Ollydbg in order to understand malware by seeing the structure of malware. Static malware analysis has an advantage that it can completely discover the purpose and functionality of malware. However, research needs time to understand the malware functionality by analysing malware structure.

In addition, most modern malwares use packer to modify itself. Packer is a program that modify other program files to compress their content. When a packer compress, encrypts, otherwise modifies an executable program, the program looks much different before it was packed. For analysing malware by Ollydbg or IDA pro, malware must be unpacked. PEid is a program that can find the signature of a known compiler or packer. Figure 2.3 gives PEid is applied to identify identifies the compiler and packer used by the malware author.

Ollydbg is a debugger that shows the assembly instruction of a program as they execute. Break-points can be setted on specific instructions to pause execution. Single step through the program one instruction at a time, or let the program run normally. Figure 2.4 show the example of using OllyDbg tool. There are four main windows in Ollydbg debugger: the code window in the center, the register window on the right, the stack window in the bottom right hand corner and the memory

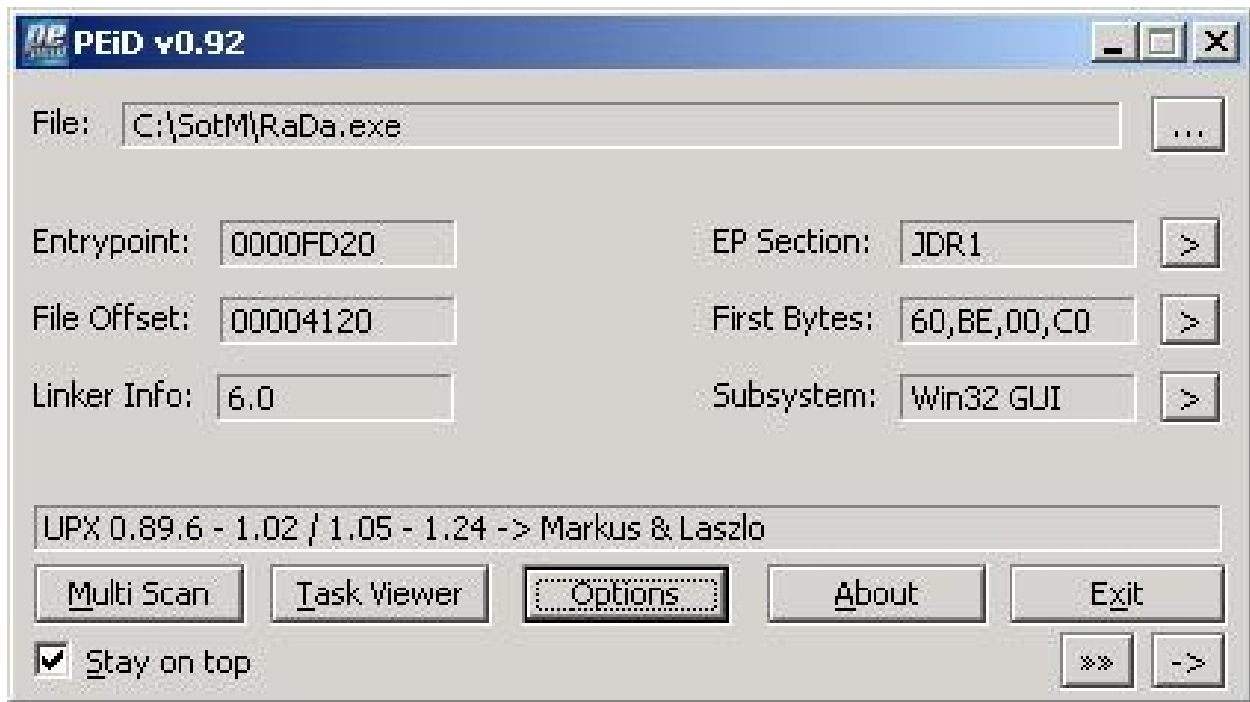


Figure 2.3: PEid identify the packer used by the malware author.

dump window in the bottom left corner.

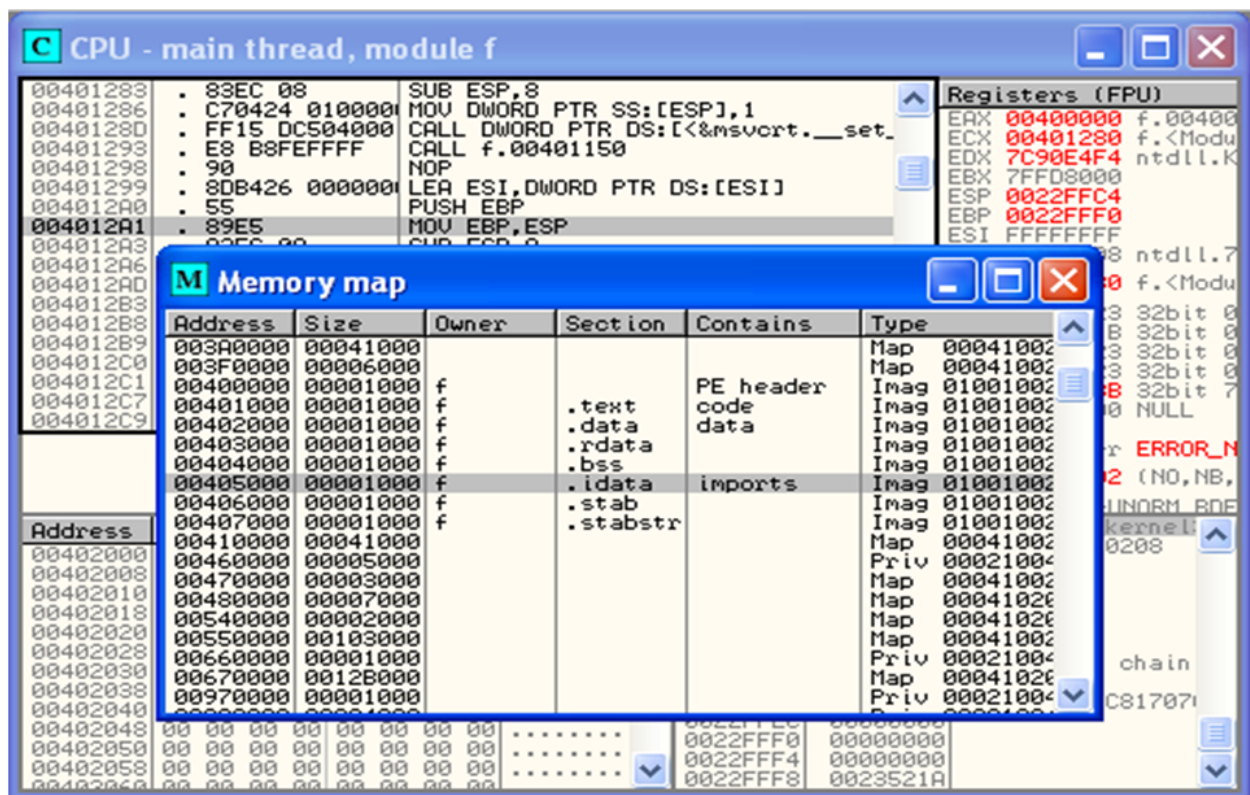


Figure 2.4: Ollydbg tool for static analysis.

IDA pro is a disassembly tool used by many researchers in different areas of reverse engineering. In addition to the disassembly listing, IDA pro also provides control-flow graphs. Figure 2.5 shows a control-flow graph provided by IDA pro.

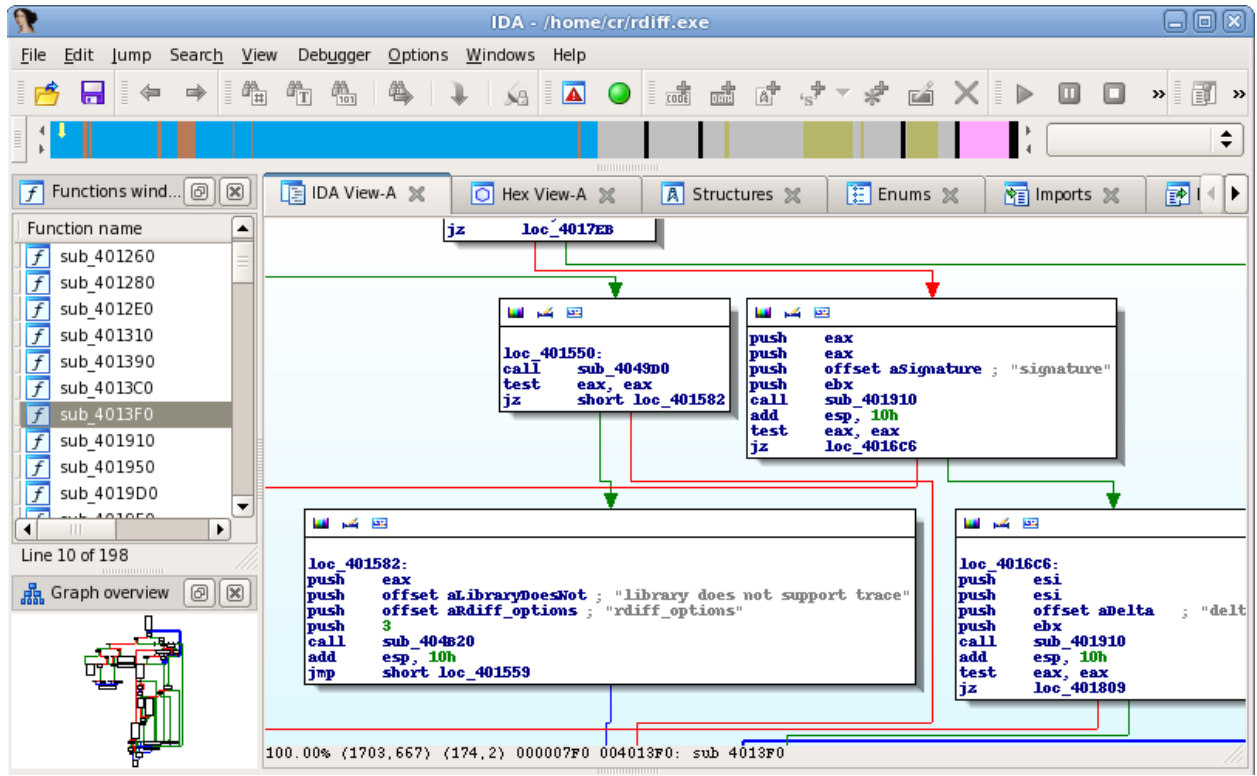


Figure 2.5: IDA pro for static analysis.

2.4 Malware categories

In general, malware is classified into a few categories or types based on behaviors, method of infection, and the resulting propagation of malware. For example, there are some malware categories: virus, trojan, worm, spyware, and rookit. The categories have specific types of malware threats:

- Virus : "Software which infects other application and use them as a spreading medium" [?].
- Trojan : "A malicious application which present itself as something else" [?].
- Worm : "Code with ability to spread from computer to computer by means of different network protocols" [?].
- Spyware : "Application aiming to havest personal information" [?].
- Rookit : "Hidden tools providing stealth services to its writer" [?].

However, the differences between the categories are a bit fuzzy, and the classes are obviously not exclusive. In addition, depending on purpose of virus industry, a unique malware can belong to rookit, virus, or spyware. The detail of malware categories is presented in the next section.

2.4.1 Use virus total to detect the name of categories.

In this thesis, MD5 hash is used to detect malware name provided by many anti-virus vendor. An MD5 hash is generated by MD5 Message-Digest Algorithm, is a widely used cryptographic hash function that produces a 128-bit (16-byte) hash value. An MD5 hash is typically expressed as a 32-digit hexadecimal number, and MD5 is not collision resistant[24].

2.4.2 Using virus total to getting vendor name

In this paper, MD5 hash of malware is used to search the name of malware by virus total. Virustotal is a web service that analyzes malware files and facilitates quick detection of viruses, worms, trojans, and all kinds of malware detected by antivirus engines. Malware's name is provided by various anti-virus vendor, but there are many names for unique malware. The Figure 2.6 reveals malware name detected by antivirus engines.

2.5 Problems of malware name

As the malware name detected by anti-virus engines is given in Figure 2.6, unique malware is not classified into unique name. For each anti-virus vendor, the name of malware detected is different to the other anti-virus vendor. The classification of each anti-virus engine is unlike others. Therefore, malware detected by anti-virus engines cannot provide meaningful characterization of malware for virus researcher. In order to overcome this problem, this paper proposed to classify the malware into families based on malware specific target and its operation behavior. As a result of malware name detected by anti-virus vendor was presented above, the new malware families are required for detect meaningful characterization of malware.

2.6 Malware families are used in this paper

This paper uses malware families which are reported by Information-technology Promotion Agency [8]. The table indicates malware families used in this paper such as Win32/Virut, Win32/Autorun, Win32/IRCbot, Win32/Gaobot, Win32/Waledac, Win32/Downadup, Win32/Sality, and W32.Mota.

Malware families	Summary
Win32/Virut	"Win32/Virut is a family of file infecting viruses that target and infect .EXE and .SCR files accessed on infected systems. Win32/Virut also opens a backdoor by connecting to an IRC server, allowing a remote attacker to download and run files on the infected computer." [12]
Win32/Autorun	"Win32/Autorun is a family of worms that spreads by copying itself to the mapped drives of an infected computer. The mapped drives may include network or removable drives." [11]
Win32/IRCbot	"Win32/IRCbot is a large family of backdoor Trojans that targets computers running Microsoft Windows. The Trojan drops other malicious software and opens a backdoor on the infected computer to connect to IRC servers. The Trojan can maintain multiple IRC server connections simultaneously to receive commands from attackers." [10]
Win32/Gaobot	"The Win32/Gaobot worm family spreads using different methods, depending on the variant. Some variants spread to machines with weak passwords. Others exploit vulnerabilities to infect machines. Once a machine is infected, the worm connects to an IRC server to receive commands." [9]
Win32/Waledac	"Win32/Waledac is a trojan that is used to send spam. It also has the ability to download and execute arbitrary files, harvest email addresses from the local machine, perform denial of service attacks, proxy network traffic and sniff passwords." [?]
Win32/Downadup	"Win32/Downadup attempts to spread to network shares by brute-forcing commonly used network passwords and by copying itself to removable drives." [?]
Win32/Sality	"Virus:Win32/Sality is a family of polymorphic file infectors that target Windows executable files with the extensions .SCR or .EXE. They may execute a damaging payload that deletes files with certain extensions and terminates security-related processes and services.
W32.Mota	W32.Mota is a worm that propagates by sending itself to email addresses gathered from the computer." [21]

Table 2.11 Malware

Antivirus	Version	Last Update	Result
AhnLab-V3	2011.08.13.00	2011.08.13	Win-Trojan/Downloader.6144.QR
AntiVir	7.11.13.37	2011.08.12	TR/Crypt.XPACK.Gen
Antiy-AVL	2.0.3.7	2011.08.13	Trojan/Win32.Agent.gen
Avast	4.8.1351.0	2011.08.14	Win32:Trojan-gen
Avast5	5.0.677.0	2011.08.14	Win32:Trojan-gen
AVG	10.0.0.1190	2011.08.14	Win32/Heur
BitDefender	7.2	2011.08.14	Trojan.Generic.304464
CAT-QuickHeal	11.00	2011.08.13	TrojanDownloader.Agent.pkw
ClamAV	0.97.0.0	2011.08.13	Trojan.Downloader-38041
CommTouch	5.3.2.6	2011.08.13	W32/Downldr2.BXKD
Comodo	9739	2011.08.14	TrojWare.Win32.TrojanDownloader.Small.OBW
DrWeb	5.0.2.03300	2011.08.14	Trojan.DownLoader.61643
Emsisoft	5.1.0.8	2011.08.14	Trojan-Dropper.Agent!IK
eSafe	7.0.17.0	2011.08.10	Win32.Agent.pkw
eTrust-Vet	36.1.8499	2011.08.12	-
F-Prot	4.6.2.117	2011.08.13	W32/Downldr2.BXKD
F-Secure	9.0.16440.0	2011.08.14	Trojan.Generic.304464
Fortinet	4.2.257.0	2011.08.14	W32/Agent.PKW!tr.dldr
GData	22	2011.08.14	Trojan.Generic.304464
Ikarus	T3.1.1.107.0	2011.08.14	Trojan-Dropper.Agent
Jiangmin	13.0.900	2011.08.13	TrojanDownloader.Agent.aavi
K7AntiVirus	9.109.5010	2011.08.12	Trojan-Downloader
Kaspersky	9.0.0.837	2011.08.14	Trojan-Downloader.Win32.Agent.pkw

Figure 2.6: Malware name is detected by antivirus engines.

Chapter 3

Related research

This chapter presents the recent approach for automatically classifying malware into malware families. For the reason that the detection based on static signature is no longer effective in chapter 2, at this time new malware classification method is required to detect malware writers used avoidance technique.

3.1 Flow graph

Another approach is using emulator to automatically unpack the packing malware, then from reverse code produce flowgraph, flowgraph matching to perform classification[29]. The system is created by the approach as the followings:

- First, System automatically unpacks malware based on application level emulation, and then uses entropy analysis to detect.
- Second, Produce control flowgraph by using graph invariant based on signature in order to measure similarity between malware.
- Next, Generic string based on control flow signature, in order to use string edit distance. Automatically unpack malware based on application level emulation.
- Finally, Malware classification uses a set similarity function and a set similarity search algorithm to represent benign and malicious classes.

The disadvantage in flowgraph approach is high cost of runtime complexity. Firstly, runtime complexity of malware classification is $O(N \log M)$ where M is the number of control flow graphs

in database, and N is the number of control flow graph input binary [29]. In addition, to identify two flowgraphs is N^3 , and N is the number of nodes in each graph. Further more, needs to unpack the sample if it was packed with an executable packer. Therefore, this approach is ineffective in malware classification system with a large number of instances.

In addition, malware classification based on flowgraph approach cannot accurately detect metamorphic malware. As presented in chapter 2, metamorphic malware can recode itself with program flow modification, function reordering. As a result, it is hard to classify malware based a set similarity between flow graph.

3.2 Optimizing decision tree in malware classification system by using Genetic Algorithm

In this time, malware classification system can use machine learning classifier. The main idea of machine learning technique classifier is search algorithm that is learnt from externally supplied instances to produce a concise model of the distribution, which then makes prediction about new instances. Current machine learning classifier in malware classification include: Naive Bayse, Support Vector machine, Decision tree, K-nearest Neighbor [14]. This approach uses combining Genetic algorithms with Decision tree. The data set is separated in training data set and testing data set. The data set includes:

```
// Malware classes for chromosome
hash_map<MalwareClass*,int> _Dataclasses;
hash_map<MalwareClass*,int> _Appsclasses;
hash_map<MalwareClass*,int> _Sysclasses;
hash_map<MalwareClass*,int> _Dosclasses;
```

Genetic algorithm is method analogous to the process of natural evolution. Genetic algorithm is used to optimize decision tree for accurately classifying malware. Malware classification system, which is implemented by the combining generic algorithm with decision tree algorithm approach, is to classify malware into two classes: benign program and malicious program, not to detect semantic characterization of malware by classifying malware into families. Therefore, the combining generic

algorithm with decision tree algorithm approach cannot use for detecting semantic characterization of malware.

3.3 Conclusion

To achieve fast malware classification, two approaches described as the above can not be used in this system. New approach for fast malware classification system will be introduced in the next chapter.

Chapter 4

Classification based on malware's meta-data using decision tree approach

This chapter describes Win32 executive files structure, known as PE files structure, that is used in classification system as malware meta-data. Afterwards, the machine learning technique called decision tree method is mentioned in classification malware system. Finally, this chapter presents classification based on malware's meta-data using decision tree approach.

4.1 PE file format[25]

4.1.1 The PE File Format

The Portable Executable (PE) file format has been designed for use by all Win32 based system. The general layout of a PE file is given in Figure 4.2. As shown in Figure 4.2 PE header consists of four parts: a MS-DOS session including DosMZ header and Dos stub; PE header; a section table; and images pages consisting of several different regions such as .text session, and .data session.

The PE file format starts with DOS MZ header. The first two bytes - "MZ" of the header is the DOS EXE signature. The first few hundred bytes of the typical PE file are taken up by the MS-DOS stub. This stub is a tiny program that displays a string as "This program cannot be run in MS-DOS mode."

The following part is PE header. The fields of this structure contain only the most basic

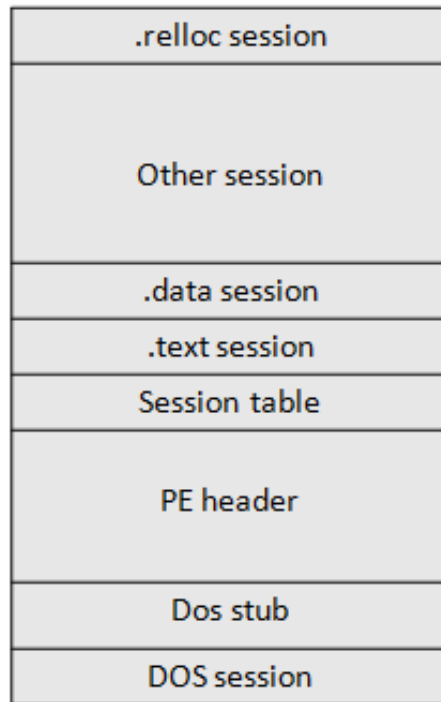


Figure 4.1: The general layout of a PE file.

information about the file such as the locations and sizes of the code and the data areas. These data structures of PE header will be explained shortly.

Following the PE header is the session table. The section table is an array of `IMAGE_SECTION_HEADERS` structures. An `IMAGE_SECTION_HEADER` provides information about its associated section such as location, length, and characteristics. The number of elements in this array is given in the PE header, known as `NumberOfSections` field. At a minimum, there are commonly at least two sections in a PE file: code and data. Figure shows a section table for a typical PE file.

That is all about the physical layout of the PE file format. The major steps in loading a PE file into memory is described as follow:

- When the PE file is running, the PE loader examines the DOS MZ header for the offset of the PE header. If it is found, it would skip to the PE header.
- The PE loader checks if the PE header is valid. Then, it goes to the end of the PE header if the PE header is valid.
- Following the PE header is the section table. The PE header reads information about the sections and maps those sections into memory using file mapping. It also gives each section the attributes as specified in the section table.

```

01 .text      VirtSize: 00005AFA  VirtAddr: 00001000
    raw data offs: 00000400  raw data size: 00005C00
    relocation offs: 00000000  relocations: 00000000
    line # offs: 00009220  line #'s: 0000020C
    characteristics: 60000020
    CODE  MEM_EXECUTE  MEM_READ

02 .bss       VirtSize: 00001438  VirtAddr: 00007000
    raw data offs: 00000000  raw data size: 00001600
    relocation offs: 00000000  relocations: 00000000
    line # offs: 00000000  line #'s: 00000000
    characteristics: C0000080
    UNINITIALIZED_DATA  MEM_READ  MEM_WRITE

03 .rdata     VirtSize: 0000015C  VirtAddr: 00009000
    raw data offs: 00006000  raw data size: 00000200
    relocation offs: 00000000  relocations: 00000000
    line # offs: 00000000  line #'s: 00000000
    characteristics: 40000040
    INITIALIZED_DATA  MEM_READ

04 .data      VirtSize: 0000239C  VirtAddr: 0000A000
    raw data offs: 00006200  raw data size: 00002400
    relocation offs: 00000000  relocations: 00000000
    line # offs: 00000000  line #'s: 00000000
    characteristics: C0000040
    INITIALIZED_DATA  MEM_READ  MEM_WRITE

05 .idata     VirtSize: 0000033E  VirtAddr: 0000D000
    raw data offs: 00008600  raw data size: 00000400
    relocation offs: 00000000  relocations: 00000000
    line # offs: 00000000  line #'s: 00000000
    characteristics: C0000040
    INITIALIZED_DATA  MEM_READ  MEM_WRITE

06 .reloc     VirtSize: 000006CE  VirtAddr: 0000E000
    raw data offs: 00008A00  raw data size: 00000800
    relocation offs: 00000000  relocations: 00000000
    line # offs: 00000000  line #'s: 00000000
    characteristics: 42000040
    INITIALIZED_DATA  MEM_DISCARDABLE  MEM_READ

```

Figure 4.2: A section table for a typical PE file.

- After the PE file is mapped into memory, the PE loader concerns itself with the logical parts of the PE file, such as the import table.

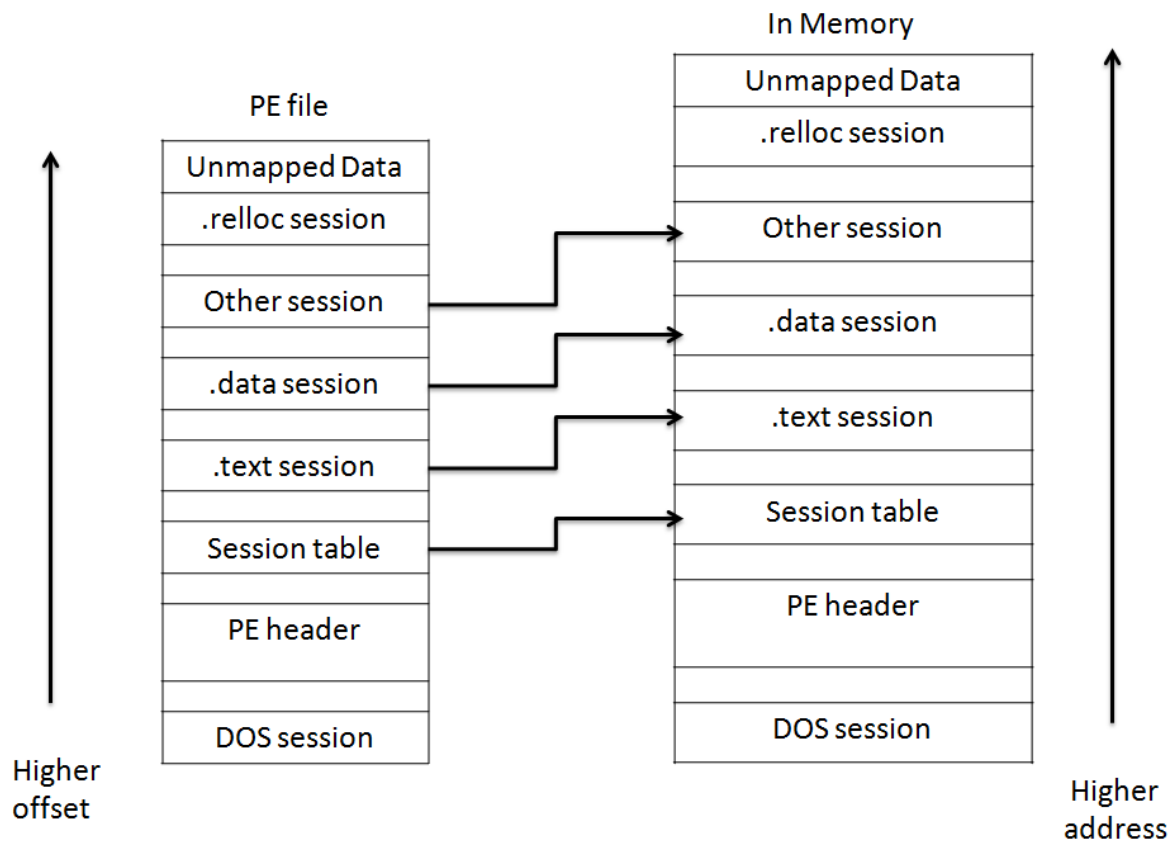


Figure 4.3: PE file format.

4.1.2 The PE Header

The PE header is second part of PE files structure. The PE header starts with two characters "PE" known the PE header file signature. The main PE header is a structure of type `IMAGE_NT_HEADERS`, which is defined in `WINNT.H`. The structure of PE header consists of a `DWORD` and two substructures composed of `FileHeader` and `OptionalHeader`. `IMAGE_NT_HEADERS` contains the three fields:

```
DWORD Signature;
IMAGE_FILE_HEADER FileHeader;
IMAGE_OPTIONAL_HEADER OptionalHeader;
```

Following the PE signature `DWORD` in the PE header is a structure of type `IMAGE_FILE_HEADER`.

```
typedef struct _IMAGE_FILE_HEADER {
    WORD Machine;
    WORD NumberOfSections;
```

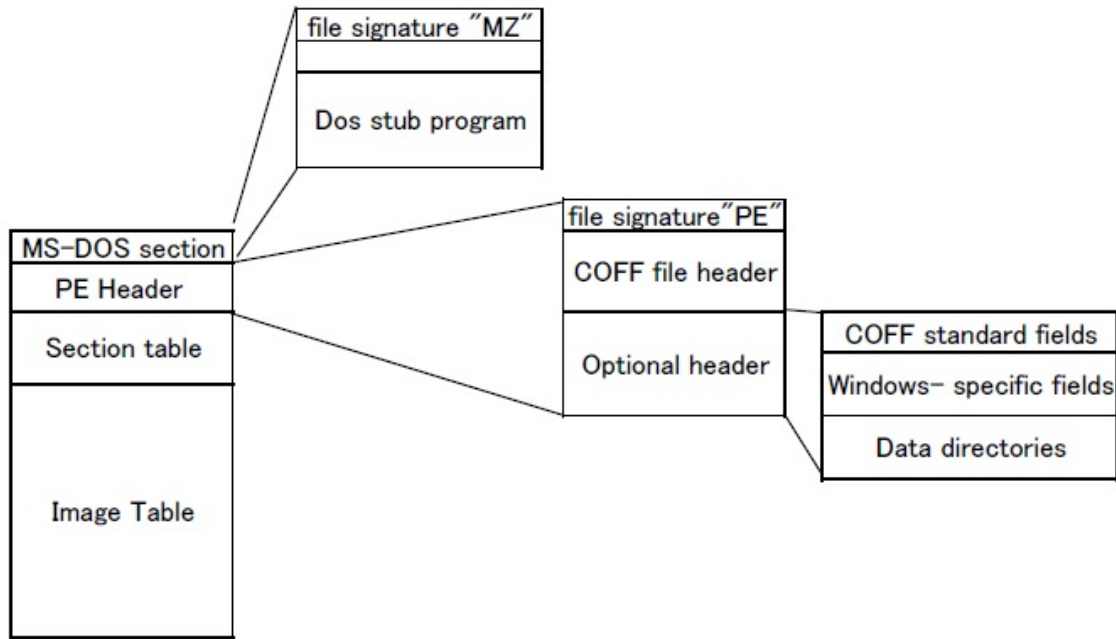


Figure 4.4: Layout a file in PE header format.

```

DWORD TimeDateStamp;

DWORD PointerToSymbolTable;

DWORD NumberOfSymbols;

WORD   SizeOfOptionalHeader;

WORD   Characteristics;

} IMAGE_FILE_HEADER, *PIMAGE_FILE_HEADER;

```

The fields of this structure contain only the most basic information about the file. The fields of the `IMAGE_FILE_HEADER` includes:

- Machine: The CPU that this file is intended for.
- NumberOfSections: The number of the members in section
- TimeDateStamp: The time that the linker
- PointerToSymbolTable: The file offset of the COFF symbol table.
- NumberOfSymbols: The number of symbols in the COFF symbol table.
- SizeOfOptionalHeader: The size of an optional header that can follow this structure.

- **Characteristics:** The size of an optional header that can follow this structure. If there are no relocations in this file, characteristics will be 0x0001. If there are no relocations in this file, characteristics will be 0x0001 . If file is an executable image, characteristics will be 0x0002. If file is a dynamic-link library, not a program, characteristics will be 0x2000.

Following the file header is the optional header which consists of three parts: COFF standard fields, Windows specific fields, and data directories. The first part includes the COFF standard field which contains the sizes of various parts of code, and the *AddressOfEntryPoint* which indicates the location of the entry point of the application and the location of the end of the Import Address Table(IAT). Then, the second part contains the information of OS. Finally, the Data directories field contains the information of Session table.

The following is the structure of optional header.

```
typedef struct _IMAGE_OPTIONAL_HEADER {
    WORD                Magic;
    BYTE                MajorLinkerVersion;
    BYTE                MinorLinkerVersion;
    DWORD               SizeOfCode;
    DWORD               SizeOfInitializedData;
    DWORD               SizeOfUninitializedData;
    DWORD               AddressOfEntryPoint;
    DWORD               BaseOfCode;
    DWORD               BaseOfData;
    DWORD               ImageBase;
    DWORD               SectionAlignment;
    DWORD               FileAlignment;
    WORD                MajorOperatingSystemVersion;
    WORD                MinorOperatingSystemVersion;
    WORD                MajorImageVersion;
    WORD                MinorImageVersion;
    WORD                MajorSubsystemVersion;
    WORD                MinorSubsystemVersion;
```

```

DWORD          Win32VersionValue;

DWORD          SizeOfImage;

DWORD          SizeOfHeaders;

DWORD          CheckSum;

WORD           Subsystem;

WORD           DllCharacteristics;

DWORD          SizeOfStackReserve;

DWORD          SizeOfStackCommit;

DWORD          SizeOfHeapReserve;

DWORD          SizeOfHeapCommit;

DWORD          LoaderFlags;

DWORD          NumberOfRvaAndSizes;

IMAGE_DATA_DIRECTORY DataDirectory[IMAGE_NUMBEROF_DIRECTORY_ENTRIES];
} IMAGE_OPTIONAL_HEADER, *PIMAGE_OPTIONAL_HEADER;

```

The fields of Optional header are described as follow:

- Magic: appears to be set to 0x010B.
- MajorLinkerVersion
- MinorLinkerVersion: the version of the linker that produced this file.
- SizeOfCode: this field matches the size of the .text section.
- SizeOfInitializedData:
- SizeOfUninitializedData:
- AddressOfEntryPoint:
- BaseOfCode:
- BaseOfData:
- ImageBase:
- SectionAlignment:
- FileAlignment:

- MajorOperatingSystemVersion:
- MinorOperatingSystemVersion:
- MajorImageVersion:
- MinorImageVersion:
- MajorSubsystemVersion:
- MinorSubsystemVersion:
- Reserved1:
- SizeOfImage:
- SizeOfHeaders:
- CheckSum:
- Subsystem:
- DllCharacteristics:
- SizeOfStackReserve:
- SizeOfStackCommit:
- SizeOfHeapReserve:
- SizeOfHeapCommit:
- LoaderFlags:
- NumberOfRvaAndSizes:
- DataDirectory:

4.2 Decision tree[27]

Decision tree is a machine learning technique for building knowledge-based systems by inductive inference from examples. Decision tree is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the Figure 4.6 from the data Table 4.5. The decision tree is created the Table from Each interior node corresponds to one of the

input variables. There are edges to children for each of the possible values of that input variable. Each leaf shows a value of the target variable.

Condition	Class	Attribute 1	Attribute	
1	A	0	0	...
2	B	1	0	...
3	B	1	1	...
...

Figure 4.5: The data table.

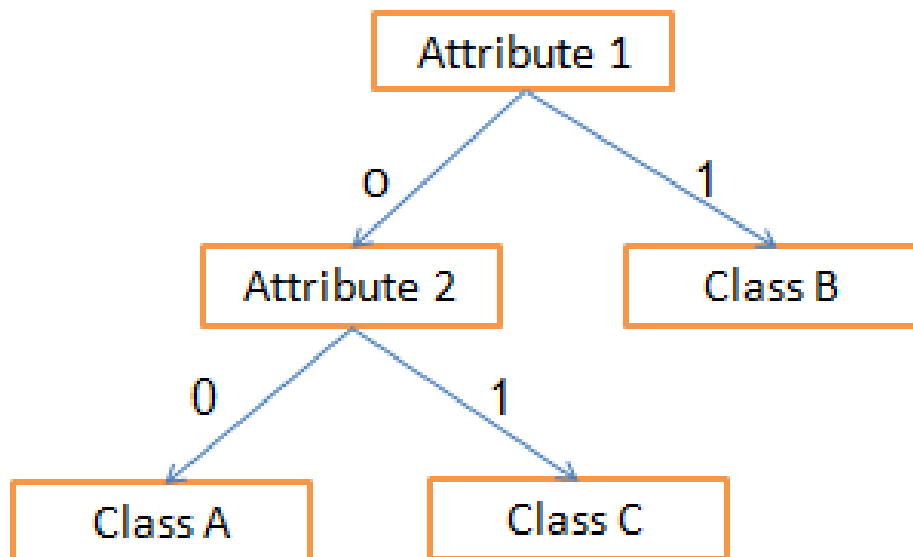


Figure 4.6: Decision tree sample.

In data mining, the decision tree technique can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form

$$(x, Y) = (x_1, x_2, x_3, \dots, x_k, Y) \quad (4.1)$$

The dependent variable, Y which is the target variable is analysed to understand, classify or generalise. The vector x is composed of the input variables, x_1, x_2, x_3 that are used for that task.

4.3 Classification based on malware's meta-data using decision tree approach

To make malware classification rapid and correct, decision tree algorithm is used in malware classification system.

As the previous section, Data comes in records of the form:

In malware classification system, Y is malware families name, $x_1, x_2, x_3, \dots, x_k, Y$ is meta-data. Malware name is provided by anti-virus engines that presented in chapter 2 cannot used in this system. The new malware families is required for detect meaningful characterization of malware. This classification system use famous malware families that reported in Information technology agency such as netsky, mydoom, autorun, mytob, and other famous malware families. Therefore, when new malware is classified into famous malware families, researcher detect some meaningful characterize of malware, and reduce time for malware analysis.

Chapter 5

Implementation

5.1 Environment

5.2 Overview

The system uses machine learning technique, named decision tree algorithm for malware classification. The goal is classifying malware fast and decision tree algorithm is used to achieve. Classification is a form of supervised learning, which requires training data, with known input/output, to form knowledge [34]. As mentioned in Figure 5.1, the system contains three following parts.

- First part: Read binary file to take meta data and input to database.
- Second part: Manually cluster malware loaded from database.
- Third part: Use decision tree algorithm to classify malware.

A vast amount of malware need analyzing, and the system uses database to store malware file's meta-data, to easily control a large number of data and to share them in the internet among web browsers. The flow of data is shown in the Figure 5.1. Firstly, binary file meta-data are read, and all of the meta-data are input into database. In the second part, the system exports meta-data in database as the training data to create the decision tree for classifying malware. Finally , the decision tree algorithm created in the second step is utilized to class unknown malware into the different malware families.

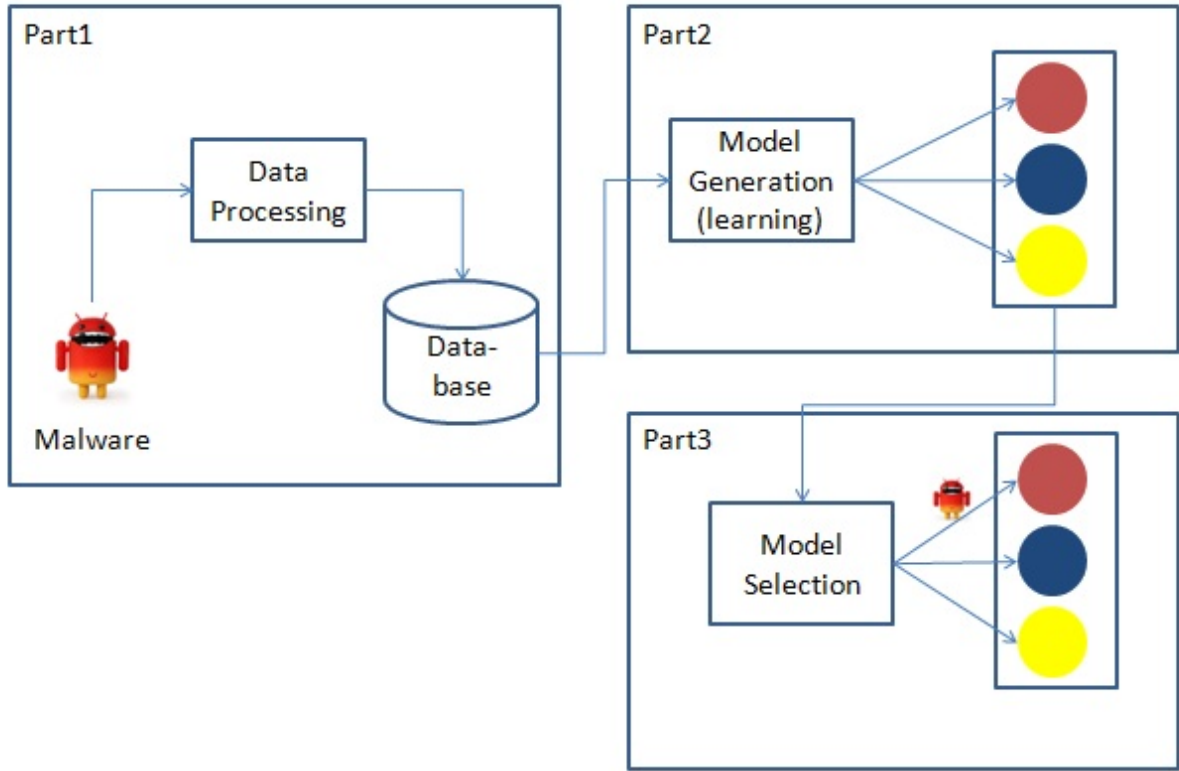


Figure 5.1: The system architecture.

5.3 Classification based on machine learning technique

5.3.1 Meta-data

In order to create a fast malware classification system, the system only uses PE header's meta-data to classify malware. PE header includes meta-data of MS-DOS section, COFF file header, Optional header, and Section header.

There is a problem that PE file's meta data value is too large to detect unknown malwares which have meta-data's differences from the training data of the system. The meta-data of PE header file are indexed to give semantic information for malware classification. For example, normally *ImageBase* field of PE header file has value of 400000h [38]. The value of *ImageBase* of malware file in the system is 0 if it has 40000, and it is 1 if it has other value, ; in this research, that value is named as semantic value. This approach uses all of PE file's meta-data, specifically *HeaderFilesize*, *AddressOfEntryPoint*, *Sizeofsection*, *Imagebase*, *Numberofsection*, *StackCommit*, *SizeHeap Commit*, *SizeImage*, *Characteristics*. The semantic value of all malware file's meta-data is calculated.

5.3.2 Create training data

Virustotal is a service that analyzes malware and facilitates quick collection of viruses, worms, trojans, and all kinds of malware are detected by antivirus vendor such as Norton and Kaspersky [?]. This research applies virustotal service to take the name of malware provided by Kaspersky and uses it to cluster malware into malware families with semantic information. Figure 5.2 indicates http post technique to automatically get name from virus total service and cluster malware into malware families which given in Figure ??.

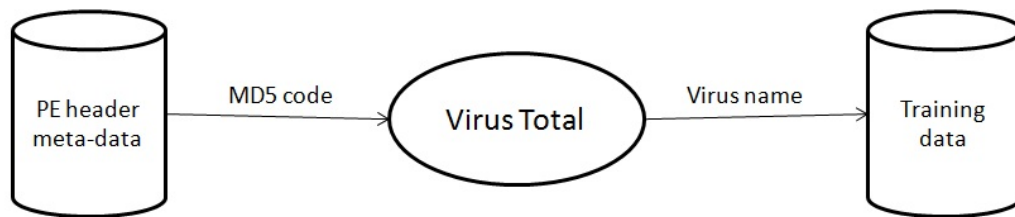


Figure 5.2: Clustering method.

5.3.3 Classification

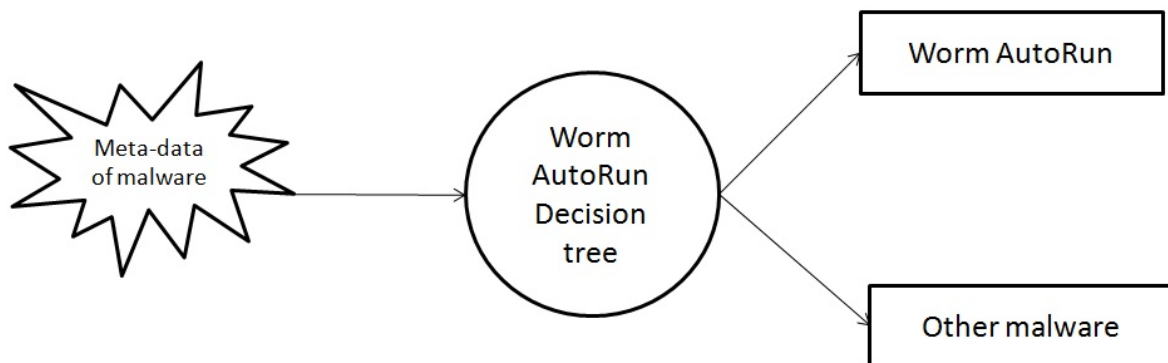


Figure 5.3: Worm autorun decision tree.

To make malware classification rapid and correct, decision tree algorithm is utilized. At that

time, it is easy to update a new malware family, the classification uses decision tree algorithm to determine each malware family. For example, based on training data from clustering part, six decision trees are created. The worm autorun decision tree is indicated in Figure 5.3.

Malware' meta-data are known as input data for each decision tree, and the decision tree determines whether the malware is of worm autorun family or not. If the input malware belongs to worm autorun, malware classification shall detect which the family that malware is, otherwise it moves to the next decision tree. Lists of malware PE header file's meta-data are Magic, MajorLinkerVersion, MinorLinkerVersion, SizeOfCode, SizeOfInitializedData, SizeOfUninitializedData, AddressOfEntryPoint, BaseOfCode, BaseOfData, ImageBase, SectionAlignment, FileAlignment, MajorOperatingSystemVersion, MinorOperatingSystemVersion, MajorImageVersion, MinorImageVersion, MajorSubsystemVersion, MinorSubsystemVersion, Reserved1, SizeOfImage, SizeOfHeaders, CheckSum, Subsystem, DllCharacteristics, SizeOfStackReserve, SizeOfStackCommit, SizeOfHeapReserve, SizeOfHeapCommit, LoaderFlags, NumberOfRvaAndSizes, Machine, NumberOfSections, TimeDateStamp, PointerToSymbolTable, NumberOfSymbols, SizeOfOptionalHeader, Characteristics. The value of each field in PE header has semantics for decision tree algorithm if it is appeared at 10% in malware training data. With this approach, the semantic value list of meta-data malware is made. For example, this is a list of meta-data created :1, 2, 6, 4, 0, 1, 1, 4, 0, 200, 4, 0, 0, 0, 4, 0, 0, 1.00E+00, 200, 0, 2, 0, 100000, 4000, 100000, 1000, 0, 10, 14c, 2, 1000, 8, 0, e0, 3, 818e. This is value before being computed :35328, 10b, 2, 0, 6200, 0, 200, b32e, b000, 9000, 400000, 1000, 200, 3, 0, 0, 0, 4, 0, 0, d000, 400, afe8, 2, 0, 1000, 1000, 10000, 0, 0, 10, 14c, 6, 0, 0, 0, e0, 10e. The malware classification part is shown in Figure 5.4. The training data from the malware clustering part is used to create the order of decision tree that makes malware classification system correct. The decision tree is given as in Figure 5.5. When PE header's meta-data of malware is inputted into the worm autorun decision tree for determining unknown malware. Then, it decides whether it belongs to worm auto run family or not.

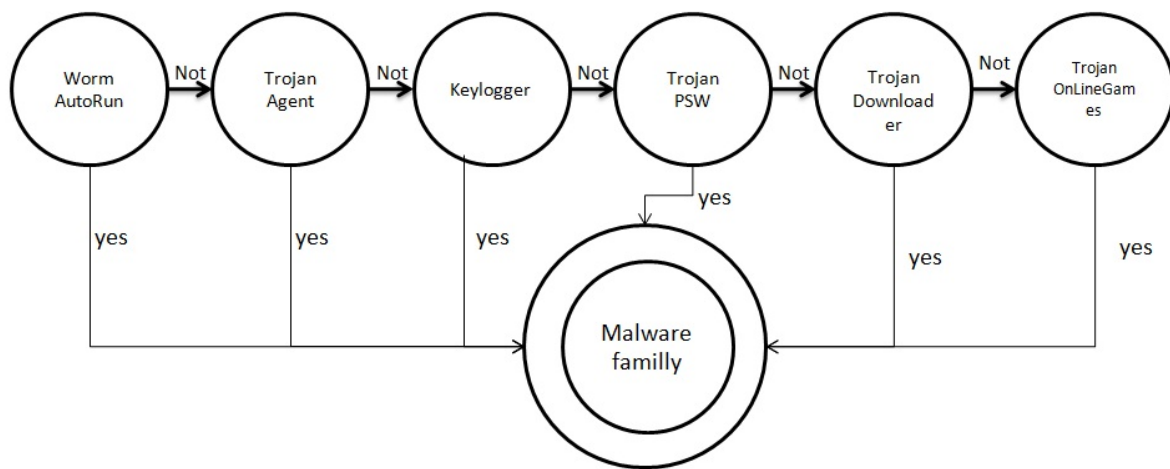


Figure 5.4: Malware classification system.

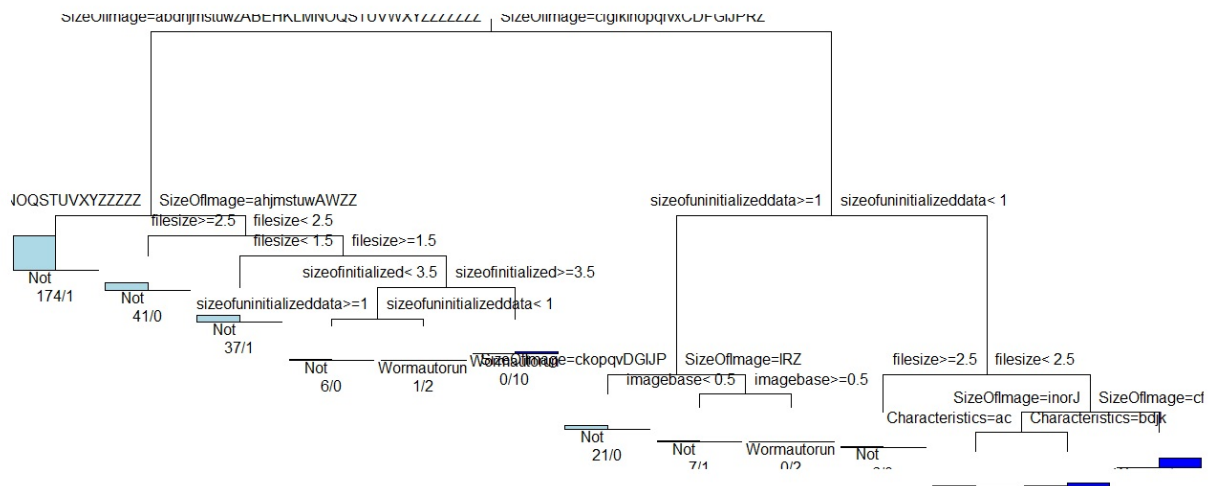


Figure 5.5: Worm autorun decision tree.

Chapter 6

Evaluation

6.1 Accuracy evaluation

In 935 malwares obtained, 635 of them are to make eight decision trees. 200 malwares meta-data in training data are applied to, in order to sort 8 decision tree, and to make the best order of decision tree as shown in Figure 6.1, this order has the best experimental result with ? malware training data. Finally, 100 remaining malwares meta-data are kept to test experimental result the system. Experiment results in the order of decision tree given in Figure 6.1.

In the decision tree order, 100 malwares are taken to check the experimental result of system, and Table 6.2 reveals that result. Some of data in axis show these total number of malware in that family, and that number is separated into groups that these malwares has been classified by the system. For example, there are four malwares in Win32/Virut family, and three of them are successfully sorted into Win32/Virut family and, while one malware is in other family. Therefore, the system recognizes the trojan agent family with 75 % of accuracy.

The system is useful to help virus researcher determine the malware family that unknown malware belongs to. The malware family contains Win32/Virut, Win32/Autorun, Win32/IRC'bot, Win32/Gaobot, Win32/Waledac, Win32/Downadup, Win32/Sality, W32.Mota, known as famous malware family. Virus researchers who know the family of malware can easily find out some semantic similarities between malwares and shows their inner similarity in behavior and static malware characteristics.

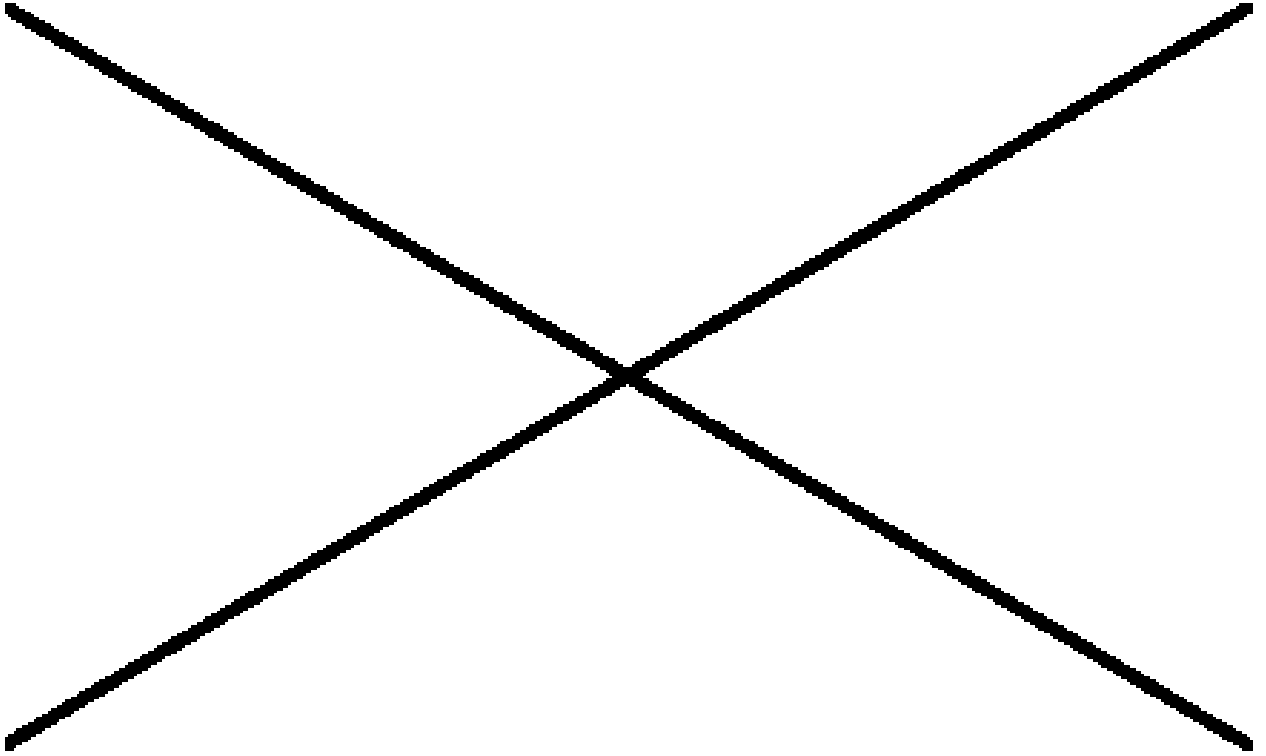


Figure 6.1: The best decision tree order.

Malware	Virut	Autorun	IRCbot	Gaobot	Waledac	Downadup	Sality	Mota	Accuracy
Virut	?	?	?	?	?	?	?	75%	
Autorun	0	0	0	?	?	?	?	50%	
IRC'bot	0	0	2	0	1	0	0	66%	
Gaobot	1	0	0	2	8	0	3	53%	
Waledac	0	0	0	0	0	0	0	0%	
Downadup	0	0	0	0	0	1	1	50%	
Sality	0	0	0	0	0	1	1	50%	
Mota	4	4	2	3	15	2	41	57%	

Figure 6.2: Experimental result

6.2 Classification efficiency

Figure shows comparison between steps to detect semantic malware characterization using Virus total with the system.

The figure 1 shows classifying time in the system. The evaluation was performed on a 2.4 HZ

core i3 laptop with 4G memory, running in ubuntu 11.4.

The result was shown in figure ???. The median time to perform classification was 0.25 seconds. The slowest sample that is required 5.12 seconds. Only 6 samples required more than 2 seconds. Processing time of followgraph approach was shown in figure ??.

Chapter 7

Conclusion

7.1 Conclusion

Analyzing a great number of new malware samples each day is a difficult problem. The fast malware classification system is successful to classify unknown malware into malware families which have semantic specifications. It is strongly believed that system is useful in malware analysis to determine malware behavior and semantic malware characteristics, and to more easily examine a large number of malwares.

However, there is a problem that the system only uses malware meta-data and cannot detect the malware family with same program structure.

Surprisingly, the system cannot be used to classify W32 malware which does not have the signature of PE file signature.

7.2 Future work

The anonymous malwares in the system are classified into malware family, and in the future work, Windows API shall be applied to determine this with similarity in program structure.

Acknowledgement

First and foremost, I am sincerely grateful to faculty members in Murai and Tokuda Laboratory, Professor Hideyuki Tokuda, Professor Jun Murai, Associate Professor Hiroyuki Kusumoto, Professor Osamu Nakamura, Associate Professor Kazunori Takashio, Assistant Professor Noriyuki Shigechika, Assistant Professor Rodney D. Van Meter III, Associate Professor Keisuke Uehara, Associate Professor Jin Mitsugi, Lecturer Jin Nakazawa.

I would be extremely thankful to Professor Keiji Takeda for his endless help and valuable comments to my thesis. Without his help, I cannot even think of writing this thesis.

I give my special thank to Doctor Masayoshi Mizutani for his sincerely help and many advices since the time I first came to the lab.

I would like to thank to Master Kunihiro Shigematsu, Yuki Uehara, my friend Toshinori Usui, for listening and giving me many useful discussion points. Their gentleness and encouragement helped me overcome hard periods in this one year.

I would also like to thank all ISC group members, Sega-kun, Alc-kun, Lushe-kun, Nobita-kun, Fuyuki-kun, Yoshi-kun, Asp-chan, Hino-kun, Rocky-kun, Vigo-kun, who have given me numerous support and always cheer me up in my hard time.

I owe my deep gratitude to members in Murai and Tokuda Laboratory who have been so friendly and kindly supporting me with my research.

Bibliography

- [1] David Zimmer. <http://www.woodmann.com/collaborative/tools/index.php/SysAnalyzer>. 21 March 2011.
- [2] Symantec Announces MessageLabs Intelligence. <http://www.symantec.com/about/news/release/article.jsp?prid=2010120701>. 12 Dec 2011.
- [3] John Bambenek. <http://handlers.dshield.org/jbambenek/keylogger.html>. 13 Sep 2005.
- [4] Malware protection center. <http://www.microsoft.com/security/portal/Threat/Encyclopedia/Glossary.aspx>. 21 Dec 2011.
- [5] Wikipedia. <http://en.wikipedia.org/wiki/Decision-tree-learning>. 21 Dec 2011.
- [6] Wikipedia. <http://en.wikipedia.org/wiki/Antivirus-software>. 14 Jun 2011.
- [7] Eugene Kaspersky. <http://www.kaspersky.co.in/news?id=207576357>. 14 Jun 2011.
- [8] Information-technology Promotion Agency, Japan. <http://www.ipa.go.jp/security/txt/list.html>. 01 Jan 2012.
- [9] Malware Protection Center. <http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Win32%2fGaobotssummarylink>. 01 Jan 2012.
- [10] Malware Protection Center. http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Win32%2fFIRCBot#symptoms_link. 01 Jan 2012.
- [11] Malware Protection Center. <http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Win32%2fAutorun>. 01 Jan 2012.
- [12] Malware Protection Center. <http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Win32%2fVirus>. 01 Jan 2012.

- [13] Qinghua Zhang, Douglas S. Reeve. *MetaAware: Identifying Metamorphic Malware*. 23rd Annual Computer Security Applications Conference 2007.
- [14] Mohd Najwadi Yusoff and Aman Jantan. *Optimizing Decision Tree in Malware Classification System by using Genetic Algorithm*. The Society of Digital Information and Wireless Communications, 2011.
- [15] MalwareChet Hosmer, Chief Scientis. *Polymorphic & Metamorphic Malware*. Black Hat, 2008.
- [16] S. B. Kotsiantis. *Supervised Machine Learning: A Review of Classification Techniques*. Informatica 31, 2007.
- [17] Paul Craig. *PE Packers Used in Malicious Software*. Ruxcon, 2006.
- [18] Panda Security Internacional. <http://www.pandasecurity.com>. 01 Jan 2012.
- [19] Malware Protection Center. <http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Name=Win32%2fWaledac>. 01 Jan 2012.
- [20] Malware Protection Center. <http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Name=Win32%2fSality>. 01 Jan 2012.
- [21] Symantec. http://www.symantec.com/security_response/writeup.jsp?docid=2004070412024899. 01 Jan 2012.
- [22] <http://news.softpedia.com/news/Kaspersky-Report-161-Million-Network-Attacks-Blocked-232812.shtml>. 14 December 2011.
- [23] <http://win32assembly.online.fr/pe-tut1.html>. 14 December 2011.
- [24] Wikipedia. <http://en.wikipedia.org/wiki/MD5>. 14 December 2011. Perdran amini, Ero Carrera *Reverse engineering on windows: Application in malicious code analysis*. Masters Thesis, Central Queensland University. 2010.
- [25] *Peering Inside the PE: A Tour of the Win32 Portable Executable File Format*. Microsoft Systems Journal. March 1994.
- [26] <http://win32assembly.online.fr/petut1.html>. 17 Feb 2010.
- [27] *Induction of decision trees*. Machine learning 1, 2007.
- [28] <http://www.viruslistjp.com/analysis/?pubid=204792101>. 17 Feb 2010.

- [29] Silvio CESARE, *Fast Automated Unpacking and Classification of Malware*. Masters Thesis, Central Queensland University, 2010.
- [30] Jingjing Yao, Qiang Hou, *Malicious executables classification based on behavioral factor analysis*. International Conference on e-Education, e-Business, e-Management and e-Learning, 2010.
- [31] Yuhei Kawakoya, Makoto Iwamura, Mitsutaka Itoh, *Analyzing Malware with Stealth Debugger*. Information Processing Society of Japan, 2008.
- [32] Kevin Coogan, Saumya Debray, Tasneem Kaochar, Gregg Townsend, *Automatic Static Unpacking of Malware Binaries*. 16th Working Conference on Reverse Engineering, 2009.
- [33] Tony Abou-Assaleh, Nick Cercone, Vlado Ke?selj, Ray Sweidan , *N-gram-based Detection of New Malicious Code*. Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004 .
- [34] Tony Lee, Jigar J. Mody, *Behavioral Classification*. Presented at the EICAR Conference, May 2006.
- [35] Yanfang Ye, Dingding Wang, Tao Li, Dongyi Ye, *IMDS: Intelligent Malware Detection System*. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.
- [36] Robin Sharp, *An Introduction to Malware*. Spring 2009.
- [37] Georg Wicherski, *peHash: A Novel Approach to Fast Malware Clustering*. December 7, 2008.
- [38] goppit, *PORTABLE EXECUTABLE FILE FORMAT*. 2011.