Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"Design and Development of Forensic Analysis Toolkit for Analyzing Malware Binary",** in partial fulfillment of the requirements for the award of degree of Master of Engineering in Software Engineering at Computer Science and Engineering Department of Thapar Institute of Engineering and Technology (Deemed University), Patiala, is an authentic record of my own work carried out under the supervision of Mr. Maninder Singh.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other University.

Bhushan Kumar Jindal

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Mr. Maninder Singh)
Assistant Professor
Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology,
Patiala- 147004.

Countersigned by

Dr.(Mrs.) Seema Bawa
Professor & Head
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
PATIALA- 147004

Dr. T. P. Singh
Dean of Academic Affairs
Thapar Institute of Engineering and Technology
PATIALA- 147004

Acknowledgement

I wish to express my deep gratitude to Mr. Maninder Singh, Assistant Professor, Computer Science and Engineering Department for providing his uncanny guidance throughout the thesis work.

I am also thankful to Dr. (Mrs.) Seema Bawa, Professor & Head, Computer Science and Engineering Department, for the motivation and inspiration she gave during the thesis work.

I would also like to thank all the staff members and my co-students who were always there at the need of the hour and provided with all the help and facilities, which I required for the completion of the thesis.

At last but not the least I would like to thank God for not letting me down at the time of crisis and showing me the silver lining in the dark clouds.

Bhushan Kumar Jindal

8043103

Abstract

Forensic analysis is a very old field of crime investigation. Lots of work has been done into development technology for this field. Earlier the chemical sciences were used to perform various kinds of analyses in case of some crime investigation. But as the way of committing crimes has changed so there exists a great need to investigate such matters with same modern technology.

Cyber crimes have now become latest kind of crimes where computer may be involved directly or indirectly. One way the computer would have been used indirectly i.e. for sending the mail etc for communication to another partner. In another way the direct involvement means attacking computer/network itself for gathering some sort of information, corrupting the computer data, causing the target network to go down or various other sort of attacks. The very common sort of attacks that are occurring in today's world are caused by malwares spreading day by day in large computer networks.

In this thesis the intent is to explain details about computer forensics. The malware binaries will be analyzed and primary aim will be to find that what sort of information can be obtained from the malware binary that would have been found on some infected system. Secondary aim would be to develop a tool that would provide some automated analysis results of forensic analysis of the malware binary. These tools may further be extended to more refined toolkit to add more analysis steps for better and elaborated results.

Table of Contents

List of Figures

List Of Tables

CHAPTER 1
Introduction

1.1 Relevance and Importance of Networks in today's life

INTERNET it is rapidly changing its presence from a medium for elites to one in common use in our everyday lives. A decade ago, the first age of the Internet was a bright light shining above everyday concerns. It was a technological marvel bringing a new enlightenment to transform the world, just as the printing press fostered the original enlightenment a half millennium ago in Renaissance times. With the development of the Internet, and with the increasing pervasiveness of communication between networked computers, we are in the middle of the most transforming technological event since the capture of fire.

We are moving from a world of, Internet wizards to a world of ordinary people, routinely using the Internet as an embedded part of their lives. It has become clear that the Internet is a very important thing. In fact, it is being used more – by more people, in more countries, in more different ways.

The network is evolving into the backbone and, in many instances, the central nervous system of everyday life. No longer just a clunky, wire-based delivery system for various passive computer applications, the network has evolved into an active, often proactive, information and communication system that touches and interacts with almost every person and department everywhere--and beyond.

According to the statistics given in a website [5,17], 1022,863,307 people are using the Internet as according to the latest figures gathered in March 2006. This figure is 183.4% more than the figure for same data in 2000.

INTERNET USAGE STATISTICS - The Big Picture

World Internet Users and Population Stats

WORLD INTERNET USAGE AND POPULATION STATISTICS						
World Regions	Population (2006 Est.)	Population % of World	Internet Usage, Latest Data	% Population (Penetration)	Usage % of World	Usage Growth 2000-2005
Africa	915,210,928	14.1 %	**23,649,000**	2.6 %	2.3 %	423.9 %
Asia	3,667,774,066	56.4 %	**364,270,713**	9.9 %	35.6 %	218.7 %
Europe	807,289,020	12.4 %	**291,600,898**	36.1 %	28.5 %	177.5 %
Middle East	190,084,161	2.9 %	**18,203,500**	9.6 %	1.8 %	454.2 %
North America	331,473,276	5.1 %	**227,303,680**	68.6 %	22.2 %	110.3 %
Latin America/Caribbean	553,908,632	8.5 %	**79,962,809**	14.4 %	7.8 %	342.5 %
Oceania / Australia	33,956,977	0.5 %	**17,872,707**	52.6 %	1.7 %	134.6 %
WORLD TOTAL	6,499,697,060	100.0 %	**1,022,863,307**	15.7 %	100.0 %	183.4 %

NOTES: (1) Internet Usage and World Population Statistics were updated for March 31, 2006. (2) CLICK on each world region for detailed regional information. (3) Demographic (Population) numbers are based on data contained in the world-gazetteer website. (4) Internet usage information comes from data published by Nielsen//NetRatings, by the International Telecommunications Union, by local NICs, and other other reliable sources. (5) For definitions, disclaimer, and navigation help, see the Site Surfing Guide. (6) Information from this site may be cited, giving due credit and establishing an active link back to www.internetworldstats.com.

Table 1.1 Number of people having Internet connections

TOP 20 COUNTRIES WITH HIGHEST NUMBER OF INTERNET USERS						
#	Country or Region	Internet Users, Latest Data	Population (2006 Est.)	Internet Penetration	Source and Date of Latest Data	% Users of World
1	United States	**205,326,680**	299,093,237	68.6 %	Nielsen//NR Jan/06	20.1 %
2	China	**111,000,000**	1,306,724,067	8.5 %	CNNIC Dec/05	10.9 %
3	Japan	**86,300,000**	128,389,000	67.2 %	eTForecasts Dec/05	8.4 %
4	India	**50,600,000**	1,112,225,812	4.5 %	C.I.Almanac Mar/05	5.0 %
5	Germany	**48,721,997**	82,515,988	59.0 %	Nielsen//NR Jan/06	4.8 %
6	United Kingdom	**37,800,000**	60,139,274	62.9 %	ITU Oct/05	3.7 %
7	Korea (South)	**33,900,000**	50,633,265	67.0 %	eTForecast Dec/05	3.3 %
8	Italy	**28,870,000**	59,115,261	48.8 %	ITU Sept./05	2.8 %
9	France	**26,214,173**	61,004,840	43.0 %	Nielsen//NR Jan/06	2.6 %
10	Brazil	**25,900,000**	184,284,898	14.1 %	eTForcasts Dec/05	2.5 %
11	Russia	**23,700,000**	143,682,757	16.5 %	eTForcasts Dec/05	2.3 %
12	Canada	**21,900,000**	32,251,238	67.9 %	eTForcasts Dec/05	2.2 %
13	Indonesia	**18,000,000**	221,900,701	8.1 %	eTForcasts Dec/05	1.8 %
14	Spain	**17,142,198**	44,351,186	38.7 %	Nielsen//NR Jan/06	1.7 %
15	Mexico	**16,995,400**	105,149,952	16.2 %	AMIPCI Nov/05	1.7 %
16	Australia	**14,189,557**	20,750,052	68.4 %	Nielsen//NR Jan/06	1.4 %
17	Taiwan	**13,800,000**	22,896,488	60.3 %	C.I.Almanac Mar/05	1.4 %
18	Netherlands	**10,806,328**	16,386,216	65.9 %	Nielsen//NR June/04	1.1 %
19	Poland	**10,600,000**	38,115,814	27.8 %	C.I.Almanac Mar./05	1.0 %
20	Turkey	**10,220,000**	74,709,412	13.7 %	ITU Sept./05	1.0 %
TOP 20 Countries		**811,986,333**	4,064,319,458	20.0 %	IWS - Mar.31/06	**79.4 %**
Rest of the World		**210,876,974**	2,435,377,602	8.7 %	IWS - Mar.31/06	**20.6 %**
Total World - Users		**1,022,863,307**	6,499,697,060	15.7 %	IWS - Dec.31/05	**100.0 %**

NOTES: (1) World Internet User Statistics were updated as of March 31, 2006. (2) Data for users in individual countries and regions may be found by clicking each country name. (3) Population numbers are based on data contained in the world-gazetteer page. (4) The most recent user information comes from data published by Nielsen//NetRatings, ITU, and other trustworthy research sources. (6) Data from this site may be cited, giving due credit and establishing an active link back to InternetWorldStats.com. (6) For definitions and navigation help, see the Site Surfing Guide.

Table 1.2 Top 20 countries in Internet usage

The tables above give us an idea how fast the Internet is growing. The number of people using the Internet is doubling every year [2,3]. More sites pop up every day. Global communication is getting more important. More and more companies connect their computer networks to the global Internet or connect multiple intranets over the Internet with the help of virtual private networks. E-Commerce is getting an important source of revenue for many companies. At the same time, computer crimes are increasing [14]. Here are some of the trends discovered by the survey.

- 85% of respondents [10] (primarily large companies and government agencies) detected computer security breaches within the last twelve months.

- 64% acknowledged financial loss due to computer breaches. 35% were willing and/ or able to quantify their financial losses. The respondents reported $377'828'200 in financial loss (in 2000 it was $265'589'940). The average total over the last three years to 2000 was $120'240'180. We are confronted with a huge increase. As in previous years, the most serious financial losses occurred through theft of proprietary information.

- For the fourth year in a row, more respondents (70%) cited their Internet connections as a frequent point of attack than one cited their internal systems as a frequent point of attack (31%). The rise in those citing their Internet connection as a frequent point of attack rose from 59% in 2000 to 70% in 2001.

1.2 Why We Need Network Security

While using the Internet, along with the convenience and speed of access to information new risks arise. Among them are the risks that valuable information will be lost, stolen, corrupted, or misused and that the computer systems will be corrupted [6]. If information is recorded electronically and is available on networked computers, it is more vulnerable than if the same information is printed on paper and locked in a file cabinet. Intruders do not

need to enter an office or home, and may not even be in the same country. They can steal or tamper with information without touching a piece of paper or a photocopier. They can create new electronic files, run their own programs, and even hide all evidence of their unauthorized activity.

It is remarkably easy to gain unauthorized access to information in an insecure networked environment, and it is hard to catch the intruders. Even if users have nothing stored on their computer that they consider important, that computer can be a "weak link", allowing unauthorized access to the organization's systems and information. Seemingly innocuous information can expose a computer system to compromise. Information that intruders find useful includes which hardware and software are being used, system configuration, type of network connections, phone numbers, and access and authentication procedures. Security-related information can enable unauthorized individuals to get access to important files and programs, thus compromising the security of the whole system. Today's commercial off-the-shelf [software] technology is riddled with holes. The sheer number of vulnerabilities is overwhelming organizations. These include vulnerabilities that allowed viruses and worms (hereafter referred to as malware) and other manual and automated attacks to inflict damages costing hundreds of millions of dollars per occurrence. Specifically: LoveLetter, a worm that severely clogged mail servers and networks in 2000; Code Red, an aggressive worm that attacked unpatched Microsoft web servers and defaced their main pages; and most recently, Nimda, a worm that spread by several different methods including email and web protocols, and searched for as many as 16 separate vulnerabilities to attack. The recent Distributed Denial of Service (DDOS) attacks are less serious but still expensive. Virus attacks, exploits directed at unpatched popular firewalls (e.g., Check Point, Cisco Pix), buffer overflows, directory traversal, other more obscure attacks against web servers, and the scope of the problem starts to become quite clear [12].

Those affected include banks and financial companies, insurance companies, brokerage houses, consultants, government contractors, government agencies, hospitals, medical laboratories, network service providers, utility companies, the textile industry, universities, wholesale and retail trades. The consequences of a break-in cover a broad range of possibilities: a minor loss of time in recovering from the problem, a decrease in

productivity, a significant loss of money or staff-hours, a devastating loss of credibility or market opportunity, a business no longer able to compete, legal liability, and the loss of life.

1.3 Various Threats To Network Security

1.3.1 Viruses

A computer virus is a special kind of computer program which:

Spreads across disks and networks by making copies of itself, usually surreptitiously. And / or can produce undesired side effects in computers in which it is active.

- There are various types of viruses:
    - Boot viruses place (some of) their code in the disk sector whose code the machine will automatically execute when booting. Thus, when an infected machine boots, the virus loads and runs. After boot viruses are finished loading, they usually load the original boot code, which they have previously moved to another location, or take other measures to ensure the machine appears to boot normally.
    - File viruses attach to 'program files' (files containing executable or interpretable code) in such a way that when you run the infected program, the virus code executes. Usually the virus code is added in such a way that it executes first, although this is not strictly necessary. After the virus code has finished loading and executing, it will normally load and execute the original program it has infected, or call the function it intercepted, so as to not arouse the user's suspicion.
    - Macro viruses are really just a type of file virus, but a particularly 'successful' type. They copy their macros to templates and/or other application document files. Although 'auto macros' were almost exclusively used by early macro

viruses (often to ensure the virus' code is the first to execute when infected templates or documents were opened), several other mechanisms are also available – in fact, some of these, such as taking over standard internal functions of the host application (say the 'File Save' command) and installing default event handlers are probably more commonly used these days.

- Script viruses also became quite successful around the beginning of this century. This was mainly due to the increase in machines running Windows Scripting Host, which was first installed by default in Windows 98 and 2000 and with Internet Explorer 5.0 and later versions. Representing new types of 'program file', but with icons more like that of 'safe' text files, standalone Visual Basic Script (VBS) and JavaScript (JS) programs became a popular target of the writers of mass mailing viruses.

- Companion viruses take advantage of features of the operating system to be executed, rather than directly infecting programs or boot sectors. Under DOS and Windows, when you execute the command 'ABC', the rule is that ABC.COM executes before ABC.EXE (in the rare cases where both files exist). Thus, a companion virus could place its code in a COM file with its first name matching that of an existing EXE file. When the user next executed the 'ABC' command, the virus' ABC.COM program would be run (usually the virus would launch ABC.EXE once its code was finished so as not to arouse suspicion). This is known as the 'execution preference companion' method, but several other forms of companion infection are also possible.

- Worms are described by some antivirus researchers as similar to viruses in that they make copies of themselves, but different in that they need not attach to particular files or sectors at all. Once such a worm is executed, it seeks other systems – rather than parts of systems – to infect, then copies its code to them in such a way as to have the code execute directly from memory. More recently the term 'worm' has been taken to mean 'a virus that replicates across a

network link', with the most common usage applied to viruses that send many copies of themselves out attached to the infected user's e-mail.

Some viruses display obvious symptoms, and some cause damage to files in a system they have infected. While one or both of these features of a virus often capture the attention of the popular media, note from the preceding discussion that neither are essential in the definition of a virus. A non-damaging virus is still a virus, not a prank and, other things being equal, viruses without obvious symptoms are more likely to spread further and persist longer than those that rapidly draw attention to themselves.

1.3.2 Trojan horse programs

Just as the mythological Trojan horse appeared to be a gift, but turned out to contain Greek soldiers who overtook the city of Troy, today's Trojan horses are computer programs that appear to be useful software, but instead they compromise your security and cause a lot of damage. A recent Trojan horse came in the form of an e-mail that included attachments claiming to be Microsoft security updates, but turned out to be viruses that attempted to disable antivirus and firewall software. Trojan horses spread when people are lured into opening a program because they think it comes from a legitimate source. Trojan horses can also be included in software that you download for free. Never download software from a source that you don't trust.

There are six ways Trojans generally capture passwords [9]

1) INTERRUPT 9h. This is a hardware interrupt that occurs every time you hit a key. The Trojan waits for an INT 9h, and then interrogates the keyboard to discover which key was pressed, saves the key, and then returns normal control.

2) INTERRUPT 16h. This is an operating system call that deals with the keyboard. Every time an application program needs a key, it calls this software interrupt. A Trojan piggybacks onto the INT 16h and saves the keys every time the program asks for them.

3) INTERRUPT 21h. Similar to INT 16h, much more widely known, but less elegant. Once again, a Trojan will piggyback INT 21h.

4) INTERRUPT 15h. Similar to INT 16h, but called from INT 9h. A Trojan will again piggyback to this BIOS function in order to capture passwords.

5) Hardware polling. This method simply polls the keyboard many times a second, and records any new keys you press as it notices you pressing them. If you type too fast, it will miss some.

6) RAM-buffer snooping. A Trojan can copy the DOS keyboard type-ahead buffer from the system area (also called segment 40:) or from the executable-programs memory regularly to disk.

Once the password is captured, the Trojan will usually either save it in RAM, or on a local or network disk (in a hidden file in a hidden directory), or even in CMOS. A purpose-written Trojan can be programmed to download passwords via modem or onto the Internet, or broadcast them to some other place for remote retrieval.

Some Trojans activate only when a password is being entered. They do this by monitoring system calls, such as the "execute program" call when LOGIN.EXE is being run. Others "read" the screen and record when they see some phrase, such as "password" appear. Many others simply record everything.

A Trojan is one of the most dangerous tools available to network intruders. The most obvious reason for the danger is that a Trojan provides the intruder with a guaranteed method of access to your network, under the guise of a legitimate user. Because the Trojan provides the intruder with a legitimate, valid username and password to access your network with, traditional methods of detecting and preventing this intrusion are useless. An intruder with a Trojan will be logging straight into your network without setting off any intrusion alarms from guessed password attempts or suspicious log-in activity.

The Trojan provides the means for the intruder to identify how to log in, and when and where it is safest to do so without being detected. Since the network has no way at all to distinguish between the intruder and the legitimate user, this access will be undetectable as well as undetected. Many network administrators believe that because they employ anti-

virus software, they are not at risk. This couldn't be further than the truth. In fact, this misconception actually amplifies the power of a Trojan to do damage. Most Trojans are not viral, and hence can never be detected by anti-virus software. Further — since a Trojan can even be a legitimate program, there is no way that anti-virus software can address this threat in the future, should they even attempt to, since the "signatures" would be copyright.

So the false piece of mind generated by a "clean" scan of a disk, system, or network is only serving to foster ignorance of the threat. Unlike viruses, no-one has a library of Trojans, and so most are unknown in nature. Further, since a Trojan is far smaller and simpler than even a basic virus, it much easier to write. Worse still, a Trojan is, by design and nature, a silent and secret threat. It does not introduce itself to its victim, and so goes unnoticed in most cases. Finally, as mentioned above, even the operation of a Trojan is generally unnoticed, reducing the chances of the Trojan ever being discovered to minuscule proportions.

1.3.3 Attacks

Including **reconnaissance attacks** (information-gathering activities to collect data that is later used to compromise networks); **access attacks** (which exploit network vulnerabilities in order to gain entry to e-mail, databases, or the corporate network); and **denial-of-service attacks** (which prevent access to part or all of a computer system).

1.3.3.1 Reconnaissance Attack

Reconnaissance is the unauthorized discovery and mapping of systems, services, or vulnerabilities. It is also called information gathering. In most cases, it precedes an actual access or DoS attack. The malicious intruder typically ping-sweeps the target network first to determine what IP addresses are alive. After this is accomplished, the intruder determines what services or ports are active on the live IP addresses. From this information, the intruder queries the ports to determine the application type and version as well as the type and version of the operating system running on the target host.

Reconnaissance is somewhat analogous to a thief scoping out a neighborhood for vulnerable homes he can break into, such as an unoccupied residence, an easy-to-open door

or window, and so on. In many cases, an intruder goes as far as "rattling the door handle" not to go in immediately if it is open, but to discover vulnerable services he can exploit later when there is less likelihood that anyone is looking.

1.3.3.2 Access Attack

Access is an all-encompassing term that refers to unauthorized data manipulation, system access, or privilege escalation. Unauthorized data retrieval is simply reading, writing, copying, or moving files that are not intended to be accessible to the intruder. Sometimes this is as easy as finding shared folders in Windows 9*x* or NT, or NFS exported directories in UNIX systems with read or read-write access to everyone. The intruder has no problem getting to the files. More often than not, the easily accessible information is highly confidential and completely unprotected from prying eyes, especially if the attacker is already an internal user.

System access is an intruder's ability to gain access to a machine that he is not allowed access to (such as when the intruder does not have an account or password). Entering or accessing systems that you don't have access to usually involves running a hack, script, or tool that exploits a known vulnerability of the system or application being attacked.

Another form of access attacks involves privilege escalation. This is done by legitimate users who have a lower level of access privileges or intruders who have gained lower-privileged access. The intent is to get information or execute procedures that are unauthorized at the user's current level of access. In many cases this involves gaining root access in a UNIX system to install a sniffer to record network traffic, such as usernames and passwords that can be used to access another target.

1.3.3.3 Denial of Service Attack

DoS (Denial-of-Service) attacks are probably the nastiest, and most difficult to address. These are the nastiest, because they're very easy to launch, difficult (sometimes impossible) to track, and it isn't easy to refuse the requests of the attacker, without also refusing legitimate requests for service.

The premise of a DoS attack is simple: send more requests to the machine than it can handle. There are toolkits available in the underground community that make this a simple matter of running a program and telling it which host to blast with requests. The attacker's program simply makes a connection on some service port, perhaps forging the packet's header information that says where the packet came from, and then dropping the connection. If the host is able to answer 20 requests per second, and the attacker is sending 50 per second, obviously the host will be unable to service all of the attacker's requests, much less any legitimate requests (hits on the web site running there, for example).

Obviously forged packets would include those that claim to come from your own hosts, addresses reserved for private networks as defined in RFC 1918 and the *loopback* network (127.0.0.0) [7].

1.3.4 Data interception

Data transmitted via any type of network can be subject to interception by unauthorized parties. The intercepting perpetrators might eavesdrop on communications or even alter the data packets being transmitted. Perpetrators can use various methods to intercept data. IP spoofing, for example, entails posing as an authorized party in the data transmission by using the Internet Protocol (IP) address of one of the data recipients.

One of the most widespread methods of Data interception is “Man in the Middle Attack (MITM)”. In cryptography, a *man in the middle attack (MITM)* is an attack in which an attacker is able to read, insert and modify at will, messages between two parties without either party knowing that the link between them has been compromised. The attacker must be able to observe and intercept messages going between the two victims.

Suppose ‘A’ wishes to communicate with ‘B’, and that ‘C’ wishes to eavesdrop on the conversation, or possibly deliver a false message to ‘B’. To get started, ‘A’ must ask ‘B’ for his public key. If ‘B’ sends his public key to ‘A’, but ‘C’ is able to intercept it, a man in the middle attack can begin. ‘C’ can simply send ‘A’ a public key for which he has the private, matching, key. ‘A’, believing this public key to be ‘B’'s, then encrypts her message with ‘C’'s key and sends the enciphered message back to ‘B’. ‘C’ again intercepts,

deciphers the message, keeps a copy, and re-enciphers it (after alteration if desired) using the public key 'B' originally sent to 'A'. When 'B' receives the newly enciphered message, he will believe it came from 'A'. A similar attack is possible, in principle, against any message sent using public key technology, including data packets carried on computer networks.

1.3.5 Social engineering

Social Engineering is an attack method used by many attackers that takes advantage of trust and complacency at work. Humans by nature are very trusting and rarely question actions that are considered normal. Another forum that Social Engineering can expose is the Computer Conference. Computer conferences are great for obtaining information. Most conferences stress openness, this within itself is not a bad idea but the problem occurs when people give too many details. Some of the information that attendee's and instructors give out could be used against them and their network(s). Information about network configuration, types of firewalls and Intrusion Detection systems were just a few items commonly shared.

Dumpster diving, also known as trashing, is another popular method of social engineering. A huge amount of information can be collected through company dumpsters. The following items as potential security leaks in our trash: "company phone books, organizational charts, memos, company policy manuals, calendars of meetings, events and vacations, system manuals, printouts of sensitive data or login names and passwords, printouts of source code, disks and tapes, company letterhead and memo forms, and outdated hardware."

Other major types of Social Engineering attacks used are *Social Engineering by Phone*, *Persuasion* etc.

1.3.6 Vandals

Web sites have come alive through the development of such software applications as ActiveX and Java Applets. These applications enable animation and other special effects to run, making web sites more attractive and interactive. However, the ease with which these

applications can be downloaded and run has provided a new vehicle for inflicting damage. Vandals can take on the form of a software application or applet that causes destruction of various degrees. A vandal can destroy a single file or a major portion of a computer system.

1.3.7 Malware

Malware is software designed to infiltrate or damage a computer system, without the owner's consent. The term is a portmanteau of "mal-" (or perhaps "malicious") and "software", and describes the intent of the creator, rather than any particular features. Malware is commonly taken to include computer viruses, worms, Trojan horses, spyware and adware [6]. In law, malware is sometimes known as a computer contaminant, for instance in the legal codes of California, West Virginia, and several other U.S. states [13].

Over the years, people have written malicious software for a number of different purposes.

Many early infectious programs, including the Internet Worm and a number of MS-DOS viruses, were written as experiments or pranks -- generally intended to be harmless or merely annoying, rather than to cause serious damage. Young programmers, learning about the possibility of viruses and the techniques used to write them, might write one just to prove that they can do it, or to see how far it could spread.

A slightly more hostile intent can be found in programs designed to vandalize or cause data loss. Many DOS viruses were designed to destroy files on a hard disk, or to corrupt the filesystem by writing junk data. Network-borne worms such as the Code Red worm or Ramen worm fall into the same category. Designed to vandalize Web pages, these worms may seem like an online equivalent of graffiti tagging, with the author's name or affinity group appearing everywhere the worm goes.

Revenge is sometimes a motive to write malicious software. A programmer or system administrator about to be fired from a job may leave behind backdoors or software "time bombs" that will allow them to damage the former employer's systems or destroy their own earlier work.

However, since the rise of widespread broadband Internet access, a greater portion of malicious software has been focused strictly on a profit motive. For instance, since 2003, the majority of widespread viruses and worms have been designed to take control of users' computers for black-market exploitation. Infected "zombie computers" are used to send email spam, to host contraband data such as child pornography, or to engage in distributed denial-of-service attacks as a form of extortion.

1.3.8 Spyware

Another strictly for-profit category of malware has emerged in spyware -- the term spyware refers to a broad category of malicious software designed to intercept or take partial control of a computer's operation without the informed consent of that machine's owner or legitimate user. While the term taken literally suggests software that surreptitiously monitors the user, it has come to refer more broadly to software that subverts the computer's operation for the benefit of a third party.

Spyware differs from viruses and worms in that it does not usually self-replicate. Like many recent viruses, however, spyware – by design – exploits infected computers for commercial gain. Typical tactics furthering this goal include delivery of unsolicited pop-up advertisements; theft of personal information (including financial information such as credit card numbers); monitoring of Web-browsing activity for marketing purposes; or routing of HTTP requests to advertising sites [15].

The following section discusses the need for forensic analysis of the malware.

1.4 Forensic Analysis of The Malware

The networks are growing and attacks are increasing, at the same time countermeasures are getting more sophisticated and widely used - firewalls pop up in nearly every company and people at home install so called “personal firewalls”. Network intrusion detection systems start their triumphal procession. And usually through these systems we get some sort of malware. Some times the anti-virus installed on the machines help to retrieve the malware that is trying to affect the system. This malware may be a virus, a trojan horse or a worm. This malware is then used for forensic analysis for getting clearer facts.

As in the military, it is important to know, who your enemy is, what kind of strategy he uses, what tools he has and what he is aiming for. Gathering this kind of information is not easy but important. By knowing attack strategies, countermeasures can be taken and vulnerabilities can be fixed [11]. To gather as much information as possible is one main goal of a forensic analysis of malware binary.

Therefore such forensic analysis of the retrieved binary may help us in getting various kind of facts in our hand. These facts can be like, who created this malware, what does this malware do, where did it come from, how it reached our network and in what ways it can be harmful to the network.

Once such information is gathered the same can be used to take the countermeasures, if it is a worm then immediately the infected system can be removed from the network. The source/creator of the malware can be tracked and caught. The malware might have been exploiting some vulnerability hence such vulnerability can be patched.

Forensic analysis of such binaries can be done by tools (computer software) specially made for such analysis. These tools can be categorized into various categories depending their usage and level at which they work. According to the level at which they work, they can provide information contained in the malware itself or from its behaviour.

CHAPTER 2
A Brief Review of Forensic Analysis

2.1 Introduction

Forensic analysis is not a new topic or technology. It has been used since many years for crime detection. In this thesis we are going to discuss about the computer forensics. But is should not be mixed with the traditional forensic analysis because there are already various traditional ways of forensic analysis other than computer forensics. The following section will discuss various kinds of forensic analyses that are already being carried out in its traditional way.

2.2 Forensic Analysis

Forensic Science is any aspect of science as it relates to the law. Typically, it includes the disciplines listed below, but just about any area of science could be called into question in a court of law [4].

2.2.1 Firearms and Toolmark Identification

Firearms and toolmark identification involves more than just guns. Also included in this broad subject area are explosives, imprint evidence and toolmark evidence.

Most physical evidence concerns itself with class characteristics and individual characteristics. Class characteristics are those characteristics that are common to a group of similar objects. For example, you buy a pair of Air Jordan sneakers. All Air Jordan sneakers have to same shape and same tread design on the bottom. These are class characteristics. Individual characteristics are those characteristics that are unique to a given object and set it apart from similar objects. You wear your Air Jordans around for a while and they get worn. The treads wear down. They get little pits and gouges in them. These little pits and gouges are individual to your shoes and no others since no one has walked over the exact same surfaces in the exact same way in their Air Jordans. These two concepts, class and individual characteristics are the most important in firearms examination.

A typical firearms examination concerns matching a bullet back to the gun that fired it, to the exclusion of all others [4].

2.2.2 Forensic Psychiatry and Profiling

2.2.2.1 Forensic Psychiatry

Forensic Psychiatry serves to define what mental illness and disorder are, what creates mental illness and disorder, how they are diagnosed and how they are treated.

In cases of mental illness, the individual does not function well in reality because of emotional fluctuations or distorted point of view interpretation.

Mental disorder is a clinically significant behavior or psychological syndrome or pattern that is associated with present distress or disability or with a significantly increased risk of suffering death, pain or disability or important loss of freedom.

Diagnosis is a medical classification that allows for communication between professionals, allows for prediction of illness course and probable outcome, and indicates the best course of treatment.

The issue of incompetency is fitness to stand trial. No one can go to trial unless they can understand what is going on and can help in their own defense. This issue can be raised at any time during the criminal process. The three bases for incompetency are emotional/psychological factors, cognitive or intellectual factors, and physical factors. An individual must undergo a competency evaluation if this claim is made. It cannot be refused.

The issue of insanity applies to the mental state of the defendant at the time at which the crime was committed. The defendant must concede guilt when the decision to enter the insanity plea is made. He/She is saying, "Yes, I committed this act, but I was so mentally impaired at the time as to not know right from wrong or what I was doing." The defendant must be competent to stand trial to enter this plea. And, finally, the defendant must prove profound defect of mental ability. Also, the illness must be directly related to the crime. You can't plead insanity if you mental defect in no way influenced the criminal action. A

compulsive hand washer cannot use that illness as an excuse for shooting twelve people from a clock tower.

2.2.2.2 Profiling

Profiling is a label given to the process by which a trained forensic psychologist sifts through the aspects of a crime scene to develop a description of the personality of the perpetrator. This personality description can include age, sex, occupation, behavioral disorders, upbringing, marital status, the type of place the perp would live in and its general condition, the type of person the perp might live with, what type of car he drives, if he has a speech impediment or acne or some other type of disability or difficulty in relating to others. They will tell you how the crime was committed. It is mind-boggling the amount of information that can be gleaned about the perpetrator of a serial crime just by looking at how that crime was committed. As John Douglas says in Mindhunter, "Behavior reflects personality." And that is what profiling is all about.

Profiling works because John Douglas and members of his team in the FBI's Investigative Support Unit have spent countless hours interviewing hundreds of convicted serial killers, serial rapists and mass murders, learning about their crimes, motives, methods and personalities. This is the database on which they draw in making conclusions from what they view at a crime scene. This stuff isn't just made-up or speculation. Nor is it an exact science. But it draws on thousands of man-hours of research and interviews and has proven to be extremely accurate.

Profiling does not produce a name. What it does produce is a detailed personality profile of a perpetrator that investigators can use to focus an investigation and pare down the list of suspects. It can also provide them with strategies on how to approach the subject during interrogation and how to break him down on the witness stand at trial [4].

2.2.3 Examination of Questioned Documents

Questioned document examination involves a great many areas of expertise. Included under questioned document examination are the following disciplines, a few of which will be hit on in this section: handwriting, typewriting, photocopying and computer printers,

forgery, paper and inks, writing instruments, computer disks, gambling machinery, stamps (as in the rubber pad kind) and the dating of documents [4].

2.2.3.1 Handwriting Analysis

Graphology, the study of handwriting to determine one's personality traits, is not handwriting analysis. It's not even considered a science; more like a parlor trick. True handwriting analysis involves painstaking examination of the design, shape and structure of handwriting to determine authorship of a given handwriting sample. The basic principle underlying handwriting analysis is that no two people write the exact same thing the exact same way. Every person develops unique peculiarities and characteristics in their handwriting.

Handwriting analysis looks at letter formations, connecting strokes between the letters, upstrokes, retraces, down strokes, spacing, baseline, curves, size, distortions, hesitations and a number of other characteristics of handwriting. By examining these details and variations in a questioned sample and comparing them to a sample of known authorship, a determination can be made as the whether or not the authorship is genuine. Figure 2.1 shows the examples of handwriting points of analysis.

The methods of forensic science vary in complexity. They may be as simple as using a magnifying glass, or as elaborate as the latest scientific instruments.

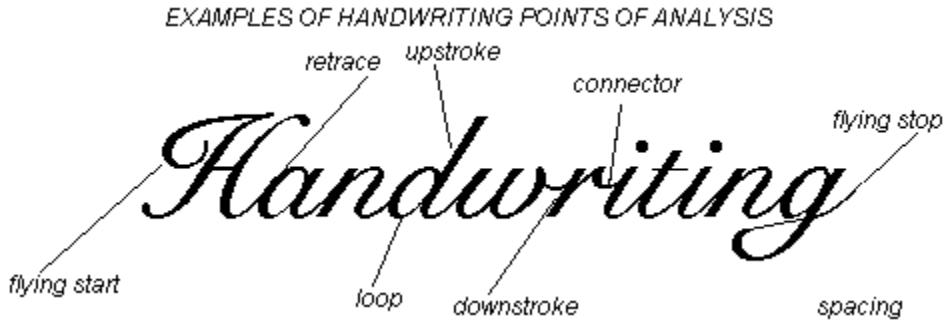

Figure 2.1: Handwriting Analysis

2.2.3.2 Typewriting

First, a review of individual characteristics. This topic is first discussed in the firearms page. If you haven't been there, you might want to go back. Individual characteristics are either inherent in the machining process of a manufactured item, or come about through the wear and tear in the use of an object. It is these individual characteristics that allow for the identification of an object to the exclusion of all others of its general type. As with typewriters, all typewriters of a particular make and model are pretty much the same but, through use, the develop defects that translate to paper when the machine is used. These defects on the typed page can be matched back to the typewriter that was used to create it.

These defects in the typeface are revealed in a number of ways. If the type bar is bent (the bar on which the letter element is attached and hammered down to the page) the letter is misaligned or 'off its feet.' Misalignments can also cause non-printing areas of a specific letter, such as losing the loop on the bottom of a 'g'. The letter can be displaced horizontally or vertically. Little clumps of plastic can adhere to the type key during manufacture and are made permanent by the coating process. This defect is called 'flashing.' As wear and tear increases, the defects become more exaggerated.

Just looking at the type style, or font, the spacing (horizontal and vertical) and type size allows for determining the make and model of the typewriter. Figure 2.2 shows the examples of typewriting points of analysis.

Ribbons are a major evidentiary component. It is possible to read a ribbon to see what it has been used to type.

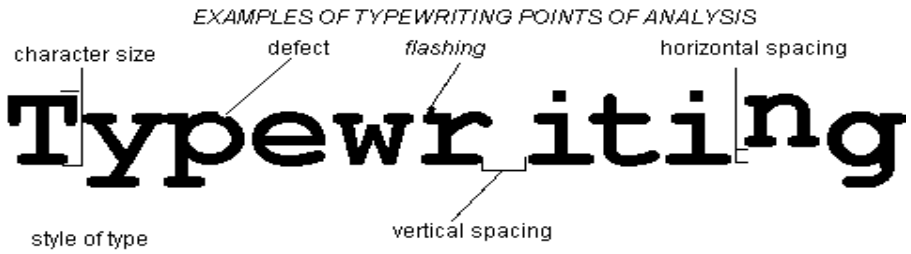

Figure 2.2: Typewriting Analysis

2.2.3.3 Photocopiers and Laser Printers

Photocopiers and laser printers use the same type of process to print a page. With a photocopier, the original document is placed on the glass platen. The document is then exposed by use of reflected light to a drum that is covered with a photosensitive material. The image of that document exists on the drum as an invisible positive photoelectric charge. Negatively charged toner, the messy black powdery stuff, is drizzled onto the drum, where it sticks to only the positively charged areas (remember magnets? Opposite charges attract, same charges repel), to create a visible image. Paper, with a positive charge, passes the drum, causing the negatively charged toner to transfer to the paper. The toner is then heat sealed to the paper, creating the printed copy. With a laser printer, the image of the original document (held by the computer in its memory) is written to the photosensitive drum by use of a laser.

The paper itself can yield many clues. Look for marks from the belts, pinchers, rollers and gears that physically move the paper through a machine. These examinations would be similar to toolmark examinations, discussed in the Firearms page.

Toner can have unique characteristics in its chemical composition. Also, look at how the tone was placed on and fused to the paper. Toner may clump up on the drum, transferring blobs of toner at a time to the printed page.

Marks on the optics (glass platen, lenses, mirrors) used to transfer or create an image on paper might contain unique defects (such as scratches) that will render anomalous markings on the printed page.

2.2.4 Personal Identification

2.2.4.1 Fingerprints

There are three basic fingerprint patterns: Loops, Arches and Whorls. Everyone falls into one of these three patterns (diagram). Within these patterns are what we call minutia points. There are about thirty different types of minutiae points, and no two people have the same types of minute in the same number in the same places on their fingertips. This is why our fingerprints are totally unique.

Your fingerprint patterns are hereditary. They are formed before you are born, while you are still in the womb, they never change through out your lifetime, and they are even around for a while after you die. So, why are fingerprints so good for identification purposes? They are totally unique, and they never change.

Your fingerprints are formed underneath your skin in a layer called dermal papillae. As long as that layer of papillae is there, your fingerprints will always come back, even after scarring or burning.

Gloves don't necessarily help you from leaving fingerprints. Surgical gloves were made to keep surgeons from infecting their patients. You can actually leave prints through surgical gloves. Surgical gloves were made to keep sterile conditions during operations. They have to fit like a second skin for surgeons to be able to pick up their instruments. They fit so tightly that fingerprints 'pass through' the latex membrane. They can also be turned inside out to yield fingerprints from the inside surfaces. Leather gloves can be treated in the same manner. Also, leather gloves can leave a print that is unique to that glove and no other (leather comes from cow skin, which is just as random as human skin). Even cloth gloves, such as mittens, can leave a distinctive print that can be traced back to the mitten that made it.

Prints are left on a surface because we are constantly secreting water and body oils and other compounds through our pores. This material is left on the surface we touch in the form of a fingerprint.

Different surfaces require different techniques for developing prints. In the movies, you usually see detectives with brushes. They are powder processing the prints. Minute particles of powder cling to the print residue as the brush passes over it. The print is then lifted with tape. Another process involves fuming. Vapors of iodine and superglue (bonds in seconds) will coalesce inside the print residue to reveal a latent print.

There are special processes that develop prints on paper, wood and cardboard. Fingerprints can be developed on objects that have been in water. Prints can be developed off of skin

(such as from the neck of a strangulation victim). There are very few surfaces on which a print cannot be developed.

Computers have revolutionized the techniques used to match fingerprints. Until recently, the old standard was the Henry Classification System; a cumbersome sequence of letters and numbers broken down into several levels of classification. It could take weeks, sometimes months to compare a suspect fingerprint to a department's print files. The advent of digital technology has changed all of that. Prints can be image scanned directly into a computer, doing away with ink and fingerprint cards. Prints can be compared at a rate of 400,000 per second. You couldn't do that in your lifetime.

It's called AFIS. Automated Fingerprint Identification System. Departments will input all the prints from arrests and all of the print cards they already have on file to create an historical record. They also input all of the prints from any unsolved crimes, in the hope that a hit might come up from a routine arrest. Local departments are linking their systems into a national database. The FBI wants national standards and a fully functional national network in place by the year 2000. With a national network, you could get busted in New York and have a print hit come up from a crime you committed in California.

2.2.4.2 DNA

A quick primer on DNA: DNA is constructed like a ladder; a ladder that has been grabbed at both ends and twisted, creating the double helix shape. The rails of the ladder are phosphate and sugar groups. They link together (sugar+phosphate+sugar+phosphate) to form the backbone. So far it's not so tough. There are four bases that form the rungs of the ladder: Cytocene, Guanine, Thiamine and Adnine. They are always in pairs and always complement each other; Cytocene is always paired with Guanine and Thiamine is always paired with Adnine. Each base forms half of the rung, meeting in the middle. Now, think of that ladder as a zipper. When DNA replicates, is unzips down the middle, separating the base pairs like a zipper. The complementary bases now attach to the opened segments to make new DNA

One DNA analysis technique looks at junk DNA. Everybody's' DNA is pretty similar. Everyone who has blue eyes has pretty much the same code for blue eyes. Everyone who has brown hair has pretty much the same code for brown hair. But these coding sequences are separated by 'junk' DNA. This 'junk' DNA is non-coding and only serves to separate the coding sequences. These 'junk' DNA sequences are totally random and totally unique to an individual. The process is extremely technical, but that is the concept, and it's really not that hard to understand.

Now, the way this works in a criminal case is thusly: The examining labs have samples of DNA, taken from a representative population group. These are entered into a database, to which the questioned DNA (that being compared for analysis in a case) is compared for frequency among the population group. That is DNA testing at its most basic.

2.2.4.3 Other Sub-disciplines

There are numerous other areas of study in the field of personal identification.

Forensic Anthropology seeks to identify remains, such as bones. From a skeleton, you can tell if the subject is human, its sex, the age of the subject when death occurred, stature, how long the remains have been in their current state and the cause of death.

Facial Reconstruction gets in the media once in awhile. The photo imposition technique involves imposing a photo of the skull over a photo of the face in question. Three dimensional reconstruction technique involves adding depth markers for tissue thickness onto a skull. The depths are an average taken from cadavers. Clay is placed over the skull, to match the depth of the markers, to simulate skin and muscle. The features are then smoothed out to make a 'face.' This has worked in a number of missing persons cases, but is considered more art than science.

Hair Comparisons are another biggie in forensic science. Hair can be determined to be human or animal. The body area from which a questioned hair came can be identified. Race can sometimes be determined. Disease conditions can be determined. And, of course, a hair found at a crime scene can be matched to the person that left it there. It's a relatively

simple comparison, involving side-by-side examination of the suspect and known hairs, similar to bullet matching [4].

2.2.5 Forensic Photography

This section is going to be pretty short. There are just a few particulars to hit on, and a couple of interesting techniques.

The first thing that needs to be done after securing the crime scene is photographing it. This creates a permanent record of the condition of the crime scene, one that is incontestable. First, take a picture that shows where the scene is; a shot with a street sign with the crime scene location in the background. Take pictures of the areas around the crime scene; alleys, dumpsters, rear areas, neighboring structures and even the structures across the street. Next, take pictures of the outside of the structure, showing points of entry and exit. Enter the structure, taking shots that show the locations and layout of the rooms. Take pictures of the whole room where the crime took place. Take close-ups of the scene or body. All pictures of items of evidence, which will be covered in the next paragraph, should be taking both with and without a scale (a small ruler showing the size of the object). Take pictures with the scale to show the size of an object. Take pictures without the scale in case its presence in the picture blocks other evidence.

What items are photographed at a crime scene? Bullet casings; photograph as a group and photograph individually. Photograph any dropped items, footprints or animal tracks. If a homicide, photograph the body or bodies. Photograph any toolmarks, bitemarks or skin impressions. Basically, anything that might be evidence is photographed.

Imprint evidence requires extra measures. Shoe imprints are photographed individually and as a series or group. Shoe imprints need to be lit from the side to show as much detail in the imprint as possible. Tire imprints are photographed from above as a whole. If the tire imprint is four feet long, then a picture showing all four feet is taken. Detail pictures are then taken showing one foot sections, each picture overlapping the one before it. This way, specific detail can be show and the overlapping pictures lined up to show the whole print. Again, all pictures are taken with and without a scale.

There is a special technique for no light situations. This technique is useful outdoors at night (perhaps a car accident scene), or in situations where the room is too big to light or there is no light available for pictures to be taken (such as a burnt out warehouse arson). The camera is set on a tripod with its shutter locked open. The photographer walks to several points in the room, popping off the flash, which is held in his or her hand. Each time the flash goes off, the film in the camera is exposed to another part of the room. The photographer does not appear as he/she is behind the flash and does not get exposed to the light when it pops off and only moves around the room while it is still dark. Remember, the film in a camera captures light. If there is no light, you can walk around in front of a camera all you want and never show up on the film.

Video is also used to film crime scenes, taking long sweeping shots that take in everything in an unbroken time frame. The problem with video is, camcorder microphones will pick up the officers talking in the background, which can sometimes be embarrassing when the tape is replayed in court.

2.2.6 Crime Scene Processing

Much of what has been discussed previously is involved in crime scene processing, so there is little need to rehash. There remain a number of points that fit into the categories of trace evidence, the homicide victim and a few administrative matters to get the nitty gritty explanation of forensic science out the way and conclude this set of web pages. This is more the loose ends section than anything, a smattering of quick explanations for topics that don't appear in the other categories. But first, some review [4].

Most physical evidence concerns itself with class characteristics and individual characteristics. Class characteristics are those characteristics that are common to a group of similar objects. For example, you buy a pair of Air Jordan sneakers. All Air Jordan sneakers have the same shape and same tread design on the bottom. These are class characteristics. Individual characteristics are those characteristics that are unique to a given object and set it apart from similar objects. You wear your Air Jordans around for a while and they get worn. The treads wear down. They get little pits and gouges in them. These little pits and gouges are individual to your shoes and no others since no one has walked

over the exact same surfaces in the exact same way in their Air Jordans. These concepts apply to all kinds of forensic evidence, from soil to glass, from rope to hefty bags.

The fracture match is another important concept, particularly when trace evidence is concerned. Tear a piece of paper in half. Hold the two halves together. This is called a fracture match. No two tears are exactly alike. One half of a tear can always be matched back to its other half. Remember that. If a half of something found at a crime scene can be matched to the other half of something found on a suspect, that's damn good physical evidence. Figure 2.3 illustrates the example of fracture match.

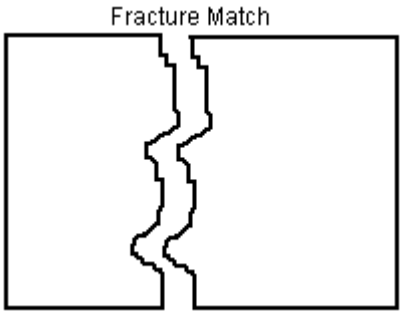

Figure 2.3: Fracture Match

2.2.6.1 Crime Scene Documentation

The first thing you do after securing a crime scene is document it. Always take pictures. They are the best record available. They show the crime scene as it was found; where objects are in relation to other objects, victims, rooms, etc. Take notes. Describe the scene, its over all conditions. Describe rooms, lights, shades, locks, food; anything that can indicate a time frame, condition of scene or that might have even the slightest evidentiary significance. Check dates on mail and newspapers. Diagram the crime scene. Take measurements. Photos are good to show where an object is in relation to another object, but measurements tell exactly how far.

2.2.6.2 Chain of Custody

Chain of Custody is of paramount importance to any investigation. It is the unbroken sequence of events that is caused by an item of evidence from the time it is found at the crime scene to the time it appears in court. Every link in this chain is documented, from

discovery at the crime scene, through evidence gathering, storage, lab analysis, return to storage, transfer to court. Every link is documented by date, time, handling individual, what was done with the evidence by that individual. If chain of custody is broken, if the evidence cannot be accounted in one step of its journey from crime scene to courtroom, it is rendered inadmissible; useless to the case.

2.2.6.3 Locard's Exchange Principal

Every time an individual comes in contact with a place or another individual, something of that individual is left behind at the place, and something of that place is taken away with the individual. If your Aunt Bertha gives you a big hug and walks away, fibers from her clothes will be on your clothes and fibers from your clothes will be on hers. Your hair is constantly falling out (circle of life, guys). You leave it all over the place. Just look around your house. You pick up carpet fibers on your shoes, dirt from the ground. Your skin flakes off. Look at the Wayne Williams case in Atlanta. He was convicted because fibers found on the body of one of his victims matched fibers from the carpet in his house. Might not sound like much, but it's GREAT physical evidence.

2.2.6.4 Entomology

To paraphrase Indiana Jones, "Bugs. I hate bugs..." But they are great evidence and can be used to determine time of death. In some cases, they can be used to determine if a body has been moved from one geographic location to another. Certain bugs incubate and hatch at certain known rates. If bugs are found on a corpse, the age of the bugs can be extrapolated backward to estimate time of death.

2.2.6.5 Blood Spatters

Blood spatters help a great deal in reconstructing a crime scene. They can be used to corroborate or disprove and alibi. They can be used to convict the guilty. There is much more to it than looking at a stain or spatter and saying, "This is where the crime took place." The patterns of the spatters and the shapes of the individual blood droplets themselves can tell how the crime was committed.

Drops falling from different heights (i.e. at different speeds) will leave different looking spatters. A drop falling from a low height of a few inches will leave a small cohesive circle. At greater heights, the circle will be larger and may even have a 'crown' effect.

Hitting a surface at an angle does even more to disrupt a blood droplet. Perpendicular impact leaves a droplet fairly uniform, as shown below. A droplet hitting a surface at an angle will bulge out in one direction, indicating the direction of travel of the droplet.

Cast off stains are a result of blunt force trauma (beating with an object such as a hammer). Pulling back from a blow produces a blood spatter that indicates direction, by creating an arc of blood droplets. You can determine the number of blows inflicted by counting the arcs. You can also determine the orientation of the individuals involved, the size of the object used and the right or left handedness of the assailant.

CHAPTER 3
Literature Review

3.1 History of Computer Forensics

The roots of computer forensics start with the first time a system administrator had to figure out how and what a hacker had done to gain unauthorized access to explore the system. This was mainly a matter of discovering the incursion, stopping the incursion if it was still in progress, hunting down the hacker to chastise him or her, and fixing the problem allowing the unauthorized access to begin with. In the beginning, the classic hackers breaking into computer systems were more interested in how things work than actually being malicious. So, collecting evidence for a hearing was not a process a system administrator needed to worry about. Just plug the hole, and often get back to personal hacking projects.

As computers evolved out of academia to businesses and government, there was more data and resources at risk. Hacker incursions became an issue handled through legal channels. Also, as computer technology advanced, it became more affordable. This allowed computers to be put not only on each employee's desk of even small business, but in people's homes. More people looking for uses for the computers lead to the increase in supply of programs. More programs made more types of information collected as possible evidence.

Evidence derived from computers has been used in court for almost 30 years. Initially, judges accepted the evidence as no different from forms of evidence they were already seeing. As computer technology advanced, the accepted similarities to traditional evidential material became ambiguous. In 1976, the US Federal Rules of Evidence was passed to address some of the ambiguities.

A lot has evolved with computers since 1976. One item of significance is the Internet. This information superhighway has become a major passage of items that fall under legal scrutiny. Another item is the amount of data an individual computer can hold. Personal computers of the early 1980's had no internal storage and the removable storage only held

360-kilobytes per diskette. Today, an average personal computer bought for teenager game playing and Internet cruising hold internally 40 billion bytes of data and removable disks hold from 2 million bytes to 2 billion bytes. Large server computers used by academia, government, and business are starting with internal storage averaging 100 billion bytes and have the expandability to use storage devices holding trillions of bytes of data.

This explosion of technology, while providing many times the computing power of the building size computers of the beginning, have made the field of computer forensics exponentially more complicated from the relativity simple tasks of evidence gathering only 5 years ago [8].

3.2 Important achievements in the discipline

Ability to intercept a Palm Pilot PDA password either by monitoring the traffic between the PDA and a workstation with a Palm Pilot cradle or by initiating a synchronization update between the password protected PDA and a second PDA.

Ability to analyze image files to detect if a message is hidden in the file using steganography. Niels Provos and Peter Honeyman at the University of Michigan have developed a process using statistically analysis of a JPEG image to detect if there is a steganographic item stored in the JPEG. Neil Johnson, a researcher at George Mason University, is working on being able to identify steganographic items in BMP and GIF images files as well as WAV and AU sound files [8].

3.3 Historic and current leaders in the field

Tracking current leaders in computer forensics is not an easy task. The people doing the cutting edge work commonly are employed by agencies like the NSA or CIA. So, even if they can gain authorization to publish their work, likelihood is it isn't published under their actual name. When these people are able to publish their wealth of knowledge, is after they have left the employer and all contractual silence is honored. Outside of the government secret agencies, the next place leaders in the field are found is in the ranks of professors of Computer Science and Engineering and law enforcement personnel [8].

These are some of the people making advancements for Computer Forensics:

• Rebecca Gurley Bace

Currently the president of Infidel, Inc., a consulting practice specializing in intrusion detection and network security technology and strategy. Before founding Infidel, Inc, she worked for the NSA for 12 years. She led the Computer Misuse and Anomaly Detection (CMAD) Research program from 1989 through 1995, as a charter member of NSA's Office of Information Security Research and Technology. She then left the NSA in 1996 to serve as Deputy Security Officer for the Computing, Information, and Communications Division of the Los Alamos National Laboratory.

• Peter Sommer

Currently serving as a Visiting Research Fellow at the LSE Computer Security Research Centre and established expert on computer security advising stock exchanges and insurance companies on systems risk. As a trained lawyer he is especially well placed to develop his current research interest in the legal admissibility of computer related evidence, especially in the context of computer crime. In December 1998 he was appointed Special Adviser in Electronic Commerce to the House of Commons Standing Committee of Science and Technology.

• Gene Spafford

A Professor of Computer Sciences at Purdue University, where he has been on the faculty since 1987. His current research interests are primarily in the areas of information security, computer crime investigation and information ethics. He is also director of the Purdue CERIAS (Center for Education and Research in Information Assurance and Security), and was the founder and director of the (now superseded) COAST Laboratory. He is also the interim Information Systems

Security Officer for Purdue University. Related to this, he is the founder and de facto director of the PCERT (Purdue Computer Emergency Response Team).

• David J. Icove

Presently employed by the Tennessee Valley Authority Police-Public Safety Service, Risk and Emergency Management Division. He also holds a position as an Adjunct Assistant Professor at the University of Tennessee in the Laboratory for Information Technologies. His primary background is working with fire, arson, and explosion cases.

• Dr. Neil F. Johnson

Currently employed by George Mason University as the Associate Directory of the Center for Secure Information Systems. His projects include steganography, covert communications, information assurance, and Cyber Warfare.

3.4 The scope of the discipline

The scope of computer forensics covers a wide field, which continues to grow as computer technology proliferates its way into every aspect of modern life. The base of computer forensics is recovering data from floppy disks, hard drives, and removable drive cartridges. Recovering data can be just finding it among the active files. Often, it will also include searching the media for files that have been deleted and been listed as unallocated space. When dealing with someone, who is actively attempting to hide information, scouring media space the operating system has registered as free or corrupted.

Within the files found on the media, the scope of what can be found continues to grow. Files early in the discipline were mostly limited to text documents, spreadsheets, and bulky images. Now on the file level, forensic complications of compression, encryption, password protection, and steganography have been added to the mix. The type of data being found in files has increased also. The operating systems now have configuration files and memory swap files. Now, practically program has its own set of temporary files. Printing is now done mainly via a queue-based system, so there are spool files with

possible evidential value. The average web browser has a history file, cookie file, a file of user saved web page addresses, and a cache of images and texts viewed.

On the hardware side recent additions include smart cards with 4 to 64 kilobytes of data space to current USB dongles with up to 64 megabytes of data space. Handheld devices like electronic organizers and personal digital assistants can have megabytes of data. Some currently found includes address books, appointment calendars, documents, e-mail, handwriting, passwords, phone book, text messages, voice messages, and web browsing associated files. Some even contain Global Positioning System connections leaving behind a trail where the PDA has been.

Another place evidential data has the possibility of being recovered is on the printers now available. Some have large caches of memory from which documents have the potential of being retrievable. Printers intended for large network setups occasionally also have hard drive type media on board for storage of files queued to print. The printer head, toner cartridge, or ink cartridge may also prove useful as physical evidence to show a printout came from a specific printer.

A branch of computer hardware, which grew out of the need to share data more quickly and the want for centralized servers to store data, is the computer network. As these networks grew and interconnected, the Internet evolved. The interconnection of all these computers opened up new routes for people to attempt to access and destroy the information stored on them. This created the need to have utilities to monitor network traffic and the people to understand what the utilities are showing them. Additional challenges are coming on-line as large wireless networks are being brought into service. In some cases, entire college campuses are being outfitted with a wireless network grid and some metropolitan areas are considering and quietly testing citywide wireless networking to offer to their populations.

The intersecting scope tree is where computers are being used so evidence is left behind. These include auction fraud, child exploitation, computer intrusion, death investigation, domestic violence, counterfeiting, email used for threats, harassment, and stalking,

extortion, gambling, identify theft, narcotics, prostitution, and piracy of software, music, images, and video [3].

3.5 Various forms of examinations conducted

When examining a computer setup, there are two general categories the examination falls into: live/real-time examination or an off-line duplication. Which of these types is necessary depends on the situation. When dealing with a network intrusion or a server and network setup, which cannot be made unavailable, live examination is the route taken. If total control can be taken of the computer to be examined, then off-line duplication is the path used.

In a live or real-time examination, often the goal is to trace and trap network activities of a system compromise. This involves initially verifying if an intrusion has actually taken place, and if so, determining how and when the intrusion happened. An additional goal, which can prove to be more difficult, is to pinpoint the location of the intruder. The primary difficulty of achieving this goal based in the fact the intruder could be anywhere on the planet. If the intruder is not someone part of the compromised network system, tracking him or her will likely require the cooperation of multiple organizations connected to the Internet as well as navigating any legal barriers as the search for the intruder crosses state and national borders. Due to the legal and human complexities of tracking down a network intruder, often they go untraced unless the network system broken into compromises a nation's security, or the intruder was able to access or destroy items of substantial monetary value. As supporting evidence of an intrusion, the audit log files of each machine compromised or an attempt at compromise, will be examined. Depending on the operating system and the settings on each system, audit log files can vary from being worthless to showing everything that was done on and to a machine.

The other situation when a live examination often takes place is when data of evidential value are stored on a network server, which cannot be taken out of service either due to significant monetary loss or risk to life. Servers falling into this category are ones for large service companies like at banks or the servers of a hospital. In this case, the files suspected

of having evidential value are copied from the server to media the investigators can take with them.

The ideal situation is one where full control of a computer system is handed over to the investigator. The first step before scouring for possible evidence is to make a byte for byte copy of the media being investigated. If the computer has a 20-gigabyte hard drive, a hard drive of at least equivalent size, if not brand, is used to copy the data from the suspect computer. In some cases, software will make a copy of a hard drive by moving the data from the suspect hard drive to burnable CD-ROMs. The idea behind this is to leave the suspect system as unaltered as possible. This leaves evidential information such as time-date stamps on files, file ownership, and last access information available to the investigator.

Once the copy of the suspect media is made, an array of methods can be used to carry out the examination. The method used often depends on the examiner and the shop employing him or her. However done, the goal of the exam is to search the media's files, unallocated space, unused space, and media formatting for information of probative value [8].

3.6 Scientific principles utilized in the discipline

The principles and methods used fall into two primary categories--one, dealing with data on a workstation or server computer, and a second for dealing with data on an active network.

Data used and held in computer systems is all stored digitally. On its basic level, all data is stored as a collection of ones and zeros. Storing these collections in specific ways lets us store documents, music, programs, and operating systems. As long as no extreme circumstances befall computer storage media, the ones and zeros are unchanged expect when altered by a computer. Cases where data can be altered without a computer are exposing media to cold below 4 degrees Celsius, exposing media to heat above 37 degrees Celsius, introducing the media into fluctuating magnetic field, or physically altering the media the data is stored on. As long as the media has undamaged, the ability to retrieve the data on it is next determined by if the format the data was written can be determined. The

format used to write data to media will vary depending on the media and the operating system or program used to write the data.

Most media specific data format differences are set my manufactures of the media, so it is fairly constant and often technical information easily available. Operating system media data formatting difference are set by the writers of the operating system. Often, these formats' technical information is also easily available. A difficulty is in unrelated operating systems often use significantly different formats [8].

On the media level, common device data formats are:

Format	**Purpose**
FAT12	• Originally a format used by DOS for devices up to 30 megabytes • Now primarily used by Microsoft on floppy disks
FAT16	• Used by later versions of DOS and Windows95a for devices up to 2 gigabytes
FAT32	• Used by Windows95b and Windows98 for hard drives • Also used by Windows 2000
NTFS	• Up through version 4 used by Windows NT • Version 5 used by Windows 2000
EXT, EXT2, EXT3	• Use by Linux
ISO9660	• Used for CDROMS
HPFS	• Used by OS/2
SYSV	• Used by Sun Solaris
UDF	• Used by DVD-ROMs
BEFS	• Use by the BE operating systems

NWFS	• Used by Novell Netware
XFS	• Used by Silicon Graphics operating system

Table 3.1: Common Device Data Formats

How ever the media data format is determined, once it is known, the examiner only needs to apply the format to the data on the media and all data stored there can be sectioned into the individual parts. These individual parts can also have specific formatting determined by the program that wrote to the media. Revealing the formatting of individual items, also known as files, is the next step to retrieving the data.

The program writing the data will determine a file's data format. The format of a data file can vary from being plain text to being encrypted with a password bypass being needed to access the file. The occurrence of difficulties like passwords and encryption will often depend on the sophistication of the computer user. While, many common programs like MS Word, MS Excel, and Quicken have individual file formats, they also have password and encryption capabilities, but many users either don't realize this or don't see a reason to use it. The commercial programs often have their file format details available to the public, their password hiding schemes and encryption algorithms have to be reverse engineered.

Another aspect of processing data on media is the files the operating system considers trashed or deleted. Unless the computer user takes specific steps, deleting a file does not actually remove the data from the media. What it does is just set a flag to the operating system signifying the space is now available for use. Some operating systems will not reuse deleted file space until all the unused free space is in use. Whether an operating system does this or not, there is great potential for retrieving probative information by collecting the 'trashed' files.

Additional information can be gained from available files besides the known data saved too them. When a file is written to a media device, it is often written as a block of data. These blocks can vary in size from 512 bytes to 64 kilobytes. The space after the end of a file not

used when a block is written out is known as slack space. Depending on the operating system, either zeros or random sections of the computer's active memory is used to fill this space. This is known as memory slack space. Some operating systems also write out multiple blocks at once, due to design or media formatting. If all the blocks to be written extents farther then the end of the file, they are just shown as used, but nothing is written to them. This space is disk slack space. Each of these can be of great value. Memory slack space can hold things like unencrypted passwords. Disk slack space can hold portions of deleted files not overwritten.

When dealing with data being collected from a network, the primary device used is known as a packet sniffer. This device can come in many forms--from a program run on a regular workstation to a dedicated piece of hardware. The principle of a packet sniffer is intercepts every piece of data crossing the network it is attached to and makes copies for analysis.

A packet sniffer's job is fairly routine since there is a limited number of formats data available for moving information over a computer network. While it is possible to create a nonstandard protocol for moving data on a network, the level of expertise necessary to do so makes it an uncommon issue.

The most common protocol encountered is called TCP/IP. The name is two of the primarily used protocols of a large suite of protocols. In the current version of TCP/IP, which is version 4, the common protocols are:

Protocol	Purpose
TCP (Transmission Control Protocol)	Used to move data in pieces, known as packets, from one machine to another. TCP specifically verifies data makes it to the destination.
IP (Internet Protocol)	The addressing scheme for machines using the protocol suite.
UDP (User Datagram Protocol)	Used to move data packets from one machine to another. UDP does nothing to verify the information makes it to the destination.
ICMP (Internet Control Message Protocol)	Used for low-level operations. This included routing information, time to get a packet from source to destination, and what gateways a packet goes through from source to destination.
IGMP (Internet Group Management Protocol)	Used for multicasting--the sending of packets to multiple destinations.
SLIP (Serial Line Internet Protocol)	Used to connect a workstation to a server via a modem.
PPP (Point-to-Point Protocol)	Used to connect a workstation to server via a modem. Newer than SLIP--provides data compression to emulate higher connection speeds and better packet error checking.
PPPOE (Point-to-Point Protocol Over Ethernet)	Similar to PPP but modem used is to connect workstation directly to a DSL connection--often as part of phone or cable television Internet service.

	part of phone or cable television Internet service.
SMTP (Simple Mail Transport Protocol)	Used to move email between machines.
SNMP (Simple Network Management Protocol)	Used for distributed network management. Allows for setting up and gathering usage statistics of network devices.
ARP (Address Resolution Protocol)	Used to map an IP address to a network card.
RARP (Reverse Address Resolution Protocol)	Used so a machine can query to find out its IP address.

Table 3.2 Some protocols of TCP/IP protocol suite.

Other common network protocols include:

IPX (Internet Packet eXchange)	Primarily used by Novell Netware servers.
SNA (System Network Architecture)	Protocol used by IBM to link mainframes together.
DECnet	Protocol developed at Digital Equipment Corporation to link machines using Digital's proprietary operating systems.
OSI (Open System Interconnection)	Protocols developed by the International Standards Organization. A complex and complete set of protocols for every kinds of network implementation. Was designed after TCP/IP, and has some similarities to it.

NetBIOS	Protocol developed by IBM. Used as initial communication protocol with Token Ring networks.
SMB (Session Message Block)	Developed by Microsoft and Intel in 1987 and used for communication with Microsoft Windows for Workgroups.
NetBEUI (NetBIOS Enhanced User Interface)	Enhanced version of the NetBIOS protocol done by Microsoft and Novell.
XNS (Xerox Network Systems)	Developed by Xerox, but did get the manufacture support expected, so seldom seen any more.

Table 3.3: Various other network protocols.

Many packet sniffers, especially commercial ones, can detect and decode most, if not all, standard network protocols. The sophisticated ones can collect packets going between two points on a network and display an ongoing listing of the information being passed. The some types of information, which can be intercepted, include:

- Files being transferred to and from the target.
- Commands being issued on the target.
- Output being returned to a machine issuing commands.
- Evidence of covert packet scanning programs running on local network machines.

3.7 Problem Formulation

What does it mean to be a Digital Forensic Analysis Toolkit? How do we categorize the different types of analysis tools? For example, an investigator can view the files and directories of a suspect system by using either specialized forensic software or by using the operating system (OS) of an analysis system and viewing the files by mounting the drive. Both methods allow the investigator to view evidence in allocated files, but only the specialized forensic software allows him to easily view unallocated files.

"Design and Development of Forensic Analysis Toolkit for Analyzing Malware Binary". This is exactly what we'll be implementing to gain an insight into the actual forensic analysis of the malware binaries and how they go about doing their job ("tracking creator's footprints and analyzing the binary for clues regarding the creator or the purpose of binary"). As the binary under investigation could possibly be a malware and may infect the computer on which it is being analyzed so here due to security reasons the binary will be analyzed without actually running it on the machine. Finally an attempt would be made to prepare a toolkit that would be useful at various levels of forensic analysis (Starting from basic file information retrieval of the malware binary, extracting the information from binary by analyzing strings contained in it and finally the hexadecimal editor will be developed for more expert users).

CHAPTER 4
Design And Development

4.1 Introduction

The design of a forensic analysis toolkit requires background information of the type of toolkit that can be made and the levels of details that each tool will provide to the user.

There are various issues related with the forensic tool designing. Two problems are discussed below that might need to be addressed while designing a forensic analysis tool.

The *Complexity Problem* in digital forensics is that acquired data are typically at the lowest and most raw format, which is often too difficult for humans to understand. It is not necessarily impossible, but often the skill required to do so is great, and it is not efficient to require every forensic analyst to be able to do so.

To solve the Complexity Problem, tools are used to translate data through one or more layers of abstraction until it can be understood. For example, to view the contents of a directory from a file system image, tools process the file system structures so that the appropriate values are displayed. The data that represents the files in a directory exist in formats that are too low-level to identify without the assistance of tools. The directory is a layer of abstraction in the file system. Examples of non-file system layers of abstraction include:

- ASCII
- HTML Files
- Windows Registry
- Network Packets
- Source Code

Similarly, the *Quantity Problem* in Digital Forensics is that the amount of data to analyze can be very large. It is inefficient to analyze every single piece of it. Data reduction techniques are used to solve this, by grouping data into one larger event or by removing known data. Data reduction techniques are examples of abstraction layers, for example:

- Identifying known network packets using Intrusion Detection System (IDS) signatures
- Identifying unknown entries during log processing
- Identifying known files using hash databases
- Sorting files by their type

It is proposed that the purpose of digital forensic analysis tools is to accurately present all data at a layer of abstraction and format that can be effectively used by an investigator to identify evidence. The needed layer of abstraction is dependent on the skill level of the investigator and the investigation requirements. For example, in some cases viewing the raw contents of a disk block is appropriate whereas other cases will require the disk block to be processed as a file system structure. Tools must exist to provide both options. The next section will cover abstraction layer properties with respect to digital forensics in more detail.

4.2 Layers of Abstraction

Layers of abstraction are used to analyze large amounts of data in a more manageable format. They are a necessary feature in the design of modern digital systems because all data, regardless of application, are represented on a disk or network in a generic format, bits that are set to one or zero. To use this generic storage format for custom applications, the bits are translated by the applications to a structure that meets its needs. The custom format is a layer of abstraction [1].

A basic abstraction example is ASCII. Every letter of the US English alphabet is assigned to a number between 32 and 127. When a text file is saved, the letters are translated to their

numerical representation and the value is saved on the media as bits. Viewing the file raw shows a series of ones and zeros. By applying the ASCII layer of abstraction, the numerical values are mapped to their corresponding characters and the file is displayed as a series of letters, numbers, and symbols. A text editor is an example of a tool operating at this layer of abstraction.

Each abstraction layer can be described as a function of inputs and outputs. The layer inputs are data and a translation rule set. The rule set describes how the input data should be processed, and in many cases is a design specification of the object. The outputs of each layer are the data derived from the input data and a margin of error. In the ASCII example, the inputs are the binary data and the ASCII mapping rule set. The output is the alphanumeric representation.

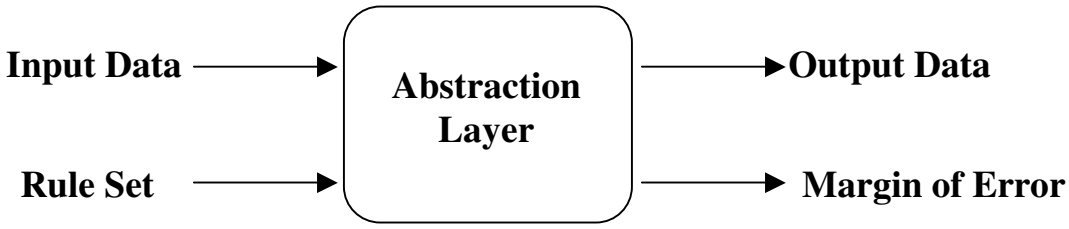

Figure 4.1: Abstraction Layer Inputs and Outputs

The output data of a layer can be fed as input to another layer, as either the actual data to be translated or as descriptive meta-data that is used to translate other input data. In the ASCII example, if the file were an HTML document then the output of the first layer, the characters, would be used as the input data to the HTML layer of abstraction. This layer takes the ASCII data and the HTML specification as input and outputs a formatted document. An HTML browser is an example of a tool that translates this, and typically the previous, layer.

Abstraction layers occur in multiple levels. The file system itself is a layer of abstraction for the stream of bytes from the disk media. Within the file system are additional layers of abstraction and the end result is a smaller stream of bytes that represents a file, which is

then applied to an application level of abstraction and it is processed further. The next section discusses various analysis categories.

4.3 Analysis Categories

The major categories of digital forensics can be defined using the notion of abstraction layers.

Physical Media Analysis: The analysis of the physical media layer of abstraction, which translates a custom storage layout and contents to a standard interface, IDE or SCSI for example. The boundary layer is the bytes of the media. Examples include a hard disk, compact flash, and memory chips. The analysis of this layer includes processing the custom layout and even recovering deleted data after it has been overwritten, [3] for example.

Media Management Analysis: The analysis of the media management layer of abstraction, which organizes storage media. The boundary layer is another collection of bytes from the media. Examples of this layer include dividing a hard disk into partitions, organizing multiple disks into a volume, and integrating multiple memory chips into memory space. This layer may not exist in all types of media, for example a database may access an entire hard disk and not create partitions.

File System Analysis: The analysis of the file system layer of abstraction, which translates the bytes and sectors of the partition to directories and files. The boundary layer is file content. The analysis in this layer includes viewing directory and file contents and recovering deleted files.

Application Analysis: The analysis of the application layer of abstraction, which translates data, typically returned from the file system, into the custom format needed by the application. Analysis in this layer includes viewing log files, configuration files, images, documents and reverse engineering executables. The input data will typically come from the file system, but applications such as databases may read directly from the disk.

Network Analysis: The analysis of the network layer of abstraction, which translates the lowest level data from a physical network or wireless network to the data that is used by an application. Analysis in this layer includes analyzing network packets and IDS alerts. Analysis of logs generated by network services, a firewall or web server for example, falls under Application Analysis.

Memory Analysis: The analysis of the memory layer of abstraction, which translates the bytes of the memory media to processes and system data. Analysis in this area includes identifying the code that a process was running and extracting sensitive data that was not stored elsewhere.

4.4 Analysis Tool Requirements

Using the previously stated information, a list of tool requirements has been generated.

Usability: To solve the Complexity Problem (data at its lowest format is too difficult to analyze) tools must provide data at a layer of abstraction and format that helps the investigator. At a minimum, the investigator must have access to the layers of abstraction that are defined as Boundary Layers. The tool should also present the data in a clear and accurate format so that the investigator does not interpret the data incorrectly.

Deterministic: To ensure the accuracy of a tool, it must always produce the same output when given a translation rule set and input.

Verifiable: To ensure the accuracy of a tool, one needs to be able to verify the results. This can be done manually or by using a second and independent tool set. Therefore, one needs access to the inputs and outputs of each layer so that the output can be verified.

4.5 Proposed Design

In the previous sections we have seen that the tool being designed should choose some abstraction layer at which the tool will work. And the maximum effort should be put at the design time itself that above given analysis tool requirements are met. We have planned to build three tools under one common integrated tool kit:

- String Extractor
- Hexadecimal Editor
- File Information Tool

The brief idea of these tools is discussed below:

String Extractor: As the binary under analysis may be containing some data in readable character also so this tool will extract all those strings and would provide that to the user. It has been decided that the tool won't extract the strings less than 5 character. The characters that are considered to be readable are all alphanumeric characters. So even these strings can give a lot of clue regarding the details of creator, the Operating system used, the language which was used and lot of other similar information can be retrieved by extracting the strings from the given binary. The strings extracted will be displayed along with the address, at which they occur, in hexadecimal format.

Hexadecimal Editor: The binary is sometimes best viewed by hexadecimal equivalent as lot of information is processed in hexadecimal format only. So this tool has been added for the help of the analyzer to check the file in hexadecimal and ASCII format at the same time. It will provide the facility of searching a given string in hexadecimal or string format. One may go to a particular byte position in the file.

File Information Tool: This is a supplementary tool provided along with the toolkit. This tool will be helpful for the user while using other tools. User doesn't need to go to the OS level tools to find the information regarding the file when he/she is doing the analysis work by using my toolkit.

Abstraction Layer: These all tools will work at application layer of abstraction. They will get a file and will provide information to the user in their own format. This information can further be used in forensic analysis purposes directly or indirectly as the need may be.

Meeting the requirements: The tools in the proposed tool kit are supposed to match all the requirements discussed in the previous sections regarding the forensic analysis tools. It will be usable, deterministic and verifiable.

Usable: For making it more useable it has been proposed that the toolkit will be developed in the Windows environment and will be GUI based toolkit. There would be a main window having option to invoke the different tools discussed above.

Verifiable: The proposed toolkit will give verifiable results making it sure that the results it provides are correct. One may open the binary in some text editor and may himself /herself verify that the information/results provided by the toolkit are correct.

Deterministic: The tools in the proposed toolkit will always be providing the same results given the same data and rule-set.

CHAPTER 5
Implementation and Results

5.1 The Toolkit

The toolkit developed has three tools. These tools can be invoked with the help of the screen given below:

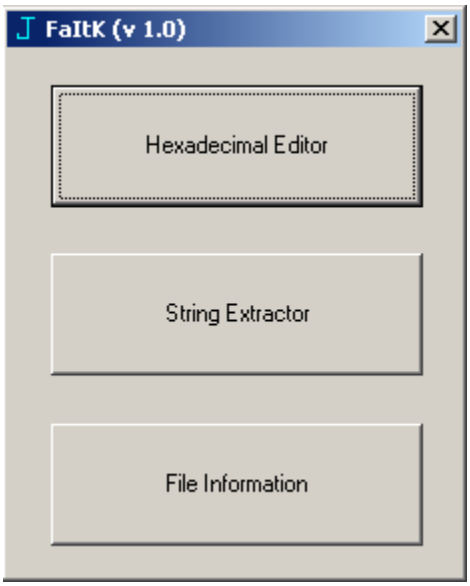

Figure 5.1: The main window of the Toolkit.

The features of the tools in the toolkit are discussed in the following sections:

5.1.1 String Extractor

It works as proposed. It can extract the strings longer than 5 characters containing alphanumeric characters. The corresponding address of the string is displayed in first column and the string itself is displayed in another column. The screenshot of String Extractor without opening any file is shown below.

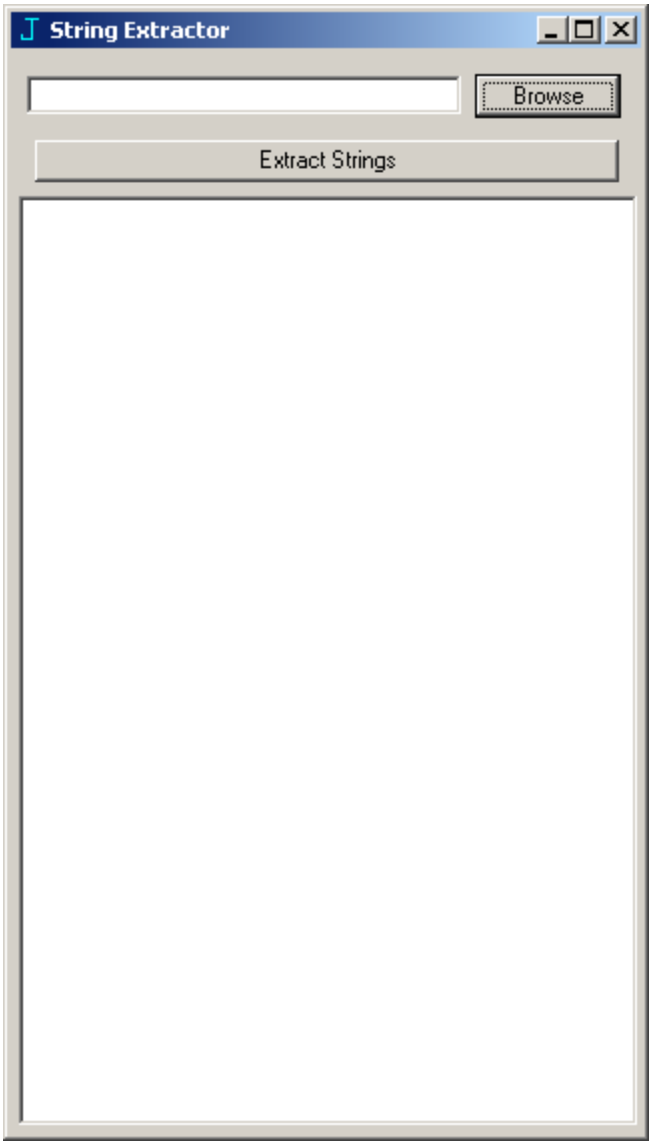

Figure 5.2: The String Extractor

5.1.2 Hexadecimal Editor

It provides the facility of viewing/editing the binary file. The file is displayed in three boxes one shows the hexadecimal address, second shows the hexadecimal equivalent of the character at the given address and third shows the actual character at that address. If the character is readable then it is displayed in its original form otherwise a smiley is displayed. The tool also provides the facility to search a particular character or string. The searching can be done on the basis of hexadecimal input also. The searched string is highlighted and corresponding hexadecimal values are highlighted. The screenshot of the tool with blank fields is given below:

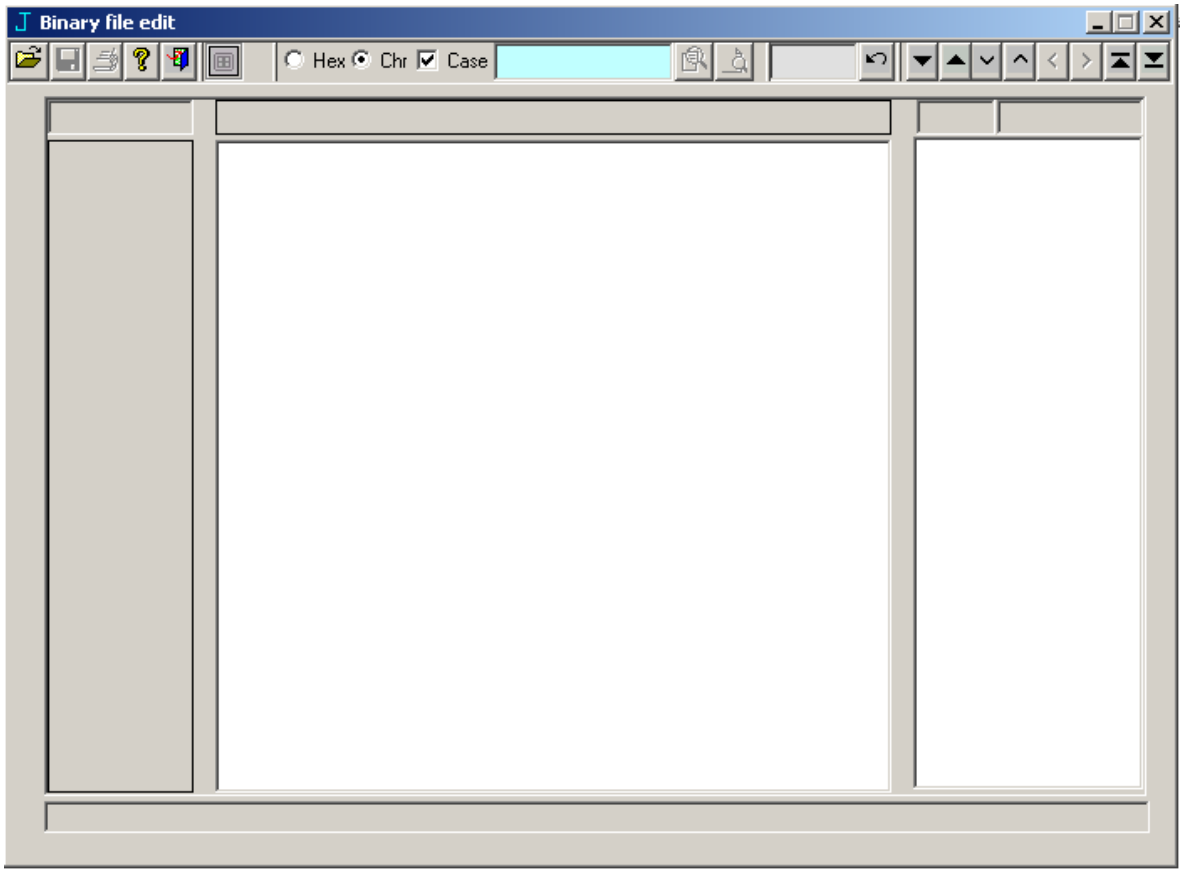

Figure 5.3: The Hexadecimal Editor

5.1.3 File Information tool

It provides the facility of checking the information about file under analysis. The user doesn't need to go to the OS level facility to check the same information regarding the file under study. It also provides the functionality to alter the attribute of the file and change various dates if needed. The screen shot of the File Information tool is shown below:

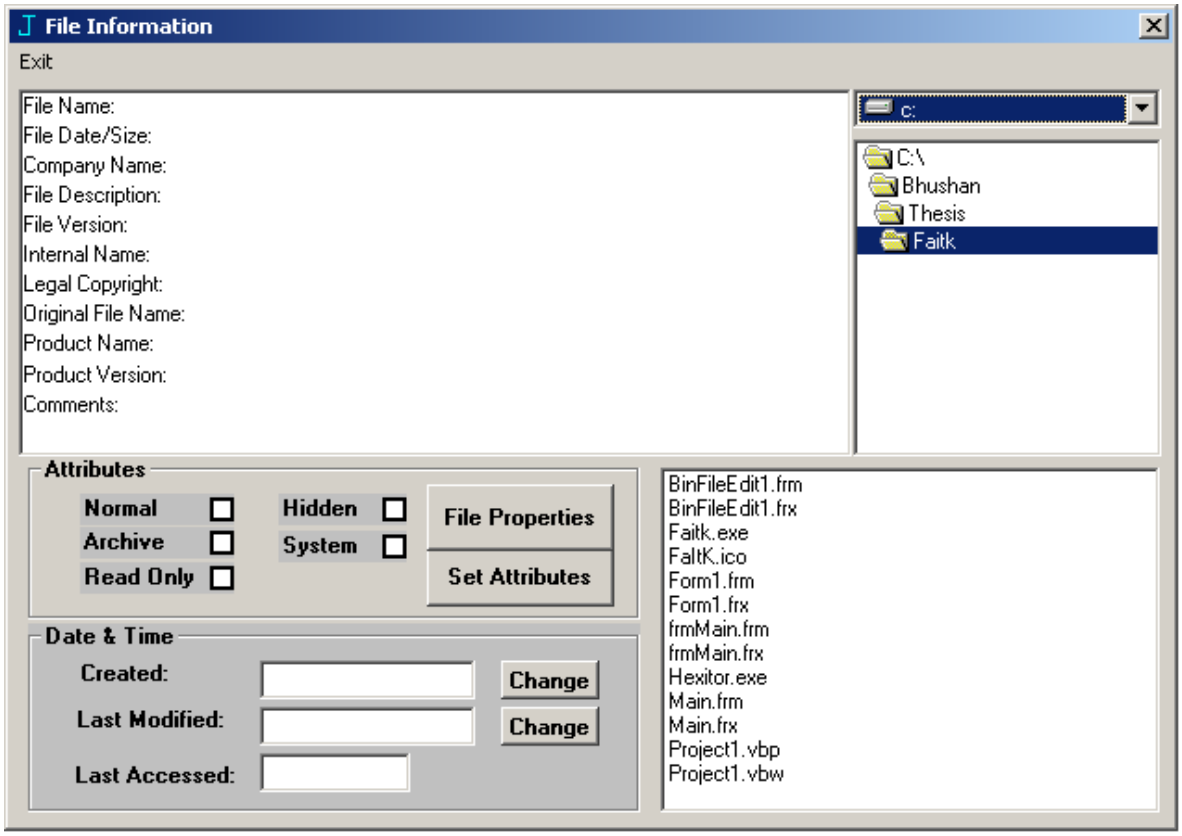

Figure 5.4: The File Information tool

The next section shows these tools along with the results.

5.2 Results

This section discusses the tools of the toolkit developed during the thesis work along with a binary file loaded into them. This binary is a Windows executable used for working on the toolkit while it was being developed and during results collection. We will see in the coming sections that what sort of information we can retrieve from the developed toolkit. The information retrieved is more useful if the user working on the information gathered knows about windows executable formats and some windows development environments.

5.2.1 String Extractor

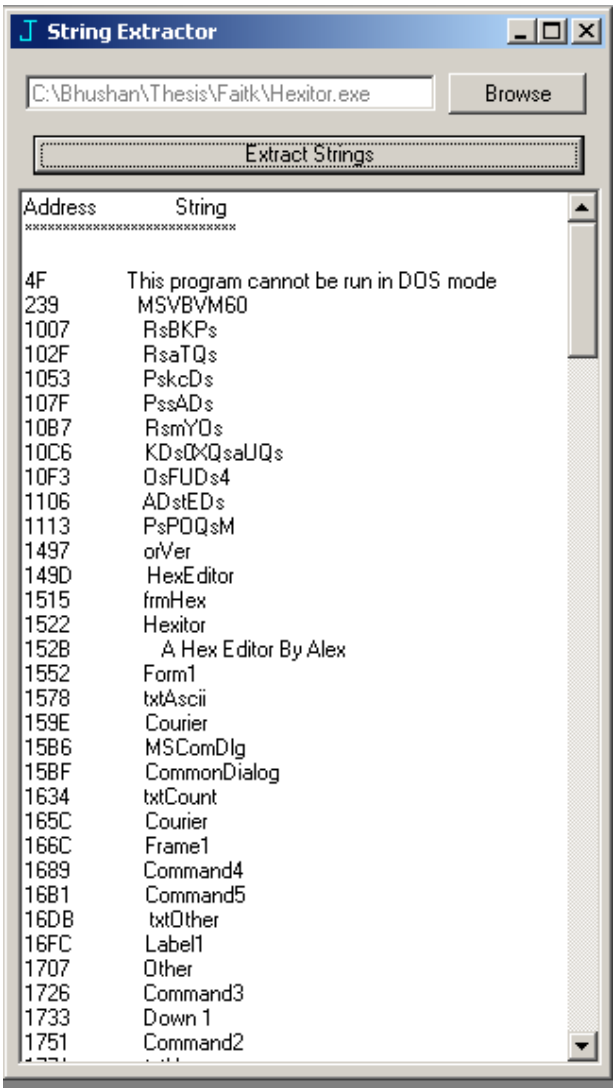

Figure 5.5: The String Extractor with loaded binary

The above screenshot shows a binary file loaded into it. The file can be selected by using the 'Browse' button. Once the filename is displayed in the textbox, clicking on 'Extract Strings' button will fill the lower text box with two-column information.

- The first column contains the address in hexadecimal format.
- The second column contains the string that exists at the shown address.

As discussed in previous section the string extractor will show the strings with minimum 5 characters.

Explanation of Results

From the information retrieved from the above tool we can make various conclusions as shown below:

- 4F → 'This program can not be run in DOS mode'
  - *Conclusion:* File is a windows executable: This string gives us the clue that this is a windows executable.
- 239 → 'MSVBVM60'
  - Conclusion: The application is developed in Visual Basic 6. The above string is found in the executables made in Microsoft Visual Basic 6.
- 1552 → 'Form1', 15B6 → 'CommonDialog', 1689 → 'Command4', 16FC → 'Label1'
  - *Conclusion:* The application has been developed in Visual Studio: The strings like 'CommonDialog', Label1 and Form1 etc. show that the file could have been possibly be made in Visual Basic.

5.2.2 Hexadecimal Editor

The following screen shot shows the hexadecimal editor after loading the binary into it.

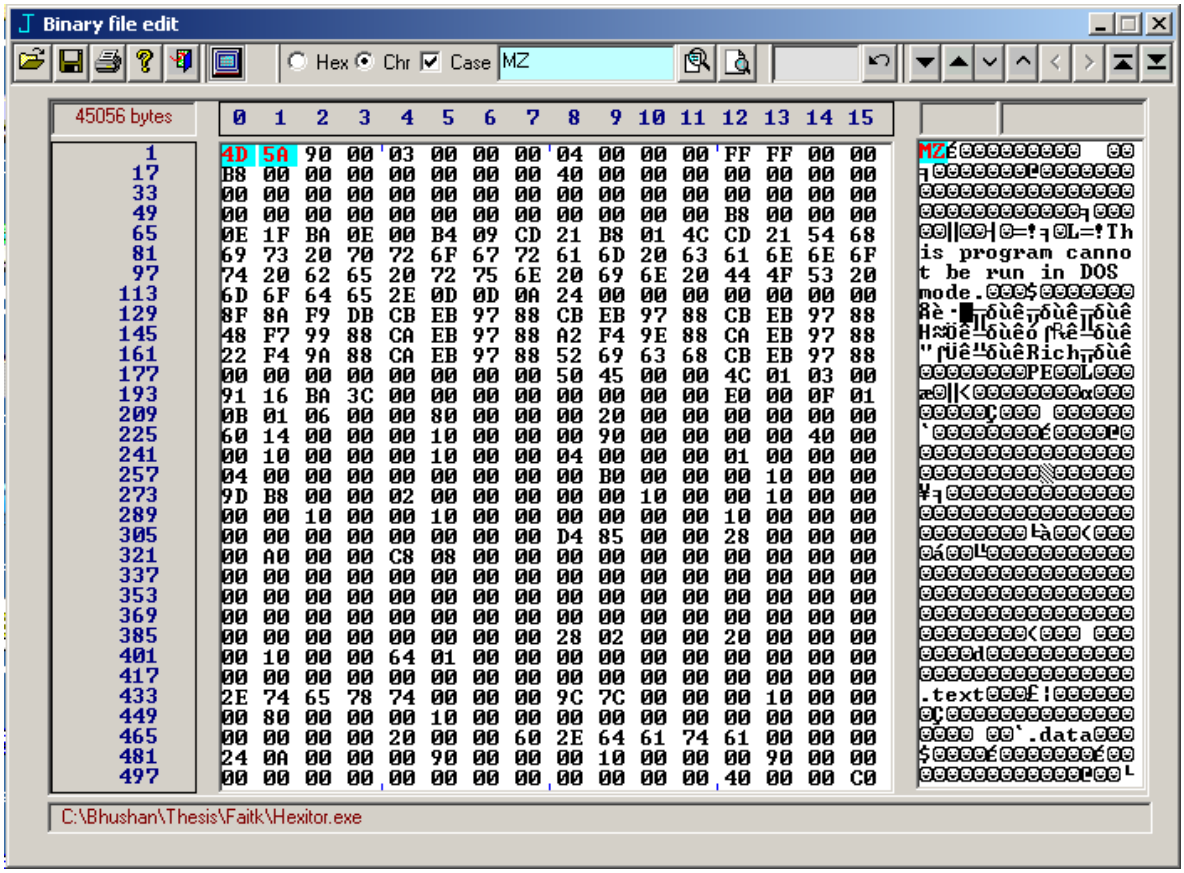

Figure 5.6: The Hexadecimal Editor with loaded binary

Editor Features and Results

It can be seen that it displays various types of information about the loaded binary:

- The Address
- Corresponding Hexadecimal Value
- Character Equivalent of that hexadecimal value

The editor also provides the facility of **searching** a particular string in the loaded binary. This screenshot itself shows the example of searching a particular string ('MZ' in this case). And from this string a person, who knows a bit about the structure of DOS executable, can conclude that it is most probably a DOS/Windows executable.

Another feature of this tool is that one may have three type of navigation within the tool:

- One line up/down
- One page up/down
- Move to first/last page

One can choose to **view/edit** the opened binary. By just clicking on the character one the corresponding hexadecimal value will be highlighted or vice-versa.

5.2.3 File Information Tool

The next screen shot is that of File Information tool. This tool displays the properties of the file along with the filename, size, company name, File Description , File version etc. It also displays the various attributes of the file i.e. whether it's a hidden file, system file or read only file etc. User can also change these attributes if so desired. Various date stamps associated with the file can also be viewed in this tool. And one may change the File Created or File Modified time stamp if so desired. If the information regarding the file is not found then the user is simply notified regarding the same.

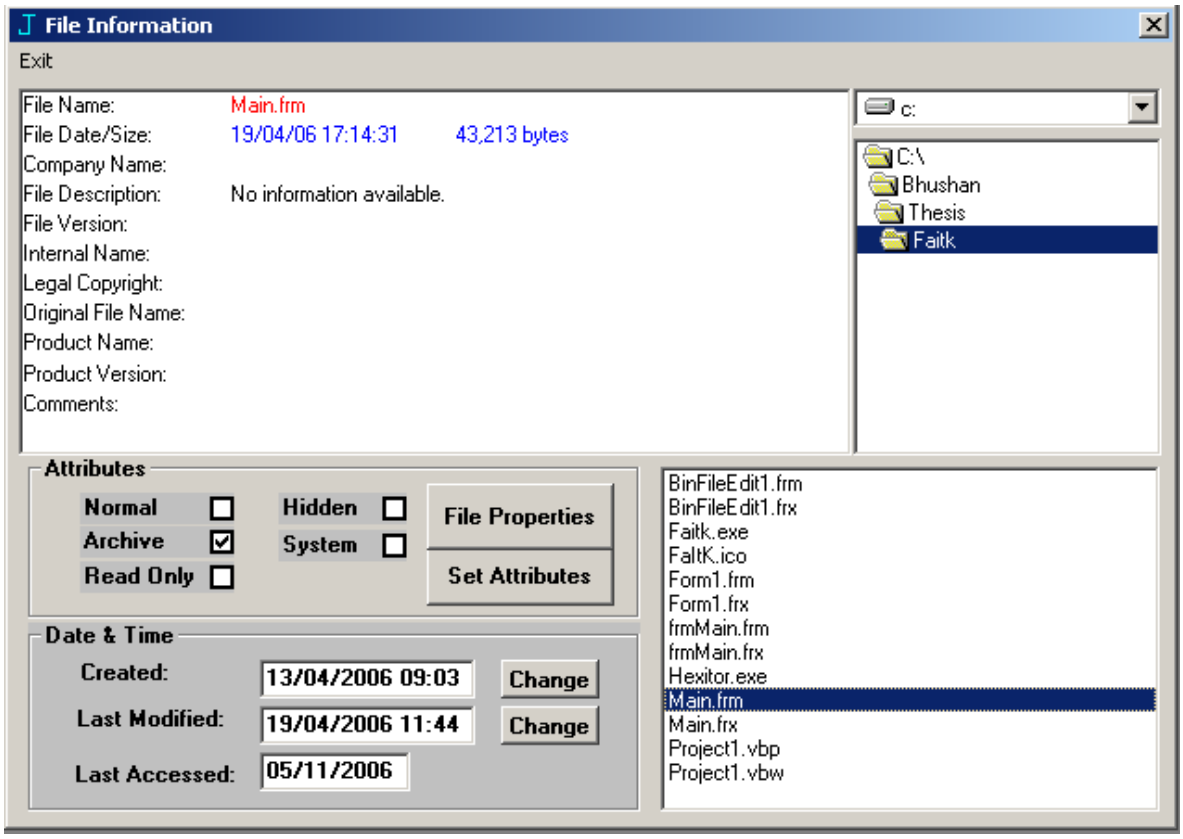

Figure 5.7: The File Information tool with a file selected.

CHAPTER 6

Conclusion And Future Scope

6.1 Conclusion

The forensic analysis is a very old field in the area of investigating the crimes. But it has been said that 'you can not win the modern battle with old weapons'. Now a days computer/cyber crime is increasing with a very steep rise. The networks are being attacked every now and then. The CERT records show that the number of hacking events being reported is increasing day by day. Various sort of viruses, trojans and worms are infecting the personal as well as organizational computers. Hence causing the data and financial loss to the attacked person/organization. As any other crime this is of utmost importance that research be done in the field of digital forensic analysis.

The today's life is fast as hence everything is supposed to be as fast as it can be. The investigation process needs to be very fast so that timely results can be given. The timely results would then help in making final conclusions regarding the crime committed. In this digital world a computer is expected to be present if the problem is to be solved in a shorter time. Here the case is that of digital forensic analysis, so one should never think of doing the forensic analysis of a computer binary without using the digital forensic analysis tools for the same.

Hence the tools proposed and developed during this thesis work will be of great use in the process of forensic analysis of a malware binary or any other binary that is expected to be causing problems. These tools can give lot of information regarding the creator, purpose and the development environment of the binary. The other kind of additional information can be found if the results are minutely checked. Further the skill of the person using this toolkit too makes a difference in the results achieved. If the person is skilled and is having enough information regarding the file structure and the development environment in which the file was developed then he/she may produce better results.

6.2 Future Scope of Work

The developed toolkit may further be enhanced by adding various tools on different abstraction layer. For example a network data monitor may be added. The toolkit will be more useful as more kind of tools are added to it in the future.

The tools may be attached with a knowledge base so that these can match various signatures already known, and hence may produce better and automated results. The knowledge base can be created by collecting information from various field experts. Once the knowledge base is attached with the toolkit, a less skilled user can also use the toolkit for forensic analysis.

In further enhancement the toolkit may be enhanced to make use of various AI techniques to automatically enhance its knowledge base.

References

[1] Brian Carrier, "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers", International Journal of Digital Evidence, Winter 2003, Volume 1, Issue 4. Available at: www.ijde.org

[2] "Digital Networks and Cultural Industries", course delivered physically on 9th December 1996. Available at: http://www.mipc.mmu.ac.uk/nqn/train1.html

[3] "The Essential Guide to Starting and Operating a Small Business in Hernando County 2004-2005", The Greater Hernando County Chamber of Commerce Business Assistance Committee. Page 16. Available at: http://www.hernandochamber.com/PDF/BusinessGuide2005.PDF

[4] "Forensic Sciences Web Pages", February 7, 1997 http://home.earthlink.net/~thekeither/Forensic/forsone.htm

[5] "Internet Usage Statistics - The Big Picture", Internet World Stats-Usage and Population Statistics, Available at: http://www.internetworldstats.com/top20.htm

[6] Malware – from wikipedia. Available at: http://en.wikipedia.org/wiki/Malware

[7] Matt Curtin, "Introduction to Network Security", March 1997. Available at: http://www.interhack.net/pubs/network-security.pdf

[8] Michael Potaczala, "Computer Forensics", Academic Computing Support, University of Central Florida, 12/6/2001 Available at: http://chantry.acs.ucf.edu/~mikep/cf/CHS5937-TermPaper.pdf

[9] Rajnish Sharma, "Design And Development Of A Linux Based Honeypot", Thapar Institute Of Engineering & Technology, Patiala. May 2005

[10] Reto Baumann, Christian Plattner, "Honeypots - A Diploma Thesis", ETH Zurich - Department of Computer Science, Feb 2002, Page 17. Available at: http://www.inf.ethz.ch/personal/plattner/pdf/diplomathesis.pdf

[11] Reto Baumann, Christian Plattner, "White Paper: Honeypots" , February 26, 2002. Available at: http://www.inf.ethz.ch/personal/plattner/pdf/whitepaper.pdf

[12] Richard Steinberger, "Proactive vs. Reactive Security". Computer Crime Research center. Available at: http://www.crime-research.org/library/Richard.html

[13] Ricardo Ochoa, "Virus/Contaminant/Destructive Transmission Statutes", NCSL, 2005. Available at: http://www.ncsl.org/programs/lis/cip/viruslaws.htm

[14] Selvadurai Jeyarajah, "Article 2", Surprise 95 Journal, Imperial College London, Volume 2. Available at: http://www.doc.ic.ac.uk/~nd/surprise_95/journal/vol2/sj1/article2.html

[15] Spyware – from wikipedia. Available at: http://en.wikipedia.org/wiki/Spyware

[16] Thomas A. Longstaff. James T. Ellis, Shawn V. Hernan, Howard F. Lipson, Robert D. McMillan, Linda Hutz Pesante, Derek Simmel, "Security of the Internet". The Froehlich/Kent Encyclopedia of Telecommunications vol. 15, pp. 231-255, New York: Marcel Dekker, 1997.

[17] "Top 20 Countries With The Highest Number Of Internet Users", Internet World Stats-Usage and Population Statistics, Available at: http://www.internetworldstats.com/top20.htm