FOI
SWEDISH DEFENCE
RESEARCH AGENCY

FOI-R—0024--SE
March 2001
ISSN 1104-9154
Scientific report

Gunnar Jervas (ed), Ian Dennis, Richard Conroy

NEW TECHNOLOGY AS A THREAT AND RISK GENERATOR

Can Countermeasures keep up with the pace?

Division of Defence Analysis
SE-172 90 STOCKHOLM

SWEDISH DEFENCE RESEARCH AGENCY (FOI)
Division of Defence Analysis
SE – 172 90 STOCKHOLM
SWEDEN

FOI-R—0024--SE
March 2001
ISSN 1650-1942

Gunnar Jervas (ed)

NEW TECHNOLOGY AS A THREAT AND RISK GENERATOR

Can countermeasures keep up with the pace?

Ian Dennis

Richard Conroy

Issuing organization	Report number, ISRN	Report type
FOI – Swedish Defense Research Agency Division of Defense Analysis SE-172 90 STOCKHOLM SWEDEN	FOI-R--0024--SE	Scientific report
	Research area code 1	
	Month year March 2001	**Project no.** E 1156
	Customers code 1	
	Sub area code 11	

Author/s (editor/s)		Sponsoring agency
Gunnar Jervas	Editor/author (Ch 1)	Department of Defence
Ian Dennis	Author (Ch 6-9)	
Richard Conroy	" (" 2-5)	**Project manager** Gunnar Jervas
		Scientifically and technically responsible Jan-Erik Svensson

Report title (In translation)

NEW TECHNOLOGY AS A THREAT AND RISK GENERATOR

Abstract (not more than 200 words)

In the past many people looked upon new technology as a means to solve problems. However, during the last few years the pace of technological change has become so fast that more and more people have started to look upon it as a risk and threat. This has to do with the fact that technology is an increasingly important competitive factor, and therefore enters the market long before all its consequences are fully understood. As a result unintentional technological errors occur, and they are becoming more serious since systems today are increasingly interconnected and interdependent. Moreover, criminal and other hostile actors try to use the growing number of weak points that are generated by this accelerating technological change.

After an introduction and overview, eight chapters follow that take up the different aspects of fast technological change and the problems that it creates. At the same time, what can be done to eliminate or reduce these problems is discussed. The conclusion is that it has become considerably more difficult to prevent all the technical errors and misuses that are generated in a time characterized by faster and faster technological change. We hope that this study will contribute to the countermeasures that are increasingly needed.

Keywords

Communication Systems, Cryptography and Steganography, Computers and Networks, Software Threats and Problems, Electronic Technologies, Managing IT-Threats and Problems, Legal problems

Further bibliographic information	Language English
ISSN 1650-1942	**Pages** p. 325
	Price acc. to pricelist SEK 225 **Security classification**

Utgivare	Rapportnummer, ISRN	Klassificering
Totalförsvarets forskningsinstitut – FOI Avdelningen för Försvarsanalys SE-172 90 STOCKHOLM	FOI-R--0024--SE	Vetenskaplig rapport
	Forskningsområde 1. Försvar- och säkerhetspolitik	
	Månad, år Mars 2001	**Projektnummer** E 1156
	Verksamhetsgren 1. Forskning för regeringens behov	
	Delområde 11. Försvarsforskning för regeringens behov	

Författare/redaktör		Uppdragsgivare/kundbeteckning
Gunnar Jervas	Redaktör/författare kap 1	Försvarsdepartementet
Ian Dennis	Författare kap 6-9	
Richard Conroy	" " 2-5	**Projektledare** Gunnar Jervas
		Tekniskt och/eller vetenskapligt ansvarig Jan-Erik Svensson

Rapportens titel

Ny teknologi som risk- och hotgenerator

Sammanfattning (högst 200 ord)

Tidigare uppfattade de flesta ny teknologi som ett medel för att lösa problem. De senaste åren har teknikutvecklingen dock blivit så snabb att den av allt fler kommit att uppfattas som en risk eller ett hot. Detta beror främst på att ny teknologi utgör ett allt viktigare konkurrensmedel, och därför kommer ut på marknaden långt innan dess konsekvenser kunnat förutses. På så vis uppstår oavsiktliga tekniska fel som kan få allvarliga konsekvenser, eftersom olika system i allt högre grad är hopkopplade och därmed beroende av varandra. Dessutom kan kriminella eller fientliga aktörer medvetet utnyttja de svaga punkter som otillräckligt utprovade system alltid innehåller.

Efter introduktion och översikt följer åtta kapitel som tar upp olika aspekter på teknologisk förändring och de problem den skapar. Dessutom diskuteras vad man kan göra för att eliminera eller reducera sagda problem. En slutsats är att det blivit allt svårare att hindra alla de tekniska fel och det missbruk som uppkommer i en tid av allt snabbare teknologisk förändring. Det är vår förhoppning att den här studien skall bidra till att motverka dessa problem.

Nyckelord

Kommunikationsproblem, kryptologi och steganografi, datorer och nätverk, mjukvaruproblem, elektroniska betalningstekniker, infrastrukturella beroendeproblem, hantering av IT-hot, legala problem.

Övriga bibliografiska uppgifter	Språk Engelska
ISSN 1650-1942	**Antal sidor:** s. 325
Distribution enligt missiv	**Pris:** Enligt prislista SEK 225 **Sekretess**

Contents

Chapter 1 – Introduction and Overview (Gunnar Jervas) 7

Chapter 2 - Communication Systems (Richard Conroy) 13

2.1 Introduction 13
2.2 Cable Communications 13
2.3 Wireless Communications 20
2.4 Threats to Communication Systems 36
2.5 Signals Intelligence (SIGINT) 41
2.6 Law Enforcement & Communication Systems 45
2.7 Future Advances in Communications Technology 48
2.8 Conclusions and Future Prospects 51

Chapter 3 - Data Security – Cryptography & Steganography (R Conroy) 59

3.1 Introduction 59
3.2 Threats to Data Security 59
3.3 Cryptography 64
3.4 Information Hiding 93
3.5 Commercial and Legal Aspects to Information Security 104
3.6 Conclusions and Future Prospects 114

Chapter 4 - Computers & Networks (Richard Conroy) 123

4.1 Introduction 123
4.2 The OSI Communications Model 124
4.3 Threat & Security Issues 150
4.4 The Internet 164
4.5 Conclusions and Future Prospects 166

Chapter 5 – Software Threats and Vulnerabilities (Richard Conroy) 173

5.1 Introduction 173
5.2 Actors, Their Motivations & Tactics 174
5.3 Penetration Techniques 183
5.4 Detection and Security 195
5.5 Conclusions and Future Prospects 207

Chapter 6 – Electronic/Cyberpayment Technologies (Ian Dennis) 221

6.1 Introduction 221
6.2 Types 223
6.3 Ideal Security Properties 232
6.4 Threats and Problems 236
6.5 Conclusion 245

Chapter 7 – Managing IT Risks, Threats and Problems (Ian Dennis) 249

7.1 Introduction....249
7.2 Understanding Potential Threats and Problems 253
7.3 Controls and Procedures to Manage Threats / Problems 261
7.4 Conclusion 270

Chapter 8 - Infrastructure's Dependence and Interdependence on Technology (Ian Dennis) 273

8.1 Introduction....273
8.2 Critical Infrastructure....274
8.3 Threats 277
8.4 Vulnerabilities....279
8.5 Prevention 297
8.6 Conclusion 297

Chapter 9 – Law Enforcement Issues (Ian Dennis) 301

9.1 Introduction....301
9.2 Types of Computer Crime 304
9.3 Gathering Evidence and Technical Challenges....311
9.4 Processing and Storing Evidence 319
9.5 An Example: The Kevin Mitnick Case....321
9.6 Conclusion....322

Chapter 1 – Introduction and Overview

During the 1990´s I got a stronger and stronger feeling – as many other people probably did – that the development was becoming more and more difficult to foresee and control. That the reason for this had to do with technological change seemed evident. Technological change had become so fast that more and more people looked upon it as a threat. This was new to me. For a long time in Sweden it had been common to interpret technological change as an instrument that could be used to solve practically any problem, including the non-technical.

As I see it today, technological change is not the only reason behind the ever more evasive and uncontrollable development. By the beginning of the 80´s a deregulation of the financial markets started in Great Britain, and spread to other parts of the world. At the same time, information technology made rapid progress. Because of these developments international capital flows - and economies in a broader sense - could be connected to each other through a much bigger, denser and faster information system than before. As different obstacles were removed and new actors entered the scene, competition accelerated. To succeed in economic life it became more and more important to be first, in regard to both material products and services. It is at the beginning of the lifecycle of a product that the price is at the highest. As soon as mass production starts the price drops and profits go down, and in a functioning market economy ineffective producers will be eliminated.

The aforementioned circumstances have led to the result that new technology is incorporated into economic life much faster than before. This means that new products enter the market long before their consequences can be foreseen, which increases uncertainty and lowers the possibilities for control. If you accept this description, the next question to consider will be, is the mentioned development a linear one? -Will this pace of change and uncertainty go on as before in the future? -Or is it still possible to counteract the negative consequences of technological change so effectively that the total result will still be positive? It is hard to deny that technological change also has a positive side.

Even if there are a number of factors behind the dynamic development during the last few years, it is evident that technological change is a very important one. It is also relatively tangible, compared to other factors of importance. Therefore in this project it was decided to start concentrating on technological development. (Later, we hope to come back to the broader picture). One important ambition has been to put the development referred to into a policy perspective - that is to say this study is not only about technological change and

the threats and risks that accompanies it, but it also deals with what can be done to remove or minimise these problems.

Below you find an overview of the subjects included in the report. Of course, further subjects could have been added, but in the course of time we have become more and more convinced that nowadays it is more important to be up to date than complete, if the latter is even possible. Our ambition has been to write a report that can be understood by people without special technological knowledge. Another aspiration has been to make each chapter readable in itself, which has led to some repetitions. For this, we apologise to those that read the whole report.

Chapter 2 was initially called Telecommunications, but after some time it was renamed *Communication Systems,* which corresponds better to the realities of today. The first part of this chapter gives a survey of the general hardware infrastructure used in these systems - both cable-based and wireless. It is emphasised that traditional cable based communication systems have come under growing pressure from wireless ones. Wireless communication systems are extremely promising in one way, but they are also the most threatening to privacy. The second part of chapter 2 examines different types of actors and their potential as well as motivations to threaten different types of communication channels. In this chapter tools for increasing security as well as attacking channels are surveyed.

In chapter 3, *Data Security – Cryptography and Steganography*, the author explains how safety can be added to channels susceptible to eavesdropping. The main security system is cryptography, the focus of this chapter. After having defined the attributes of a "good" cryptosystem the author describes and discusses different types of symmetric as well as asymmetric ciphers. Key management, which is a big problem, is explored in some detail. Chapter 3 also deals with the art of hiding information. In cases when it is not desirable to encrypt information you can for example send it through covert channels. In steganography you disguise the message in linguistic or technical ways so it cannot be seen. During the last few years the art of hiding information has drawn an increasing interest from both academic and commercial sectors.

Chapter 4 – *Computers and Networks* – first deals with network problems starting from the OSI (Open Systems Interconnection) Communications model. This model divides a network into seven layers, starting with a physical one (coaxial cables etc) at the bottom and then escalates up to an application layer, seen by the user at the top. The author considers each of these layers, mainly from a security point of view. He identifies different threats to computers and their connections to networks as well as how you can diminish the impact of these risks. It is concluded that even if you apply detection systems, firewalls

and so on some risks will always remain and have to be taken care of, especially when you plan for critical services. Finally, the future of the Internet is taken into consideration. An effort to develop Internet 2 is already under way, mainly in the US. Its capacity will grow and there will be new and better protocols. However, as capabilities to attack also develops, new threats will always have to be dealt with.

In chapter 5, *Software Threats and Vulnerabilities*, it is documented that companies during the nineties have met with an increasing number of financial losses as a result of computer problems, both from accidents and from malicious damage. A 1999/2000 CSI/FBI survey found that 90 percent of the respondents, mostly large corporations and government agencies, had detected security breaches during the previous year. A steadily increasing trend has been remote attacks, usually directed at the Internet connection. This chapter concentrates on remote attacks. The first section deals with who the attackers are, their goals and ambitions. Another section describes a number of common techniques used to penetrate systems. The last part of the chapter accounts for products and procedures to deal with these problems as well as entirely new types of attacks.

It is concluded that effective education of all computer users, including administrators, must be a priority to improve the impact of software tools.

Chapter 6 - *Electronic/Cyberpayment Technologies* - reminds us that electronic payment technology is now used by more than 30 000 financial institutions worldwide. Its fast development offers a lot of convenience, but also big risks, as new forms of crime follow soon after new types of payment are introduced. Some of the new technologies offer unprecedented opportunities for misuse such as money laundering, extortion and electronic theft. The chapter first describes different types of electronic payment systems that are now available. Next, the ideal security properties of financial data transmissions and emerging payment technologies are examined. Third, threats and problems like interception, hacking, denial of service, money laundering, etc. are discussed. The chapter ends by considering what can be done to address these issues and concludes that the future of cyberpayment on the whole looks bright.

Chapter 7, *Managing IT Risks, Threats and Problems*, reminds us that most organisations share a common problem: they seldom fully understand the trade-off between convenience and IT security. One reason for this is that IT people more often are employed for their technical skills than for their understanding of organisational goals. Chapter 7 considers how to manage these and related problems. The author recommends a method that begins with identification of different threats and problems, thereafter ranking them, both depending upon likelihood and severity. Then he examines the controls and procedures that can be used to manage the more serious ones. It should be remembered that it is

seldom practically and financially viable to handle all the threats and problems – if they have low probability and impact it is often prudent to ignore them. The author also discusses how procedures and controls used can be audited, both by using internal employees and external consultants. It is pointed out that the process outlined needs to be iterated and developed at periodical intervals.

In chapter 8 – *Infrastructure's Dependence and Interdepencence on Technology* – consideration is given to what extent infrastructure is dependent on technology and how much different types of infrastructure (Computer networks and the Internet, electricity, gas etc) are dependent on each other. A general trend is that technological developments such as computerisation and standardisation during the last few years have made infrastructural vulnerabilities much more acute. As a result there is an increasing need to look at what constitutes critical infrastructure, how to make priorities and how to develop a working crisis management system. As interdependencies between different infrastructures – national as well as international - have increased considerably, there is a need for a very broad approach.

Chapter 9, *Law Enforcement Issues,* points out that computer related crimes have a tendency to develop just as fast as information technology does. Thus, the means and ways of law enforcement should grow at least at the same speed, which has turned out to be a difficult goal to reach. The problem is made even worse by the fact that there are big differences in local laws. This makes it possible for advanced criminals to route their data traffic through intermediary countries, where laws on technology crimes are very lenient or non-existent. In a systematic account the author first discusses different types of computer crimes such as breaches of logical security, computer virus releases and copyright infringement. Then he takes up the tools and techniques available to law enforcement personnel for gathering evidence and the technical and legal challenges that will meet them in trying to do that. In this connection the author stresses that, because of the complexity of the issues, it is advisable to use combined technical and legal teams from the outset of an investigation.

***

In producing this report, my thanks goes to Ian Dennis and Richard Conroy, who wrote four chapters each. They are both outstanding technicians. Without their competence and stamina it would not have been possible to produce this study. At The Swedish Defence Research Agency (FOI) I have relied on a reference group made up of Henrik Christiansson, Erik Anders Eriksson, Jan-Erik Svensson and Staffan Molin, all senior researchers and technical specialists at the Agency. They have gradually read and made comments on the chapters of this report. Another reason for relying on these three people is that they also should be able to transfer knowledge from the project to FOI as a whole. Even if

it is too early to evaluate the results, I already have noticed dissemination of the knowledge within the organisation.

Chapter 2 - Communication Systems

2.1 Introduction

In this chapter, an in-depth study of different communication systems will be presented, with analysis of their strengths and weaknesses. Consideration is given to how these systems can be used to monitor and collection information covertly and how future developments will affect the security of these systems. We will begin by looking at cable communication systems in use; electrical and fibre optic cable based networks are ubiquitous in modern life and transmit voice, data and television. The legal and illegal techniques used to penetrate and abuse these systems are described with comments on the future of cable within the rapidly growing personal communications marketplace.

The staggering growth of the communications sector has in large part been due to the demand for free-space and satellite systems. The second section looks at the current technology behind radio frequency communication systems, cellular phones and satellite systems. Of particular interest are the cryptosystems used to protect these communication channels, with a detailed description of inherent weaknesses and attacks that can compromise security. The future of wireless communication systems is extremely promising though potentially also pose the greatest threat to privacy.

The second half of the chapter considers the groups and individuals with the potential to threaten communication channels and tries to identify their motivations. Tools are described both for increasing security and for attacking channels. The threat of national signal intelligence gathering is described in more detail to illustrate the methods and tools used in well-resourced attacks.

Beyond the next generation of hardware and software there are several advances which will significantly alter the security of communication systems. In the last section we will consider some of these advances and the implications they have for the planning communication systems.

The aim of this chapter is to survey the general hardware infrastructure used in communication systems and complements Chapters 3 & 5, which deal with data and software security issues respectively.

2.2 Cable Communications

Communication by electrical cables began in earnest with the work of Samuel Morse in the 1850's, who invented the electric telegraph and the Morse code,

used to convey a message in terms of long and short signals over copper cables. These quickly superseded previous systems, because of superior construction and simplicity, pioneering nearly instantaneous communication around well-developed countries. Twenty years later, the telephone, which conveyed speech over the same telegraph wires, was developed by Alexander Graham Bell. Bell also showed that light could be used to transmit speech, a proof-of-principle experiment for optical fibre communication.

Access to electrical cable communication systems is now ubiquitous in developed countries, both through subscriber lines and anonymously through public payphones. Although the technology controlling the system has developed beyond Bell's recognition, the concept of a narrowband, continuous circuit between individual users carrying speech or data has not changed in 150 years.

However with a move towards broadband and burst applications, such as video conferencing, high definition television, interactive television, networked programs and digital information systems, the demands on cable systems are now becoming more diverse. Although the majority of interconnects in first world countries now use optical fibre, the final link to home users is almost exclusively still electrical cable. This provides the main bottleneck to higher data throughputs to end-users and one of the main points of contention between service providers and users. This *local loop* remains monopolised in many countries, for example the UK and Germany, hindering progress towards broadband services but permitting easier access for government agencies. As illustrated in the USA, with up to 60% of households now receiving cable television and a similar percentage using the Internet, the demand for high-speed local-loop and broadband access will force a change in policy, with consequences for information gathering and security.

A change is already visible. The introduction of ISDN (Integrated Services Digital Network) and DSL (Digital Subscriber Line), with the prospect of more advanced ATM systems in the future will compete with the rapidly growing wireless networks [1]. The use of more efficient packet switching networks instead of circuit connections will introduce a number of new challenges, for example in permanent network presence and billing according to bytes transferred. Only with integration to other services, such as cable television and Internet connections, will the telephone remain in its current form; this revolution will undoubtedly happen in the next decade.

There are several important differences between the two major types of cable: electrical and fibre-optic. Electrical cables are inherently insecure because the flow of current, and hence the transmitted signals, can be detected easily using various methods. In contrast, signals in an optical fibre are difficult to detect and

tampering can be easily detected. From a practical perspective, copper cables suitable for high frequency signals are expensive, weighty and lossy and therefore are more expensive to operate, with the need for more repeaters and environmental support. Optical fibres are a fraction of the weight, less susceptible to environmental variables, potentially pose lower health threats and can transmit bandwidths several thousand times that of electric cable without significant loss. Electrical cables are however easily interfaced to existing technologies, pose fewer technical problems and already exist in nearly every modern building.

Currently the most common networks are electrical and optical telecommunications networks and cable television. The next subsections will describe the main features and vulnerabilities of these three systems, with comments on their future development and security issues.

2.2.1 Cable Telephone

Telephone systems are notoriously vulnerable to a wide range of attacks. The size and complexity of these systems make them difficult to update rapidly with new technology, aided by intransigence to anything that would hinder law enforcement or national security agencies. The fact that so many people use cable telephone systems exclusively to communicate makes it a boon for information gathering agencies. In this subsection we will first consider the impact of wiretapping and bugging of telephone systems before considering the response of underground movements, *phone phreaking*.

2.2.1.1 Wiretapping

Telephone conversations are extremely vulnerable because they are transmitted in the clear and therefore it is trivial to reconstruct them from the transmitted signal. *Wiretapping*, using a physical device to intercept the signals on a cable, has been used by all intelligence agencies from the start of widespread public use of phones. In most countries possession of wiretaps and bugging devices is illegal and a court order is required to place a wiretap with one notable exception, the USA.

In the US, the Digital Telephony Bill explicitly allows for authorisation from other sources. The FBI has required telephone companies to establish the capability to monitor 1% of the engineered capacity of their phone systems, and the ability to locate and monitor traffic under the 1994 Communications Assistance for Law Enforcement Act (CALEA). The heavy onus on telecommunication companies to provide this level of *point-and-click wiretapping*

for the security agencies and the negative publicity causing a shift in public opinion, has caused SIGINT gathering plans to backfire. There is currently a legal challenge to this Act by The Electronic Privacy Information Center (EPIC) and the American Civil Liberties Union, supported by several telecommunication groups. As part of this backlash against indiscriminate wiretapping, the Internet Engineering Task Force has rejected proposals to include new standards in Internet protocols that would facilitate surveillance [2].

In 1997, approximately 2.5 million conversations were intercepted on 1186 lines in the USA, with approximately 20% of them providing incriminating evidence [3]. This does not include approximately 800 wiretaps the Americans used for foreign intelligence surveillance or consensual tapping. Altogether there was a 60% rise in the number of wiretaps in the US over a ten-year period, and most western countries have a similar growth pattern. This number substantially rises in other areas of the world, to a speculated 200,000 wiretaps in Mexico [4], where monitoring of political opponents, human rights groups and journalists is widespread. The East German Secret police was also notorious for wiretapping, employing 10,000 people to conduct and listen to wiretaps before the Berlin Wall fell [5].

There have been many scandals caused by wiretapping. The French Socialist party lost the 1991 elections partly because of a leak the week before the election. Telephone conversations intercepted by a special counter-intelligence unit responsible to Socialist President Francois Mitterrand were leaked to the media causing increased suspicion in all quarters. The French government had been in trouble the previous year with the European Court of Human Rights for illegal wiretapping and there is strong evidence there is still substantial wiretapping activity by French intelligence agencies [6].

To detect these activities there are many products available on the marketplace both for tapping, bugging and detecting lines that have been compromised [7]. A wide variety of products are available that can alter a phone's behaviour, including attachments which will make it look as though the phone is not connected to incoming calls, but will allow outgoing calls, attachments to stop the phone ringing and secret microphones which can be activated by an incoming call or which can be used after a call has ended, so as premises and people can be monitored [8]. Counter-surveillance equipment includes line monitors, which monitor the line voltages and currents and detects any frequency variations. These variations indicate cross talk with a R.F. transmitter (bug) and frequency scanners/analysers can then be used to locate the R.F. emissions.

In addition scramblers, from simple voice changers through to more secure systems, are widely available providing security and anonymity if required by a

user. One of the most interesting cases of government interference to maintain access to telephone calls was when the US government paid AT&T to retrofit a limited number of their phones with Clipper, the US escrowed encryption system, instead of releasing a DES-based telephones. The high profile of government intervention and the widespread use of wiretaps and bugging equipment have increased public suspicion encouraging the use of other channels for conveying sensitive information. Where telephones need to be used, the underground movements have developed a range of measures, *phone phreaking*, to combat surveillance.

2.2.1.2 Phone Phreaking

In the 1960's it was discovered access could be gained to the inner working of electro-mechanical telecom systems using various pieces of circuitry, an activity often referred to as *phone phreaking* [9]. These circuits are referred to as *boxes* and came in a range of colours that indicated their use: a blue box generated a range of tones, often 2600Hz, which gave the user free access to the toll trunks for long distance calls; aqua boxes inhibited the 'lock-in' tracing used by the FBI; black boxes stopped callers to your number being charged; red boxes fooled payphones into thinking money had been inserted etcetera. Details of the design of these boxes are freely available on the Internet [10], though are mainly for historical interest because digital ones have almost exclusively replaced the vulnerable electro-mechanical exchanges.

Not to be hindered by these changes, other features of the phone system have also been used extensively by underground organisations for fraudulent use of telecom networks. Call-back features like ANAC, which give the phone number of the telephone in use, number loops where multiple numbers are connected together, customer name and address numbers (CNA) and auto-disconnectors, have all been discovered and used. Useful numbers are discovered with scanners, which auto-dial sequences of numbers to discover what is at the other end. A wide variety of tools are available on the Internet for discovering and exploiting these vulnerabilities [11]. The scanners are similar to the network scanners used to discover potential Internet computer targets and have been fought reasonably successfully by the telecom companies through the implementation of misuse tracking software [12], which can detect extended use of these scanners and unusual behaviour. Using anonymising services, for example re-routing calls through PBXs (private branch exchanges) of large organisations, scanners are still used to a lesser extent. They can also be used for hacking voice-mail and other services by guessing PIN numbers. Access to and monitoring of exchanges can also provide such information directly, though is generally combated by good physical security of exchanges. Physical interception, by wiretapping, remains a constant threat however.

Modern digital exchanges with features such as conference calling, call-waiting and caller line identification (CLIP) can also be exploited. However monitoring of unauthorised use has got more sophisticated, using real-time automatic number identification (ANI) and call rejection / operator intervention to counteract beige boxing. Despite these advances, cable telephone fraud remains a serious problem, costing an estimated $40 billion a year worldwide [13].

The future of the current cable telephone infrastructure is uncertain given the rapid growth of cellular and broadband cable services, which can provide enhanced and more flexible services. Unless the pricing structure is changed to reflect this and more flexible services, such as permanent line presence and high-speed asymmetric data access are introduced, this sector of the communications market will stagnate. It is more likely that integrated service providers will become more prominent offering voice, data and entertainment services through a single portal.

2.2.2 Cable Television

Cable television, formerly known as Community Antenna Television (CATV), has existed since the late 1940's as a means of sharing the reception of a powerful master antenna with subscribers. Terrestrial and satellite receivers, which send up to eighty television channels to subscribers' homes, have now replaced these antennas. The cable used is typically copper coax using analogue protocols to transmit the images, though within the next ten to twenty years there will be a migration to digital technology as digital terrestrial and satellite television become popular.

The cables provide a means for two way communication between the subscriber and provider, opening the possibility for interactive television and controlled access such as subscription channels and 'pay per view' stations (Pay-TV). Cable television also provides both distribution (one-to-many) and targeted (one-to-one) channels, widening the scope of services available. Cable systems now pass by 92% of American homes and a growing percentage in other countries [14], providing perhaps the most concrete example of *digital convergence*. Digital convergence is the concept of all home information systems, such as television, telephone and Internet access, being provided through a common, high-speed portal. Cable operators have already experimented with providing phone services and Internet services on the back of cable TV, making use of the fact that it provides a bandwidth nine hundred times greater than a normal, twisted pair phone cable. Cable modems [15], using this technology, are becoming a popular alternative to telephone modems, because of their higher data speeds and the option to have a permanent presence.

These extra, premium services, which are the main selling point of cable television, are also the main target for abuse of cable television, with a $3 billion black market. Whereas phones actively use the phone network, satellite and cable receivers are generally passive and therefore it is much more difficult to detect fraudulent activity. Therefore greater security of communications is required to ensure proper access to these premium services, in itself an advantage over telephone cables. Methods to get unauthorised access are often found however because of the commercial value of these services using a variety of techniques, some of which are described later in Section 2.3.3.2 which deals more in depth with satellite television.

Cable subscription is likely to increase over the next decade as more of these premium services become available and compete with existing dedicated systems, such as the phone network. Ultimately there will be a convergence between these technologies to provide a single multi-purpose cable for digital, secure communications to meet the demands of both consumers and providers.

2.2.3 Optical Fibre Communications

There are several technical limitations to the bandwidth of frequencies that can be carried on an electrical cable; typically using a carrier frequency of several gigahertz, near the upper limit for an electrical cable, approximately 1GB/s can be transmitted. At these frequencies the signal is highly attenuated, requiring regular electrical amplification over long distances where the signal can be easily and discretely wiretapped. The advantage of optical fibre is therefore immediately apparent: the carrier frequency is increased by four orders of magnitude and transmitted at the speed of light, increasing bandwidth by at least the same margin. Moreover it is extremely low-loss and more difficult to intercept without detection.

The success of optical fibre is highlighted by its dominance in all areas of interconnection, from local area networks to long haul trans-oceanic links. For example, the latest trans Atlantic optical cable, FLAG Atlantic-1 has a bandwidth of 80GB/s (1 telephone circuit uses 8kB/s, equating to 1 million simultaneous conversations), with the proposed TAT-14 having a bandwidth of 1080GB/s, more than 21000 times the capacity of the last trans-Atlantic electrical cable laid in 1983. Further details of the technologies involved in fibre communications can be found in various places [16].

Optical fibres are considered one of the strongest links in a communication system and therefore are rarely the subject of attacks; electronic devices such as repeaters, routers which radiate large quantities of RF energy and their controlling computers are generally targeted first, unless there is a strategic

advantage to attempting to tap the signal from a fibre. There is evidence however at least the US is developing techniques for tapping optical fibres, in particular submarine ones. Fibres can be made more tamperproof than electric cable and their low-loss makes them extremely sensitive to an increase in loss caused by tapping. In addition OTDR (optical time domain reflectrometry), a common fault-checking tool, can detect faults or splits in the fibre preventing unauthorised insertion of a tap and repeater. Nevertheless an extremely sensitive, efficient and fast detector could make use of the evanescent field of the fibre to tap information without detection. The technology involved is considerable and therefore it is more likely alternatives would be sought, such as the co-operation of the PPTs in the countries controlling the fibre or wiretapping the underwater repeaters.

Attacks on optical fibres are possible. The sensitivity of optical fibres makes them very sensitive to the injection of light to scramble the signal, providing disruption. Therefore it is more probable that fibres traversing unsecured areas will be destroyed or disrupted rather than modified for interception.

The future for optical fibres looks impressive. New technologies such as all optical switching and amplification combined with improvements in multi-plexing will improve delivery, potentially with local-loop to end users becoming all-optical. This will replace old circuit-switched networking technology with much more efficient hybrid circuit-packet switching systems, benefiting both providers and users.

Quantum cryptography using optical fibres, potentially the ultimate means of local key distribution, will become a commercial reality in the next decade; quantum technologies will be described in more detail in the penultimate section.

2.3 Wireless Communications

Wireless communication began more than a century ago with the work of Marconi. At that time laying a several thousand kilometre submarine cable was unfeasible, whereas bouncing relatively low frequency (short wave) radio waves off the ionosphere to cover the same distance was a much cheaper and practical solution, permitting reception in a wide area, or footprint. As with the telephone system, the principles of wireless communications have changed little in the last century, though recently there has been a rapid expansion of services and technologies applied to making systems more portable and flexible.

The range of wireless services is immense and pervades most of our lives. The carrier frequency, and thus the bandwidth, can vary enormously depending on

the application required; very low frequency radio waves can be used for underwater communication with submarines, while at the other end of the electromagnetic spectrum, microwave links provide communication channels for satellite and line-of-sight terrestrial communication. Broadcasts are made by a diverse group of individuals, commercial companies, and by the military and diplomatic service, at frequencies defined by their requirements and international / national law.

In this section we will consider three of the main categories of wireless systems: radio, cellular phones and satellite communications. All of areas are growth sectors in the telecommunications marketplace. In particular the cellular phone market has boomed, attracting customers with its miniaturisation and flexibility while increasing the range of services available.

This rapid expansion of the user base has attracted the need for interception, disruption and destruction. A number of basic attacks are common to all wireless systems. The crudest, taken from the military, is to destroy the communications link. This requires the physical destruction of the transmitter or receiver by conventional techniques such as sabotage, bombing or projectile weapons or by more unconventional techniques such as weapons that create massive electromagnetic pulses [17].

Disruption, usually temporarily, is a more common technique, using a high power, jamming broadcast to prevent transmission and/or reception. Such an attack can be conducted remotely, however the position of the jamming signal is easily detectable and the source can be neutralised. It can also be defeated using techniques such as spread spectrum and frequency hopping which push the power requirements of the jamming transmitter to unfeasible levels.

A similar form of attack, which can be more effective, is to saturate the communications channel with traffic. The relatively low bandwidths available and the increase in time for messages to be broadcast can be problematic for portable systems, which are limited primarily by the lifetime of their power supplies. Good traffic management and resource planning can minimize the impact of such an attack, however the threat can be difficult to trace and diagnose.

The biggest advantage and conversely its most serious shortcoming is the ability to receive the transmitted signal within a footprint, often extending over a large area, in part determined by the carrier frequency used. Any receiver, friendly or unfriendly, within this area will receive the signal, making passive interception extremely difficult to detect. Therefore for broadcasting and receiving messages of a sensitive nature, it is necessary to use strong cryptographic protocols to

increase channel security. This was discovered too late, and to the embarrassment of the telecom companies, with analogue cellular phones [18].

The increasing coverage of wireless devices combined with further miniaturisation and an increasing range of services provide an attractive future for the industry. This is highlighted by the ten figure sums paid for the frequencies allocated to the next (third) generation of mobile phones in European countries [19]. Digital radio and direct satellite communication, both with interactive feedback through landlines, will provide complementary and potentially competing technology for the provision of lifestyle and entertainment services. Integration with other services is proposed, for example with electronic cash, to expand the user base by making your mobile phone a potential replacement for your wallet.

2.3.1 Radio Frequency Communication

Radio communication began over a century ago though was quickly superseded in technological terms by telephone and television; more recently satellite communication systems and mobile phones have pushed radio technology to higher data bandwidths. Radio has however enjoyed a resurgence lately, with growth in a number of areas. The underlying driving force for this resurgence is the application of digital technology and the competitive data rates available compared to standard telephone lines. In addition, the broadcast nature (one-to-many) is attractive for some services.

By replacing a modem with a terminal node controller (TNC), and a telephone with a radio set, it is possible to surf the Internet over the airwaves. As with the Internet, computerised radio systems were developed for ARPANET in the 1960's. Packet radio [20], as the public version has become known as, is a fast growing area because of the low cost involved, and the previous speed limited of 9600bps is being pushed back consistently. Currently there are a range of projects using and developing MB/s and GB/s links [21], with slower speeds still available for wider area and lower frequency links.

The second advantage of packet radio is that multiple users can share one channel with a range of at least 100km, depending on the carrier frequency and the presence of repeater nodes [20]. Time division can permit one-to-one and one-to-many broadcasts, analogous to cabled computer networks. The frequency bands also allow multiple channels, akin to wavelength division multiplexing in optical fibres. Spread spectrum techniques can reduce the power levels sufficiently to bypass regulatory agency licensing while improving signal strength, privacy and multiplexing at the technological cost of synchronisation.

The primary problems with packet radio are variable attenuation and multi-path interference of the signal, unidirectional transmission or reception, and the potential for man-in-the-middle attacks.

The military, the original customers of packet radio systems, have a substantial lead on the technology in this field [22], though the indications are that this technology is flowing more rapidly to the public arena. This will manifest itself in the appearance of mesh networks and hierarchical networks which can provide a greater variety of services and greater interconnection with other services. Improved techniques for managing noise and interference, including broadband amplifiers, adaptive notch filters and improved signal processing with advanced aerial design and multi-antenna reception systems will help to make these developments possible [23]. The multi-antenna reception is an important development for use in heavily built up areas, where most of the potential consumers live. The commercial availability of VLSI chips, millimetre and microwave technology will permit personal mobile networks, and integration to the level proposed with Bluetooth, described later.

The other large area of radio frequency wireless communications is in personal, portable communications equipment. For example, two common items, cordless phones and radio pagers use radio frequencies (RF) to transmit their signals. The unencoded transmissions of cordless phones can easily be intercepted at short range using a suitable antenna. The frequency schemes of various phones are readily available on the Internet [24], and with a scanner, phone calls made on these frequencies can be easily intercepted or made. Most cordless phones use analogue technology with no security features and this is an obvious area for future research in order to improve home security. The DECT [25] (Digital Enhanced Cordless Telecommunication) protocol, using a 64-bit session key, is a promising development which will permit multiple handset interfacing, including cordless and mobile phones with a base station which can be connected to PSTN and ISDN systems.

Radio pagers, commonly used for passing short messages to their owners, are also vulnerable to interception. Programs are available on the Internet which can decode the plaintext signals to recover the transmitted text using a scanner and computer [26]. There is widespread interception of pager messages by law enforcement agencies [27], SIGINT agencies [28] and underground groups [29] because of the weak analogue protocols used.

It is difficult to assess the future of pagers, given the diversification of services offered by other technologies, in particular the short message service (SMS) on cellular phones, considered in the next section. The simplicity and portability of these devices however works in their favour, as does their integration to provide short messages to roaming computers and individuals over a wider range than

mobile phones. With over 40 million pagers in use in the USA alone, the limited numbers of ways in which the system can be abused by the user is attractive for employers to prevent abuse, significantly more difficult to do with mobile phones. Therefore it is probable pagers will remain popular, however there is a need for the implementation of digital and secure protocols.

2.3.2 Cellular Phones

The growth in the use of mobile phones in the last decade has been staggering, mirroring independently and exceeding the growth of the Internet, with over 300 million users worldwide, a figure expected to reach in excess of a billion in the next decade [30]. This growth has been made possible by a range of technologies to allow mass production of miniaturised, low power phones and by heavy investment by telecommunication companies to install base stations which interface to land-lines. Particularly in Europe, with its relatively high concentration of people, has the impact of cellular, or mobile phones been most felt, with more than 50% ownership in some countries [31].

The first generation of phones used analogue technology to send and receive calls, but has been superseded by second-generation digital phones, currently dominating the market. The third generation of wide access digital phones offer a greater range of services, including the Internet, and is beginning to reach the marketplace. In this subsection we will consider each of these generations, their strengths and weaknesses, and the future of cellular phones.

2.3.2.1 Analogue Mobile Phones

Cellular phones, developed out of an increasing demand for flexibility in phone systems, in particular for communication while moving and for remote areas. The earliest systems, in the late 1970's, used analogue radio technology to communicate with base stations that patched the user into the cable telephone network. These systems were assumed to be reasonably secure because the technology required at the time to intercept the transmissions was expensive and complicated to use. Similarly there were few users due to the expense, no-frills and complication of the systems. Therefore they did not attract much attention until the mid-to-late 1980's.

As the user base increased in the mid-1980's, the billing system quickly became integrated into the handsets through a 32-bit serial number burned into the phone and the phone number issued to the phone. Initially the default was to permit all calls through the base stations except from a list of known bad numbers. This was eventually changed so that positive identification was required before the

call could be connected. Although the later seems a more logical starting position, the initial small user base and the difficultly in implementing such an authentication system led to the former being a compromised starting position for the sake of speed and ease of use. This loophole was rapidly and extensively exploited until it was closed with 'tumbling' phones that would generate random serial numbers each time a call was made and because of the default negative authentication the call would always be connected and could not be cut off.

By the late 1980's radio frequency scanners had become widely available, typically for a few hundred dollars, and they could be used to intercept the unencoded cellular phone transmissions at a distance of several miles. In the USA alone, there were more than 10 million scanners compared to the 50 million cellular phones, indicating that nearly everyone who had a mobile phone was likely to have had a call intercepted. Scanners could also be used to carry out more sophisticated attacks, such as position location and tracking, though there is not much evidence of this in the public domain.

The other main use for scanners was to obtain valid serial numbers that could then be programmed into unauthorised phones and used, or as it is commonly called 'cloning'. Black boxes were widely available on the black market to automate this process, with airports and motorways as favourite places to scan for serial numbers. Often cloning was combined with tumbling and roaming away from the home service area to make detection harder.

Although a number of technical countermeasures, such as traffic analysis, RF fingerprinting and PIN codes have been used, they have not provided any robust defence against the rising tide of phone fraud with analogue phones. Legal measures, such as banning scanners and cloning, requiring tamperproof phone serial numbers and stricter pending legislation have had limited success in curbing fraud. Fraudulent use of phones is now widespread and is estimated to cost nearly $1 billion per annum, with approximately 5% of all calls made fraudulently, rising to 60-70% in some areas.

Therefore the cellular phone companies were forced into digital technology using cryptographic authentication to combat this growing problem. This reactive development of the security architecture rather than proactive development has been a side effect of the rapid growth of the cellular phone industry and the realisation of the target-rich and huge economies of scale afforded to attackers. A partial implementation, Cellular Digital Packet Data (CDPD), was experimented with, but it was successfully broken [32] before fully digital phones were implemented.

2.3.2.2 Digital Cellular Phones

Moving to digital technology permitted a number of other advances in addition to security. The flexibility of phones in terms of range, clarity and lifetime was increased with digitisation through improved technology and error correction. Interfacing to laptop computers was also simplified with the introduction of digital cellphones, increasing the mobility, scope and range of information services available.

The services offered by these phones are comparable to small portable data assistants, including address books, schedule managers, games and short messaging service (SMS). Of particular interest financially to the phone manufacturers has integration to other services, specifically electronic mail and Internet access. Access to additional electronic services has already been introduced with varying success. Wireless Application Protocol (WAP) has been pushed heavily in Europe but to date has failed to catch subscribers' imaginations, whereas i-mode in Japan has succeeded in signing up 10 million subscribers in a 14 month period [33]. Both services offer Internet access using modified HTML protocols, though the small text area, difficulty in entering messages and low data rates (9.8kB/s) has put off European customers, a phenomena not experience in Japan due to cultural and lifestyle differences. The use of a packet switching mobile phone network, the rarity of home computers connected to the Internet and the high charges of NTT have also contributed to the Japanese market growth. Of concern is that WAP security is provided by the WLTS (Wireless Layer Transport System), which although based on the well-studied TLS protocol, has been shown to have a number of weaknesses [34]. In the case of both WAP and i-mode, the enhanced access and security given by the next generation of phones will rapidly supersede these services.

Worldwide, several digital protocols are currently in existence, the most prevalent of which is GSM, used by over 80 million subscribers in Europe. The security for GSM is comprised of several algorithms. A3/A8 is used for authentication and session key generation. Service providers have the freedom to use their own algorithm within this; the most commonly used is COMP128 [35], which has been broken and is a point of vulnerability [36]. A5 is used for the encryption of the radio link and comes in three flavours of increasing strength, the strongest being the A5/1 algorithm, which uses a 64-bit key. Even with a brute force attack, computing the secret key from the phone transmissions is not impossible as the 64-bit key is normally crippled to 54 bits, at the request of law enforcement agencies. It has been reported that real-time decryption of the A5/1 algorithm is possible using a single PC within two minutes of the beginning of a phone call using a number of weaknesses, though this has not yet been demonstrated in the field [37].

A more concrete example of their vulnerability is the cloning of a GSM phone, using a simple hand-held electronic organiser to provide the SIM information. It should be noted however that it needs to clone the information from a pre-existing SIM card, rather than being able to reverse engineer numbers from the emissions picked up with scanners, as for the analogy phones [38]. The secret key can also potentially be recovered from the SIM in a number of other ways to make all sessions insecure. These include using smart-card hacking techniques, monitoring the RF frequencies emitted by a phone, mimicking a base station [39], and asking the card nicely!

In the USA, the most common systems are TDMA (time division) and CDMA (spread spectrum). All the North American systems use roughly the same security framework, relying heavily on cryptography for the radio-link security. There are four main security algorithms employed, CAVE for authentication and key generation, an XOR mask for voice encryption, CMEA for control channel encryption and ORYX for wireless data encryption. Of these, CAVE is the only one to remain unbroken. This is largely attributed to the closed-door design policy originally used, and the approach of security through obscurity.

The largest weakness in the mobile phone system is however the microwaves links between the base stations and the network. They do not use the same security and encryption protocols and therefore are potentially easy to intercept with an antenna in the beam path. Although there are no reports of major attacks on these links from the civilian sector, governments certainly monitor such links [40].

History is likely to repeat itself in the next decade, with digital scanners becoming less expensive and computer power increasing to allow real-time decryption of current algorithms. Digital cellphones are also generally dual-mode and can drop back into analogue mode outside the range of digital base stations, compromising security and offering another form of attack, by flooding local digital base stations. Using a fake mobile base station to do this and provide real-time decryption also poses a significant threat given the relatively low cost, $10,000 [41]. Perhaps the largest threat is that the telecommunication companies will not use open standards for the third generation of mobile phones and similar vulnerabilities will be found.

The larger user base has also shown concern about security and possible backdoors into their systems for use by law enforcement agencies who have the ability to monitor hundreds of calls simultaneously. As with the cable telephone network, the underground organisations are only now beginning to understand the full potential of digital cellular phones. It has been reported that these phones contain many hidden features, including the ability to turn phones into digital

scanners, which in turn could be used to monitor the whereabouts of other users in the same cell [42].

Although substantially more secure than analogue phones, the potential for fraud and abuse of the digital cellular phone network is growing daily. This is however being combated by intelligence traffic monitoring [43] and improved security, which should curb fraud at an acceptable level for service providers, significantly below that of analogue phones. The improving data capabilities and potential for third party encryption techniques to be used with the phones is another potential boon for the industry and the security and privacy of subscribers. The demand for further services, in particular interconnectivity will soon see these phones superseded by third generation mobile phones.

2.3.2.3 Third Generation Cellular Phones

The strength of interest in running the third generation of mobile phone networks caught many by surprise, and generated substantial windfalls for European governments [44]. The GSM protocol of second-generation phones will be replaced with UMTS [45] (Universal Mobile Telecommunications System) giving faster, higher-capacity connections, enabling voice, video and Internet services to be sent to mobile phones. The use of the General Packet Radio Service (GPRS) will provide greater integration with other wireless services and the advantages of a packet switched network. Better data storage and interfacing abilities will also lead to greater integration with other services. Indeed, there is a growing market in m-commerce (mobile commerce) similar to e-commerce, making use of these facilities.

In the short term the data rates of these phones will probably be limited to 56kB/s to reduce radiation exposure [46] however within 2-3 years this will probably be increased up to a maximum of 2MB/s, depending on demand and technical constraints [47]. These bandwidths will permit streaming video, CD quality music and fast Internet surfing.

However the power of these phones is making a number of groups uneasy [48]. Tracking software can locate a phone to within 15 metres, useful for emergency services and location based services, could also be used for unsolicited advertising, services and tracking without user control [49]. The need for the telecommunication companies to recoup their investment, as much as $600 per head just for the license, will put pressure on them to make as much money as possible.

Using mobile phones for commerce, either as electronic cash machines or to purchase items over the network, is also of concern because of the gap between mobile phone encryption techniques and the protocols used by banks and

financial institutions. This requires intermediate decryption and hence vulnerability of the transmitted financial data. The encryption and security protocols themselves have been developed in private and as mentioned before, if services are no available at a higher level, lower level weaker standards will be used. However vulnerability to attacks such as fake base stations should be lower. A preliminary analysis of ciphering in GPRS and UMTS has already been reported [50].

The Finnish government are already investigating the migration of its FINEID smart card program to include SIMs [51]. The need for secure interactions with databases to access sensitive material and government encouragement/legislation will provide both opportunities and threats. *Cryptonomicon* paints one future with offshore data havens where information can be stored away from the prying eyes of governments, security agencies through to insurance companies. Such easy, and potentially unauthorised access to wide ranging information is of understandable concern and will need to be addressed in order to win over public confidence.

The rate at which new technology is being released in the marketplace is also of concern, particularly where propriety, private protocols have been used. Currently the lead-time between development and commercialisation is 2-3 years in contrast with the 35 years for radio and 12 years for television. Inevitably such a frantic pace will lead to errors and weaknesses.

Undoubtedly 3G phones will be scrutinised as much as the general public buys them. Their extended range of features will be a boon for all interested parties: law enforcement agencies, users and underground groups alike. Existing and new vulnerabilities will be demonstrated, but fraud is unlikely to be achievable to same level as with analogue phones and kept to an acceptable level for the telecommunication companies. Security concerns amongst users are likely to increase as integration of services, in particularly financial ones, increases. Similarly the economic potential will give rise to phone viruses and worms, which can infiltrate phone handsets to access sensitive information.

The future for cellular phones is very promising with a clear roadmap of future advances, in contrast to cable telephones. Security and health risks are the major concerns of users while providers are concerned about increasing revenue through offering premium services. The coverage of mobile phones will never be universal however and therefore there will always be a niche market for satellite communication systems.

2.3.3 Satellite Communications

The use of satellites for communication purposes has rapidly grown over the last 30 years. The range of users varies from individuals, to companies and the military. The impact of satellites is easy to forget given all the recent advances in technology, however the first satellite, Telstar, launched in 1962, provided the first broadband intercontinental communications channel. Satellites today remain the preferred means of broadband communication and play on their strength of reaching parts of the world other communication systems cannot.

Satellites are now routinely used for relaying television and radio stations, phone and paging calls, data transfer, video conferencing and information gathering. Other applications include asset tracking (e.g. cargo ships), distance learning, remote data collection (e.g. marine weather stations), search and rescue, and communication links for oil rigs and expeditions. The 'parking lots' used in space for the satellites are dependent on these applications:

1. for continuous coverage in a single area, such as needed for television, a geostationary orbit (GEO) is used at a distance of 35786km; large (>30cm) dishes are required for reception and larger for transmission (>1m), however the direction remains fixed without need to switch source

2. if two way, low-power communication is required, for example for satellite phones, low earth orbits (LEO) at a distance of 3000-4000km are used; compared to GEO the signal strength is three orders of magnitude greater, permitting smaller aerials (~cm size) however handover between satellites which are only "visible" for twenty minutes per orbit can be problematic

3. for low-power receiving systems such as the global positioning system (GPS) intermediate circular orbits (ICO), or medium earth orbits (MEO) at a distance of 10000km. Providing complete global coverage is cheaper than LEO with similar sized aerials if reception only is required

4. highly elliptical orbits (HEO) are used by the Russians to provide coverage for their northern cities which cannot economically be connected by cables. They provide a compromise between GEO and LEO systems

5. sun-synchronous orbits and dawn-to-dusk orbits are special LEO/MEO orbits with 24 hour periods to provide repeated daily coverage, particularly useful for scientific data collection.

The broadcasting footprint can be varied from an area the size of the USA to less than one hundred kilometres, depending on the application's requirement, again affecting reception and transmission powers required. Commercial frequencies vary from the L-band (0.39-1.55GHz) through to the K-band (10.9-36GHz) again tailored to the application requirements; for example satellite phones use the lower frequencies, television the higher frequencies. The International Telecommunications Union (ITU) keeps track of satellites in orbit [52] some of which it is possible to buy/rent the use of bandwidth on [53].

The primary advantages of wireless and in particular satellite communication systems are that transmission costs are not distance sensitive and point-to-multipoint broadcasts are possible with a large bandwidth and tends to be very reliable. Timing delays (0.5 seconds) for the uplink and downlink can be significant for some applications and the requirement for licensing by regulatory agencies are again problems common to wireless applications, particularly to satellite systems. The most significant disadvantage is of course the cost of launching a satellite, though in real terms this has decreased by two orders of magnitude in the last thirty years.

The high costs involved in providing the infrastructure and reliance of satellite systems led to it being reported in the British press in 1999 that a UK military communications satellite was hacked and used to blackmail the UK government [54]. However these reports are unsubstantiated and unlike to have happened because of the control systems used. Of greater threat is access to the uplinks and transponders on commercial communications satellites. A more concrete example of this threat was the arrest of Jason Diekman for hacking into the NASA satellite control system amongst others [55]. Although no lasting damage was done, the prospect of holding a several million-dollar satellite for ransom is a significant threat. As the integration of different media increases, the increasingly common use of satellites, for example in providing Internet connectivity [56], carries with it a greater risk of hacking of these services, which would in turn provide a powerful broadcast medium for any group wishing to use it.

The cost of launching a destructive or disruption attack is prohibitive for most groups and individuals and therefore the most common attack is however passive interception and unauthorised decryption, in particular for satellite television. In the next two subsections we will consider two of the major growth areas in satellite communication systems, phones and television, and prospects for their future growth and security issues.

2.3.3.1 Satellite Phones

The prospect of global personal communications using small, hand-held terminals was unthinkable until the last decade and was consigned to the realm of science-fiction. The cost of providing such a system has dropped in real terms while demand has risen, leading to a number of systems either in planning or in the early stages of operation. To provide such a system, many low earth orbit (LEO) satellites are required, for example the ill-fated Iridium system used 66 such satellites (though originally planned to have 77 and named after the 77^{th} element).

A number of features were included in the Iridium satellites such as inter-satellite links, phased array antennas, and data and pager services to increase their attractiveness and competitiveness with terrestrial mobile phones. Current satellite phones can be used with both satellite and terrestrial cellular systems (through the GSM protocol) though they cost in excess of $1000, significantly more than cellular phones. The other major player, Globalstar does not have inter-satellite links, and therefore can use less (48) satellites, at the cost of an increased demand on land-side technology and with less security because of the two-way link to space. It will however offer a better data/voice mix than the poor 2400 b/s offered by Iridium.

To handle data communication Motorola, amongst others, have invested in several data specific satellite constellations. The M-Star system, consisting of 72 satellites will offer a 160Gb/s service while the Celestri network of 63 satellites will offer a broadband services such as video-on-demand and act as a second generation Iridium. However analysts believe the demand to launch so many satellites into low earth orbits cannot be met by available launch systems, and therefore some compromises will have to be made [57]. Indeed Iridium LLC collapsed in 1999, ICO Global Communications was restructured and Globalstar is currently undergoing financial hardship [58].

Funding however continues to be forthcoming for many new LEO projects. In the last three years, Craig McCaw, Bill Gates and Ed Tuck have teamed up with Boeing to plan the ten billion dollar, 288 satellite, Teledesic constellation in the Ka-band, which will offer an "Internet in the sky" using advanced, smart, routing technology and with channel rates up to 1.2Gb/s [59]. Launching is scheduled to begin in 2002 and services begin in 2004, though any financial intervention of McCaw in ICO Global Communications and/or Iridium LLC may lead to a monopoly of commercial satellite services in the near future.

With no complete, functioning system in place yet, it is difficult to assess the risks and security of personal satellite communication systems. However with the fuss being made by national security agencies over the technology and trying

to bind national laws and restrictions onto the service providers, the systems will inevitably be compromised in the same way as cellular phones. In the US the FBI have held up licenses being granted to a number of companies while wiretapping issues are dealt with. The problem arises because of the borderless nature of the system, but it is not a borderless world from a legal perspective, a common problem with Internet issues as well.

The military have a number of dedicated satellite communication systems for command and control purposes [60]. Portable systems are typically supplied with modular encryption units, such as VINCENT (Voice Encryption Terminal) used by the US. Interestingly there is also evidence that the military use civilian satellite phones for battlefield operations [61].

There will undoubtedly always be a demand, albeit limited, for satellite phones and data connections, particularly in large developing countries [62] who do not have the resources to develop a complete cable network. Security is comparable to mobile phones and it must always be assumed channels can be intercepted, in particular by intelligence agencies. The small size of the user base has limited interest in research into cloning and hacking of phones, compounded by the technical difficulties added by the use of CDMA (code division multiple access). The encryption protocols, as with cellular phones, are proprietary and therefore potentially weak and can be easily be intercepted passively. The flexibility of design should however be noted as it can be adapted to 3G cellular phone technology and new protocols without hardware changes on the satellites.

The whole face of satellite communication may be revolutionised if one of the large conglomerates can successfully entice users for their LEO systems after ironing out the existing problems and offering services comparable to terrestrial wireless services. With a large increase in the user base, for example using 'Internet in the sky', closer inspection of the security systems is inevitable with security breaches. The cost of any intrusions could be significant for the operators and therefore it is essential there is proactive planning.

The investment of these conglomerates may also be misplaced with the rapid increase in bandwidth of terrestrial services. One possible alternative route to supplying mobile Internet is to use GEO satellites with cellular phones, because for most users the downloaded information is greater than uploaded information. Data can be multiplexed on top of TV signals, as with cable modems, providing high bandwidth data streams. Indeed, some satellite transponders act as dedicated network relays and for private video conferencing. The future of LEO satellite systems will either take off dramatically or fold as Iridium has. Identifying a substantial target audience, as with satellite television discussed next, is a vital prerequisite.

2.3.3.2 Satellite Television

From the first satellite broadcast of Elvis singing live, there has been an explosion of extra-terrestrial television, forming a hundred of billion dollars business. The primary attraction is access to hundreds of million customers with a single transmitter, substantially less than covering the same population with terrestrial transmitters. The higher carrier frequencies used also permit multiplexing, providing up to a hundred channels from a single satellite. In order to increase revenue, operators encrypt premium channels, which can be decrypted using a smart card purchased from the operators for a monthly fee.

These extra, premium services are the main target for abuse of cable television, with a $3 billion black market. It forms one of the biggest battlegrounds between cryptographers and cryptanalysts. There is a lot of commercial value in pay-TV, and whereas phones actively use the phone network, satellite and cable television receivers are generally passive, and therefore it is much more difficult to detect fraudulent activity. Operators have found however that electromagnetic pulses can disable counterfeit chips while not affect genuine decoder chips and have made several high profile prosecutions [63].

The Pay-TV control access control system varies according to receiver technology used. The majority of receivers are still analogue based, and basic techniques, such as removing the horizontal and vertical synch signals, are used to scramble the signal. These however can be easily recovered to restore the picture. More complicated schemes have been employed, using smart cards to provide a control key that is used to decode the signal, which has been scrambled using a digital frame buffer. VideoCrypt and EuroCrypt are two such systems, and although more difficult to analyse and reverse-engineer the decryption key, there is still widespread fraud of these systems. Detailed description of these systems and how they can be attacked can be found without much difficulty on the Internet [64,65]. These attacks vary from simply adding extra conductive tracks to the smart cards, to using computers to capture data sequences and replay them to the decoder, and to more hi-tech solutions such as focussed ion beam etching. Many devices to bypass the encryption using these techniques can be purchased both through black market magazines and the Internet.

For the emerging digital TV systems, the broadcast signal is digitally modulated, encrypted and multiplexed, as for example in the DVB, or DSS/VideoGuard systems. These systems haven't fully been put to the test yet, however it can be assumed that with a multi-million dollar market, all digital decoders will be reversed engineered with weeks of release to look for vulnerabilities. There is already substantial information available on the Internet on the DSS/Videoguard system with hacks [66].

Ultimately, as with computer encryption the power of the protocol depends on the size of the key space available and flexibility of the system. Customisation and obscurity have been shown to provide little benefit from such sustained attack. For proper protection of premium cable and satellite services, publicly scrutinised algorithms will have to be employed after extensive testing of possible cloning and bypassing techniques of the encryption.

The satellite television is well-established and keeping pace with terrestrial technology developments. One area in which it cannot compete however is in truly interactive services where some use of terrestrial channels is required. Ultimately an integrated entertainment and information cable portal is the largest threat to satellite television operators, however licensing agreements with cable operators and the demand for television in sparsely populated areas will ensure future income. In contrast perhaps the most exciting development for mobile interactive services will perhaps be in short range systems described in the next subsection.

2.3.4 Other Wireless Services

A number of other wireless services exist for a range of applications. Of particular interest for computer networks are IrDA [67] (Infrared Data Association) and Bluetooth [68].

IrDA is a master/slave protocol for infrared communication links in a 1 to n (1 master device, n client devices) fashion. Initially operating at 115.2kbits/s, the latest version operates at 16Mbits/s using a rather complex layered protocol stack. This system is implemented on many laptop computers and personal data assistants for communication with a base computer and therefore the transmissions are likely to contain sensitive information. There are no security procedures, in part because tapping is a non-trivial affair (interception needs to be within a few metres and within a 30 degree angle from the transmitted signal). Using a mirror or similar device however sufficient signal could be recovered to intercept data.

Bluetooth is an open specification for wireless communication of data and voice. It is based on a low-cost, 2.4 GHz short-range microwave radio link, and supports point-to-point and point-to-multi-point connections. Bluetooth, which offers a transmission rate up to 1 Mb/s, is targeted for computing and communication devices like desktop computers, printers, fax machines, mobile phones etc. The protocol features stream cipher encryption with key lengths from 8-128 bits and 128-bit symmetric key, challenge-response user authentication. However, encryption key length is preset in each individual device, and cannot be overridden by the user. Furthermore, each device has a

random number generator of its own. Key management is left up to calling applications. Bluetooth is effectively an extension of the IEEE 802.11 standard for wireless LAN that supported 40-bit RC4 encryption.

The use of Bluetooth is likely to influence many other spheres of life. It has been proposed that transponders could be included in road side signs to provide traffic and local information or to stream information and music to people travelling underground. Large consumer electronic companies like Ericsson and Electrolux believe it will also be included in all household electrical items, from fridges to washing machines to monitor and control the home environment [69]. Incorporation into PDAs, laptop computers, phones and gaming consoles is already planned, suggesting Bluetooth may become one of the most widely used wireless protocols, bypassing the cost and inconvenience of cables.

The increasing demand for personal roaming access to services is likely to see a rise in usage of these systems, in particular Bluetooth. As with other wireless systems, interception is always a concern and therefore it is essential that strong encryption protocols are in use, which provide sufficient security. The entire spectrum of treats to communication systems will be considered more fully in the next section.

2.4 Threats to Communication Systems

In this section we will consider the main threats to communication systems in terms of the types of attacks that may be performed and potential attackers. Apart from military and intelligence network channels, most channels are inherently insecure and it should be assumed that all communications could be intercepted. In addition, and more obviously, a communication channel could be disrupted or destroyed by an attacker. We will first consider the groups and individuals who may be motivated into attacking a communications network, before considering the techniques they are likely to use.

2.4.1 Actors

The groups and individuals posing a threat to communication channels can be broken down into a number of major risk levels:

1) **Common Consumer** – this group represents the general public and their access to devices and techniques to bypass protection and / or gain unauthorised access to services. Primarily this group is passive, using technology passed down from higher levels, such as smart card cloning and phone phreaking, with the main risk being financial loss to service

providers. Active attacks are rare while frequent technology changes and usage monitoring generally deal with passive attacks.

2) **Amateur** – enthusiasts with recreational interest in telecommunications pose a greater threat, developing techniques to counter protection. Generally their budgets are small, limiting their scope, however they typically have a wide knowledge [70] and access to common equipment. Primarily their targets are equipment modification and passive interception either for personal interest and / or small-scale financial reward.

3) **Restricted Professional** – individuals who make a living from the telecommunications industry pose several threats. The first is that sensitive information regarding the telecom networks can be unwittingly or deliberately disseminated to unauthorised entities; the former by a careless employee not shredding sensitive documents, the later by a disgruntled employee. An employee with a grudge could also potentially booby-trap or trojanise software controlling networks for his or her own purposes. An employee can also use their technical knowledge to gain personal advantage, for example free-access to pay services for themselves and others. A more grey area is where a former employee goes into business for themselves, for example as a surveillance expert, utilising their knowledge. In this case they have access to the full range of production equipment and have an excellent working knowledge, making them a competent threat. They primarily commit active attacks, which can be detected by implementing good management and codes of practice.

4) **Professional Organisation** – multi-national companies often employee or have internal departments responsible for gathering intelligence. These groups have access to substantial budgets and the latest equipment and the potential to develop their own specialised equipment. They offer a formidable threat backed with substantial resources generally targeted at other organisations that pose a financial threat. Again the attacks are primarily active in order to penetrate a hostile environment, however their secondary task can also be to monitor and audit internal communication network usage.

5) **Intelligence Agency** – the highest risk is posed by governmental intelligence agencies that have essentially unlimited resources, access to cutting edge research technologies and can operate above the law to some extent. There is strong evidence that both random and targeted SIGINT operations are carried out by these organisations, the targeted operations being very intensive against groups and individuals posing a national security threat (e.g. terrorists).

Rogue motivations of individuals and small groups also need to be considered. For example, revenge for a perceived injustice (e.g. sabotage), terrorism of individuals or minority groups (e.g. malicious phone calls), poverty (e.g. stealing equipment), corruption of employees by moral or financial means (e.g. moles leaking sensitive information) and poor security procedures (e.g. careless disposal of sensitive information) all need to be considered when assessing risks and threats of communication systems.

2.4.2 Attacks on Communication Systems

Having identified the groups and individuals likely to pose to a communication system, it is important to consider the techniques they will use, the frequency of attack and the impact of the attack. These attacks include:

1) **Destruction** – This is the most basic form of attack, destroying the physical means to communicate. The risk and low intelligence returns of carrying out such attacks limit its use general to a last resort, or as an emotional or warfare response. For example destruction has been used in cases of industrial espionage to prevent a competitor gaining an upper hand in a crucial deal and during international wars to prevent communication between enemy units. Destruction can also be of limited value because of fast replacement of hardware or alternative routing of communications. However it can be beneficial to mask other activities that would normally be detected by SIGINT. The threat, mainly from professional organisations, such as terrorist groups, and foreign powers, can be countered by adequate physical security at communication installations and system management or resources to provide alternative systems. Small-scale destruction and sabotage can take place, particularly with cable networks, where access to cabling is easier than to protected switching and broadcasting hardware. Use of line testing equipment and multiple routes can minimize the risk of data loss, however it is more likely eavesdropping would be goal of an unauthorised entity with access to data cables.

2) **Disruption** – Disruption of a communications system is generally more productive than destruction on a small scale. Disruption can take several forms, from physical devices, such as large amplifiers to drown out the signal, to software-based denial-of-service attacks. Disruption can be as effective as destruction particularly when the origin and nature of the attack cannot be traced with the possibility for disguising the attack as a normal system failure while carrying out the attack from a remote location. For example a computer controlling network traffic could be sent a 'ping of

death' causing it to go offline, during which time traffic is no longer routed or an alternative unauthorised server could provide rerouting of data. Hardware disruption is normally a resource consuming activity and therefore limited to professional and national agencies. However software attacks, in particular remote denial-of-service attacks have been growing in frequency from some threat groups. Preventing disruption requires careful management of resources, using techniques such as spread spectrum broadcasting and firewalls to minimize the impact of disruption attacks on normal data flow.

3) Active (Physical) Interception – Where undetected access is available to networking hardware or a communication channel, an unauthorised entity is likely to employ a device to intercept transmission in the first instance to gain intelligence. Hardware devices used include wiretaps, bugs, rebroadcast antennas, re-router amplifiers and laser vibrometers; software programs used include packet sniffers, network analysers and traffic analysers. Possession and use of hardware devices is illegal in most countries and therefore typically only used by professional level groups and individuals. Detection of hardware wiretapping is difficult, particularly where the cables cross through public areas. Software wiretapping tools are readily available, for example on the Internet, and therefore are in more widespread use while being equally difficult to detect in a high data volume network. Network traffic and channel monitoring are the most effective ways of detecting interception, as any intelligence gained needs to be passed back to the infiltrators in some way.

4) Passive Interception – When a channel's data flow 'leaks' into a larger footprint covering public areas, interception is possible without the need for physical intervention, minimizing the risk of detection. For wireless communications this can take the form of a satellite antenna in a public place but within the footprint of the satellite transmission; in computer networks this is equivalent to installing packet sniffing software on a friendly machine. The most effective tool in combating passive interception is encryption, because decryption to recover the plaintext of the channel requires the decryption algorithm and keys used and / or an understanding of other languages, acronyms or abbreviations. Hardware techniques for encoding the information for transmission such as spread spectrum, multiplexing and burst transmissions can also provide additional protection. The latter, favoured by the military for delayed, long distance communications, is equivalent to the software concept of steganography, where the information transmitted is disguised by other traffic or noise. Passive interception is the most common threat and used by all threat groups. It is also the most difficult to detect because of its passive nature and is difficult to protect from in law.

5) **Equipment Modification** – Easier access to plaintext data and services can often be achieved by modifying existing equipment for the communication system. This can take the form of smart card cloning, protection circuit bypassing, disabling of encryption and alteration of signal levels for increased interception range. Modification requires technical knowledge of the communication system and the protocols it uses and therefore limited to groups and individuals with sufficient familiarity and resources. Protection against modification is a complicated battlefield, as illustrated by smart cards, in particular those giving access to pay TV services and those used in mobile phones. The cloning of these cards and accessing their secret keys has been cunningly achieved using a variety of techniques, reducing the complication to the ability of amateurs in some cases. This is inspite of, or perhaps prompted by, attempts to introduce tamperproof equipment, frequent key changes and price differentials.

6) **Sensitive Information Revelation** - The leaking of information regarding the operation of a network can be extremely damaging. This can be either deliberate (e.g. corruption) or accidentally (e.g. social engineering) done by employees or by groups or individuals who have researched the hardware used (e.g. weaknesses in the encryption algorithms used). Mass media stories revealing abuse of network services, such interception of analogue cellphone messages and phone cloning, can have a negative effect in public perception and affect the user base.

Assessing the impact of these attacks and the frequency at which they are likely to occur provides a good basis for developing security and disaster management protocols. By factoring in availability of resources, threats and risks can be minimized to an acceptable level. Proactive planning to include transparent security measures with proven strength and regular assessment of new risks needs to be undertaken in order to maintain customer confidence.

New systems need to take into account their target customer base and their level of sophistication, a level that is consistently rising, to ensure minimal abuse. There is strong reasoning in employing third parties for penetration and unauthorised use testing for identifying problems and weaknesses before they become public knowledge.

There are however some risks that cannot affordable be covered and therefore situation management planning is essential. One such risk is from passive interception by a skilled opponent with substantial resources. In the next section we will consider the threat posed by national intelligence agencies.

2.5 Signals Intelligence (SIGINT)

The covert interception of communications is the biggest threat posed to the security of a message. Communications intelligence (COMINT), the branch of SIGINT dealing with the interception of communication signals, is practiced by almost every advanced nation, providing intelligence on diplomatic, economic and scientific developments. This section gives a review of the capabilities and constraints on COMINT activity and the threat it poses to communications security.

The annual global expenditure on COMINT is believed to be approximately $20-30 billion, mainly from the English-speaking nations as part of the UKUSA alliance. The highly automated system developed by UKUSA, often known as ECHELON [71], is capable of covertly intercepting space-borne communications, undersea cables and microwave links. Public awareness and discussion of this interception has increased rapidly following a 1997 STOA report [72]. The system has existed since the 1970's and significantly expanded its fields of operation between 1975 and 1995.

The tasking requirements is the main limitation to the information which can be gathered; in the 1960's and 1970's the UK and the USA were primarily interested in monitoring domestic political opposition figures, but by the 1990's this had spread to targeting narcotics trafficking, money laundering, terrorism and organised crime.

The UKUSA secret agreement, signed in 1947, came out of WWII collaboration between the UK and USA on global communication monitoring. Three other English-speaking nations, Canada, Australia, and New Zealand joined the agreement as second parties. It was not publicly acknowledged until March 1999 when the Australian Government confirmed that its Defence Signals Directorate (DSD) "does co-operate with counterpart signals intelligence organisations overseas under the UKUSA relationship" [73]. The NSA underpins much of the agreement through its worldwide field stations, such as the largest at Menwith Hill in England.

The exposure of the ECHELON system has concerned many nations, particular in Western Europe to the extent that a German MEP is suing the countries in operating the system [74]. The nations involved argue it is a necessary tool in combating organised crime and terrorism [75] and is not used for economic or political advantage or to infringe civil liberties.

The future of the ECHELON system is assured as more information is exchanged electronically and therefore can be intercepted. The driving force, the USA, will ensure through political means access to worldwide sites from which

to listen in order to maintain its dominance in world affairs and technological development. The impact of ECHELON may however been countered by other developed nations through the introduction of strong cryptography for all communication channels, for example in mobile phones where European manufactures dominate the market. In doing this however, they face the dilemma of also weakening their own intelligence gathering agencies.

There are at least 30 other nations involved in SIGINT activities, most notably the Russians through the FAPSI (Federalnoe Agenstvo Pravitelstvennoi Svyazi i Informatsii) with 54,000 employees. China also maintains a substantial SIGINT system, with at least two stations monitoring Russia in collaboration with the USA. Recently many Middle Eastern and Asian nations have been investing heavily in SIGINT, in particular Israel, India and Pakistan.

With the high cost and negative impact of conventional warfare in the current political and economic climate, investment in SIGINT is seen by many nations as an investment in the battlefield of the future. For example the investment of the Indian government in information technology and incentives for foreign companies to setup high technology businesses on the subcontinent is part of a strategic development to become the dominant technological force in that part of the world. The ease of access to networks and the information they carry combined with the desire for intelligence both within and outside its boarders, will prompt the development of significant SIGINT infrastructures in all developed countries.

The development of any SIGINT system comprises of six different areas: planning, access, collection, processing, production and dissemination. In more detail:

1) **Planning** – even with extensive facilities available, SIGINT operations are prioritised at the planning stage. The planning stage also takes into account customer requirements, such as the speed at which the information is required, as well as logistical problems such as computing resources and how to gain access to the communication links. The range of operations varies from monitoring of individuals, though to monitoring economic factors such as essential commodity prices, industrial and technological strength and political concerns such as arms control and negotiating positions.

2) **Access** - Access to the desired communications medium is the essential first step. There are several main communication systems used:

a. **Radio** - Historically long-range, high frequency radios were used, where the signal is reflected from the ionosphere to give communication over up to several thousand kilometres. Interception is therefore relatively simple, with only a suitably quiet area of land

required. The USA and UK in particular operate a number of sites worldwide used to collect HF signals, often containing diplomatic intelligence.

b. **Microwave** - Microwave links were introduced in the 1950's to provide inter-city communication using antennae of 1-3m separated by 30-50km. Although a directional system, there is always some leakage that can be intercepted. A USA satellite system, CANYON comprised of at least 7 satellites, collected spillage from microwave links, particularly in Russia, where the permafrost restricted use of underground cables. The system was so successful that it has been replaced by improved systems called MERCURY and RUTLEY.

c. **Cables** - Submarine cables provided the first high-bandwidth, intercontinental communications medium. Modern systems can now carry up to 5Gb/s (approx. 60,000 simultaneous telephone calls). Although in theory secure, access to these cables can be achieved in a number of ways. In western countries, laws were modified to give security agencies unlimited root access to the cables. Major access operations, such as the NSA's SHAMROCK collection activity from 1945-1975 on ILCs (International Leased Carriers) show the length and breadth to which access has been given. For hostile countries, the US Navy found out as early as 1971 it was possible to tap Soviet underwater cables using inductive coils and recording pods, which could be placed by drones. Potentially some of these are still in place, although the optical fibres which are now replacing the copper cables are not susceptible to the same eavesdropping, however the electro-optical repeaters used may be susceptible. This poses one of the biggest problems to the future of SIGINT – the collection of intelligence from optical fibre communication systems in foreign countries.

d. **Satellite** - Commercial satellite (COMSAT) ILC interception is practiced by many countries worldwide using ground-based systems. These are primarily used to gather intelligence from communication satellites and formed the basis of the ECHELON system developed by the UK and USA. This is the largest area of growth in SIGINT, with most of the western countries operating or building stations for this purpose. An interesting problem to be faced by SIGINT agencies is how to collect information from low or medium orbit systems such as Iridium which use 66 satellites to allow personal communication systems to/from anywhere in the world. Each satellite however covers only a small area and moves very fast and therefore the signals are difficult to intercept.

e. **Remote Sensing** - The USA, almost exclusively, has also been using satellites since 1960 to collect SIGINT information; some, such as Boeing's Trumpet, with antennas approaching the size of a football field. Their targets have included VHF and UHF radio, telemetry, cellular phones, paging signals, and mobile data links. A review of some of the technology used in SIGINT satellites can be found in Herskovitz's 'A Sampling of SIGINT Systems' [76].

f. **Internet** - The Internet is the fastest growing medium of information exchange. As in other areas, SIGINT activities have kept apace with civilian developments. Many of the major Internet Exchange Points (IXPs) are controlled by, or are accessible by government controlled agencies. Sniffer programs that filter out irrelevant information and collect the few percent, such as e-mail, file transfers and virtual private networking that is potentially interesting are used at these points. These programs can also monitor the flow of traffic to and from particular computers. SIGINT stations also employ 'bots'; programs which wander across the Internet gathering potentially useful information. ECHELON is also known to employ a computer system called CARNIVORE to collect information from the Internet [77].

g. **TEMPEST** - Electromagnetic emissions from computers are also processed for short-range monitoring of computer usage where no other method can be successfully employed. Such TEMPEST (Transient Electromagnetic Pulse Emission Standard) activity is effective over several hundred metres and can be demonstrated using simply hardware, available off the shelf [78]. There is strong evidence that both the UK and USA have employed such technology to access otherwise unobtainable information, often from foreign embassies [79].

3) **Collection** – The rapid flow of information can be collected in a number of different ways according to the aims of the operation. Filters can be used to look for specific channels being used, or information being transmitted. Alternatively, high-speed recorders can be used to trawl information to be analysed off-line if there is no specific target in mind. Currently monitoring can handle the flow of data, however as the bandwidth of applications dramatically increases with the explosion of the Internet, the size of the equipment required may make collection a more specific activity.

4) **Processing** – the rendering of the collected information into a useful form to be analysed is normally automated with some human intervention. This can require a number of operations such as deciphering, translating, filtering and identifying. This process can potentially be the most resource intensive.

The ECHELON system makes use of automated "watch lists" which include names, addresses, topics of interest, telephones numbers and other criteria. If any matching information is found it is forwarded automatically as raw intelligence. In particular, local dictionary computers, for local languages and dialects are invaluable to the intelligence gathering community. There is strong evidence these are used extensively on Internet and telex messages, however the evidence for voice capability is less strong, though rumoured. The ability to analyse voiceprints, "topic spotting", and traffic analysis for verbal communications can however provide as powerful information. The increasing difficulty of rendering information into a useful intelligence form, will lead to greater emphasis being placed on obtaining the information at the source or destination, rather than while it is being transmitted.

5) **Production** – once the data is in a useful form it has to be analysed, evaluated and interpreted into intelligence. The interpretation of the gathered intelligence again depends on the customers' requirements. Production and processing are often overlapped until sufficient filtering at the processing level can reduce the often manpower intensive demands at the production level. The development of "smart" filters for all channels is therefore imperative in high volume and random data collection, such as the ECHELON system.

6) **Dissemination** – as the output of the SIGINT process, the nature of the disseminated information can give valuable information about the collection process. The output can be fed back into new aims for the operation, completing the intelligence gathering cycle.

There is a quandary to be faced: should individuals have the right to privacy or should intelligence agencies have access to information in order to combat illegal activities. The battle of encryption highlights this quandary with individuals employing strong encryption, steganography and chaff and winnowing in order to minimize the impact of COMINT. Widespread use of these techniques would indeed reduce the effectiveness of networks such as ECHELON, but at what cost? Is this a price worth paying? In the next section, we will consider governmental response to communication systems and the laws applied.

2.6 Law Enforcement & Communication Systems

Governments tread a fine line between privacy and law enforcement. In this chapter and later chapters, several examples are given where governments have

introduced or modified laws in order to protect their right to monitor suspected criminals. Often these laws try to push back the rights of personal privacy and in most democratic countries a balance is maintained by negative public and commercial reaction to overbearing legislation.

The most debated issue of the last decade has been the ownership and use of strong encryption. Encryption of communication channels minimizes the impact of wiretapping by intelligence and law enforcement agencies; yet to prevent public use of encryption is, and has been proven to be, extremely difficult. For example, in the early 1990's the NSA forced AT&T to change its secure telephone system to use Clipper chips, which had been manufactured by NSA. When this proved unpopular, the US government proposed "key escrow" to maintain access to encrypted messages. It applied pressure to EU nations in particular to adopt similar "key recovery" schemes, however to date this has appeared to have a negative effect.

A further strand of the US pressure came in the form of the annual ILETS (International Law Enforcement Telecommunications Seminar), founded by the FBI. Although under the guise of wanting to improve intelligence gathering for law enforcement purposes, it left the European representatives with no doubt that the main purpose of introducing "key recovery" was to gather foreign intelligence [80]. The distinction between law enforcement requirements for monitoring specific, or small groups of communication channels under judicial review is quite different from the widespread trawling and broad monitoring employed by security agencies who are not concerned about the whether the parties being monitored have been involved in illegal activities.

The ILETS meetings however have provided a generic set of requirements for legal interception in a document called "IUR 1.0", adopted by the EU in 1995 and several non-EU nations. The major telecommunication companies have also been asked to incorporate features at the design stage to allow for legal interception, indicating the importance with which intelligence gathering is regarded. IUR was updated in 1998 to take account of satellite personal communication systems, such as Iridium, and additional security requirements for network operators and service providers. These amendments were however rejected by the Police Co-operation Working Group (ENFOPOL) and are currently still being redrafted into an acceptable form [81].

The US intelligence agencies appear to be keen to blur the distinction between intelligence gathering and law enforcement for their own purposes, however this has not been effective at home or abroad to the extent hoped for. Since the 1970s the US has treated economic intelligence "as a function of national security enjoying a priority equivalent to diplomatic, military and technological intelligence" [82].

Similar policies exist in the other UKUSA countries, who have all demonstrated a means by which intelligence information can be passed to national companies where it has been of economic advantage. There are several well-documented examples involving multi-billion dollar contracts: the US Raytheon Corporation succeeding Thomson-CSF as the supplier of a rain-forest surveillance system after the NSA assisted in exposing bribery of Brazilian government officials [83]; a similar case gave Boeing and McDonnell Douglas a $6 billion contract from the Saudi national airline in 1995, after the NSA intercepted faxes and phone calls which found that Airbus agents were offering bribes to a Saudi official [84]; information has also been used by the US government to strength its trade negotiating position, such as with the Japanese over cars and emission, and the GATT trade negotiations [85].

These high profile cases and a number of other cases where personal confidentially has been breached by intelligence and law enforcement agencies as well as the increasing sophistication of civil liberty groups for generating publicity and public awareness has led to greater pressure against intrusive legislation. The changing economics of the telecommunications marketplace, because of the deregulation of the large state-run monopolies, has also proved resistant to intrusive legislation, primarily because they are left to foot the bill. The increasing mobility of businesses provided by e-commerce and the prospect of m-commerce, combined with developing nations' desire to enter the IT revolution pose a significant threat to over zealous legislation in developed countries because of the loss of business.

These factors do not appear however to be preventing a number of countries from legislating intrusive measures. A recent case is the passing to law of the Regulation of Investigatory Powers (RIP) Bill in the UK. An in-depth analysis of the RIP Bill can be found here [86], and including some of the issues it raises such as prosecution of those not willing to provide decryption keys even when it may be self-incriminating, the right of employers to secretly spy on employees and seizure of data with court authority. The Electronic Privacy Information Center's survey of international encryption policy [87] places the UK legislation on par with that in non-democratic countries and going against the trend set in other first world countries.

The growing importance and diversity of services and providers combined with public access to security measures such as encryption are key issues in determining law enforcement legislation for communication systems. The globalisation of systems make national regulation restrictive and therefore laws governing communication systems will becoming increasingly made at an international level. Importantly it will always have to tread a fine line between the rights of citizens and the fight against crime while keeping an eye to future developments in communications technology. Many countries have reacted,

rather than been proactive, to the growing importance of data and communication in the last two decades and therefore it is constructive to consider some of the medium term advances, which may change the future of telecommunications.

2.7 Future Advances in Communications Technology

There is a bewildering array of new ideas that may strongly influence the future of communication systems. Many of these are software algorithms and protocols to meet new demands. The growing use of wireless devices and optical interconnects will undoubtedly change public use of communications and data. However as with automobile technology, there is a vested interest in the major producers of these goods to slow the rate of change in order to maximise financial gain.

There is however two concepts related to the quantum world that may have a drastic effect on how we perceive, receive and process information. A short review of these two technologies, quantum computing and quantum cryptography follows with an indication of the effect they are like to have and the time scales involved before useful systems will be commercially available.

2.7.1 Quantum Computing

One of the largest threats to current encryption schemes is the prospect of a quantum computer. Computers based on integrated circuits perform calculations in serial, with limited ability to perform calculations in parallel. The rapidly increasing number of integrated circuits on a chip, proportional to computing power, has doubled approximately every 18 months or so (Moore's Law) for the last 30 years. This rapid increase in power has met the requirements of most applications. We are now approaching a fundamental, quantum barrier however. The prospect of quantum computers introduces a number of new ideas and potentially new applications to the world of computing.

The heart of a quantum computer uses quantum mechanical states of particles to store information in quite a different way to the conventional '1's and '0's used by standard computers. These states can be a superposition of '1' and '0' as well as either '1' or '0', making it possible to hold more than one bit of information per state in conventional terms, or in quantum terms each state holds a qubit of information. Preparing a large number of particles which can interact in this superposition of states and them imposing boundary conditions before examining the states can perform a calculation, which is equivalent to parallel

processing; for one hundred particles this would mean the equivalent of 2^{100} calculations in a single cycle, giving massive parallelism.

This would be ideal for factoring large numbers, severely weakening the effectiveness of current encryption algorithms, the importance of which has already been illustrated by the fact that algorithms have already been proposed for how this could be done [88]. Other schemes have also been described to allow better database searching, important in the rapidly growing area of data warehousing.

Although a functional quantum computer still lies beyond the grasp of current technology the last decade has seen a number of significant advances. The single-most problem is maintaining the quantum coherence between the prepared states and this can be destroyed by the slightest disruption, from atom collisions, to changes in the temperature or light the system is exposed to. Inevitably errors will also be introduced into calculations through these environmental changes, however it is impossible to use current error correcting techniques to detect them. A number of groups have been working, essentially from the ground up, on the requirements for a quantum computer, with moderate success.

Simple quantum gates have been demonstrated using cold, trapped, ions and NMR, however this technology could not reasonably be scaled to the point where it would provide useful computing power. This potential has not been lost on government agencies though, with DARPA and Los Alamos National Lab both spending significant resources in trying to build working devices.

The realisation of a quantum computer would change the face of computing overnight and a drastic rethink of many computing activities would be required. The availability of such logic and the ability to process optical information in such a way would also have a profound effect on communication systems and protocols used; quantum computer are however more than a decade away from being in a useful form.

2.7.2 Quantum Cryptography

The use of quantum computers by cryptanalysis to break current encryption techniques may not be such a threat if the power of the quantum world can be harnessed successfully for encryption before then. Currently quantum cryptography is further advanced than quantum computing and laboratory trials are already underway to test its feasibility in real-world situations.

Quantum cryptography potentially offers the ideal means of secure communication. According to current fundamental physics it can form an

unbreakable cipher, and can also be used to detect if anyone is eavesdropping on the line, while taking advantage of the high speed, flexibility and security afforded by optical fibre.

The idea of quantum cryptography dates back to the 1960's but was not explored further until the 1980's. Bennett and Brassard at IBM's Thomas J. Watson Laboratories laid down the theoretical background in 1984 followed by a practical demonstration of the principle in 1988 [89]. This has been followed by several other demonstrations, including one over 23km of optical fibre between Geneva and Nyon by the University of Geneva, and one over a free space link of 1km by Los Alamos, aimed at providing secure links to satellites. The theory of quantum encryption has also developed significantly to include error correcting and eavesdropper detection algorithms and different quantum schemes that permit cryptography. The two main schemes investigated to date are the partial indistinguishability of non-orthogonal (polarisation) state vectors and quantum entanglement, also used in experiments on quantum teleportation.

Quantum cryptography does however have a number of problems [90]:

1) **Fundamental Issues** – quantum cryptography has made no significant contribution to cryptography, other than to lift ideas from public-key cryptosystems. More work is required to investigate quantum signature and authentication schemes. Secondly, and more importantly, the unconditional security of a number of basic protocols, such a bit commitment, one-way identification and one-out-of-two oblivious transfer, have been shown to be impossible in a series of no-go theorems.

2) **Technological Issues** – Currently the limited distance over which quantum cryptography can be carried out and the extremely low data rates and post-processing overheads make quantum key distribution infeasible for commercial markets.

3) **Commercial Issues** – the size, cost and difficulty in integrating quantum cryptographic machines with existing technologies would act as prohibiting factors to fast commercial implementation of quantum cryptography, however these issues can be addressed by longer term development.

4) **Security Issues** – while the concept of quantum cryptography has been developed, none of the systems demonstrated so far has actually been secure because of loopholes in the implementation, some of which may prove non-trivial to remove.

5) **Psychological Issues** – the vested interest and distrust of conventional cryptographers, combined with governmental desire to suppress widespread use of such technology may further limit its implementation.

All however is not so bleak. The field of quantum cryptography is still relatively new and a number of developments such as quantum teleportation, quantum repeaters, and multiple level systems may address the previous issues, in combination with miniaturisation and further testing and development. Ultimately a work international standard on quantum cryptography and the availability of physical systems for quantum and cryptographic analysis would help to strength the case for wider implementation.

2.8 Conclusions and Future Prospects

In this chapter commonly used civilian communication systems have been described assessed for their security and applications. Prospects for each of the systems has also been outlined with forthcoming technologies, which may impact communications technology.

Traditional cable based communication systems have come under growing pressure from wireless services, in particular for voice and data applications. This has been in part due to the intransigence of large telecommunication monopolies to recognise and respond to the wireless sector. With the introduction of the next generation of mobile phones, the wireless sector will outstrip equivalent cable systems in all areas, a challenge which needs to be met by cable operators.

In both the wireless and cable sectors there is likely to be significant integration of currently separate services to consolidate their importance. In particular the integration of financial services, such as e-banking and m-commerce will provide significant opportunities in the next decade.

All of the hardware described has been vulnerable to a greater or lesser extent to monitoring and interception by unauthorised persons. As this process becomes more automated and access to greater numbers of channels becomes available there will be an inevitable shift towards increased intelligence gathering by security agencies, including indiscriminate trawling. To some extent this can be counteracted by encryption techniques, however even these may not be sufficiently strong or available to make significant difference.

The globalisation and integration of communications systems is ideally illustrated by the development of low earth orbit satellite systems. Voice, fax, data and video will all be available on the same channel with a larger bandwidth than existing telephone lines. The broadcast nature of satellites may also open

up possibilities such as more public broadcasting or on the other hand limiting the bandwidth of uplinks to limit user feedback. The potential monopolisation of the satellite industry by the US should also be of concern to European nations as undoubtedly they will be subjected to close scrutiny by the NSA.

Optical fibres and their associated technologies provide the most exciting prospects for the telecommunication industry with the advances in hardware, software and conceptual ideas. All optical processing is a Holy Grail, which is painfully being realised, though technologies such as quantum cryptography may see military if not commercial field usage within the next decade.

Communications hardware is only one small part of the picture however and these advances must be considered in parallel with developments in computer systems and their associated software as inevitably they will begin to lead the demands on communication networks of the future.

References and Further Reading:

[1] Infosyssec Portal for Telecommunications Tutorials

http://www.infosyssec.net/infosyssec/teletut1.htm

[2] George Leopold, *Digital Wiretap Law Draws Court Challenge,* TechWeb

http://www.techweb.com/wire/story/TWB19991118S0009

[3] *The Nature and Scope of Governmental Electronic Surveillance Activity*, Center for Democracy & Technology

http://www.cdt.org/digi_tele/wiretap_overview.html

[4] Mark Fineman, *Latest Mexico Wiretap Scandal Spurs Move to Curb Widespread Practice*, LA Times (June 17, 1995)

[5] Comments of Hansjorg Geiger, German Federal Commission for the Stasi Files (April 14, 1993)

[6] Dave Banisar, *French Wiretapping Scandal Leads to Electorial Defeat*, Privacy Times

http://www.eff.org/pub/Privacy/Surveillance/Foreign_and_local/France/fr_wiretap_scandal.article

[7] e.g. Granite Island Group

http://www.tscm.com/

[8] e.g. Telecommunications Surveillance

http://seussbeta.tripod.com/Tap.html

[9] *The Alt.Phreaking FAQ 1.2*, The Ocean County Phone Punx

http://www.19f.org/archive/FAQ.html

[10] Phone Phreaking Box Guides

http://www.textfiles.com/phreak/

http://www.linuxsavvy.com/staff/jgotts/underground/boxes.html

[11] Wardialer Guides

http://packetstorm.securify.com/wardialers/

http://www.infosyssec.net/infosyssec/telephon1.htm

[12] Richard Thieme, *Off With His Hands - Hacking Culture and the Hunger for Knowledge*, ThiemeWorks

http://www.thiemeworks.com/write/archives/hands.htm

[13] *Telephone Fraud*, Bellsouth Neighborhood Watch

http://www.bellsouth.com/neighborhood_watch/cn_nw2_telefraud.html

[14] *How a Cable System Works*, Continental Cablevision

http://www.geocities.com/SiliconValley/Park/3254/cablsys.htm

[15] Guides to Cable Modems

http://uk.search.yahoo.com/search/ukie?p=cable+modems&y=y

[16] Guides to Underwater Telecommunication Cables

http://www.spie.org/web/oer/august/aug99/subcable.html

http://www.alcatel.com/submarine/refs/

http://davidw.home.cern.ch/davidw/public/SubCables.html

[17] Guides to EMP Attacks

Ian Sample, *Just a normal town...*, New Scientist, 167:2245, 20 (2000)

http://www.dallas.net/~pevler/

http://www.infowar.com/mil_c4i/mil_c4i8.html-ssi

[18] Cellular Phone Interceptions

See Boehner v. McDermott, 191 F.3d 463, 465 (D.C. Cir. 1999) (describing the taping of a cell phone call including Newt Gingrich);

Office of the Independent Counsel, Referral to the United States House of Representatives pursuant to Title 28, United States Code, § 595(c) § I.B.3 ("The Starr Report")

Paul Vallely, *The Queen Brings Down The Shutters*, The Independent, Aug. 19, 1996

[19] News stories about third generation mobile phones

http://uk.search.yahoo.com/search/news_ukie?p=third+generation+mobile+phones&b=6&h=s

[20] Greg Jones, *Introduction to Packet Radio*

http://www.tapr.org/tapr/html/Fpktfaq.html

[21] Packet Radio Primers

http://www.stack.net/~victor/hamradio/packet/packet.html

http://hydra.carleton.ca/articles/hispeed.html

[22] Phillip M. Feldman, *Emerging Commercial Mobile Wireless Technology and Standards: Suitable for the Army*, Rand Corporation Report - MR-960-A, 1998

http://www.rand.org/publications/MR/MR960/

[23] Future Developments in Radio Technology

http://www.cwc.oulu.fi/home/

http://bwrc.eecs.berkeley.edu/Presentations/

[24] Scanner Frequencies

http://www.angelfire.com/sc/scannerpost/scannerfrequs.html

[25] DECT Forum Website

http://www.dect.ch/

[26]Peter Baston's POCSAG decoder

http://www.users.rapid.net.uk/carl/pocsag.htm

[27] *Interception of Pager*, Annual Report of the Interception of Communications Commissioner for 1998, UK

http://www.homeoffice.gov.uk/oicd/crica.htm

[28] Duncan Campbell, *Interception Capabilities 2000*, Vol. 2 of *Development of surveillance technology and risk of abuse of economic information*, European Parliament Publications - No EP/IV/B/STOA/98/1401

http://www.cyber-rights.org/interception/stoa/ic2kreport.htm

[29] Erik Bloodaxe, *The Wonderful World of Pagers*, Phrack Magazine, 5:46

http://www3.l0pht.com/~oblivion/blkcrwl/cell/pager/p46-8.html

[30] *WWW goes MMM*, HP Computer News

http://www.hpcn.com/english/themen4_99/e005_499.html

31 Stories related to the widespread usage of mobile phones

http://biz.yahoo.com/bw/000912/ny_scarbor.html

http://www.cellular.co.za/news_2000/news-04202000_britains_mobile_phone_boom.htm

[32] Bruce Schneier, *Cryptanalysis of the Cellular Message Encryption Algorithm*

http://www.counterpane.com/publish.html

[33] Peter Hadfield, *Sayonara WAP*, New Scientist, 168:2261, 39 (2000)

[34] Markku-Juhani Saarinen, *Attacks Against The WAP WTLS Protocol*
http://www.jyu.fi/~mjos/wtls.pdf

[35][Extracts From] *TECHNICAL INFORMATION: GSM System Security Study*, RACAL RESEARCH LTD.

http://jya.com/gsm061088.htm

[36]Lauri Pesonen, *GSM Interception*

http://www.tml.hut.fi/Opinnot/Tik-110.501/1999/papers/gsminterception/netsec.html

[37] Cryptanalysis of GSM's A5

http://cryptome.org/a51-crack.htm

http://cryptome.org/a51-bs.htm

http://jya.com/crack-a5.htm

[38] J. Sandberg, *Flaw Is Found in Digital Phone System That May Let Hackers Get Free Service*, The Wall Street Journal, April 13, 1998, pp. A3, A12.

http://www.jya.com/gsm-cloned.htm

[39] Mika Müller, *Intruder Scenarios In Telecom Networks*

http://www.tml.hut.fi/Opinnot/Tik-110.501/1999/papers/scenarios/scenarios.html

[40] *UK 'monitored Irish phone calls'*, BBC News - Friday, July 16, 1999

http://news.bbc.co.uk/hi/english/uk_politics/newsid_395000/395843.stm

[41] *GSM Cloning*, ISAAC Group, University of Berkeley

http://www.isaac.cs.berkeley.edu/isaac/gsm.html

[42] John Markoff, *Hacking Chips on Cellular Phones*, Wired magazine (Mar 1993)

http://hotwired.lycos.com/collections/hacking_warez/1.01_hacking_chips_pr.html

[43] e.g. Watch Dog Fraud Detection System, Basset Telecom Solutions

http://www.nofraud.com/

[44] Jennifer Schenker, *Third Generation Gap*, Time Europe, Vol. 156 No. 4 (2000)

http://www.time.com/time/europe/magazine/2000/0724/telecoms.html

[45] Primers about Universal Mobile Telephone Service (UMTS)

http://www.infowin.org/ACTS/ANALYSYS/PRODUCTS/THEMATIC/MOBILE/Evolution_Cellular_Systems/ariks.html

http://www.cellular.co.za/umts.htm

[46] David Concar, *Get Your Head Around This...*, New Scientist speacial investigation into mobile phones

http://www.newscientist.com/nsplus/insight/phones/mobilephones.html

[47] Paul Marks, *Your Everything*, New Scientist, 168:2261, 42 (2000)

[48] Orla Ryan, *Big Brother or friendly helper?*, BBC News Online

http://uk.news.yahoo.com/001018/79/amlwu.html

[49] Mark Schrope, *You are here*, New Scientist, 168:2261, 44 (2000)

[50] Antti Vähä-Sipilä, *Ciphering in GPRS and UMTS*, 8309700 Advanced Topics in Telecommunications, Report (April 2000)
http://www.apparatus.org/~avs/gprs-umts-crypto-revised.pdf

[51] *Finnish citizen card and electronic identification*, IST99 Helsinki

http://www.ist99.fi/finland/fined.html

[52] *Space Network List,* International Telecommunications Union

http://www.itu.int/itudoc/itu-r/space/snl/index.html

[53] e.g. Nordic Satellite Company Services

http://www.nsab.se/services/

[54] Jane Wakefield, *Our satellites are hack proof*, ZDNet News - March 1st 1999

http://www.zdnet.com/zdnn/stories/news/0,4586,2217730,00.html

[55] Arthur Gordon, *Orange County Man Arrested for Hacking Into NASA Computers*, internet.com News - September 22nd 2000

http://la.internet.com/news/article/0,2325,5321_466731,00.html

[56] e.g. Europe Online

http://www.europeonline.com/

[57] *Motorola plans third satellite network*, EE Times, Issue 959, June 23rd 1997

http://www.techweb.com/se/directlink.cgi?EET19970623S0017

[58] *Loral struggles to decouple from satellite fiasco*, Yahoo News October 30th 2000

http://uk.biz.yahoo.com/001030/80/ant3e.html

[59]G. Harry Stine, *LEO systems take flight -- Satellite carriers poised to launch new services*, Internet Week, Issue 684, October 6th 1997

http://www.techweb.com/se/directlink.cgi?INW19971006S0023

[60] *Space Policy Project – Desert Star*, Federation of American Scientists

http://www.fas.org/spp/military/docops/operate/ds/communications.htm

[61] Mark Robichaux, *Desert Storm Demand Buffets Satellite-Phone Firm*, The Wall Street Journal, 1 February 1991, page B2.

[62] e.g. Blue Sky Satellite Communcations

http://www.blueskysat.com/homepage.htm

[63] Craig Jarvis, *Cable thief sentenced to 5 years*, United Network of Communications Law & Enforcement

http://www.unclenet.org/unclenet/fraud01.htm

[64] Markus Kuhn, *Attacks on Pay-TV Access Control Systems*, Cambridge University

http://www.cl.cam.ac.uk/~mgk25/vc-slides.pdf

[65] FTP Site of TV Decryption Information

ftp://ftp.informatik.uni-erlangen.de/pub/multimedia/tv-crypt/

[66] Primers on DSS Cards

http://www.hackerscatalog.com/dssfaq1.htm

http://www.iol.ie/~kooltek/hasdss.html

[67] IrDA Website

http://www.irda.org

[68] Bluetooth Website

http://www.bluetooth.com

[69] Paul Marks, *Brave New Fridge*, New Scientist, 168:2261, 61 (2000)

[70] Books on Eavesdropping Technology

http://www.spybusters.com/Books.html

[71] Duncan Campbell, *Development of Surveillance Technology and Risk of Abuse of Economic Information*, Scientific and Technological Options Assessment for the European Parliament, PE 168.184 (April 1999)

http://www.europarl.eu.int/stoa/publi/default_en.htm

[72] *An appraisal of the Technologies of Political Control*, Steve Wright, Omega Foundation, European Parliament (STOA), 6 January 1998

[73] Statement by Martin Brady, Director of DSD, 16 March 1999. broadcast on the Sunday Programme, Channel 9 TV (Australia), May 1999

[74] *German EU Delegate Sues 'Unknown' Over Echelon*

http://slashdot.org/yro/00/10/16/1152252.shtml

[75] *UK's Straw Defends New Anti-Terrorism Bill*, Intel Bulletin 019991216 (Dec 1999)

http://www.spytechagency.com/Intel%20Bulletin/intel%20bulletin%20019991216.htm#01

[76] Don Herskovitz, *A Sampling of SIGINT Systems*, Journal of Electronic Defense EW Reference and Source Guide", Supplement, Jan. 1998, 30-36. Journal of Electronic Defense, Jul. 1998, 51-59.

[77] Details of the Carnivore System

http://cryptome.org/carnivore-demo.htm

http://www.securityfocus.com/frames/?content=/templates/article.html%3Fid%3D97

[78] W. van Eck, *Electromagnetic radiation from video display units: an eavesdropping risk?*, Computers & Security, vol.4, no. 4, pp. 269-286 (Dec 1985)

[79] P. Wright, *Spycatcher – The Candid Autobiography of a Senior Intelligence Officer*, William Heinemann, Australia (1987)

[80] Kurt Westh Nielsen & Jérôme Thorel, *Interview with David Herson - SOGIS*

http://www.ing.dk/arkiv/herson.htm

[81] Duncan Campbell, *Special Investigation: ILETS and the ENFOPOL 98 Affair*

http://www.telepolis.de/tp/english/special/enfo/6398/1.html

[82] *Dispatches : The Hill*, Channel 4 Television (UK), 6 October 1993

[83] Raytheon and Brazil Sign SIVAM Contract

http://www.raytheon.com/c3i/c3iproducts/c3isivam/news/news001.htm

[84] Scott Shane and Tom Bowman, *America's Fortress of Spies*, Baltimore 3rd December 1995

[85] e.g. *Company Spies*, Robert Dreyfuss, Mother Jones, May/June 1994; Financial Post, Canada, 28 February 1998

[86] Regulation of Investigatory Powers Information Centre, Foundation for Information Policy Research

http://www.fipr.org/rip/index.html

[87] *Cryptography and Liberty 2000 - An International Survey of Encryption Policy*, Electronic Privacy Information Center

http://www2.epic.org/reports/crypto2000/

[88] e.g. AT&T papers relating to quantum computing

http://www.research.att.com/~shor/papers/index.html

[89] C.H. Bennett, C. Brassard, A. Ekert, *Quantum Cryptography*, Scientific American, Vol. 269, pp. 26-33 (October 1992)

[90] Hoi-Kwong Lo, *Will Quantum Cryptography ever become a successful technology in the marketplace?*, quant-ph/9912011

http://arXiv.org/abs/quant-ph/9912011

Chapter 3 - Data Security – Cryptography & Steganography

3.1 Introduction

Secret codes and self-destructing messages are familiar to all children and adults, in particular from spy movies and comics. It is perhaps surprising how often we use these ideas in our everyday lives without realising, from using mobile phones through to bank services and watching satellite television. Our use is not as dramatic, but the purpose is the same: that information sent can only be read by those people it was intended for.

Everyday large amounts of information are exchanged by various means of communication, or *channels*. Some of these channels have an inherent amount of security to protect information. For example using the postal service, the recipient knows to a greater or lesser extent whether the information is authentic (e.g. signature from sender) and whether or not it has been viewed by unauthorised people (envelope seal broken). Other means of communication, for example the telephone, do not have these safeguards and rely on trust between the users and service providers.

This chapter will explore the protocols used to transmit electronic information, or data, and how security can be added to channels susceptible to eavesdropping. The basis of most secure systems is the application of *cryptography* and this will be the focus of the chapter. The other major and complementary method of securing information, *steganography*, or information hiding, will also be considered. The chapter begins by considering the origin of threats and the risks they pose. The current legal and commercial impact of information security will be considered. The chapter will conclude by looking at future advances that may significantly change information security.

In the previous chapter, we have considered the risks and threats to the hardware of a communications system and mentioned some of the ways in which penetration can be detected and prevented. Increasingly the threat comes from passive interception and software tools that can be operated remotely. The starting point for any analysis of data security is therefore that a motivated opponent has access to the communications channel without detection.

3.2 Threats to Data Security

With access to a communications channel, an attacker can perform one of a number of basic attacks. These types of attack are common not only to the world

of information technology, but also to the world in general. In the first subsection, we will consider these attacks and their effect on data being transmitted over a network. In the second subsection, the motivations and identities of the individuals and groups wishing to intercept data will be considered. Using this information, we will consider techniques in the remainder of the chapter that can be used to counter these attacks.

3.2.1 Attacks on Communication Systems

Apart from military and intelligence network channels, most channels are inherently insecure and it should be assumed that all communications can be attacked. Steps can be taken to determine hardware penetration as outlined in the previous chapter, however for large or multi-level networks or general users, these are not viable options. Instead, adding security to the data transmitted provides the most effective deterrent to most attacks.

The primary classes of attack on a data channel include:

1) **Destruction** – This is the most basic form of attack, destroying the logical connection of a communications channel. This can be done by disabling either the sender or receiver, or the route used to send data. Software attacks, for example those described in Chapter 5, enable this to be done remotely, potentially causing data loss and disruption of services. Using authentication, acknowledgement and traffic management controls can minimize the impact of destructive attacks. The flexibility in network designs and the increasing use of security software to prevent such blatant attacks has led to other attacks being used in preference.

2) **Disruption** – Disruption of data systems is generally more potent than destruction. Disruption can take several forms, from software-based denial-of-service attacks to corruption of routers and name servers. Disruption can be as effective as destruction particularly when the origin and nature of the attack cannot be traced. There is also the possibility for disguising the attack as a normal system failure and for disguising an active interception. For example a computer router controlling network traffic could be sent a "ping of death" causing it to go offline, during which time traffic is no longer routed or an alternative unauthorised server could provide rerouting of data. Hardware disruption is normally a resource consuming activity and therefore limited to professional and national agencies. However, software attacks, in particular remote denial-of-service attacks, have been growing in frequency from all threat groups. Preventing disruption requires careful management of resources, using

techniques such as firewalls and network analysers to minimize the impact of disruption attacks on normal data flow.

3) **Passive Interception** – The passive interception of data as it is transmitted across a network can be easily achieved using programs such as packet sniffers. These programs can be configured to monitor either all network traffic or to watch for particular keywords and signatures. Detecting the presence of such a program is extremely difficult, particular when the data being transmitted traverse public networks, such as the Internet. Therefore, it is important that data confidentiality is maintained. Passive interception is the most common threat and used by all threat groups for intercepting everything from passwords to credit card numbers and propriety data. It is difficult to prevent passive interception, even within the law, the most active deterrent being strong encryption.

4) **Active Interception** – If access cannot be gained to a channel through passive means, a more risky active attack may be used. This would involve either hijacking a connection by impersonation, or installing Trojanised software on the computer of the sender or receiver. Active interception can provide a way to bypass encryption protocols used, best illustrated by the man-in-the-middle attack described later. Protection against penetration is often the weakest link when encrypted channels are used, favouring this form of attack. The use of authentication and system security tools can minimize the risk of an active attack, while auditing and monitoring software can be used to detect, analyse and prosecute attackers.

5) **Modification** – Modification of a data channel is an important subset of active attacks. Most computer protocols have strict requirements on the way in which data is transmitted. Modification of the data, either through deletion, addition or changes in the packets being transmitted can cause disruption or destruction of the channel and data being sent. For example a "ping of death" uses an abnormal packet to cause a remote machine to crash instead of send a normal acknowledgement. To prevent modification attacks data authentication and integrity checks are required. Encryption protocols described later can introduce these to minimize the impact of modification attacks, a goal that will eventually be extended to services. The most common service attack is modification of http requests to a web server in order to gain access to the operating system or stored data. The rapid growth of the services offered by a web server has not been matched by security procedures and therefore there are modification attacks being discovered with high frequency for the majority of web server software.

These attacks are extremely effective on an unsecured, unencrypted data channel. In a physical analogy the sender, messenger and receiver are all vulnerable to attack. Using encryption only makes the message that the messenger carries difficult to read; the message can still be decoded or prevented if the sender or receiver is attacked instead or the messenger is corrupted in some way. This illustrates the importance of first armouring the sender and receiver to ensure their safety from attack and protecting the data channel, ensuring the whole process is difficult to attack. The effect of destructive and disruption attacks can be minimized by adequate resource planning.

In the next subsection we will consider the groups and individuals who attack data and their motivations for doing so. Consideration needs to be given to data security both when it is transmitted and stored. In this chapter we will concentrate primarily on the aspects of data security while in transmission, though many of these concepts can be extended to secure data storage, a subject covered in more detail in Chapter 5.

3.2.2 Actors

The groups and individuals posing a threat to communication channels can be broken down into a number of major risk levels:

1) **General Users** – this group represents the public and their use of data. In most cases the general public are the victims of data crime, with their channels monitored and their computers probed. They pose a threat in emotionally motivated circumstances, destroying and corrupting data and where security is sufficiently weak where commonly used tools can give access to sensitive data. Basic security measurements, frequent data backups and user management can prevent these attacks from having any significant impact.

2) **Amateur** – when a general user is motivated and invests time in an attack they pose a greater threat. A basic search on the Internet will reveal many tools that can be used to carry out generic attacks on both computer systems and data channels. The common, generic nature of the attacks means they can be detected by most security software. The impact of these attacks can be significant though because of the number of people attempting them is large compared to any other group. Any vulnerability that is not dealt with sufficiently quickly will be exploited down to this level as tools become available on the Internet. Therefore, regular updating of software is required and network-monitoring software installed to detect unusual activity in order to prevent these attacks.

3) **Restricted Professional** – a computer security professional or hacker is typical of this class of attacker. They have extensive experience and knowledge of computer systems and protocols. This is the primary level at which exploits and weaknesses are identified. Additionally an understanding of encryption protocols allows decryption of weak ciphers and strong ciphers that have been compromised. This research element provides extra versatility and sophistication in the attacks used. Typically destructive and disruptive attacks are not used, instead favouring passive and active interception attacks, because of their potential for information gathering. In contrast to amateurs, targets are generally not randomly chosen, instead specific systems and channels are targeted because of their content. It is difficult to defend against a well-motivated restricted professional without significant investment in security measures. Unauthorised active intrusion can however always be detected in some form and with sufficient management of the encryption of channels the threat posed by passive interception is minimized.

4) **Professional Organisation** – multi-national companies often employ or have internal departments responsible for gathering intelligence. These groups have access to substantial budgets, the latest equipment, and the potential to develop their own specialised equipment. They offer a formidable threat backed with substantial resources generally targeted at other organisations that pose a financial threat. Again the attacks are primarily active in order to penetrate a hostile environment, however their secondary task can also be to monitor and audit internal network usage.

5) **Intelligence Agency** – the highest risk is posed by governmental intelligence agencies that have essentially unlimited resources, access to innovative research technologies and can operate above the law to some extent. It should be assumed they have access to all known exploits and know of extensive ways in which encryption systems can be weakened. Even with strong encryption, there is strong evidence that intelligence agencies have sufficient resources to decrypt messages by brute force if so motivated. The best protection against this kind of opponent is to use information hiding and encryption algorithms with minimal data exchange and limited, random connection times to networks.

In order to counter the threat posed by these attackers, an assessment of risks needs to be made according to the sensitivity of the data to be protected and the resources available for security. The minimum level for an organisation should include network security software and encryption of sensitive data both in storage and in transmission. With increasing resources, more active measures such as network monitoring, penetration testing and protocol analysis can assist in defending against attacks.

The next two sections will consider techniques that can be employed to improve the security of data. The first of these, cryptography, changes the data into an unreadable form, which ideally can be only then be decrypted by the recipient. The second, complementary method is steganography, which attempts to hide the data amongst other larger data files. An analysis of these two methods and the security that they offer will be given, followed by an overview of data security from a legal and commercial perspective.

3.3 Cryptography

Cryptography has been primarily developed as a government tool for the protection of national secrets and strategies and therefore has a long and fascinating history. One of the most comprehensive reviews was published by Kahn [1] before the advent of personal computing and covers all the major uses of cryptography up until then.

There are many types of cryptographic systems in existence today offering different levels of information protection, and are often broadly classified as *strong* and *weak*. The distinction is drawn in the length of time and amount of resources that would be required to theoretically crack the algorithm used. Strong cryptography is generally considered to be where the resources required are excessively large and the length of time taken to retrieve the plaintext is well in excess of the time in which the information would hold any value. Weak cryptography is where this is not the case and the algorithm used is vulnerable to cryptographic attack within the time that the information is useful.

This distinction is of course constantly changing as faster computers are developed and new mathematical methods are developed for analysing algorithms. For information that can remain sensitive over several decades, this is a particularly important consideration in system design.

In general, strong cryptography uses the same building blocks as weak cryptography and therefore it is beneficial to consider some of these simple ciphers to understand the impact of strong cryptography and its strengths and weaknesses.

3.3.1 Cryptosystem Basics

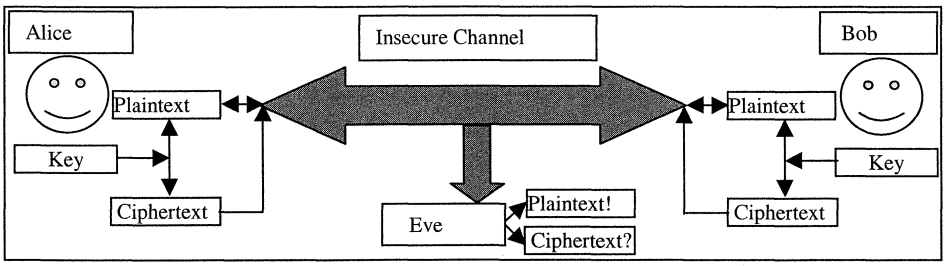

Figure 1 – Basic Cryptosystem

Suppose Alice wishes to send Bob a message through an unsecured channel then Eve, a third-party, may be able to intercept the message, either passively or actively. If the message is sent as *plaintext*, information that can be understood easily, then Eve can tap the channel and passively listen to the messages and / or actively change the messages in some way without being easily detected.

Encryption is the process of encoding a plaintext message to obtain *ciphertext*. The reverse process, to recover the plaintext message, is called *decryption*. Ideally only the person to whom the message is sent to can decipher the message and determine whether it is authentic. A system to encrypt and decrypt is referred to as a *cryptosystem*, and is illustrated above. The purpose of encryption is to add *confusion* and *diffusion* to the message to prevent it from being analysed successfully.

Many encryption algorithms use a *key*, so that the ciphertext depends on both the original plaintext and the key. This provides an extra level of security because even if the eavesdropper knows the encryption protocol used, the key, and hence the message, remains a secret between the two communicating parties. A key also permits the plaintext message to be encoded in different ways by just changing the key. Therefore, a number of different parties can use the same protocol with independent keys and to only be able to decipher the messages sent to them.

Defining a "Good" Cryptosystem

In broad terms, the objectives and requirements of a cryptosystem are to provide:

1. **Privacy** – the information sent or stored is kept secret and confidential from all but those authorised to see it. It protects against disclosure to unauthorised identities.

2. **Authentication** – corroboration of the source of information and identification of an entity is highly desirable through the use of digital signatures, certification and time-stamping. It provides assurance of someone's identity.

3. **Integrity** - the data sent cannot be altered in any way by an unauthorised person without detection. It protects against unauthorised data alteration.

4. **Non-repudiation** – any entity cannot deny previous commitments or actions. It protects against originator of communications later denying it.

In addition, there are a number of other desirable properties for a good cryptosystem including [2]:

1. The amount of secrecy needed should decide the amount of labour appropriate for the encryption and decryption

2. The set of keys or the enciphering algorithm should be free from complexity, which would make the legitimate encryption or decryption process too lengthy

3. The implementation of the process should be as simple as possible

4. Errors in ciphering should not propagate and cause corruption of further information in the message

5. The size of the enciphered text should be no larger than the text of the original message

The use of all these properties can help in assessing the suitability of a cryptosystem. The only *provably secure* cryptosystems use *one-time pads*, unfeasible for all but military and diplomatic applications. Therefore, a more reasonable approach is to define a system that is "secure enough" or *effectively secure*, which does not have this key distribution problem. Security is then defined in terms of the probability of recovering the plaintext from a given ciphertext by an unauthorised entity under different scenarios. Compromises are also made on according to the software or hardware used and the way in which the data and keys need to be transmitted.

Ultimately it is the implementation of a cryptosystem rather than its design which is often the limiting factor. The Enigma code used by the German armed forces during the Second World War was thought to be effectively unbreakable because of its huge key space. However, carelessness in the form of repeated

messages in two different cryptosystems, sending the same message at the same time every day and poor key security permitted first the Poles and later the Allies to decrypt a significant percentage of messages. Therefore, implementation and management of cryptosystems is as important consideration as the algorithm to be used.

Often decisions are based on the resources currently available to the most significant threat group. The methods of attacking a cryptosystem will be described in the next subsection as a precursor to understanding how the strength of cryptosystems is defined and how some systems can be significantly weakened.

Attacks on Cryptosystems

The goal of *cryptanalysts* is to recover plaintext from ciphertext and investigate the effective security of the algorithms used. They generally use a variety of techniques to weaken or investigate the effectiveness of a cryptosystem through:

- **Impersonation** – a third party pretending to be an authorised entity. This is an active attack and normally requires extensive manipulation of the network to highjack or spoof a session.

- **Integrity Violation** – the data flow in the channel is altered by an unauthorised entity. This form of active attack is not as powerful as impersonation, but can be effective in creating a denial of service, preventing communication, or interpreting system implementation.

- **Illegitimate Use** – an unauthorised person gains access to the system using a backdoor, system vulnerabilities, insider attack etc. This is the most common form of attack on the Internet, though does not immediately imply access to encrypted data.

- **Information Disclosure** – encrypted information can be decrypted by an unauthorised entity. Once system security has been breached, further information can be gathered. This is generally a passive attack and difficult to detect.

From the gathered ciphertext and any plaintext disclosed by system weaknesses, a cryptanalyst can use a number of basic attacks on the cryptosystem typically following one of these patterns:

1. **attempt to break a single message** – this is the most difficult attack, often relying on a *brute force attack*, and generally impossible within a reasonable period of time for a strong cryptosystem

2. **attempt to recognise patterns in the ciphertext** – with enough ciphertext using the same key it becomes easier to determine patterns in the messages, and therefore decipher the messages

3. **attempt to find a weakness in the cryptosystem** – this is one of the most common forms of attack as a good analysis of the cryptosystem can reveal weaknesses which dramatically reduce the number of keys which could have been used to encrypt the message

For a cryptosystem without weaknesses or an unknown system, there is a range of attack strategies based on the amount of information the cryptanalyst possesses:

1. **Ciphertext Only** – the analyst only has ciphertext with which to work. The decryption has to be done based on mathematical and statistical analysis.

2. **Probable Plaintext** – additional information can make decryption significantly easier, for example if the structure of the messages is known to be a standard memo format

3. **Full Plaintext** – if the analyst is in possession of the ciphertext of a plaintext message then all that needs to be identified is the algorithm or key which has been used to transform one to the other

4. **Chosen Plaintext** – if the analyst has infiltrated the sender's system, then he can change data and examine the effect on the encryption process, which in turn can reveal useful information about the algorithm and key used.

5. **Chosen Ciphertext** – in this case the analyst has access to the algorithm and ciphertext. By running massive amounts of plaintext through the algorithm and analysing the results, it is possible to deduce the key used. This approach will fail however if two or more distinct keys can produce the same ciphertext from different plaintexts.

With these attacks go a good knowledge of language structure and computer protocols as well as statistical and mathematical tools and lots of ingenuity and luck. Perhaps the most important asset though is computing power, and this has increased dramatically over the last twenty years, typically by an order of magnitude every five years.

The basis of many of the current strong cryptosystems is that it would take an unfeasibly long time to carry out a brute force attack with current techniques and technology. Encrypted data can remain sensitive for many decades and it is not uncommon for encrypted channels to be logged for later interpretation. Therefore, risk assessment has to consider these factors in the design of a cryptosystem.

Generally, however an attacker will try other methods before a brute force attack, such as analysing the use of the cryptosystem. The story of the German Enigma machine during World War 2 is a good example of how a good cryptosystem can be weakened by poor implementation and management [3]. Similarly the majority of Internet security problems are with access and control authentication because these are the most vulnerable to attack.

With these attacks in mind, we will now consider some basic ciphers to understand where vulnerabilities arise in algorithms. The mathematical nature of cryptanalysis permits a rigorous investigation of systems, the details of which are beyond the scope of this chapter, but can be found in numerous books [4]. There is a distinction between symmetric (a single private key) cryptosystems and asymmetric (two keys, one public and one private) cryptosystems, which will be discussed in more detail later. Both systems use mathematical operations to encrypt and decrypt data; in the asymmetric case the same algorithm is used but with the conjugate key and the symmetric key the operations are applied in reverse using the same key. In the following subsection we will consider the building blocks of a symmetric cipher.

3.3.2 Basic Symmetric Ciphers

In building a good cryptosystem using the criteria outlined in the last sub-section, two basic building blocks are used which provide diffusion (*transposition*) and confusion (*substitution*) of the plaintext message. These basic cipher techniques are mixed and repeated according to whether the information needs to be sent in the form of a continuous stream or in the form of discrete blocks.

Substitution Ciphers

Substitution ciphers map the normal (plaintext) alphabet to an altered (ciphertext) alphabet to provide confusion. This was used almost exclusively in historical cipher systems, and dates back to the Egyptians who used non-standard hieroglyphics more than 3000 years ago.

The simplest substitution or *monoalphabetic cipher* is often called a Caesar cipher and is commonly found in many children's comics and books. It works by mapping each plaintext letter to a different letter of the alphabet. It is obviously a very simple encoding process, and can be quickly decrypted. It is worth noting that hypothetically this is quite a secure encoding message scheme, as a brute force attack would have to examine 26! ($\sim 10^{26}$) possible decipherments. By today's standards however, it is very weak and can be easily decrypted using a number of basic analytical methods based on language structure.

A refinement of the monoalphabetic substitution cipher is to use several of these substitutions in a rotating sequence to help minimize any structure in the ciphertext. This introduces the possibility of a key to denote the sequence in which the substitutions are applied. Although must stronger, such a cipher is also not immune to breaking. A novel approach to security is to use a key which itself is a message and can be used a dummy for the real message even if the ciphertext is decrypted. An in-depth description of techniques for analysing polyalphabetic substitution ciphers is given by Pfleeger [5].

It is possible to implement a provably secure cryptosystem using only a substitution cipher. Gilbert Vernam at AT&T showed that using a random number sequence from a magnetic tape mixed with the keys on a teletype to give the ciphertext is secure under the conditions outlined previously. This can also be done with binary information using an exclusive-or (XOR) between the random binary stream and the binary message. Again, the only problem with this type of cipher is the distribution and management of the one-time tapes.

One way around to solve this distribution and management problem is to use a pseudo-random number generator. These are generally not real random numbers, but numbers with a very long period. The general method of generating them using a computer is using the following equation: $n_{i+1} = c * n_i + b \bmod(w)$ where c and b are constants and w is a large integer and n_i and n_{i+1} are the numbers in the psuedo-random sequence. For short messages this is reasonably secure, though with enough ciphertext it is possible to determine the sequence and calculate the values of c, b & w, through a probable word attack. It is important that the random numbers used don't have some underlying structure, for example, telephone numbers or letter sequences from a book; otherwise, they are vulnerable to analytic techniques.

All the substitution ciphers described in this sub-section are linear in complexity and in implementation time. This means the length of time taken to encrypt or decrypt the message are linearly dependent on the number of characters in the message. This is not true for many of the commonly used algorithms used today,

which are nonlinear in complexity with message length and therefore can require significant lengths of time to encode/decode long messages.

Transposition Ciphers

The second goal of a good encryption system is to provide diffusion of the message throughout the ciphertext. This can be achieved by rearranging the letters of the message in a predetermined way, using a *transposition* or *permutation*. The purpose of rearranging the letters is to break any patterns in the message that may assist a cryptanalyst in deciphering a message.

One of the most basic transpositions is the *columnar transposition*. In this case, the plaintext is arranged in as the rows of a 2D array which are then read off in column format. The complexity of this cipher is different from that of the substitution ciphers, because it requires the whole message to be present before encryption or decryption can take place, and relies on the ability to perform the transposition on the whole ciphertext at one time. Therefore, storage is a more significant problem and limits the length of messages on which such transpositions can be performed.

Again, the structure of the English language can be used by a cryptanalyst to decipher transpositions. By applying columnar transpositions of different column size two or more times to a message, however greater security can be achieved. There is however in general a mathematical relationship between the plaintext position and ciphertext position, which can be discovered by looking for unusual digrams, such as QU and XE. Thus, the major disadvantage of this scheme is that the same mapping is applied to all letters of the plaintext.

Independently both transposition and substitution ciphers are weak, however when applied together in a repeated combination they form the basis of the majority of modern strong cryptosystems. The application is usually the main determinant in which cryptosystem is used; some applications require a stream of data to be sent, others can work with blocks of data.

Stream and Block Ciphers

In computer systems there are generally two types of communication channels, *block mode* and *bit streams*. The bit mode provides a stream of bits that are generally not related, for example the keystrokes from a keyboard. Block mode is where a definite amount of related data is transferred in well-defined blocks. Both of these channels can be encrypted using different techniques, each with its merits and disadvantages.

Stream Ciphers

Stream ciphers have the advantage that only one symbol is being encrypted at a time, so there are no transmission or reception delays and the production of the ciphertext is generally fast. The second advantage is that errors do not propagate because of this separate encoding, though conversely, there is no error correction or data integrity, which would prevent malicious data attacks on the stream. Therefore, "bit-flipping" attacks can be quite effective in corrupting data streams. The low diffusion, or the order in which the data is being transmitted remains the same, can also be used to decipher the stream. Only substitution ciphers can be used for bit streams because of the 1D nature of the data.

An example of a stream cipher is the RC4 algorithm, which uses a 2048 bit key to generate an 8-bit output. It was originally kept secret by RSA Data Security, Inc. (RSADSI), but was successfully reverse-engineered in 1994 to give a simple algorithm [6], which can be easily memorised and implemented with fast operation [7]. It is used in a number of applications, such as Netscape, Microsoft Internet Explorer, Lotus Notes, Microsoft Windows, Microsoft Access, Adobe Acrobat, Oracle Secure SQL and many others, but usually implemented in such a way that the key stream can be recovered and encrypted messages decrypted. This illustrates the problem of treating a cryptosystem as a black box and the dangers of using a bit stream.

Stream ciphers are particularly useful for real-time applications where data levels are dependent on users and applications. The fast encryption and decryption times required for real-time applications does however limit the complexity and security of algorithms compared to block ciphers, which are more suited to long term storage and large data transfers.

Block Ciphers

Block ciphers in contrast to stream ciphers encrypt a block of data. Both substitution and transposition ciphers can be used to encrypt the data providing greater security, diffusion of the plaintext through the ciphertext and the ability to check for errors and the integrity of the data. However, in general, they are more complex algorithms making encryption slower, and a single error bit will cause the loss of a whole data block.

To compensate against errors, data integrity can be protected in a number of ways. If each block is encrypted independently (Electronic Codebook or ECB) this is susceptible to cut & paste attacks, or in other words block replacement. By adding interdependence, or chaining part of the previous block into the

current block (Cipher block chaining or CBC), data integrity is greatly improved.

Symmetric block ciphers are the basis of most common cryptosystems, and include DES, AES, RC5 and RC6, described later. Their fast data throughputs and security features make them attractive for most applications.

In the next subsection we will consider some of the cryptosystems currently in use and their applications, strengths and vulnerabilities.

3.3.3 Modern Cryptosystems

There are two main types of cryptosystems currently in use. Traditional cryptosystems where the sender and receiver share a common key are referred to as *symmetric* or *private key* systems. The development of or *asymmetric* or *public key* systems however have had a more significant impact on personal computing in the last decade as the mathematical algorithms and computing power has made them realisable for individual users.

The rapid development of computer hardware and software and their now widespread use in commerce has lead to demands for information security systems to protect sensitive information. IBM led the way in the 1970's with the development of several cryptosystems including what was to become the Data Encryption Standard (DES) for the USA. DES became an official standard for unclassified information in 1977 and still remains one of the best known, scrutinised, and used cryptosystems, with widespread commercial usage.

This section will begin by looking at private key systems, in particular the Data Encryption Standard (DES), which has become a universal symmetric key system, used in many commercial transactions and non-secret governmental work. Questions about the algorithm have always been raised however and recently support for it has been withdrawn in certain areas. The contenders for the replacement of this standard, the Advanced Encryption Standard (AES), will also be examined with a number of other more modern symmetric key systems that show promise.

The major problem with private key systems however is key management and this problem can be addressed using public key cryptosystems, which allow the secure distribution of private keys. We will look at two algorithms used in public key cryptography, the first of which, the Merkle-Hellman knapsack, is a good example of a solid system, which was found to be susceptible to cryptanalysis. The second, the Rivest-Shamir-Adelman (RSA) algorithm appeared at approximately the same time, however it has withstood extensive

attack, with no serious flaws being found. It can however be costly and resource intensive to implement.

All of these modern algorithms however rely on the same concept of a *hard* mathematical problem. A hard problem is one where the number of possible solutions for the key is large and an exhaustive search of the key space is expected to be infeasible. Most of these problems are NP-complete problems with a key space that is at least exponentially dependent on the length of key, though there are some areas of mathematics such as Galois fields and factoring large numbers which are not NP-complete, but which are considered difficult and used as the basis for secure encryption. In general however, there is no guarantee that a simple solution to these problems will not be found, or that algorithm and computation systems for analysing these problems will not make them breakable in the future. This is the biggest problem concerning current crypto-algorithms and needs to be taken into account when sending data that could remain sensitive for many years.

The other important property necessary for use in cryptosystems is that these problems should be easy to compute. The more complicated the algorithm then the en/decryption process will be slower, limiting potential applications. There is also a trade off between speed and redundancy; increased message redundancy can lengthen the odds of successful decryption at the expense of increased byte size. This can permit a message that can reveal one of a number of messages depending on the key used for the decryption, useful for multiple recipients of different trust levels.

Symmetric (Private) Key Cryptosystems

Most of the historical ciphers used a common key for both the sender and recipient, or what is now known as a symmetric key system. Symmetric key systems are vulnerable if the key is discovered from either party and this forms part of the weakness with the management of the secret keys. It does however provide a more immediate level of authentication than asymmetric systems, because only someone with the same key can send a message which can be decrypted successfully.

Once the key has been compromised, all the messages sent are vulnerable to be deciphering and bogus messages can be sent to either party that will appear authentic. Thus *key management*, the delivery and security of the secret key, is one of the major concerns of any symmetric-key cryptosystem. This can be done in a number of ways, for example using a secure channel, or by fracturing the key and sending it by a number of different means to complicate interception.

Traditionally this is why symmetric key cryptosystems have been limited to large organisations that can distribute keys with sufficient security.

A related problem is the number of keys required in a multi-user system. To illustrate how this problem snowballs with increasing network size, each new user needs a unique key to correspond with all other users and vice-versa, so for a 100 user network 4950 unique keys would be required to give secure and independent communication between each of the users. It is desirable to generate keys at sufficient time intervals to prevent unwanted interception and for large amounts of data this can be within a day. Therefore, the distribution of keys in a secure fashion can be a significant problem even on a local area network.

The main advantage of symmetric key cryptosystems is the speed of encoding and decoding because shorter key lengths are required to offer equivalent cryptanalytical difficulty to longer key length asymmetric key systems. For comparison, symmetric systems are about three orders of magnitude faster than public key systems for the same level of security. This is why most high speed communications links and hardware cryptography units use symmetric keys because computer data throughputs can be de/encrypted in real-time, making it transparent for users.

Data Encryption Standard - DES

In the early 1970's the US government recognised a need for a secure encryption technique which could be used by the public for sensitive information, in particular by the banking community. In 1972 the U.S. National Bureau of Standards (NBS) issued a call for algorithms which would fulfil the following desirable criteria:

1. It must provide a high level of security
2. It must be completely specified and easy to understand
3. The function of the algorithm is to provide security; the security should not depend on the secrecy of the algorithm
4. It must be available to all users
5. It must be adaptable for use in diverse applications
6. It must be economical to implement in electronic devices
7. It must be efficient to use

8. It must be able to be validated

9. It must be exportable

These criteria are similar in goal to those outlined for a "good" cryptosystem earlier, and indicates some of the forward thinking of the NBS at the time for example in its implementation in hardware (6), in public scrutiny (3) and world-wide marketing (9). Response was poor and a second call was issued before it was decided to develop the "Lucifer" algorithm submitted by IBM, which had been working on it for several years. IBM further developed Lucifer under licence to form what became known as the Data Encryption Standard (DES) or more accurately the Data Encryption Algorithm (DEA within the USA, DEA-1 outside the USA). DES was officially adopted as a federal standard in1976 and later as an international standard by the International Standards Organisation (ISO).

What is interesting is that DES was constructed so as not to be vulnerable to differential cryptanalysis, a method of looking for correlations between the input and output of each round. This concept has only been in the public domain since 1990, though there is strong evidence that IBM and NSA knew of differential cryptanalysis at the time of DES development. This is highlighted by the fact that similar Feistel (or product) ciphers until them were all susceptible to analysis using differential techniques. With the decision by the US government to relax its cryptographic export rules recently, this has led many people to believe the National Security Agency (NSA) has developed mathematical and computer systems capable of breaking current algorithms with relative ease, though it will be many years again before they are in the public domain.

Since its release there has been growing concern about the DES on a number of levels. Initially there was a suspicion that the NSA had introduced a trap door function into the algorithm to allow covert monitoring of traffic. A Congressional committee however exonerated the NSA of improper involvement in a classified hearing though curiously the NSA has withdrawn further support for DES recently. This is perhaps because computing power has grown to a sufficient extent that DES can be cracked by brute force. In 1998 a specifically designed computer costing less than $250,000 was used to crack a DES challenge in less than 3 days [8], while the following year a distributed network program working on a similar challenge took 22 hours and 15 minutes [9]. Theoretically, application specific ICs (ASICs) can examine more than 100 million keys a second per chip and they cost $10 which ordered in 5000+ quantities, meaning than for:

- $50,000 = 500 billion keys/sec = 56 bit DES in 20 hours = 40 bit key in 1 second

- $1M = 10 trillion keys/sec = 56 bit DES in 1 hour = 40 bit key in 50ms

- $10M = 100 trillion keys/sec = 56 DES in 6 mins = 40 bit key in 5ms

This is well within the expenditure of many security organisations and an estimated value for DES is now put at 8 cents; if your secret is worth more than 8 cents its not worth encrypting with DES [7]. Indeed in 1998 the German Courts ruled that DES was "out of date and unsafe" for financial applications.

There are a number of possible ways of strengthening the existing DES algorithm. Interestingly the original Lucifer algorithm used a 128-bit key and this level of security can be approached by using a triple-DES algorithm, giving an effective key length of 112 bits. With such a scheme, typical software can achieve encryption rates of approximately a MB/sec, which is several times slower than the data rate from a hard-disc or network information and therefore real-time usage would be difficult to implement for high bandwidth applications. Such developments are however more likely to be superseded by the next generation of symmetric key standards.

Advanced Encryption Standard – AES

In 1997, the National Institute of Standards and Technology (NIST) announced the AES development effort with a call for algorithms. The algorithms were to be unclassified which could be publicly disclosed and available royalty-free world-wide using a symmetric key cryptosystem of at least 128, 192 and 256 bits. By the end of August 1998 fifteen candidates were chosen for further testing and from these five were short-listed in April 1999. The five finalists were:

1. **MARS [10]** – this algorithm was developed by IBM and was 'tweaked' after the first round to improve the random number generation. It provides relatively fast encryption and decryption though with a relatively slow key setup.

2. **RC6 [11]** – RC6 is the RSA Labs entry into the AES competition, and is the successor for the RC5 algorithm. In was developed in collaboration with Ron Rivest and is the fastest of the algorithms for encryption and decryption.

3. **Rijndael [12]** – The strange name for this algorithm is derived from the names of the authors, Joan Daemen and Vincent Rijmen, at the University of Leuven in Belgium. The speed of the algorithm appears to be affected by key length giving variable performance.

4. **Serpent [13]** – this internationally developed algorithm was developed with a long term view, using prudence and experience to provide the best security against short-cut attacks, but at a cost of speed, making it the slowest algorithm.

5. **Twofish [14]** - Twofish is a 16 round algorithm developed by Bruce Schneier and colleagues at Counterpane. It is reasonably fast, though an attack against the algorithm has already been recorded [15].

The NSA was asked by NIST to help test the algorithms both in software and hardware implementations and to investigate their strength. On 2nd October 2000, NIST announced that the winner of the proposed AES is Rijndael. Rijndael was chosen for its flexibility, low overhead implementation and speed in both hardware and software. The raw key space for a 128-bit key is approximately 10^{21} times greater than for DES and therefore unlikely to be attacked in the near future without significant improvement in computing power or if a weakness is found in the algorithm. From a legal perspective there is no plans currently to limit export of AES implementations. Further details of the AES competition can be found on the Internet [16].

With so much commercial and critical infrastructure using DES, the crossover to AES will be a slow process. The speed of its implementation will be a good guide to commercial attitude towards encryption and data security and in addition, a guide to international fears of US monitoring and decryption capabilities. The choice of a foreign algorithm the use of open design and testing procedures and the relaxing of export licensing for encryption algorithms are all welcome developments.

Other Private Key Cryptosystems

A number of other common symmetric key algorithms exist which include:

- **RC2** – This is the companion to the RC4 stream cipher – again it was a RSADSI trade secret, but was reverse-engineered and posted to the Internet in 1996. It uses a 1024 bit key and has a relatively low data throughput rate.

- **IDEA** – This cipher has undergone several name changes as it has developed – originally as PES, IPES and now IDEA. It is incorporated

into PGP and uses a 128-bit key but its use is currently covered by a patent. Again it is relatively slow at data encryption.

- **Blowfish** – This is a 448-bit key system, which was optimised for high speed operation on 32-bit processors. The data throughput can typically reach several MB/s with current technology and is amongst the fastest block protocols.

- **CAST** – Carlisle Adams and Stafford Tavares of Nortel designed this system with a 128-bit key. Although it has a good formal basis, it is a relatively new cryptosystem and has not been as extensively tested for weakness as some of the others. This has similar performance to Blowfish and both can cope with slow network throughput in real-time.

- **Skipjack** – After the failure to gain widespread support for Clipper, this algorithm was declassified in 1998. It uses an 80-bit key in 32 rounds, though is considered too weak for long term security; it also has slower data throughput than the more modern algorithms.

- **GOST** – This is a Russian algorithm developed as an equivalent to DES. The specification is not completely in the public domain, but uses a 256-bit key in 32 rounds. This has similar data rates to 3-DES and is not recommended as an algorithm to implement.

The vulnerability of all the symmetric key systems described in this subsection to new mathematical and computation techniques cannot be overstated. In the short term, the development of new mathematical techniques and problems will open up new encryption algorithms based on similar "hard" problems. Some of these techniques are discussed in Section 3.3.5 along with new protocols such as quantum cryptography, which may eventually overcome all these vulnerabilities.

The impact of AES will act as a good gauge to business concerns over encryption and data privacy. Currently there are no alternatives to the symmetric key systems described, with all algorithms using transpositions and substitutions. The major developments are likely to continue to be increasing the key length used and the complexity of the rounds used in enciphering the plaintext. Unless there are significant advances in realising a quantum processor or in cryptanalysis then current symmetric algorithms are likely to remain effective for at least the next decade.

The complementary asymmetric cryptosystems are not as mature as symmetric systems and therefore will face more radical changes over the next decade. Currently the majority of innovative mathematical research and interest is on asymmetric key problems, the class of cryptosystem we will consider next.

Asymmetric (Public) Key Cryptosystems

The primary advantage of public key systems is that a user can send a secret message to any other user without the need for a secure channel for key distribution. This offers the possibility of a network managed key distribution system, where private, symmetric keys can be distributed using the public key system to allow the use of faster private key encryption protocols.

The most common implementation of an asymmetric key system is in *public-key cryptography*. In this case one of the keys for encrypting information is distributed, the *public-key*, and a corresponding *private-key* is kept secret for decryption. Therefore anybody with a copy of your public key can send you information which only you can decipher. This significantly decreases the number of keys required for a multiple user system.

Cryptosystems that use different keys to encipher and decipher messages are a relatively recent concept in cryptographic history. This concept was publicly introduced in 1976 by Diffie and Hellman [17], though there is evidence that the British Secret Service invented it a few years earlier [18].

The primary disadvantage of a public key system is that if it is compromised it can be extremely damaging. The problem arises because the key in a public key system is typically used hundreds or thousands of times. Private key systems however generally use the key once per message limiting data risk to that message if the key is compromised. The breach of a public key pair also has implications in other areas such as impersonation and fraud and therefore needs to be taken very seriously in mass usage schemes where such encryption would be treated as a black box. This leads to a desire for much greater key lengths, typically 512 bits for real time applications, 1024 bits for short term (1 year) use and 2048 bits for longer-term usage. It also requires users to keep their private key securely so as not to be compromised. Theoretical discussions have already taken place about how viruses and sniffers could be written to detect private keys held on computers and distribute them to unauthorised people or interfere with authentication [19]. Undoubtedly as encryption becomes more widespread and brute force attacks become more difficult, alternative tricks such as these will become a reality.

The second issue is that there is an inherent trade off in all encryption systems and in particular in public key cryptosystems: the longer the key used to encrypt the message, the longer in theory it will take to decipher the message using brute force. However messages encrypted with longer keys also require more processing and therefore limit the throughput of the cryptosystem. As mentioned previously, in a public key cryptosystem it typically takes a thousand times longer

to encode a message to the same level of security as in a private key system, limiting the use to non-real-time and short message situations.

Even with these disadvantages, public key cryptosystems are commonly implemented, in particular for sending e-mail. Some common examples of asymmetric cryptosystems are Elgamal (named after its inventor Taher Elgamal), RSA (named after Ron Rivest, Adi Shamir and Leonard Adleman), Diffie-Hellman and DSA (Digital Signature Algorithm – invented by David Kravitz). We will now look at some of these systems in more detail.

Merkle-Hellman Knapsacks

In 1978 an encryption algorithm based on the knapsack problem was reported [20]. The knapsack problem uses two large, relatively prime numbers to calculate an encryption and decryption set related by a modular operation. A brute force to determine the two numbers will yield the plaintext, but for 200 binary digit relatively prime numbers this would take 10^{47} years if one possibility could be computed every microsecond [5].

However, the interceptor does not need to solve the basic problem to break this encryption system. Shamir [21,22] identified a series of flaws in these transform knapsacks that reduces the key search time from being exponentially dependent on key length to polynomially dependent, making it easily computable within a reasonable period. Although other transform knapsacks have been proposed, they do not seem secure enough compared to other commonly used algorithms and can be difficult to implement.

There is a useful moral to be learnt from the knapsack story. Although some algorithms may appear to offer high security because of the long key lengths, the only true test is for the algorithm to be scrutinised worldwide by cryptanalysts. Security through obscurity is a poor second choice.

Rivest-Shamir-Adelman (RSA) Encryption

A second public key system was announced in the same year as the knapsack problem based on another underlying hard problem by Rivest Shamir and Adelman [23]. To date no serious flaws have been identified after extensive analysis suggesting a fair degree of confidence, though like all cryptosystems this is under the caveat of current methods and technology. It provides both signature checking and encryption using the same algorithm, though is currently covered by a patent which expired September 2000.

On face value the RSA algorithm is similar to the knapsack problem. It uses two large prime numbers as the keys that are complementary inverses through a modulus operation. The mathematical difficulty comes in factoring large numbers, which is not known to be NP-complete and therefore at the least brute force decryption is exponential in time with key length.

There is no quick way in determining whether a large number is prime, instead a number of heuristic algorithms are used to generate keys with a given probability of being prime [24]. Phil Zimmerman detailed a method for computing a RSA cryptosystem efficiently in 1986 [25], which later formed part of the Pretty Good Privacy software package.

Pretty Good Privacy (PGP)

Phil Zimmerman released PGP in 1991 in response to an attempt by the US government to gain access to all information channels, free of encryption. PGP is a hybrid cryptosystem, using both symmetric and asymmetric ciphers. The plaintext of the message is first compressed if applicable, reducing its size and decreasing any structure. A random number, the *session key*, is then generated which is used to encrypt the compressed plaintext using a conventional symmetric key algorithm to create the ciphertext. The session key is then encrypted with the recipient's public key, and this is added to the ciphertext to form the complete message.

When the message is received by the intended person the message is split again into the two parts. The private key can recover the session key using a public-key algorithm. The session key can then be used to decrypt the ciphertext, and the plaintext retrieved after decompression. The combination of both encryption methods combines the convenience of public key encryption with the speed of symmetric key encryption. The symmetric ciphers available in PGP are CAST and IDEA, working with 128 bit keys and Triple DES, which uses a 168-bit key. These ciphers are used in cipher feedback mode on 64-bit blocks. Although Triple DES is the best studied of these ciphers, it is also the slowest. Key management is achieved by storing public keys and private keys in the form of *keyrings* on the users hard disc. The private key is encrypted in the private keyring using a hash of a *passphrase*, generally a sentence created by the user containing numbers letters and punctuation, which provides, in theory, more security than a password. To date no serious weaknesses have been found in the cryptosystem and it is one of the most widespread public key systems.

The story of PGP is interesting on a number of levels. It is a good example of the ethos of the Open Source movement, battling large organisations (in this case the US government and MIT which holds the patent for the RSA algorithm

in the US), providing source code for public scrutiny, and using publicity to good effect. It is also a good indication of the current trends of software development and attitudes on the Internet to privacy and non-disclosure issues. PGP is now developed worldwide and is being pushed hard to become the *de facto* encryption plug-in for the common e-mail readers. Its flexibility, and multi-platform implementation is one of its key attractions, including such extras as digital signatures and fingerprints.

Digital Fingerprints, Signatures & Certificates

In addition to encryption, cryptographic algorithms can also be used for two other complementary procedures. These procedures, fingerprints and signatures, are not used to convey a message but to provide integrity and authentication checks respectively. They are both based on the concept of hash functions.

Hash functions have been used in computer security since the beginning, in particular to store passwords. For example, in Unix-like operating systems the hash value of the password is stored and when the password is entered the hash value recalculated and if it matches the stored hash value, then access is granted. This prevents all passwords being made vulnerable by access to the stored file; instead, a brute force attack needs to use a dictionary search with each hash value computed and compared against the stored string.

Hash functions are usefully employed in data integrity and authentication where they provide an authentication against tampering and a check on who encrypted the message. Each hash value, or message digest, generated for a message can be encrypted with the message to provide verification, or stored or sent independently to act against non-repudiation. These properties make these techniques particularly useful in a commercial or legal framework.

Digital Fingerprints

The concept of a fingerprint is to reduce a variable length input (the plaintext or ciphertext message) to a fixed length output, typically 128 or 160 bits, essentially acting as a checksum for the message. From the fingerprint is not possible to deduce the input, it is difficult to recalculate the fingerprint given a change in the input and without the key and algorithm and it is difficult to find two inputs that give the same output. The flexibility of fingerprinting also permits it to be used on any kind of input or message and is routinely used to protect plaintext and binary files from corruption while they are distributed, for example on the Internet.

The application of the strength of the encryption algorithm is the key to the power of fingerprinting and signatures. Both systems can make use of other details, for example time-stamps, to provide a legal framework in which electronic data can be filled in the same way as traditional paper for use as commercial and legal evidence.

Digital Signatures

Diffie and Hellman, the inventors of public key cryptography, also introduced the idea of digital signatures. Digital signatures enable the recipient to verify the authenticity of the message and verify the message is intact, providing authentication, integrity and non-repudiation. A digital signature works by encrypting a digital fingerprint using the senders private key. The signature can be decrypted using the senders public key by the recipient and used to verify the authenticity and integrity of the message. As long as a secure hash function is used, then falsification of a digital signature is as difficult as brute force decryption. Digital signatures can be particularly useful in authenticating and validating other users public keys.

Digital Signature Algorithm (DSA)

Along with DES, NIST also commissioned an algorithm for digital signatures, the Digital Signature Algorithm (DSA) [26]. The base algorithm of DSA is a variant of the Elgamal algorithm, which was designed specifically for signatures only, though the same algorithm can be used for encryption as well.

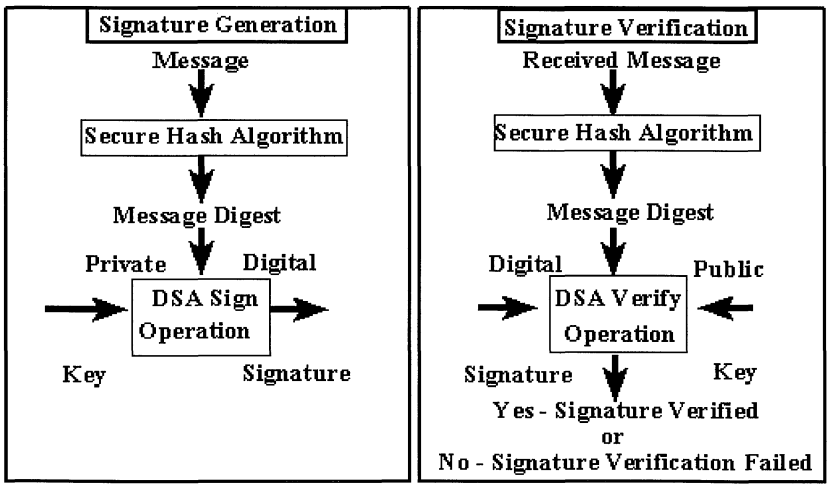

Figure 2 – Flow diagram for digital signature algorithm

The implementation of the DSA algorithm is illustrated above. Verification of the sender's identity comes from a comparison of their message digest found by applying their public key to the signature with an independently computed digest by the receiver. Any changes in either of the keys or the message itself will lead to verification failure. Therefore, both the integrity of the message and the authentication of the sender are preserved. DSA is widely implemented, for example it is included in the PGP package, for signing messages and is not known to suffer from any significant problems.

Digital Certificates

Digital certificates have been developed as a way of helping to establish public key authentication for use in encryption and verifying signatures. The idea is that a certificate acts as a form of credential, such as a driving licence or passport. It can contain a variety of information in addition to the public key and the user's identification, though requires some form of endorsement to be valid. This endorsement is given in the form of digital signatures applied to the certificate by trusted parties. Thus, it provides a means of validating the owner of a public key.

The need for certificates arises from a weakness of any cryptosystem in that messages can be intercepted and retransmitted by an intermediary (man-in-the-middle attack) without detection.

A certificate can also be used to contain other information. There are two main types of information, certificate usage and certificate constraints. The usage information provides extra information about the user whereas the constraints limit the applicability of the certificate. For example, the usage information can include e-mail addresses, key usage period, the CA (Certification Authorities) policy and other general information not found in the basic certificate information. The constraints include whether the certificate is a CA certificate, and any name or policy constraints that need to be applied.

Additional security for certificates can be provided through time stamping, which can be used to prove a document existed at a certain time. A time-stamped countersignature can prove the document was valid at a given time though currently multiple time-stamped signatures are only available using PGP and S/MIME signed data.

Sometimes in management decisions it is necessary for a decision to be made by a number of people. In this case there are protocols for key splitting, where any

subset of a number of users can be required to enter their key in order to sign or encrypt a document on behalf of an organisation.

The primary issue with digital signatures is again public key distribution. The development of a system to disseminate valid keys is one of the key challenges for the acceptance and growth of E-Commerce this decade. Governments are attempting to legislate in some countries, while in others commerce or the computer fraternity is taking the lead. The most contentious issue is who should sign certificates and how should they be distributed.

3.3.4 Key Management

As mentioned, the most difficult aspect of a cryptosystem is the management of the keys for the encryption algorithm. These keys generally fall into two classes, short-term session keys which are used once and then discarded and long-term keys, such as public keys, which are used many times over an extended period. The short-term keys are generally generated automatically by the software as required and then discarded at the end of the session, with no user input. Long term keys however have a higher level of authentication and confidentiality and are usually generated explicitly by a user.

In particular, key management needs to address a number of problems:

- **Distribution of Keys** – a way in which to securely distribute your own public key and obtain someone else's is required

- **Establishing a Secure Channel** – it is important that all parties using an encrypted channel have confidence in who the other parties are and to whom the messages are being sent

- **Key Storage** – the storage of both public, private and secret keys is an important issue so as not to weaken a cryptosystem through system penetration

- **Revocation** – if a key is compromised there needs to be a method of revoking the key and determining which data is still valid.

Keys should however in general be generated by each user and not accepted from a third party to prevent the private key being falling into the hands of unauthorised third parties.

The establishment of keys can now be more easily achieved using the advantages of public key cryptography which permit secure transfer of keys

within the limits of verifying the other user. There are three general trust models for establishing key validity:

- **Direct Trust** – this is the most basic trust system where an individual user validates the key personally

- **Hierarchical Trust** – for large organisations a tree structure can be created, where a meta-introducer, or CA, certifies trusted introducers and other certificates to a user pool, in a tree structure. A certificate's validity can then be established by tracing back to a root certificate issued by the meta-introducer.

- **Web of Trust** – this model is based on the notion that the more information the better, and is cumulative in nature, encompassing both of the other models. As users trust and validate keys then can be signed, published and collected. The level of trust and validity of each signed certificate can then form a complete directory of public-keys.

Key storage is a very careful balance between security and backup requirements. For example it may be desirable to backup a key in case of data loss or for long term storage. However each extra copy of the key is an additional point of failure in the cryptosystem and therefore care should be taken in deciding how to store keys. Often this is done by splitting the key in much the same way as having multiple signatures required to validate a document. These key shares can be distributed amongst trusted third parties, such as family, friends, safe, bank, solicitor etc. and be used to reconstruct the key in case of the loss of the original.

Of these problems, perhaps the most difficult one to deal with effectively is key revocation. One approach is to have a certificate revocation list (CRL), equivalent to the credit-card blacklists. In practice, these would be difficult to implement because of performance issues and being vulnerable to attack, for example from denial-of-service attacks.

Several other methods have been suggested though all suffer from failure to give complete revocation promptly and secure. The Online Certificate Status Protocol (OCSP) is perhaps the most promising though requires a trusted third party to query certificates validity at suitable periods and answer queries about their validity. Alternatives include self-revocation and issuing anti-CRLs, however none of these systems addresses all the problems raised.

Certification Authorities

Some organisations also employ Certification Authorities (CA) to validate certificates and sign genuine ones and only use these to communicate with outside agencies. This introduces a *trusted third party* into the process. For larger organisations one CA cannot perform the whole process and therefore may need to require on *trusted introducers*, who can validate keys, but not create new trusted introducers. An alternative approach can be to create Registration Authorities who are responsible for user processing and checking and act as trusted intermediaries to the CA.

A CA authority works by receiving the public key of a user, which is signed using the private key. This provides a secure means of send the public key to the CA and that the user holds the private key. The CA can then authenticate the user through other means and sign the key with its own signature and return the result to the user. The user can then verify the CA's signature and after this can issue their public key themselves or through the CA.

A CA generally works with a profile as defined by one of many available standards to configure these extensions and tailor the certificate for a particular environment. For example ISO 15782 deals with certificate management for banking, the US DoD use MISSI, the USS Federal Government use FPKI, the Australian government PKAF and Microsoft their own profile. In general, these policies define a number of conditions that may include:

- Obligations of the User
- Obligations of the CA – CA security
- Certification Publication and Issuing Details
- Updating of Keys
- Confidentiality Policy
- Applications for which the certificate may be used / not used
- CA liability
- Financial Responsibility – indemnification of the CA
- Compliance Auditing
- Security Auditing
- Disaster Recovery

Example of Certification – X.500 Directory

There was a desire to construct a global directory which would give a distinguishing name to everyone on earth. Concerns about the misuse of the directory were to be addressed though the use of X.509 certificates which would provide access control. The implementation of such certificates however is a lot more

straightforward in government and military structures, though not so easily achieved with individuals and small businesses.

For example in constructing the X.500 directory you would typically require at least three levels of CA, from the department one in which a person worked, to the company one to the national one. In practice however such a structure is flat with no hierarchy and CA certificates are limited mainly to web usage and adding plug-ins to software.

It is hoped however X.500 directories may still succeed in being repositories of public keys based on the system outlined, however within many large organisations alternative directory schemes already exist and the change over to an X.500 directory structure may be long and slow. The use of LDAP servers (Lightweight Directory Access Protocol) on the Internet to access X.500 directories is growing rapidly and can be accessed through the Internet using many common programs (for example MSIE, Netscape etc) and therefore is likely to grow in usage for public activities. However, a better structure needs to be developed to address a number of the shortcomings and structural problems.

The main shortcomings of the X.509 protocol are that the full infrastructure does not exist to issue and revoke certificates, with workarounds having been introduced to hide the problems and progress towards a more complete standard going extremely slowly. The second major problem is that certificates are issued based on identity and not on the key and therefore revocation because of change of location, or the ownership of multiple keys becomes difficult. The other difficulties with X.509 certificates are that they do not fully distinguish between authentication and confidentiality keys, such that authentication keys can validate user issued confidentiality keys which can be used on a short term basis, and with the growing amount of information included with certificates so that they almost become dossiers on the user.

Some of these shortcomings are addressed by PGP Certificates. PGP certificates are key-based, not identity-based and therefore multiple certificates can be issued to one person, or a key shared between multiple people. Authentication keys can also be used to certify confidentiality keys, removing two of the largest problems of the X.509 protocol. Further, the Simple Public Key Infrastructure (SPKI) binds a key to an authorisation or capability (i.e. the power), which can be distributed directly or via a directory and which contains a minimum of information.

3.3.5 Future Advances in Cryptography

Symmetric key cryptography is a mature technology, having existed essentially unchanged for hundred of years. In order to keep pace with the development of mathematical and computational trends, symmetric algorithms have been given an increased complexity in the design phase and longer keys, though still use the basic building blocks of transposition and substitution. These schemes are very successful in securing data and unlikely to change dramatically in the near future.

In contrast, asymmetric systems are still evolving after only twenty-five years in existence. A number of new algorithms are being developed around mathematically hard problems, which will provide a new level of difficulty beyond the analysis of existing problems. The main drive for these algorithms is to reduce the key length and number of mathematical steps required as this ultimately determines the encryption speed. Cryptanalysis of algorithms is still at a relatively early stage and it is highly probably new techniques will be discovered. In the longer term asymmetric key algorithms, if optimised, may compete with symmetric encryption in some applications. In any case its future is assured from the growing interest in digital signatures, fingerprints and certificates for business and personal applications over networks.

The biggest paradigm shift in cryptography over the next decade will probably come out of the application of quantum systems to cryptographic problems. Two quantum systems in particular will effect cryptography at either end of the spectrum: quantum cryptography will provide a provably secure, commercially viable technique for distributing information; in contrast quantum computers will provide a means to reduce intractable, hard problems to ones within a realisable time scale for current algorithms and key-lengths.

We will now consider some of these changes in more detail and the impact they are likely to have on current use of cryptography.

3.3.5.1 New Asymmetric Algorithms

Current asymmetric algorithms are based around factoring of prime numbers and discrete logarithms, both of which are not proven "hard", whereas both the knapsack problem and lattice problem have been proven "hard". The difficult in proving the "hardness" is a concern because a shortcut may exist in search for the correct key, reducing a brute force attack to a realistic time frame.

Therefore there is a search for other mathematically problems which can be proved to be "hard". One such problem is in the field of *elliptical curves*

(Elliptical Curve Cryptography - ECC) working with discrete logarithms [27]. Algorithms have been developed around more basic mathematical operations, such as addition and subtraction, and over a harder key space like the group of elliptical curves realising such advantages. This would reduce the key length for a public key system of the same level of difficulty by nearly a tenth.

Private Key (DES)	Public Key (RSA)	Public Key (ECC)
40 bit		
56 bit	400 bit	
64 bit	512 bit	
80 bit	768 bit	
90 bit	1024 bit	160 bit
110 bit	1702 bit	196 bit
120 bit	2048 bit	210 bit
128 bit	2304 bit	256 bit

Table 1 – Relative Strengths of Different Cryptosystems (after [28])

The above table shows the advantage of ECC, especially where high data throughput is required or resources are limited. As with all new cryptosystems however, this strength is dependent on no weaknesses being found through developments in mathematics or weaknesses in the algorithm implementation. In such a new field, these developments pose two major vulnerabilities. The third problem arises from all the patents that have been issued on related technology.

If more simplistic calculations with shorter keys are possible for an asymmetric key system this would be of great commercial interest, allowing real-time transmission of encrypted one-to-many services. The increase in speed is also more generally attractive for hardware and real-time applications that are currently limited.

3.3.5.2 Quantum Cryptography

The ultimate goal of a cryptosystem is to be provably secure. Currently the only schemes in operation use one-time pads, a cumbersome method requiring secure distribution of the pads. In the last few years, there has been an explosion of theoretical and experimental innovations in the understanding of quantum physics. Quantum physics allows the construction of completely new forms of logic gates, provably secure cryptosystems, quantum bits (qubits) and a sort of "teleportation", potentially revolutionising information theory.

Classical information theory agrees with everyday intuition. In quantum theory, "entanglement" of different systems permits two or more states to be encoded on one system and these systems may be non-local. Entanglement also offers the potential for teleportation by interacting the object to be cloned with one of two entangled photons [29]. The provable security arises because an eavesdropper cannot interfere with the entangled system without being detected, due to the quantum mechanical laws governing the process.

There are at least three main types of quantum cryptosystems that solve the key distribution problem securely; these are cryptosystems use random one-time pad encoding based on:

- Two non-commuting observables [30]
- Quantum entanglement and the Bell Theorem [31]
- Two non-orthogonal state vectors [32]

There are however a number of major drawbacks to quantum cryptography currently. The first problem is the range over which information can be sent. Each qubit of information is transmitted using a single photon, which can be absorbed or scattered, the probability of which increases with distance. Free-space trials have been successful over one kilometre and potentially a distance of up to 50km is possible in optical fibre with acceptable data loss [33]. Quantum systems will not be able to replace traditional systems however because it is not possible to build repeaters to amplify and clean the signal or multiplex the signal to groups of users.

The second problem is that an efficient source of single photons and a high speed, efficient detector is required. Presently both the emitter and receiver are slow and inefficient. To some extent, these requirements are being addressed by the large corporations who have already begun to heavily invest in quantum cryptography [34]. A further problem is that high stability of the system is required to key timing and coherence together in the system. Typically, the complete one-time key needs to transmitted within minutes before the system needs to be reset within the stabilised environment of a university laboratory. The practical problems of a commercial environment are yet to be fully understood and a more active method of stabilising the whole communication system need to be found.

These problems currently limit the information rate to the tens of bits per second level for key lengths up to several thousand bits with sufficient implementation difficulty to not be an attractive commercial prospect. These corporations hope these issues can be addressed to allow them to tap into the very lucrative

financial and commercial intranets where security is a high priority. A complementary discussion of the issues of quantum cryptography can also be found in Chapter 2.

3.3.5.3 Quantum Computing and Cryptanalysis

As well as providing a provably secure means of communicating, the quantum world also promises the prospect of systems capable of highly-parallel computations [35]. A single computation on 'n' qubits is equivalent to 2^n serial computations on current computers. Therefore, it does not require many entangled qubits in order to reach the computational levels required for a brute force attack of current cryptosystems [36].

Theoretical algorithms have already been published for cryptanalysis using a quantum computer [37], closely followed by experimental developments. The current limit is 4 qubits in a rather unfriendly configuration, however the commercial and governmental/military [38] pressure and financing to realise a useful system is likely to provide simple systems within the next decade. More exotic computational devices, for example based on DNA [39], will also revolutionise the way in which computations are carried out in the longer term.

If a viable quantum computer were to become available, this would pose a significant threat to all existing commercial cryptosystems. A re-think of the use of cryptography would be required and new methods sought. For this reason, the disclosure of a working machine is likely to come some time after its development, particularly if it is developed by the military or intelligence agencies first. There is a theory that such machines already exist and this is why the US government has relaxed controlling on cryptography and is not afraid to introduce the standard use of longer keys. In any case, it will be at least a decade before the availability of such machines commercially.

Cryptography will have to look to new ideas, such as quantum cryptography, over the next decade to ensure data security for the later half of this century because the effective security of current cryptosystems is likely to be limited to one or two decades. The dramatic increase in data flow does open another possibility for some applications, which is to "hide" sensitive data within larger data files. We will consider this concept further in the next section.

3.4 Information Hiding

3.4.1 Introduction

There are alternative techniques to encryption for improving the security of data transactions. In some cases it is undesirable to use encryption, either because the application does not permit its use, the users do not have sufficient equipment to decrypt the message, or the information is to be covertly transmitted. In this case, information hiding can be used providing a different, complementary technique.

Steganography is the historically used word to describe information hiding and is derived from the Greek translation of "covered writing". As technology has advanced, information hiding now covers a vast array of techniques from invisible ink and microdots to spread spectrum communications. Information hiding is also referred to (especially in military circles) as *transmission security* (TRANSEC). The important distinction is that cryptography scrambles a message so it cannot be understood, whereas steganography disguises a message so it cannot be seen. The most important factor in successfully using steganography is to stay one step ahead of those trying to intercept your message by devising a protocol that cannot be detected. Most commonly, it is used in a complementary fashion with encryption to provide a secure and covert means of communication.

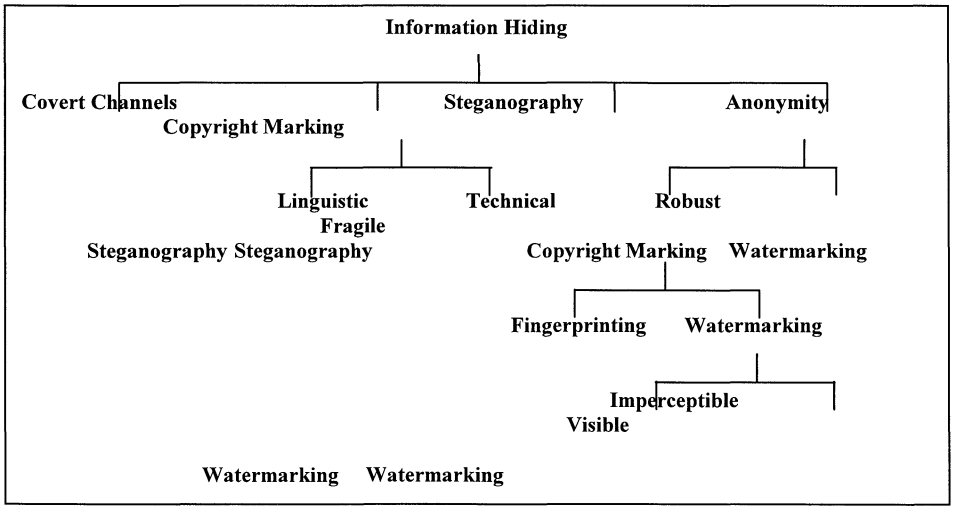

Figure 3 - Classification of Information Hiding Techniques based on [40].

The art of information hiding has also drawn interest from the commercial sector, primarily because it introduces the possibility to watermark products, such as images, music and software. This is reflected in the number of academic publications on the subject of information hiding increasing from two in 1992 to more than 100 in 1998 [41]. The driving force is to provide copyright protection in the form of digital media, which support easy duplication. The military, the criminal fraternities, and Internet users drive research in a slightly different direction by placing great value on unobtrusive or anonymous communications, and these are the two areas this section will concentrate on.

3.4.2 Historical Overview

Throughout history, people have hidden information for a multitude of reasons using a variety of methods. The earliest recorded users of steganography were the Greeks [42]. Interestingly, some of the Greek methods were also used successfully by German spies at the beginning of the 20th century [43].

In this century, invisible ink [1] and more recently by changing type [44] have been favourite tools for embedding messages. A pilot project found that by shifting text lines up or down by a pico as a binary notation was robust enough, even after multiple photocopying, to decipher and could not be noticed by most people [45]. It is rumoured that during premiership of Margaret Thatcher in the UK, she arranged for all the word processors of cabinet ministers to have their identities encoded in such a way to stop leaks. The ancient Chinese used a variation of this idea in the form of a paper mask to send and receive covert messages. This concept was reinvented in the Cardan (1501-1576) and is generally now known as a Cardan grille. Recently a British bank tried using this idea to conceal customers' PIN numbers, though the poor implementation led to a rather weak system [46].

One of the most interesting codes developed was based on musical scores, where each note corresponded to a letter, used by J. S. Bach [47] and John Wilkins. Wilkins also described a system whereby the geometry of shapes could be used to hide secret messages.

The Second World War saw the introduction of a number of new techniques, including the use of radio broadcasts to send null ciphers or unencrypted messages and microdot technology, originally used during the Franco-Russian War of 1870-1871 [1] and during the Russian-Japanese war of 1905 [48]. Since then the diversity of information techniques has greatly expanded, particularly with the proliferation of personal computers and the large volume of data communicated.

One of the first examples of the success breaking of a secret message was by the secret police of Queen Elizabeth I of England who added to an enciphered postscript to a letter from a Mary Queen of Scots to a group of assassins. The return mail, detailing the names of the assassins was successful intercepted and used to arrest and execute the assassins and Mary. This illustrates well the vulnerability of steganography when the method for information hiding is discovered.

Recently the latest limit of this process has been demonstrated with DNA strands and with dye molecules [49], a reduction in size of more than one thousand times from microdots and introducing the possibility of innocent human vehicles carrying messages in their bloodstream. The majority of information hiding systems however are computer based and this will be the focus of this section.

3.4.3 Steganographic Systems

One of the most common faults with information hiding schemes it that they rely on *security through obscurity*. This assumption, that the enemy will remain ignorant of the system, can cause dramatic failure. A classic example is where censors intercepted a cable-gram during WWI saying 'Father is dead' which they modified to 'Father is deceased' [1]. The reply was a giveaway: 'Is Father dead or deceased?'

A good information hiding system therefore has much in common with a good cryptographic system and should fulfil the same requirements as posed by the Kerckhoff Principles [50]. The most important of these is that it should be assumed that the enemy knows the method used to hide the message and therefore security must lie in the choice of a secret key. The problem of designing a good system is also closely related to many other problems in analogue systems, such as spread spectrum communication channels and hidden communication in background noise. The problem with digital systems is they introduce an element of precision and repeatability, which can often be easily detected again the background.

There are two commonly used approaches to information hiding using computer data. One replaces high entropy noise with the coded message, such as using the least significant bit of digital images, effectively *embedding* data within a *cover-image*. The second method attempts to strip all identifying information from the message, so that it resembles a random string of bytes that can then be transmitted. The latter can be mixed with the output from a pseudo-random generator using a key to increase its security. An extension to this scheme was proposed by Rivest and is known as *chaffing and winnowing* [51]. This system has several

interesting advantages. The message is sent in the form of blocks, down to 1 bit in size, which have a message authentication code (MAC) attached, computed from the message block and a secret key. The blocks are then randomised with more blocks, which have bad MAC blocks for the secret key. Therefore, the message is still transferred in plaintext, however in order to sort the wheat from the chaff it is necessary to know the secret key. This system also permits authentication and parallel streaming of message to different secret keys, though this scheme has a number of vulnerabilities, such as being able to frame an innocent part by generating a secret key to fit a given message.

A good computer program currently encodes a message into large text files (*cover-text*) commonly used graphics (*cover-image*) or audio files (*cover-audio*). Additional security can be provided by encryption of the message using a *stego-key* that can limit detection and recover of the embedded data to the relevant parties. These camouflage techniques rely on the fact that the embedding causes no change in the perception of the cover data and therefore is a function of the human or computer perception system.

To show the potential for information storage, a standard image of 640 x 480 pixels in 256 colours contains about 300k of information. By using 1-bit of each byte, it would therefore be possible to conceal approximately 36k of information in this image. The potential would increase dramatically for 24 bit images, which are more forgiving because of the greatly increased palette. Therefore certain graphic file types, such as 24 or 32 bit BMP files are preferred to 256 colour or greyscale GIF files, while heavily compressed files like JPGs are virtually impossible to use because of their lossy nature. For music files, minimal compressed formats like RAW and AU are preferred to heavily compressed ones like MP3.

Ideally, the information should be embedded in such a way as to raise least suspicion. Generally, a complete image is not suitable, because widely varying colours, or solid expanses of colour can be markedly changed by information hiding. Therefore, only a user-defined part of the host document, or a random scattering of altered bits should be used in a good system. This random scattering can be achieved by introducing a stego-key defining the pattern. Similarly, in audio files both the frequency and temporal domains can be used with care [52]. A human will only hear the louder of two tones close in frequency, allowing frequency masking, similarly it takes a while before a quiet tone can be perceived after a loud tone has been heard. A similar though more robust system uses echo hiding which encodes echoes of less than a millisecond, which cannot be perceived, using the cepstral transform [53]. These techniques are often borrowed from those used in lossy compression algorithms because the information can be removed without significantly compromising sound or visual impact, though are also susceptible to these very algorithms.

As mentioned, the most common and simplest approach is to hide the information using the least significant bit. This is, however it is vulnerable to even the slightest image manipulation. Two other techniques exist which are more robust and less susceptible to image processing, though generally limited to higher quality pictures. The first is masking and filtering, which works in a similar method to conventional watermarking. The information is applied using a predefined mask and can be recovered conversely using a filter in such a way that the concealed information becomes more a part of the host image through the manipulation of more bits.

The most robust method is to use algorithms and transformations, such as redundant pattern encoding, encrypt and scatter, using spread spectrum methods. These techniques are much less susceptible to image processing and can even withstand lossy image compression to some extent, though they trade off robustness against message size. For example in scattering algorithms, one bit of the embedded message will be spread over a selection of n pairs of pixels, where one pixel's luminosity is increased and one decreased, maintaining the average luminosity. In spread spectrum techniques, the robustness is achieved by shaping the embedded data to emphasise it in the most significant components of the data, so it will not be removed by the compression algorithms [54]. This is equivalent to hiding the information in the name of the main character in a novel, so that alteration to remove or change the name becomes impossible without substantially altering the content. The large bandwidth of the cover data in images and sound files makes spread spectrum ideas very attractive with much work having been done in this area [55,56,57,58]. Some of these transform steganographic schemes have been designed specifically to operate on compressed formats such as JPG [57] and MP3 [59] or other compression techniques such discrete cosine transforms, wavelet transforms and discrete Fourier transforms.

One major disadvantage with some systems is that they require the original data file to assist with the decoding. The existence of this file can substantially aid any interception and message recovery, and most information hiding packages will shred a cover data file after use.

A comparison of available steganography programs was recently carried out by Johnson and Jajodia [60]. In their tests, one of the best software programs for embedding into image and audio files was S-Tools, publicly available on the Internet [61].

3.4.4 Anonymity

There is also in existence a different type of information hiding service which is growing rapidly with the expansion of the Internet. Anonymous remailers and proxys allow users to send and receive information stripped of identifying marks. These are attractive for a number of reasons, such as the filtering of spam (junk) mailings, commercial advertising, and for the protection of identity, for example in digital voting. However, these services are often targeted by minority groups who use the services to espouse extreme messages, causing legal action against the providers.

There is sufficient interested in developing legitimate anonymous services nevertheless, particularly with the concept of electronic voting [62] and digital cash. Some governments have already committed themselves to providing such services, though this would represent a substantial test of the secrecy and privacy of any network. The guarantee of anonymity must be at many different levels, from the actually system used to the ways in which information is transmitted; there is no point having a strong anonymous digital cash system if the purchasers identity is then given away through the email message used to order the transfer.

3.4.5 Covert Channels

Covert channels are important for a number of interest groups, in particular the military and criminals. The interest is in using communication paths that were neither designed nor intended to carry information to minimize risk of discovery. One example, now entering the commercial sphere is the use of ionised trails from meteors entering the earths atmosphere to send burst radio messages. The transient and near random nature of these events make them difficult to intercept. More commonly we are used to interference in our day to day life due to cross-talk between electrical devices which is often heard as noise on a radio or television receiver.

One of the growing security concerns is that the information on a computer or video screen is also emitted as electromagnetic waves that can be intercepted. EMSEC (Electromagnetic Emission Security) was first mentioned in open literature as early as 1967 though was the late 1980's before it became widely publicised [63]. EMSEC is commonly divided into three types: Tempest (sometimes claimed to stand for Transient Electromagnetic Pulse Emission Standard) which is the interception of stray RF emissions from equipment, Hijack where sensitive information has contaminated another electrical signal assessable to an attacker, and Nonstop where information has accidentally modulated secondary emissions. Other threats also exist which do not fall into

these topics, for example the use of electromagnet radiation to probe structures and to decipher the returned signal. Although initially a military problem, the range of interested parties in these covert channels has increased to include banks, where ATM machines are vulnerable [64] and smart card system providers [65].

A relatively simple system can be constructed (for approximately $20) which can reconstruct the video signal at a distance of several hundred metres [66] which has led to so-called Tempest certified hardware becoming available to protect against such eavesdropping [67]. Simple software has been demonstrated which can hide information in the video signal that can be easily reconstructed with modified TV receivers [68] and there are grounds to believe that more sophisticated programs using spread spectrum techniques exist. This opens up the possibility of software piracy detector vans, similar to TV detector vans.

With the increase in computer interfacing and the availability of software controlled receivers interception becomes a more significant problem. Already there is software available on the Internet which can be used to decode messages to pagers using standard equipment. There has been speculation that there has been at least one case of a Tempest virus where the virus created a background electromagnetic signal that could be picked up some distance away [69]. Indeed the use of stray electromagnetic radiation stretches back in military circles much further. In his book *Spycatcher*, Peter Wright recounts the origin of Tempest attacks on cipher machines in order to determine the French position, when Britain was negotiating to join the European Community [70].

In addition to the video display, many other computer peripherals can be actively or passively monitored such as printers, serial cables, phone lines, power lines, network cables and coax cables. Therefore it becomes a substantial task to minimize EM leakage from a networked computer system.

Software, hardware and procedural measures can be used to minimise risk [71]. Methods include using filtered fonts, which effectively filter the frequency spectrum emitted, to reprogramming the keyboard scans in a random fashion to avoid attacks on the input side as well. Separation of red equipment (carrying confidential data) from black equipment (such as radio modems) is actively carried out in many sensitive establishments and where both exist together, particularly thorough testing is carried out (U.S. standard NACSIM 5100A, N.A.T.O. standard AMSG720B). The strength of testing reflects that long-term cross-correlation can give useful information even for very weak traces of the displayed information in electric, electromagnetic and / or acoustic emissions. Further information on the impact of Tempest can be found on the Internet [72].

Covert channels can also exist in many forms of software. A subtle, *subliminal channel*, exists in digital signatures, where a message can be embedded into the random key message used in these schemes. A more comprehensive discussion of the history of subliminal channels can be found in [73].

However, in secure computer system design covert channels are of particular concern. Covert channels can exist in software that unwittingly pass information between different security levels or transmit computer viruses or Trojan horse programs. Common examples including timing variations and error messages in communication protocols and operating system call interfaces [74,75]. These concerns are discussed further in the chapter dealing with computer system security.

3.4.6 Digital Fingerprinting and Watermarking

The proliferation of the Internet is major concern on one hand and a major market in the other for many entertainment industries. The ease with which music, videos and software can be copied has led to great interest in hidden marking techniques to identify and limit copying. The cost of counterfeiting was estimated to be more than $24 billion in 1995 [76], including the use of computers to assist in counterfeiting money and bankcards. There has been a rapid increase of new technologies to combat traditional forgery, for example optically variable devices such as holograms, inks and interference coatings, tamper-evident features such as laminates, reactive inks and sub-graving, machine read-able magnetic strips and in-built smart chips.

The problems of implementing a secure watermarking system are well illustrated by the DVD consortium, which has called for proposals for a copyright marking scheme to enforce copy management. It is proposed that hidden information within the media can indicate to the hardware whether copying is permitted. For example, TV broadcasts can be marked for copy once only, while commercial music or videos can be marked never copy and compliant hardware would act on these instructions [77]. Unfortunately, like many commercially supplied systems, DVD relies on being tamper-resistant, which is very hard to achieve in practice. It is perhaps significant though that some copyright leakage is acceptable to many businesses, For example in the words of Bill Gates: 'Although about three million computers are sold each year in China, people don't pay for the software. Someday they will though. As long as they're going to steal it, we want them to steal ours. Then we'll somehow figure out how to collect sometime in the next decade' [78].

There are also more benign plans to introduce marking schemes where by a user viewing a product on TV can simple mark that product for purchase or monitor-

ing applications such as checking that adverts have been played on radio or television.

Copyright marking has to be robust against possible attacks in order to provide protection; currently though insufficient research has been done in order to define 'robust' and depends on the application and the lengths to which marking is considered valuable. Some systems currently employed use visible digital watermarking similar to traditional watermarking [79], though most systems concentrate on invisible or transparent digital watermarks, which have wider applications. Fragile watermarking may also have limited use, for example in authenticated camera images (eg. police speed cameras) used as courtroom evidence where it can be easily shown whether the image has been tampered with.

Digital copyrighting can be used in two complementary fashions, one to *watermark* the data to identify the owner of the information and the second to *fingerprint* or label the information to identify the original customer. Often these are combined to aid in determination of where illegal copies originated. An alternative method occasionally used is to deliberately introduce mistakes into databases and lists which can easily be used to identify illegal resellers.

One of the difficulties with copyrighting is that it is long lived, for example up to 70 years, and therefore gives a long time in which the watermarking or fingerprinting can be analysed and removed. It is therefore important to understand the limitations of the currently available systems and the need for a basic theory of steganography to allow theoretical proof of robustness.

3.4.7 Limitations of Information Hiding

As with cryptography there are many over-inflated claims concerning information hiding software and digital watermarking/fingerprinting. As mentioned earlier the definition of robust is an important criteria and currently this varies greatly. A good discussion of this problem and an attempt to formalise analysis of steganography is made by Petitcolas [41]. There are two lines of attack on information hiding, one is to prevent the transmission of the message by scrambling the data or blocking the channel and the second, and more lucrative is to be able to recover the embedded message.

Scrambling the embedded message is currently relatively easy. Watermarking schemes that are robust against compression, noise addition, low pass filtering, rescaling and cropping have been demonstrated, however other processing such as rotation have often been ignored. Steganographic systems have similar

limitations though are often subjected to a different type of attack in order to extract the information rather than mangle it.

Three types of steganographic attack have been identified by Craver et al. [80] to diminish or remove digital watermarks and fingerprints. The first are *robustness attacks*, aimed at deleting as much evidence of the watermark as possible, the second type are *presentation attacks* which modify the content so as the detector cannot find the watermark, and the final group are *interpretation attacks* which aim to prevent assertion of ownership.

The two most basic types of attack are to change the amplitude spectrum of the data for example by adding random noise, or to change the frequency spectrum by adding timing jitter or to adjust the length. A more complicated attack on the robustness is to perform several minor manipulations, for example slightly stretch, shear, shift, bend and rotate, such that the image appears unaltered to the user but is sufficient to defeat most watermarking schemes. StirMark is a generic tool for carrying out such changes and is freely available [81]. An alternative approach is to estimate the watermark and then try to remove it. This can be successful especially with less complicated cover files where there are only limited possibilities for embedding the watermark. Removal of echo hiding is more difficult and requires an iterative approach to assist in minimising the echoes.

Presentation attacks are an alternative approach. By changing the mechanism by which the audio or image is displayed it is possible to remove any trace of watermarking. In the case of images, Petitcolas and Anderson have proposed a mosaic attack, where an image is subdivided into smaller subimages, which can then be transmitted independently. These subimages can be reassembled in a web page and displayed effectively as one image to the end user. The final approach is to introduce some dispute as to the ownership of the data. The most obvious way to do this is to add a second watermark though this can be extended to several other ideas [82]. Similarly, one of the concerns of a weak fingerprinting system is that it can be maliciously used to attribute pirating to another a source if discovered, an undersirable situation. Non-repudiation is one of the major difficulties with watermarking and fingerprinting and currently there is no obvious technical solution.

As with most cryptographic attacks however most real attacks on information hiding systems have come from opportunistic exploitation of loopholes and vulnerabilities found by accident.

A suggested working definition of what is meant by a robust watermarking system was outlined recently [41]:

- Marks should not degrade the perceived quality of the work and be embedded into the data rather than the header or wrapper.
- Detecting the presence and / or value of a mark should require knowledge of a secret. The mark access should be asymmetric, so that the data remains well hidden, though extraction with the correct secret is not difficult.
- If multiple marks are inserted, then they should not interfer with each other; moreover if different copies of an object are distributed with different marks, then different users should not be able to process their copies to generate a new copy that identifies none of them.
- The mark should survive all attacks that do not degrade the work's perceived quality, including resampling, re-quantisation, dithering, compression, and especially combinations of these. The embedded data should be self-clocking so that any missing bits, due to data manipulation do not destroy the message.

Although there are no currently available marking schemes fulfilling these criteria, with time and research it is believed these points can be addressed to provide a robust marking system. However, it is 'a wrong idea that high technology serves as a barrier to piracy or copyright theft; one should never underestimate the technical capability of copyright thieves' [83]. The wide variety of cover data sources and applications ensures that such strength is not always required and currently available systems are perfectly adequate.

3.5 Commercial and Legal Aspects to Information Security

Data security plays a major role in many aspects of daily life from the mail we receive to the financial transactions we make. There is always the assumption that privacy is guarded by law and the best security procedures are used. These public expectations are not always met for a variety of reasons, primarily cost, legal reasons and law enforcement. In the next two subsections, we will consider first the legal perspective of data security and then its commercial reality to illustrate the two main controlling forces.

3.5.1 Legal Regulations

Worldwide there are greatly differing attitudes towards regulation of data security. Policies are divided into two main categories: the obligations of those storing third person data to ensure privacy and secondly the regulations governing which security methods can be used. An example of the first type of legislation is the UK's Data Protection Act (1998) and the second type the UK's

Regulation of Investigatory Powers Act (2000). These two forms of regulation are indicative of the tightrope many governments walk between ensuring sufficient privacy for individuals and maintaining the ability to monitor potential infringements of the law.

This tightrope is passed down into the commercial world where regulations control export licensing, and government agencies, in particular law enforcement and intelligence agencies, apply pressure to companies to make security protocols sufficiently weak to permit data monitoring.

In this chapter, we have considered the risks and threats to data security and therefore will concentrate in this section on agreements and legislation, which has threatened data security, rather than enforce it. Reviews of data protection laws and the future role of data protection can be found here [84,85,86].

3.5.1.1 Introduction to International Policy

There is a philosophical question as to why a user would wish to be able to send data securely. If everyone was a law-abiding citizen, then any information exchanged should be open to scrutiny if desired. The argument offered by many governments is that secure encryption offers a method by which criminal elements, the so-called *four horsemen of the infopocalypse*: terrorists, paedophiles, drug dealers and criminals, can network and exchange information, obviously highly undesirable. The solutions offered vary from export restrictions on cryptographic products to government controlled keys and backdoors into communication systems.

The counter argument is that in comparative terms, this is equivalent to citizens only being allowed to use postcards or envelopes which any government licensed agency can look inside with no evidence of tampering in the regular mail service. To impose such powers retrospectively on the postal service would be impossible to enforce, however only a relatively small percentage of the population use electronically mail regularly to communicate; therefore in principle such measures are easier to pass into law. Effectively it is an invasion of personal privacy ironically defended in part by data security laws at the other end of the spectrum.

It has been argued strongly in western countries that traditional interception in the case of suspected crime was resource intensive and strictly controlled under law and therefore was self-regulating. However, electronic interception is not subject to the same conditions. The ability to intercept and analyse electronic information has been greatly enhanced by the many computer advances such as speech recognition and pattern searching, automating the monitoring process.

Many large multi-national organisations operate special departments to advise on cryptography because often it is not a straightforward analysis of the law that counts, but the interpretation of the law. This can lead to a substantial difference between claimed policy and actual policy.

Wassenaar Arrangement

The 'Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies' [87] was adopted by the countries involved in COCOM (Coordinating Committee for Multilateral Export Controls) as its replacement in 1996. The provisions of the Wassenaar Arrangement are largely the same as the COCOM and have been adopted by 31 countries worldwide.

In 1991, COCOM decided to allow export of mass-market cryptographic software (including public domain software). Most member countries of COCOM followed its regulations, but the United States maintained separate regulations. The General Software Note (applicable until the December 1998 revision) excepted mass-market and public-domain crypto software from the controls. Australia, France, New Zealand, Russia, and the US deviate from the General Software Note and control the export of mass-market and public-domain crypto software. Export via the Internet does not seem to be covered by the regulations. There is a personal-use exemption, allowing export of products "accompanying their user for the user's personal use" (e.g., on a laptop).

In September 1998, negotiations in Vienna did not lead to changes in the crypto controls, although it was apparently considered to restrict the General Software Note and possibly also to ease controls for key-recovery crypto. This was realised in December 1998, which resulted in restrictions on the General Software Note and in some relaxations:

- all crypto products of up to 56 bits are free for export,
- mass-market crypto software and hardware of up to 64 bits are free for export
- the export of products that use encryption to protect intellectual property (such as DVDs) is relaxed
- export of all other crypto still requires a license.

Interestingly, the new Cryptography Note which has replaced the General Software Note does not mention mass-market asymmetric crypto products; but it would be fair to assume that asymmetric mass-market products of similar strength to 64-bit symmetric crypto would also be exempt. There was no change in the provisions on public-domain crypto, so that all public-domain crypto

software remains free of export restrictions. Nothing was said about electronic exports (e.g., via the Internet), which consequently remain unclear.

The Wassenaar provisions are not directly applicable: each member state has to implement them in national legislation for them to have effect. The role of international agreements will however play an increasing role and acts as a good indicator for the future. The globalisation of markets and crime requires greater co-operation between separate states, best dealt with by uniform agreements. Therefore, the Wassenaar agreement or a modified replacement will take the lead in dictating global law and overseen by the major players, in particular the USA. Some countries consider these agreements to "interfere" in national law making policy, however are often persuaded to sign up with political pressure.

3.5.1.2 USA Cryptography Regulations

The US government has the longest, and perhaps the most complicated involvement in crypto policy. In the 1970's the National Security Agency (NSA) became involved in trying to block public funding for cryptographic research through the NSF and in classifying any patents using cryptography. Their role greatly increased in the early 1980's when control of computer security was moved from the National Bureau of Standards (NBS) to the NSA with jurisdiction given later over private databases. The late 80's and early 90's saw their role reinforced by trying to block the publication of strong algorithms, while trying to promote their own encryption algorithms.

In 1991, Senate Bill 266, an anti-crime bill in the USA, was used as a backdoor by the US administration to try and gain access to information channels by mandating service providers and manufacturers to "ensure that communications systems permit the government to obtain the plain text of voice, data and other communications when appropriately authorised by law". In response, Phil Zimmerman published PGP to make secure encryption a possibility for all. The following year the measure was defeated though this was only the opening move.

The Digital Telephony bill was successful in 1994 after being defeated in 1993 and mandated phone companies to install wiretapping ports at digital switching centres, to create the new technology of *point-and-click wiretapping*. This make interception technically a lot less complex and the year after the FBI disclosed plans that would require the phone companies to provide the facility to simultaneously tap one percent of all phone calls. This second step was however defeated in the US Congress in 1995, though provides an indication of the interests and agenda of the security services.

The second front in the battle to monitor electronic information was opened by the US government in 1993 when it announced a bold new encryption policy initiative. At the centre of this initiative was the *Clipper Chip*, a hardware device containing a new, classified NSA encryption algorithm. This chip could be incorporated into anything from phones and faxes to televisions and computers. The key for the encryption process was programmed when the chip was manufactured and the key kept by the government in escrow. This alarmed many of the interested parties and assurances by the government that Clipper was only an option and the key would only be used to decrypt messages "when duly authorised by law". The backlash was sufficiently strong (80% of Americans opposed it) that the only major company to incorporate Clipper was AT&T, who retrofitted it into a limited number of their phones after being paid by the government to not release their DES-based telephones. Although Clipper was officially adopted as the Escrowed Encryption Standard (EES) in 1994, the only visible sign of its existence today are the declassified Skipjack and KEA algorithms used.

The US government followed up by outlawing the distribution of encryption software in the Anti-Electronic Racketeering Act in 1995. This was reinforced the following year by an attempt to persuade OECD nations to adopt a similar policy, but the other nations were not willing to adopt such restrictions. Two blows occurred in 1996, one was the revelation that the 64 bit key for Lotus Notes, used by many organisations including the Swedish government, was partially escrowed by the NSA who held 24 bits of it. The second blow was from the NRC report in May 1996 which recommended the dropping of crypto restrictions and make DES exportable, while making crypto policy debated in public.

Since then, the US government has placated the largest and loudest opponents of restrictions though a number of compromises, such as case-by-case export licenses and exemptions and dispensations being granted. The major opponents now are the software companies and user groups. Unusually the export controls do not exist as a conventional law, instead are enacted each year by a presidential declaration of a national emergency. Their effect is interesting however in that strong encryption software is available to anyone requiring it within minutes on the Internet, but that most crypto implemented worldwide uses weak, crippled or compromised software. Evidence of US involvement in security breeches, for example in the CIA hacking of the European Parliament computers in 1996 and the Swedish government use of Lotus Notes, does not endear the US position to foreign governments, who feel the dominance of US software companies opens up the possibility of backdoors and compromises.

The Economic Strategy Institute estimated the US crypto controls would cost US industry $50 billion dollars in lost revenue in 2003, primarily from encryp-

tion dependent industries, such as those involved in e-commerce [88]. The interaction between business and government is an interesting and complicated one, particularly in the USA, with government concessions on issues like criminalizing reverse engineering of propriety algorithms met by concessions by industry to assist law enforcement and intelligence agencies. The development of the AES as an open source, widely available standard is promising and may indicate a policy shift in the USA.

In the next subsection we will consider this interaction between government and business further after first examining other cryptographic regulations enforced worldwide.

3.5.1.3 European Union Cryptography Regulations

Regulation of cryptographic products (under the heading of Dual-Use Goods) within the European Union has been in place since the middle of 1995 [89]. In general, a license is needed for the export of crypto hardware and software outside of the EU, with the exception of mass-market and public-domain software. For a transitional period, the Regulation also requires a licence procedure for intra-Community trade of encryption products. Export to seven "friendly" countries (Australia, Canada, Japan, New Zealand, Norway, Switzerland, USA) appears to be less restricted, through close military and governmental collaboration. Amendments since has relaxed the rules governing mobile phones and similar equipment using crypto and user accompanied software for personal use.

The publishing of 'Towards A European Framework for Digital Signatures And Encryption' [90] by the European Commission (EC) in 1997 notes that the dual-use Regulation left room for national implementation and that, consequently, "a large variety of domestic licensing schemes and practices exist. These divergences can lead to distortion of competition." The Commission was of the opinion that the Dual-Use Regulation should be adapted in view of the requirements of the cryptography market. It advised that the EC should:

- progressively dismantle intra-Community controls on commercial encryption products (i.e. not necessarily for very advanced encryption);
- launch a discussion on the scope and interpretation of certain provisions, such as the General Software Note (stipulating that public-domain software is not subject to controls);
- deal with problems like intangible means of transmission (such as fax or email).

The Dual-Use Regulation was to be replaced by a new regulation by 1 January 1999, setting up a Community regime for the control of exports of dual-use goods and technology [91]. The Dual-Use Regulation was considered not to have sufficiently stimulated a convergence of national policies and practices; it was complex and "too cumbersome to be useful in practice". The main change for cryptography proposed it that for exporting crypto products within the EU, export licenses will be replaced by a simple notification. In addition, the controls would now also include export through intangible means. To date this has not occurred, though the EU will soon discuss the December 1998 changes in the Wassenaar Arrangement in order to implement them. A number of nations however have already outlined differing positions to that of the amended Wassenaar Arrangement and there is the potential that the Danish will call for a debate on it [92].

With regard to eavesdropping, the 1995 European Council Resolution on the lawful interception of telecommunications (96/C329/01) contains a requirement for network operators and service providers, if they use encryption, to provide intercepted communications to law-enforcement agencies "en clair", similar to the Digital Telephony Bill in the USA. The illegal interception of encrypted signals, specifically protecting pay-services like satellite television, is detailed in the green paper on "Legal Protection for Encrypted Services in the Single Market" adopted by the EU in 1997. The paper proposes harmonization of national laws to prohibit the manufacture, sale, importation, possession, and promotion of illicit decoders, as well as unauthorized decoding.

The question of certification and key exchange/recovery was first covered by the draft proposal for the establishment of a Europe-wide network of Trusted Third Party Services (ETS) published in 1996 and 1997. The network would be established for providing certification services by private TTP's. Although primarily meant for establishing an infrastructure for the use of public key encryption, the proposal may also try to address the legal access problem, e.g., through key recovery. The ETSI (European Telecommunications Standardisation Institute) is currently developing a standard for Trusted Third Parties, which would include lawful access to encrypted data.

With the release of the Communication from the Commission: "Towards A European Framework for Digital Signatures And Encryption", the European Commission chose a direction away from key recovery. The Communication stresses the economic and social importance of cryptography: "the public needs to have access to technical tools allowing effective protection of the confidentiality of data and communication against arbitrary intrusions. Encryption of data is very often the only effective and cost-efficient way of meeting these requirements." The EC appears wary of key recovery issues and regulation should be required to be limited to what is absolutely necessary. However, at a

conference in January 1998, the EU Ministers of Justice and Home Affairs agreed that law enforcement agencies must have access to keys or plaintext, under certain circumstances.

The difficulty with regulation is highlighted by the fact that the French government classified Netscape as being in the second most dangerous weapon category, under a decree dating from before World War 2. The revelations regarding the Echelon surveillance system led to the French government removing all controls on crypto products in 1999. Further revelations about the Echelon system and the collusion of EU member states has prompted a lawsuit by a German MEP [93], fuelling the attraction of strong cryptography for EU states keen to protect business and diplomatic initiatives from US eyes. The UK remains the only EU state toeing the US line with attempts to introduce key escrow and control of cryptographic technology. This decision will inevitably cause the UK to be isolated from sensitive business and political decisions, which could be made elsewhere in the EU.

3.5.1.4 Worldwide Cryptography Regulations

A complete survey of cryptographic regulations in the majority of developed countries can be found in [94,95]. It is worth noting the current trend is generally towards relaxation of restrictions on cryptography after a decade of attempts to restrict its use. The use of steganography is less tightly regulated in that it requires a definition of how hidden the data is and is not as widely used as cryptography, primarily because it is only used for small amounts of information transmitted in an unobvious way. Similarly, the use of digital signatures is considered to be of benefit to all and therefore is covered by different legislation, a review of which can be found here [96].

Regulation and restriction of data security protocols is akin to regulation of the nuclear industry. The countries with these commodities try to force other countries, particularly small ones, into agreements to limit the spread of this commodity in order to maintain their advantageous position. This forces countries into expensive development of their own propriety systems, which can be incompatible, or to enter into a licensing agreement to acquire the technology under limited conditions. Typically, if a country feels threatened it will resort to propriety systems to defend its own security, whereas in a secure position it would opt for the most compatible and inexpensive system. This shapes the future technologies adopted by countries worldwide and why universal cryptographic regulation will be impossible to adopt.

The growing importance of e-commerce and potentially of m-commerce will help to relax regulation, as consumers demand adequate security for their

financial transactions and service providers diversify. Relaxing of regulations also gives a guide to the interception abilities of law enforcement and intelligence agencies, who will undoubtedly remain able to intercept desired communications.

The impact of future developments to revolutionise cryptography is also likely to force governments into reactive measures to protect their own interests. Proactive steps should be taken to take account of likely advances and determine the level at which they should be regulated. Data privacy will remain a hotly contested issue over the next decade with governments walking an ever-thinning tightrope.

3.5.2 Commercial and Financial Information Security

The use of cryptography to protect data for financial purposes is now common practice in many areas. This includes areas as diverse as satellite television, software evaluation, cellular phones and DVD players. In all these areas there has also been confrontation when individuals and groups have examined the cryptosystems used and found ways in which they could be weakened.

The rapid growth of Internet and mobile services has also prompted the adoption of cryptosystems for secure financial transactions. In particular the use of digitial signatures and certificates Although the protocols themselves have been well-analysed, the implementation and management of systems are often weakened by poor business practices. This has lead to a number of scares on the Internet, in particular over the safety of on-line shopping.

At first glance, business appears to have taken a useful tool and crippled it through misunderstanding. In this subsection we will look at how businesses have addressed the use of cryptosystems and their interaction with users and the government.

3.5.2.1 Reverse Engineering of Cryptosystems

Businesses who invest money and effort in a product understandably wish to ensure that only subscribers have access to the product and that subscribers believe the product to be secure. Their preference is to use propriety cryptosystems developed in-house, often using snake-oil [97] for advertising.

In order to protect their data, large businesses have built intellectual property fences around these cryptosystems, with the help of government policy. These fences have effectively made it illegal to reverse engineer the protocols. In

return for government support, businesses weaken algorithms for law enforcement and intelligence purposes.

Protection of algorithms is taken very seriously. For example, the UK intelligence service prevented Dr Shepherd from Bradford University from publicly disclosing his knowledge of the A5 algorithm used in GSM mobile phones [98], and there is currently a large court battle between the Motion Picture Association of America (MPAA) and a group of individuals. MPPA is trying to hold the group responsible under the Digital Millennium Act for distributing code to by-pass the CSS cryptosystem used on DVDs [99]. Legal action between businesses over reverse engineering of other algorithms has also been reported, for example one case has been fought over copyright infringement [100]. Further details of UK cases involving data security can be found here [101].

The need to prosecute these cases is in part due to the weakened implementation of the cryptosystems used. With use of open standards for strong cryptosystems and intelligent application of these systems would have prevented many of these costly lawsuits and not affected public perception. Unfortunately, the future does not look promising with many businesses still favouring security through obscurity.

3.5.2.2 Applications of Digital Certificates and Signatures

Over the last five years there has been a growing interest in introducing legislation to cover digital signatures as a legal means of validating and authorising a document to allow the reader to act on it as a statement of the signer's intent or consent. A signature can provide a number of different functions in daily life, including identification, involvement, association, verification and endorsement of the signed information through to proving someone was at a given place at a given time.

The important feature of a digital signing system must be that the process is conscious and in full control of the signer to be legally binding and fair. Often a traditional document is used to back up this process and provide extended support for non-repudiation.

In the USA, Utah was amongst the first states to introduce a digital signature act in 1995. It was relatively cumbersome relying on CAs and the X.509 protocol. Later legislation in other states reverted to the concept of "you can't refuse a signature just because its digital" without legislating heavily on the responsibilities or the use of CAs. For example in California, the basic concept is that any agreed-upon mark can be used as a digital signature. Internationally, both Germany and Italy have complicated systems relying on licensed CAs. The

German implementation stretches to over 300 pages detailing everything from random number generators to signature algorithms. The Italian system requires CAs be ISO9000 compliant with everything certified to various ITSEC levels restricting the technology which can be used but addresses most of the digital signature issues.

Pan-nationally, the UN has drafted articles on electronic signatures that are rather vague in most respects and technology-independent, though which draw a distinction between an "electronic signature" which indicates approval and "enhanced electronic signature" which is unique, verifiable and under the sole control on the signer. The EU has issued a directive of electronic signatures, which defines an electronic signature as linking signer and data created by means solely controlled by the signer. It recognises that a regulatory framework is not required for closed systems, that signature products be made free available with the EU and signatures can't be denied recognition just because they are digital. Further, it makes accreditation and licensing voluntary and non-discriminatory, with recognition of certificates from non-EU countries issued under equivalent terms.

A more thorough discussion of the different laws governing digital signatures in particular can be found on the Internet [96].

3.5.2.3 Electronic Banking and Commerce

Cryptography is used widely in banking and commerce to protect data. Financial institutions have taken the lead in many respects, with most favouring DES and moving to AES in the near future, for their internal networking. Interaction with customers is less well secured using a variety of methods, with perhaps on-line banking being the largest threat. Sensible use of the secure socket layer (SSL) protocol is implemented by all banks; however there always remains the threat of software vulnerabilities, which have given access to credit card numbers and personal transactions [102,103].

As the Internet revolution matures, the implementation and stability of software should improve to the extent that crime can be kept to a manageable level in a well-run organisation. The attractiveness of on-line services and their vulnerability to remote, anonymous attacks will mean that crime will never go away, however with planning of resources and using strong cryptosystems, this threat can be minimized. These issues will be considered more thoroughly in Chapter 6.

3.6 Conclusions and Future Prospects

In this chapter, we have examined how data security can be improved using cryptography and steganography. Most communication channels are insecure and are therefore liable to be compromised by an attacker. A range of groups and individuals were identified who pose a threat to data security and their goals and motivations examined.

The most common way of improving the security of a channel is to use cryptography. In the third section, the basics of cryptography and cryptanalysis were described to provide a backdrop to a discussion about the requirements of a strong cryptosystem. A number of modern cryptosystems, both traditional symmetric schemes and more recent asymmetric schemes, were analysed for the security they offer. The major logistical problem of cryptosystems, key distribution, remains a problem for the widespread implementation of cryptography and a number of ideas were explored. Current cryptosystems will have a shelf-live of approximately a decade at present; the future of cryptography was considered at the end of section three.

A complimentary technique, steganography was described in the fourth section. In this case, data is disguised rather than made unreadable. Information hiding is important commercially in providing digital fingerprinting and formed the basis of analysis in the section. Importantly, the limitations of steganography were considered, and the ways in which it could be combined with cryptography to provide a robust security system.

The commercial and legal impact of data security was investigated in the fifth section. The effectiveness of cryptography in ensuring confidentiality is illustrated by the restrictions which have been used to try and limit the widespread implementation of strong cryptosystems.

It is worth noting that when cryptography is employed correctly it is generally the least weakest link, and therefore unlikely to be attacked unless no other alternatives are possible. Deploying cryptography on a large scale, to be used by people who will treat it as a black box however is not an easy task and one of the challenges for the future of data security.

References and Further Reading:

Whitfield Diffie, Susan Landau, *Privacy on the Line: The Politics of Wiretapping and Encryption*, MIT Press (1999)

Electronic Frontier Foundation, John Gilmore (Editor), *Cracking DES: Secrets of Encryption Research, Wiretap Politics & Chip Design*, O'Reilly & Associates (1998)

Warwick Ford, Michael S. Baum, *Secure Electronic Commerce: Building the Infrastructure for Digital Signatures and Encryption*, Prentice Hall (2000)

Randall K. Nichols, *ICSA Guide to Cryptography*, McGraw Hill Text (1998)

Bruce Schneier, David Banisar, *The Electronic Privacy Papers: Documents on the Battle for Privacy in the Age of Surveillance*, John Wiley & Sons (1997)

Simon Singh, *The Code Book : The Evolution of Secrecy from Mary, Queen of Scots to Quantum Cryptography*, Doubleday (1999)

William Stallings, *Cryptography & Network Security: Principles & Practice*, Prentice Hall (1998)

John R. Vacca , *Satellite Encryption*, Academic Press (1999)

Philip Zimmermann, PGP : Source Code and Internals

[1] David Kahn, *The Codebreakers*, Macmillan, New York (1967)

[2] C. Shannon, *Communication Theory of Secrecy Systems*, Bell Systems Technical Journal, 28, 656-715 (1949)

[3] David Kahn, *Seizing the Enigma*, Arrow (Reissued 1996)

[4] Bruce Schneider, *Applied Cryptography*, John Wiley and Sons (1995)

C.A. Deavours, L. Kruh, *Machine Cryptography and Modern Cryptanalysis*, Artech House (1985)

[5] C. P. Pfleeger, *Security in Computing*, Prentice Hall International Editions (1989)

[6] Mark Chen, *Cyberspace: Pandora's Mailbox, RC4 a secret no longer*

http://www.zmag.org/zmag/articles/chen.htm

[7] P. Gutmann, *Encryption and Security Tutorial*, University of Auckland

http://www.cs.auckland.ac.nz/~pgut001

[8] EFF DES Cracker Project, Electronic Frontier Foundation

http://www.eff.org/descracker.html

[9] DES-III Effort, distributed.net

http://www.distributed.net/pressroom/press-des-iii.html

[10] The MARS Cipher Homepage, IBM

http://www.research.ibm.com/security/mars.html

[11] RC6 Block Cipher, RSA Laboratories

http://www.rsasecurity.com/rsalabs/rc6/index.html

[12] The block cipher Rijndael, University of Leuven

http://www.esat.kuleuven.ac.be/~rijmen/rijndael/

[13] Serpent Cipher Homepage, Cambridge University

http://www.cl.cam.ac.uk/~rja14/serpent.html

[14] Twofish: A New Block Cipher, Counterpane Security Inc.

http://www.counterpane.com/twofish.html

[15]Lars R. Knudsen and Vincent Rijmen, The Block Cipher Lounge – AES

http://www.ii.uib.no/~larsr/aes.html

[16] Advanced Encryption Standard Development Effort, NIST

http://csrc.nist.gov/encryption/aes/

[17] W. Diffie, and M. Hellman, *New Directions in Cryptography*, IEEE Trans. Info. Theory, 22:6, 644-654 (1976)

[18] J.H. Ellis, *The Possibility of Secure Non-Secret Digital Encryption*, CESG Report, (January 1970)

http://www.cesg.gov.uk/about/nsecret/

[19] Bruce Schneier, *Attacking Certificates with Computer Viruses*, Cryptogram (April 1999)

http://www.schneier.com/crypto-gram-9904.html#certificates

http://www.newscientist.com/ns/19990313/newsstory3.html

[20] R. Merkle,and M. Hellman, *Hiding Information and Signatures in Trapdoor Knapsacks*, IEEE Trans. Info. Theory, IT-24-5, 525-530 (1978)

[21] A. Shamir, and R. Zippel, *On The Security of Merkle-Hellman Cryptographic Scheme*, IEEE Trans. Info. Theory, IT-26-3, 339-340 (1980)

[22] A. Shamir, *A Polynomial Time Algorithm for Breaking the Basic Merkle-Hellman Cryptosystem*, Proc. Crypto '82, Plenum Press, 279-288 (1982)

[23] R.L. Rivest, A. Shamir, and L.M. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Communications of the ACM (2) 21 (1978), 120-126.

[24] R. Solvay and V. Strassen, *A Fast Monte-Carlo Test for Primality*, SIAM Journal Comp., 6, 84-85 (1977)

[25] P. Zimmerman, *A Proposed Standard Format for RSA Cryptosystems*, Computer, 19:9, 21-34 (1986)

[26] *Digital Signature Standard*, Federal Information Processing Standards

http://www.itl.nist.gov/fipspubs/fip186.htm

[27] M.J.B. Robshaw and Y.L. Yin, *Elliptic Curve Cryptosystems*, Technical Note, RSA Laboratories (1997)

[28]

[29] Bennett, Brassard, Crépeau, Jozsa, Peres and Wootters, Phys. Rev. Lett., 70 (1985)

[30] S. Wiesner, SIGACT News 15, 78 (1983); original manuscript written circa 1970

C.H. Bennett and G. Brassard, in Proc. IEEE Int. Conference on Computers, Systems and Signal Processing, IEEE, New York (1984).

C.H. Bennett, F. Bessette, G. Brassard, L. Salvail, and J. Smolin, *Experimental quantum cryptography*, J. Cryptology 5, 3 (1992).

[31] A.K. Ekert, Phys. Rev. Lett. 67, 661 (1991); A.K. Ekert, J.G. Rarity, P.R. Tapster, and G.M. Palma, Phys. Rev. Lett. 69, 1293 (1992)

[32] C.H. Bennett, Phys. Rev. Lett. 68, 3121 (1992)

[33] Quantum Cryptography Primers

http://www.qubit.org/

http://www.quantum.univie.ac.at/research/crypto/

[34] David Mowbray, *Quantum Dots Spot Single Photons*, Physics World, 27 July 2000

[35] Quantum Computing Primers

http://www.rdrop.com/~cary/html/quantum_c_faq.html

http://www.banished.demon.co.uk/quantum/Qindex.htm

[36] David Deutsch and Artue Ekert, *Quantum Computing*, Physics World (March 1998)

http://www.qubit.org/intros/PhysicsWorld/PhysicsWorld.html

[37] Peter Fordham and Stephen Perrott, *An Introduction to Quantum Computing and its Consequences for Cryptography*, Surveys and Presentations in Information Systems Engineering

http://www-ics.ee.ic.ac.uk/surp00/report/pjf98/

[38] John Mades, *Quantum Computers and Their Impact on DoD in the 21st Century*, MSc Thesis, Naval Postgraduate School

http://www.cs.nps.navy.mil/people/faculty/rowe/madesthesis.htm

[39] DNA Based Computation Primers

http://dna2z.com/dnacpu/dna.html

http://www.abcnews.go.com/sections/tech/DailyNews/dnacomputer000112.html

[40] R. J. Anderson, *Information Hiding*, First International Workshop, Vol. 1174 of Lecture Notes in Computer Science, Isaac Newton Institute, Cambridge, England.

[41] F. A. P. Petitcolas, R. J. Anderson and Markus Kuhn, *Information Hiding – A Survey*, to appear in Proc. IEEE, Special Issue on Protection of Multimedia Content (May 1999)

[42] A. Tacticus, *How to Survive Under Siege*, Clarendon Ancient History Series, Oxford, England, Clarendon Press (1990)

[43] B. Newman, *Secrets of German Espionage*, London, Robert Hale Ltd., (1940)

[44] J. Brassilet et al., *Document Marking and Identification using Both Line and Word Shifting*, *Proc. Infocom95*, IEEE CS Press, Los Alamitos, Calif., 1995.

[45] J. Brassil, S. Low, N. Maxemchuk and L. O'Garman, *Electronic marking and*

identification techniques to discourage document copying, Infocom, 1278-1287 (1994)

[46] R. J. Anderson, *Why Cryptosystems Fail*, Communications of the ACM, Vol. 37, no. 11, 32-40 (1994)

[47] F. L. Bauer, *Decrypted Secrets – Methods and Maxims of Cryptology*, Berlin, Germany, Springer Verlag (1997)

[48] G. W. W. Stevens, *Microphotography – Photography and Photofabrication at Extreme Resolutions*, London, Chapman & Hall (1968)

[49] DNA Cryptography and Steganography

http://www.cs.duke.edu/~reif/BMC/reports/BMC.FY99.reports/BMC.tasks/Task3.html

[50] A. Kerckhoffs, *La Cryptographie Militarie*, Journal des Sciences Militares, Vol. 9, 5-38 (1883)

[51] R. L. Rivest, *Chaffing and Winnowing*, MIT Lab, USA (April 1998)

[52] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, London, Academic Press (3rd Ed.) (1989)

[53] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, *The frequency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking*, In M. Rosenblatt, ed., Symposium on Time Series Analysis, pp. 209-243, New York, New York, USA: John Wiley & Sons, Inc., 1963.

[54] M.D. Swanson, B. Zu, and A. H. Tewfik, *Robust Data Hiding for Images*, Proc. 7th Digital Signal Processing Workshop (DSP 96), Leon, Norway, IEEE (1996)

[55] J. R. Smith and B. O. Comiskey, *Modulation and information hiding in images*, Workshop on Information Hiding, University of Cambridge, UK (1996)

[56] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, *Secure spread spectrum watermarking for images, audio and video*, International Conference on Image Processing (ICIP'96), pp. 243-246, IEEE, (1996)

[57] F. Hartung and B. Girod, *Watermarking of MPEG-2 en-coded video without decoding and reencoding,* In M. Freeman, P. Jardetzky, and H. M. Vin, eds., *Multimedia Computing and Networking* 1997, vol. 3020, pp. 264-273, SPIE, (1997)

[58] L. Boney, A. H. Tewk, and K. N. Hamdy, *Digital watermarks for audio signals*, International Conference on Multimedia Computing and Systems, pp. 473-480, IEEE (1996)

[59] Fabien A. P. Petitcolas, MP3Stego, Cambridge University

http://www.cl.cam.ac.uk/~fapp2/steganography/mp3stego/

[60] N. F. Johnson and S. Jajodia, *Exploring Steganography: Seeing the Unseen*, Computer, 26-34 (1998)

[61] S-Stools – Steganography Toolkit

http://indyunix.iupui.edu/~emilbran/stego.html

[62] J.C. Benaloh, *Verifiable Secret-Ballot Elections*, Ph.D. thesis, Yale University, New Haven (1987)

[63] H. J. Highland, *Electromagnetic Radiation Revisited*, Computers & Security, Vol. 5, 85-

184 (1986)

[64] D. E. Denning, *Information Warfare and Security*, Addison Wesley, 189 (1998)

[65] S. Chari, C. Jutla, J.R. Rao, P. Rohatgi, *A Cautionary note regarding evaluation of AES candidates on smart-cards*, 2nd Advanced Encryption Standard Candidate Conference, Rome, Italy, 133-147 (1999)

[66] W. van Eck, *Electromagnetic radiation from video display units: an eavesdropping risk?*, Computers & Security, vol.4, no. 4, pp. 269-286, Dec. 1985.

[67] D. Russel and G. Gangemi, *Computer Security Basics*, Chap. 10: TEMPEST. Sebastopol, California, USA: O'Reilly & Associates (1991)

[68] M. G. Kuhn and R. J. Anderson, *Soft tempest: Hidden data transmission using electromagnetic emanations*, Information Hiding, 124-142 (1998)

[69] E.R. Koch and J. Sperber, *Die Datenmafia: Computerespionage und neue Informationnskartelle*, Rowohlt

[70] P. Wright, *Spycatcher – The Candid Autobiography of a Senior Intelligence Officer*, William Heinemann, Australia (1987)

[71] R. J. Anderson and M.G. Kuhn, *Soft Tempest – An Opportunity for NATO*

http://www.ftp.cl.cam.ac.uk/ftp/users/rja14/nato-tempest.pdf

[72] Joel McNamara, The Complete, Unofficial TEMPEST Information Page

http://www.eskimo.com/~joelm/tempest.html

[73] *The history of subliminal channels*, IEEE Journal of Selected Areas in Communications , vol. 16, no. 4, pp. 452-462, May 1998

[74] V. Gligor, *A guide to understanding covert channel analysis of trusted systems*, Tech. Rep. NCSC-TG-030, National Computer Security Center, Ft. George G. Meade, Maryland, USA, Nov. 1993

[75] B. Lampson, *A note on the connement problem*, Communications of the ACM, vol. 16, no. 10, pp. 613-615, Oct. 1973

[76] I. M. Lancaster and L. T. Konntnik, *Progress in counterfeit deterrence: The contribution of information exchange,* In van Renesse, 134-139 (1989)

[77] J. Dittmann, P. Wolhmacher, P. Horster and R. Steinmetz, Multimedia and Security – Workshop at ACM Multimedia '98, vol 41 of GMD Report, Bristol, United Kingdom, Sept. 1998

[78] The Bill & Warren Show, Fortune, 44 (20th July 1998)

[79] G. W. Braudaway, K. A. Magerlein and F. Mintzer, *Protecting Publicly Available Images with a Visible Image Watermark*, In van Renesse, 126-133 (1989)

[80] S. Craver, B.L. Yeo, and M. Yeung, *Technical trials and legal tribulations*, Communications of the ACM, 41:7, 44-54 (Jul 1998)

[81] Fabien A. P. Petitcolas, Image Watermarking - StirMark

http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/

[82] *Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications*, IEEE Journal of Selected Areas in Communications,vol. 16,no.

4, pp. 573-586 (1998)

[83] J. Gurnsey, *Copyright Theft*, Aslib Gower (1995)

[84] Alan F. Westin, *Data Protection in the Global Society, Visions for Privacy in the 21st Century: A Search for Solutions*, Conference Papers, Victoria, British Columbia, May 9 - 11, 1996

http://www.privacyexchange.org/iss/

[85] European Union Working Party on the Protection of Individuals with regard to the Processing of Personal Data

http://www.europarl.eu.int/dg2/hearings/20000222/libe/arts/art29/en/default.htm

[86] OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data

http://www.oecd.org/dsti/sti/it/secur/prod/PRIV-EN.HTM

[87] Wassenaar Organisation

http://www.wassenaar.org/

[88] E.R. Olbeter, C. Hamilton, *Finding the Key: Reconciling National and Economic Security Interests in Cryptography Policy Executive Summary*, Economic Strategy Institute (April 1998)

http://www.econstrat.org/crypto.htm

[89] EU Council Regulation (EC) No. 3381/94 (amended by Regulation (EC) 837/95 of 10 April 1995) and EU Council Decision No. 94/942/CFSP (last amended by Council Decision 98/232/CFSP)

[90] EU Commission on Ensuring Security and Trust in Electronic Communication, *Towards A European Framework for Digital Signatures And Encryption,* COM (97) 503

http://europa.eu.int/ISPO/eif/policy/97503toc.html

[91] EU European Internet Forum, *Ensuring Trust and Security in Electronic Communication*, COM (1998) 257
http://158.169.51.11/eif/policy/policy.html

[92] Update on PGP-export restrictions, Cryptome

http://jya.com/wass-dk.htm

[93] *German EU Delegate Sues 'Unknown' Over Echelon*

http://slashdot.org/yro/00/10/16/1152252_F.shtml

[94] *Cryptography and Liberty 2000 - An International Survey of Encryption Policy*, Electronic Privacy Information Centre (EPIC)

http://www2.epic.org/reports/crypto2000/

[95] Bert-Jaap Koops, *Crypto Law Survey*

http://cwis.kub.nl/~frw/people/koops/lawsurvy.htm

[96] Simone van der Hof, *Digital Signature Law Survey*

http://rechten.kub.nl/simone/ds-lawsu.htm

[97] Matt Curtin, *Snake Oil Warning Signs: Encryption Software to Avoid*

http://www.mindrape.org/papers/cryptography-faq/snake-oil
[98] John Young's Archives, Crack A5
http://jya.com/crack-a5.htm
[99] DeCSS Central – DVD CCA Lawsuit(s)
http://www.lemuria.org/DeCSS/cca.html
[100] David Swarbrick, *Encryption as confidentiality marker*, UKCrypto Mailing List
http://haig.cs.ucl.ac.uk:80/staff/I.Brown/archives/ukcrypto/0399-0699/msg01076.html
[101] David Swarbrick's law-index and law-bytes
http://www.swarb.co.uk/
[102] Graeme Wearden, *Egg wasn't the first, according to the National Crime Squad*
http://www.zdnet.co.uk/news/2000/33/ns-17480.html
[103] *Banking error puts Roger Moore's royalties on internet*, Ananova
http://uk.news.yahoo.com/001109/4/aomt5.html

Chapter 4 - Computers & Networks

4.1 Introduction

Everyday there are new security scares about the Internet. On occasions they are well founded and have required a significant re-think on how computer and network security is organised. Many of the problems that have arisen are because of the very rapid growth of networking and services, without full thought being given to the implications and compromises made to implement them. Many of the higher-level applications used assume that lower level protocols are secure, but a lack of integrity and confidentiality at the lower level make both vulnerable to attack.

Networking is a highly diverse subject, meeting many different requirements. A flexible system is necessary in order to implement all these requirements. This chapter will characterise the risks and threats to a network using the Open Systems Interconnection (OSI) communications model. This commonly used model divides a network into seven layers, from the physical layer detailing the hardware through to the application layer, which is front-end seen by the user. Each of these layers will be considered in the first section of this chapter from a security point of view. Although a useful and comprehensive model, some computer systems, especially proprietary and military networks, are not easy to analyse in this form, but reference to them has been included at relevant points. Even the commonly used TCP/IP protocol suite, which forms the basis of the Internet as is examined in detail in this chapter, in not directly translatable to the OSI model. This is the reason why inconsistencies may appear between the general descriptions given for each layer and the discussion of implemented systems.

The second section of this chapter identifies particular threats to the hardware of a computer and its connection to a network. Active and passive attacks are detailed and steps described which can minimize the impact of these threats. Inherently however, because of the protocols and applications used, an off-the-shelf computer and operating system connected to a network will always remain vulnerable. Services can be disabled and intrusion detection systems and firewalls can be implemented to provide protection and detection, however the threat remains and this must be taken into account in planning computer networks, especially for critical services.

The final section of the chapter will consider the future of the Internet and the technology and that will be introduced over the next decade. The impact, in particular of new networking protocols, is significant because it addresses a

number of the key weaknesses in the current IP/TCP protocols used for Internet and Ethernet traffic.

4.2 The OSI Communications Model

In this section the threats and risks to a computer network will be assessed according to the Open Systems Interconnection (OSI) communications model, illustrated below, in Figure 1. This divides networking into 7 distinct layers, ranging from the physical hardware through to networked applications. The seven layers are also subdivided into two main categories: upper layers (4-7), which deal with applications issues and generally are only implemented in software and are technology independent while the lower layers (1-4) handle data transport issues, implemented in both hardware and software and are technology dependent.

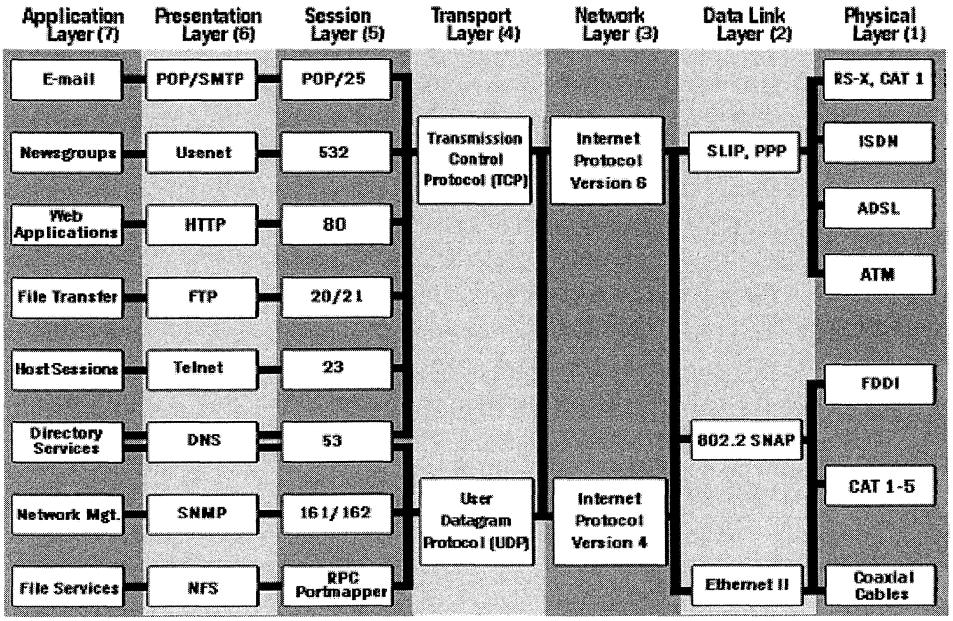

Figure 4 - OSI Communications Model

Typically when information is transmitted from one application to another across the network, it first passes down through the layers until the data link and physical layers, which transmit it across the network. The data is then reassembled at the remote computer as it passes back up through the layers to the receiving application. Each layer is normally composed of three elements: the *service user*, the *service provider* and the *service access point* (SAP). In this context, the service user is the OSI layer that requests services from its equivalent layer on the remote computer, the service provider. The sender's

computer first checks whether the remote computer is a service provider by sending a request to the service access point. For example this could be a request to a WWW server from a user to send a web page. These are virtual concepts, but are useful for visualisation of the roles played by the different protocols.

Each layer uses control information to communicate with its peer layer in other computer systems; this control information usual takes the form of headers and trailers on packets of data sent. Each layer encapsulates the header, data and trailer from upper levels, into a data packet to which it can add a header and trailer. In effect these shells of control information add significant redundancy to data packets sent over the network, however they ensure error-control, delivery, authentication and integrity at all levels. Control information also permit multiple, parallel connections to a single service.

In the next five subsections we will consider each of the layers and the role they play in networking. Vulnerabilities and security features exist in all of layers and we will look at these and future advances which will affect current network technology.

4.2.1 Physical Layer

In the OSI communications model, the physical layer, also known as the *bit pipe*, supports the electrical or mechanical interface to the physical network medium. For example, this layer determines how to put a stream of bits from the upper (data link) layer on to the pins for a parallel printer interface, an optical fibre transmitter, or a radio carrier. For networking this is generally performed by a network card or modem, which generates a stream of electrical voltages corresponding to data bits.

We will first look at how bit pipes are defined. Many different media are used to convey data, many of which we considered in Chapter 2. We will consider the public telephone and data networks in more detail again from the perspective of being the physical layer of a network. We will also consider several commonly used standards in this layer including ISDN, ATM, and FDDI.

4.2.1.1 Categorisation of Bit Pipes

The ANSI/EIA (American National Standards Institute/Electronic Industries Association) Standard 568 [1] is one of several standards that specify "categories" (the singular is commonly referred to as "CAT") of twisted pair cabling systems in terms of the data rates that they can sustain. These categories are:

Category (CAT)	Maximum Bandwidth	Application
1	<1 MHz	Modems (0.3-56Kbps) RS 232 & RS 422, ISDN & ASDL AppleTalk
2	4 MHz	1.544 Mbps T1 / DS1 Frame Relay (0.056-1.544Mbps) 1 Base 5 (IEEE 802.3) 4 Mbps Token Ring (IEEE 802.5)
3	16 MHz	6.312Mbps T2 / DS2 Digital Subscriber Line (DSL) (0.5-8Mbps) 10 Mbps Ethernet (IEEE 802.3)
4	20 MHz	10 Mbps Ethernet (IEEE 802.3) 16 Mbps Token Ring (IEEE 802.5)
5	100 MHz	100 Mbps TPDDI (ANSI X 319.5) 100 Mbps Fast Ethernet (IEEE 802.3) 44.7Mbps T3 / DS3 Cable Modem (0.5-52Mbps) 155 Mbps ATM
5e	100 MHz	Same as CAT 5 plus 1000BaseT
6	250 MHz	1000BaseT

Table 2 - Comparison of the Maximum Data Rates of CAT Standards

CAT3 and CAT5 are the most popular standards, with CAT5 being incorporated into the specifications for short distance GB/s connections. These standards do not however include specifications for optical fibre and free space communications, which are defined by their own particular and non-universal standards.

4.2.1.2 Public Switched Telephone and Data Networks

The standard telephone system operating in most countries is referred to as the *Public Switched Telephone Network* (PSTN) and is managed by *local exchange carriers* (CELs). Currently they use a proprietary CAT5 switching infrastructure, which is leased from its manufacturers, who have a vested interest in making such technology prohibitively expensive. To date they have prevented service providers from differentiating themselves by the services they can offer, or from constructing competing technology.

The second substantially smaller strand of the public switched network, is the *Public Switched Data Network* (PSDN), a packet switched network generally using the X.25 protocol, consisting of network *points-of-presence* (POPs) and remote access devices. It is rapidly growing however, driven by the demands of

the Internet, intranets of large organisations, virtual private networks (VPNs), and remote access. However, the PSTN, a circuit switched network, remains the principal means for delivering data services. According to Dataquest, 46.5 million analog modems will be sold in the year 2000. And nearly all personal computers purchased today come equipped with a 56K modem, preserving the role of the PSTN as the main route for data.

With more than $250 billion dollars invested in the telephone infrastructure in the USA alone, it is unlike these two technologies will coalesce in the near future, however many industrial pundits claim that packet switching will eventually replace circuit switching after a period of co-existing, as the traditional infrastructure is slowly side-lined [2]. This change will come about because of the enhanced services, flexibility and speed that technologies like ATM can offer. Integration with services as diverse as television, telephones and computer will be possible on this infrastructure, approaching the goal of digital convergence. The construction of the telephone network is also like to mirror how computer networking strategies have developed from a large centralised service using proprietary hardware and software, to a distributed network using new, generic, open technologies to dramatically reduce the cost of market entry. This will have a knock on effect to users, reducing over-all costs, while enhancing system flexibility and functionality.

4.2.1.3 Physical Layer Standards

FDDI (Fibre Distributed-Data Interface) is a standard, from the American National Standards Committee X3-T9, for data transmission on fibre optic lines in a local area network that can extend in range up to 200 km (124 miles), can support thousands of users and used to interconnect LANs using other protocols. An FDDI network contains two token rings, one for possible backup in case the primary ring (100 Mbps capacity) fails, offering an overall capacity to 200 Mbps. FDDI-II is a version of FDDI that adds the capability to add circuit-switched service to the network so that voice signals can also be handled. Work is underway to connect FDDI networks to the developing Synchronous Optical Network (SONET), which in turn is part of broadband ISDN (BISDN).

SONET is the U.S. (ANSI) standard for synchronous data transmission on optical media; the international equivalent of SONET is Synchronous Digital Hierarchy (SDH). Together, they ensure standards so that digital networks can interconnect internationally and that existing conventional transmission systems can take advantage of optical media through tributary attachments. They define standards for a number of line rates up to the maximum line rate of 9.953GB/s. Actual line rates approaching 20GB/s are possible.

ATM (asynchronous transfer mode), not to be confused with automated teller machines, is a dedicated-connection switching technology that runs on top of SONET and FDDI, which organizes digital data into 53-byte cells or packets and transmits them over a medium using digital signal technology. Individually, a cell is processed asynchronously relative to other related cells and is queued before being multiplexed over the line. ATM is designed to be easily implemented by hardware (rather than software), making faster processing speeds possible, giving speeds of up to 10 GB/s. The interest in moving to ATM is its support and integration of analog, baseband, and broadband services, with smart traffic control giving bandwidth on demand by varying the transmission protocol used; effectively ATM can operate as both a circuit switching network and a packet switching network and in intermediate modes depending on requirements. ATM, has also introduced a number of other important concepts such as switched/permanent virtual paths and circuits and parallel virtual machines, which will effect the future operation of networked services.

Integrated Services Digital Network (ISDN) is a set of CCITT/ITU standards for digital transmission over ordinary telephone copper wire, as well as over other media. It is aimed at the integration of digital and analog services on one connection, providing the capability of voice and data channels simultaneously. There are two levels of service: the Basic Rate Interface (BRI), intended for the home and small enterprise, and the Primary Rate Interface (PRI), for larger users. The BRI consists of two 64kB/s channels and one 16kB/s channel. The PRI, in contrast is aimed at service providers and larger organisations, consisting of 24 channels (1.5Gbps) in the US or 31 channels (1.92Gbps) in Europe. With the prospect of "fibre to the home" connections, broadband ISDN has already been planned which will encompass frame relay service for high-speed data that can be sent in large bursts over a Fibre Distributed-Data Interface (FDDI), and the Synchronous Optical Network (SONET). BISDN will support transmission from 2 Mbps up to much higher, but as yet unspecified, rates.

4.2.1.4 Threats and Security of the Physical Layer

Some of the main threats and risks to the physical layer were already outlined in Chapter 2. Eavesdropping, through wire-tapping, is the main threat to security at the physical level for cable systems, with substantial evidence this has been practiced both at the national level by security agencies and at the personal level by opportunistically placed equipment. For free-spaces systems the threat is more real because of the greater ease with which signals can be passively intercepted.

The face of physical networking will however change dramatically in the next decade. The use of optical fibres to support the higher bit rates demanded by

users, and the physical difficult in intercepting data at close to the limit of electronics, will make these interception techniques obsolete to a greater or lesser extent. The remaining two threats of disruption and destruction of the physical layer will remain however.

Greater competition for the growing data marketplace will have an effect on the services providers and what they provide. The opening up of monopolies held by national telecommunication companies has revolutionised the telephone network and likely to revolutionise the data network. The flexibility of the new protocols, like ATM, to multiplex everything from analogue broadband to high-speed data channels is part of this revolution which brings greater connectivity to every aspect of our lives, and potentially greater risks.

4.2.2 Data Link Layer

The data link layer, which is subdivided into two sublayers, the *Media Access Layer* (MAC) and the *Logical Link Control* (LLC), provides error control and synchronisation for the physical layer. Different data link layer specifications define different network and protocol characteristics, including physical addressing, network topology, error notification, sequencing of frames, and flow control.

The most basic security programs work by using the packets received in this layer. For example a packet sniffer intercepts packets at the data link layer, which can logged in a raw format for later analysis. Physical addressing is usually set in hardware in contrast to network addressing defined in software. Programs, for example arpwatch [3], can track the "physical address/network address" pairing and watch for impersonation.

There are two primary classes of networks defined at the data link layer level: *wide area networks* (WANs) and *local area networks* (LANs). LANs are the most common, with *Ethernet* dominating over *Token Ring* LANs because of its greater flexibility. WANs are dominated by X.25 networks current, though they are being replaced rapidly with ATM, SLIP and PPP data link layer protocols are also commonly used, primarily for serial line access, in particular modems. The hardware and firmware used to implement these protocols consist of repeaters, bridges, routers, switches, hubs and gateways. We will now consider all of these technologies in more detail.

4.2.2.1 Ethernet LANs

The most common form of local area networking (LAN) is *Ethernet*. As with most basic computer terms, the word was coined at Xerox PARC in the early

1970's and introduced a number of basic network features, such as collision detection, listen before talk and multiple access, which is also why the Ethernet channel access protocol is called Carrier Sense Multiple Access with Collision Detect (CSMA/CD). Currently the most widely used version of the Ethernet standard (IEEE 802.3) is still the original specification with a maximum data rate of 10MB/s, though this is currently being replaced by 100MB/s technology, with 1GB/s technology in the development stage. The flexibility of Ethernet is illustrated by the variety of cabling supported: thick coaxial, thin coaxial, twisted pair and optical fibres. It also supports both baseband protocols (such as 10BASE2 and 10BASE-T), which use only one logical channel on each physical channel, and broadband protocols (such as 10BROAD36), which support several services like voice and video on a single cable. It can also support different topological designs, such as bus and star designs, depending on networking requirements.

In the Ethernet standard the LLC sublayer manages services, both static connection-orientated service and dynamic connectionless services. Static services are vulnerable to attacks because they cannot be re-routed in the case of disruption and can be sensitive to the order of packet arrival. Resources are also consumed in providing the service and tied up when the service isn't being used to its full capacity. However for applications that do not tolerate delays and packet resequencing, like streaming audio and video, static services have to be used.

The MAC sublayer of the data link layer manages protocol access to the physical network medium, enabling unique identification. ARP (Address Resolution Protocol) is the protocol for mapping an Internet Protocol address (IP address) to a physical machine address that is recognized in the local network. This relationship is generally cached in memory and therefore it can potentially be corrupted by an attack. An example would be an attack to reconfigure the ARP table in a router to redirect data. Impersonation attacks are also possible, by two machines replying to the same broadcast requests for information. Caches can also be shared and therefore can be poisoned by sharing a bad ARP cache. Problems can also be caused by sending wrong answers to broadcast requests causing services to be disabled and diskless work stations to malfunction. This weak mapping is a common attack point and the subject of many exploits [4].

There are also early and proprietary versions of Ethernet, like Novell networks and Ethernet Version 2, which use a similar topology and protocols but different frame formats. Ethernet version 2 was designed before the IEEE specifications were published and only differs marginally. Novell networking similarly use a different header format where the logical link control information was blanked, leading to the nickname '802.3 raw'.

Ethernet will continue to dominate LAN design because of its flexibility and its ability to support a wide range of bandwidths. It meets the demands of most networked applications, however it is relatively inefficient and the MAC sublayer is particularly vulnerable to attacks. Unfortunately there does not appear to a sufficiently attractive alternative for LANs, even on the horizon.

4.2.2.2 Token Ring

The other major form of LAN networking is token ring networking. In this case a special frame, called a token, rotates around the ring when no stations are actively sending information. If a station wants to transmit on the ring, it must capture this token frame. The owner of the token is then the only station that can transmit on the ring, in contrast to the Ethernet topology where any station can transmit at any time. The transmitted data is received by each node on the ring and error-corrected if required. When the data reaches the destination node it transmits an acknowledgement to the sender, who in turn transmits a 'network free' token if it has finished.

Unlike Ethernet, two token ring networked machines cannot be directly connected together; a multistation access unit acts as a hub controlling the logical ring, which in fact is connected physically in a star arrangement with link distances up to approximately 150 metres. The main disadvantage of ring networks is illustrated by the rapidly decreasing packet frequency as the number of machines in the ring increases, from the maximum of 4MB/s or 16MB/s.

Token rings are relatively inflexible in topology and bandwidth, making them unattractive for dynamic networks. In particularly the decrease in bandwidth with an increase in the number of workstations is a major issue. Therefore the use of token ring networks is likely to decrease for all but specialised applications.

4.2.2.3 SLIP & PPP Networking

SLIP (Serial Line Internet Protocol) [5] and PPP (Point-to-Point Protocol) [6] are both protocols used for communication between two machines, primarily on a serial line. Relative to the OSI reference model, both protocols provide a layer 2 (data-link layer) service, generally in conjunction with TCP/IP networking protocols, described later.

PPP is usually preferred over the earlier de facto standard (SLIP) because it can handle synchronous as well as asynchronous connections. PPP is also a full-duplex protocol that can share a line with other users and it has error detection that SLIP lacks. In addition PPP can be used on various physical media,

including twisted pair, fibres and satellite transmission. It also uses a variation of High Speed Data Link Control (HDLC) for packet encapsulation.

The non-permanent and slow nature of serial lines, in particular modem connections has led to several features being included such as demand-dial, redial, call-back, tunnelling, byte-stuffing, compression and filtering which are generally there to benefit the users. The majority of users are home users and they utilise services provided by ISPs (Internet Service Providers) through the use of PPP and SLIP which provide a seamless interface to the Internet. The flexibility of who can connect to dial-up lines and call-back facilities however inevitable leads to a number of abuses [7].

Conversely, some ISPs have been accused of poor service, including lack of connections, censorship, lack of security and providing a haven for *spammers*. As demand for the Internet grows and users become savvier to the growing range of ISP services, these problems should diminish. In common with Ethernet, the inclusion of security features at this level is only now beginning to be tackled, hopefully leading to more secure and robust networking.

4.2.2.4 WANs - X.25 Standard, Frame Relay and ATM

For organisations in diverse locations, local area networks are not appropriate and therefore wide area networking (WAN) needs to be used. The X.25 standard defines a telephone network for data communications, providing users with WAN connectivity across public data networks. Initially it was developed to increase subscriber numbers to public data networks and aimed to eventually become a global standard. It was developed for the old analogue and copper-based PDSN and PTSN, which was prone to high error rates. It defines a specification for layers 1-3 of the OSI model to allow point-to-point interaction at speeds from 9.6-64Kbps, with error and flow control, though not supporting connectionless services.

X.25 networks are still in use, though are gradually being replaced by digital technology such as frame relay and ATM technology which can carry ISDN. Frame relay is particularly suited to narrowband, circuit orientated ISDN connections whereas ATM is suited to high speed, packet-switched BISDN.

Frame relay is similar to X.25 in functionality and format, but with higher performance and efficiency, exploiting advances made in WAN technology. It is a form of packet switching based on the use of variable length link layer frames while eliminating the network layer, providing the streamlining and also acting as the distinguishing characteristic with Ethernet networks. It intrinsically assumes that flow and error control do not need to be implemented in the lower levels.

ATM, also called *cell relay*, is similar in concept to frame relay, and has been described previously. It is effectively the big brother of frame relay, supporting data rates several orders of magnitude higher, but using fixed cell sizes to reduce overheads. It has now become one of the major protocols used on the Internet backbone and will play a growing role because of its flexibility.

4.2.2.5 Repeaters, Bridges, Routers, Switches, Hubs and Gateways

Connecting workstations using the same protocol is relatively easy using a hub. However different types of WANs and LANs use unique and incompatible signalling systems, and therefore cannot be directly connected. Instead there are a number of methods of linking networks together. We will now consider some of these core elements for connecting computers together on a network.

Hubs are designed to help expansion of LANs within a given site. Hubs provide multiple ports from a central position for network access and can be chained together to provide larger networks, limited only by timing requirements [8]. There are two main types of hubs, repeater hubs and switching hubs. Unlike a repeater hub, whose individual ports combine together to create a single large LAN, a switching hub makes it possible to divide a sets of ports into multiple LANs that are linked together by way of the packet switching electronics in the hub. This allows a large number of individual LANs to be linked together and provide greater security against packet sniffing because of filtering at the hubs.

Bridging joins networks at the data link layer (layer 2) and acts as a traffic director. It retains a list of the MAC numbers for all the active stations on the network and uses this to determine the path of packets. These lists have become more complicated to reflect a number of attacks, for example jamming networks by injecting incorrect packets which bounce backwards and forward between bridges, sending packets to non-existent services, and poisoning the MAC table.

From a functional point of view, *switching* is exactly the same as bridging. However switches use specially designed hardware called Application Specific Integrated Circuits (ASICs) to perform the bridging and packet-forwarding functionality (as opposed to a central CPU and special software). Consequently, switches are much faster than bridges. Firmware implementation is generally more difficult to corrupt, and packet switches provide greater security by targeting packets to the correct machines, reducing the effectiveness of packet sniffers and increasing the detection of impersonation attacks.

In contrast to bridging and switching, *routing* works at the network layer (3), and separates physical networks into different logical networks. The sending workstation determines if its outgoing traffic is destined for a local or remote network, and for the later it sends the frame directly to the router. The router

examines the frame to determine the final destination, and using a routing table, forwards the frame to the destination network. Routing can either be a dedicated piece of hardware or can be performed using a computer as a server with two network cards, which can be used to actively filter information passing through it, forming a safety barrier, or *firewall*.

In general, routing is preferable to bridging because of the separation of networks, the ability to intelligently filter and control frames, and its flexibility in dealing with distance networks. There is however a number of network protocols, such as Microsoft's NetBEUI, which have no layer 3 network addressing and therefore cannot be routed.

Conversely routing must be used for frames going between different network topologies, for example IPX, AppleTalk or Token Ring networks connecting to an Ethernet network. *Gateways* are special purpose devices that can convert one protocol stack to another and are often used in conjunction with routers. These gateways are expensive to maintain as protocols are updated and changed, though a necessity for some interconnects. Exploits for routers are commonly available on the Internet [9] and therefore it is essential that care is taken in maintaining them.

In addition to these devices there are a number of other devices commonly used, for example electrical repeaters (level 1 devices) used in long distance interconnects. These repeaters use amplifiers, to clean up and retransmit attenuated signals and can leak significant RF radiation, which can be easily intercepted by eavesdroppers at close range. Similarly the other network connectors are susceptible to hardware attack, especially remote units that often have local port access for diagnostic purposes. More details of these attacks is given later and in Chapter 2.

A number of attacks on network hardware are possible. Traffic re-direction attacks are common to bridges and routers. The basic attack involves Eve telling Alice and Bob that a convenient root between their sites passes through her router, which she can then listen to the traffic passing through. This source routing was originally assumed always to be coming from an honest source, and although it continues to be used for diagnostic purposes, it is advisable that these features are disabled. These ICMP exploits will be described later in more detail.

There are no simple or single solutions to promoting better security at the data link level. Intelligent packet-switching hubs are one solution, which can inhibit packet sniffing on local hubs to combat local attacks. Access lists, firewalls and packet filters incorporated into routers and gateways can protect LANs from attacks originating from a WAN. Furthermore, built-in hardware encryption support would provide more secure point-to-point communications. However

the main disadvantage in implementing solutions at this level is that it generally requires regular maintenance of hardware and firmware, that is more costly and inefficient than software implementation.

4.2.3 Network Layers

The *network layer* provides routing and related functions that enable multiple data links to be combined into an internet. This is accomplished by the logical addressing (as opposed to the physical addressing) of devices. The key concerns addressed by this layer are the syntax of the data, control semantics and timing to insure reliable communications. In transmission it is responsible for encapsulating the data packets with flow control and error checking information, while in reception it reassembles the data packets using this information, which is then stripped off before the data is passed to the upper layers. The network layer also supports both connection-oriented and connectionless services from higher-layer protocols and multiplexing of different services and protocols. The most commonly used network layer protocols are the Internet Protocol (IP) and the Internet Control Message Protocol (ICMP), which form the backbone of the Internet. Other proprietary protocols include the Datagram Delivery Protocol (DDP) part of AppleTalk, IPX part of Novell networking, DECNET and SNA. We will now consider IP and ICMP in more detail.

4.2.3.1 IPv4

Internet Protocol (IP) is the central, unifying protocol in the TCP/IP suite, the protocol forming the backbone of the Internet. Currently IP (IP version 4) is a very simple protocol making very few guarantees, ignoring important considerations such as integrity and delivery. IP does however handle addressing, fragmentation, reassembly, and protocol multiplexing.

The IP datagram, attached to the front of a data packet consists of the sender's and receiver's addresses, length of packet and checksum, and identification bits denoting the service and version. Addressing originally supported five different types of network (with an address space of 256^4), with class B defining all the Internet IP addresses. In contrast to MAC addresses, IP addresses correspond to a virtual network connection, rather than a physical piece of hardware. This makes impersonation of an IP address significantly easier than a MAC address. It also permits multiple hosts to share a single connection. IP also supports local broadcast, multicast and subnet addressing through the address structure. This address structure interacts with MAC addresses through the ARP (address resolution protocol).

The simplicity of the current version of IP reflects the design considerations in 1981. It was never envisioned that IP would be used on a network the scale of the Internet or for the diversity of purposes. The lack of integrity and authentication controls is the major security weak point of IPv4 [10]. Spoofing of network services is the most common attack and can be achieved in several ways: address spoofing, data spoofing and connection hijacking. In more detail the attacks which can be made directly or indirectly to IPv4 include:

- *IP spoofing* (also known as address spoofing) is an attack designed to take advantage of trust relationships between hosts (eg. with .rhosts under UNIX). In this attack, the intruder will attempt to impersonate some trusted Internet host in order to gain access to the victim system. The procedure involves disabling the host that is being impersonated, (for example by TCP SYN flooding), guessing the TCP sequence numbers, and sending fake datagrams to the victim host. The attacker would then attempt to open some back door to provide an easier port of entry to the victim's system. Spoofing involves manipulation of the TCP/IP stack in order to forge the IP address, a function denied to all but "root" accounts. Windows machines are typically more at risk to this kind of attack because of their unrestricted user privileges, in contrast to Unix-like operating systems.

- If an attacker can somehow "see" the datagrams, they can either substitute their payload, or inject false datagrams into the traffic in order to carry an integrity attack. This is called *data spoofing*. In some cases such integrity attacks might lead to opening a back door to the system, and in others they may be directed at corrupting data. Integrity attacks can be combated to a degree by providing checksums with the transferred data (or digital signatures), however, this is only possible in a limited number of cases, and hence they provide a high threat. For instance, with a growing number of users and companies taking their business (at least partially) to the web, integrity of the information presented there is crucial. An intruder might easily launch some sort of data spoofing attack when the user is uploading data to a remote web server, and change, for instance, some vital piece of information. Downloading is just as important, as the downloaded data might contain malicious software, and the likes. The consequences of such actions are impossible to predict.

- There is also a variation of spoofing, which aims at taking over connections rather than forging them or substituting payload data. Called *connection hijacking*, the attacker attempts to take over an established connection. This form of attack works well in systems that only authenticate users during login procedures (for example telnet or

ftp sessions). The attacker waits while a legitimate user performs an operating system level authentication (usually some sort of a password scheme). After the user logs in, the attacker takes over a connection and the system sees him/her as a legitimate user.

- A different hijacking variation is *connection scavenging* [11]. It involves utilizing a dropped dial-up connection before the remote server notices that a connection has been broken. Because a dial-up host has no means of notifying the server that the connection has been lost (i.e. no alternative route exists to route the packets), an attacker may continue using a session opened by someone else. With an increasing number of dial-up users, this might become a serious problem.

- Disclosure attacks provide an attacker with some data about the network, or data flowing through the network. Packet sniffing refers to a disclosure type attack, possible when the attacker is logically located on the path between two communicating hosts, or in some cases, when he is located on the same subnet. The attacker can listen to all the datagrams as they are sent on the network and process their contents. Sniffing is a passive attack, and it is considerably difficult to detect it. It poses a problem because there are several publicly available software packages, which automate many tasks so that even novices can use them effectively. Sniffing however is time consuming. An attacker must usually sniff for a considerable amount of time before the attack yields useful information, however the type of data acquired may be highly useful. For instance by spying on a telnet session being open, the attacker will easily learn the user name and password of a legitimate user.

- Analysis of data flows can give information about what it contains, or what it is supposed to mean. This is true even for encrypted data. Frequency, size, and other attributes of transmissions can be logged to assist in this process. This attack is very difficult to utilize properly because it involves a significant amount of guesswork. It does, however, carry a large risk as it is very difficult to protect against using current practices.

- Several features provided by IP allow attackers to craft their datagrams in such a way as to avoid detection by some intrusion detection systems (IDS) or bypass packet-filtering firewalls. These attacks are called *evasion attacks*. The most commonly utilized features are datagram fragmentation and tunnelling. These are powerful attacks because they allow seemingly normal data to pass through a firewall, which would otherwise prevent it from doing so.

- *Operating system fingerprinting* is a common technique for determining the OS type running on remote machine by testing for patterns in behaviour of the TCP/IP stack. It is made possible by several overlapping factors such as certain ambiguities in protocol definitions, complexity of TCP/IP stack software, differences in the software among revisions etcetera. There are serious implications of the attacker knowing the OS of the remote host. Many attacks exploit software bugs in services, or applications running on attacked hosts. Attacker's ability to determine the type and version of OS will help identify the likely targets for such attacks. Fingerprinting has a considerable advantage over determining service software versions by scanning ports. In order to scan a port, the attacker must connect to it, and listen to a reply. Any IDS system configured to watch for port scans will immediately notice that someone is probing the system, and alert the administrators.

- Along with confidentiality, authentication and integrity, non-repudiation is one of the cornerstones of a secure communications system. Due to the considerable ease with which IP datagrams can be spoofed, forged, or modified, it is difficult to place liability for an attack on a particular host/user, since the IP address used in the attack might have been stolen or subverted. Even if the attack was actually carried out by that host/user, he/she can always claim otherwise, and it is hard to prove the claim to be false.

- DNS security is a large area in itself and is discussed elsewhere in more detail. There are a few possibilities however to utilize spoofing attacks to feed bogus DNS data to hosts. By spoofing replies an attacker can provide false host-to-address mappings. At best, this might prevent the host from communicating with the rest of the network, at worst, it opens doors for much more serious attacks.

These weaknesses seriously compromise the Internet Protocol and hence all the layers above it, which use it. A number of projects, for example Kerberos [12] and SKIP [13], have tried to compensate for these deficiencies, however their difficult in implementation has led to limited use. The pressure of the shortage of network addresses, and a need for universal action to be taken on security issues are the two major impetuses for developing a new version, IP version 6.

4.2.3.2 IPv6

IPv6 (or IPng – IP next generation) offers a number of benefits reflecting how Internet usage has changed. The address length has changed from 32 to 128 bits to support more devices on the network (IPv4 numbering will run out within 5

years), with the formal inclusion of features such as authentication, anycast and multicasting. Support for multicasting, the broadcast of one signal to be received by multiple computers, has also been incorporated into the Internet backbone to allow audio and video presentations using QoS parameters. IPv6 also streamlines the excessive routing overheads of IPv4 using Common Interdomain Routing Protocol (CIRP).

The security portion of IP (IPsec) is the other major inclusion. It significantly reduces the demands and changes of individual users computers and removes the onus of security for the applications layer. IPsec provides support for public key exchange, authentication and encryption, which can be upgraded and changed without affecting the overlying applications, removing the need for frequent, costly updates. The advantages of including security at the network layer include:

- Encryption and authentication in the physical layer is only possible on hop-by-hop basis, hence at every router along the path of a datagram, security fields would have to be recalculated. That, combined with various underlying physical networks of the Internet, would probably provide an unreliable, computation intensive, and incompatible solution.

- There are several attacks (for instance DoS attacks against TCP, such as forging RESET command), which can only be combated on a lower layer.

- There are at least two independent transport protocols, TCP and UDP, and there is only one common network protocol, IP.

- Applying security features at application level has many benefits, however it is difficult to make it transparent to the end user. Also the number of applications that would have to be modified in order to use the new features is prohibitive, and if an attempt was ever made to carry out software modifications, the result would most likely to lead to incompatible solutions.

Using these arguments IPsec incorporates mechanisms for authentication, integrity control, and confidentiality of IP datagrams. The use of strong cryptography provides these properties but brings its own, for example in how to exchanges keys securely. The issue of key distribution and the regulations governing the export and import of strong cryptographic systems will be discussed more fully in the next chapter.

IPsec is not without its problems [10]. Implementation of the new IP has been rather slow for a number of reasons in addition to the key distribution problem. Encryption and protection of packets prevents "legal" monitoring; government

agencies are understandably reluctant to block a source of intelligence and analysis. Application layer firewalls essentially would also become severely inhibited without access to all the private keys of the machines within its domain because it would be unable to analyse the payload of packets.

Although spoofing, evasion and analysis attacks can be combated very effectively by IPsec, a number of new problems are likely to arise. These include bugs in the new software required, particularly in the transition from IPv4 to IPv6 and the hybrid networks used in this period. The new IP also remains vulnerable to denial-of-service attacks caused by buffer overflows, a problem which is unlikely to go away. The increase in name space is also a concern and likely increase the overheads and vulnerability of DNS; it is probable that there will be another overhaul of the naming scheme of the Internet within the next decade in order to address these problems.

IPv6 is generally considered a "good thing" and will eventually replace IPv4. The transition will be slow and painful and will need to address some important issues such as end user awareness of what is required in order to maintain good security. The advantages however significantly outweigh the disadvantages and will significantly decrease the range of attacks and increase the complexity required for successful intelligence gathering

4.2.3.3 ICMP & ARP – The Support Protocols

While IP handles the basic problem of routing data through the Internet, there are a number of auxiliary management issues that need to be addressed. Within the Internet separate protocols are used to perform these tasks and keep IP focused. ICMP (Internet Control Message Protocol) is the most common of these and is used to report errors and carry out simple measurements. The types of errors reported are the exhaustion of the hop count, or an invalid header parameter, or an unreachable destination. The measurements are simple, echo this packet or echo and timestamp this packet. It is also used in common ping and traceroute programs.

ICMP is the source of eternal arguments, because originally it was designed for assurance and debugging. It is now the most common way of scanning networks and identifying vulnerable hosts. Echo reply requests, time-to-live packets, and unreachable messages can all give away valuable information about hosts to an attacker. As mentioned earlier, interception can be assisted using the redirect requests of ICMP. More disruptive attacks have also been discovered, for example the *ping of death* which cause computers to crash by sending them echo requests with a larger data structure than expected and *ping flooding* a brute force denial-of-service attack [14]. A variant of ping flooding, *smurfing* [15]

spoofs ICMP packets from the victim's machine and sends them to a broadcast address, essentially amplifying the request, and causing the resulting flood of responses to crash the victim's machine. Further ICMP exploits can be found here [16].

Some of the other protocols that support IP in the network layer are:

- **ARP** - Address Resolution Protocol is used to map IP addresses to IEEE 802 addresses. The sought after IP address is broadcast and the station with that address responds with its own ethernet address.

- **RARP** - Reverse ARP, does just the reverse of ARP!

- **BOOTP and DHCP** - used to distribute network configuration information to a station as it boots up.

As with ICMP, there are a number of ARP, BOOTP and DHCP attacks [17]. Spoofing of ARP packets, denial-of-service and flooding of ARP tables have all been used to good effect. Remote booting of workstations using information transmitted over a network is extremely risky and generally should not be used because of this threat.

Defence against exploits of the ICMP and related protocols are best achieved using a firewall designed to reject these packets and to disable response on machines wherever possible. By default however these protocols are enabled and not blocked. This is one of the reasons why these exploits are so popular as they are generally an indication of poor security and can be easily carried out. Fortunately these protocols are also being updated with IPv6, removing many of these vulnerabilities. In any case, if they are not critically required, they should be disabled as they remain a potential security risk.

4.2.4 Transport Layer

The *transport layer* implements reliable data transport services that are transparent to the upper layers. Transport-layer functions typically include flow control, multiplexing, virtual circuit management, and error checking and recovery. Some transport-layer implementations include Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Name Binding Protocol, and OSI transport protocols.

The two most common transport layer protocols, UDP and TCP, provide assurance of data integrity on top of the network layer. TCP is primarily for applications that require all data to arrive in the correct order and therefore is heavily controlled. UDP, with fewer overheads, is for applications where it is

not important that all datagrams arrive and that they are in the correct order. We will now consider these two protocols in more detail and the TCP/IP suite used on the Internet.

4.2.4.1 Transmission Control Protocol (TCP)

The Transmission Control Protocol (TCP) is a connection-oriented protocol that specifies the format of data and acknowledgments used in the transfer of data. TCP also specifies the procedures that the computers use to ensure that the data arrives correctly. TCP allows multiple applications on a system to communicate concurrently because it handles all multiplexing of the incoming traffic among the application programs. TCP is also responsible for verifying the correct delivery of data from client to server; data can be lost in the intermediate network due to a number of effects. TCP adds support to detect errors or lost data and to trigger retransmission until the data is correctly and completely received. *Sockets* is a name given to the package of subroutines that provide access to TCP/IP on most systems.

TCP provides minimal confidentiality, authentication and integrity, and therefore can be easily attacked. The common attacks made possible by the current version of TCP include:

- **SYN Flooding** – At the beginning of a connection, a synchronisation packet (SYN) is sent to the service provider. SYN attacks take advantage of a flaw in how most hosts implement this three-way handshake. When Host B receives the SYN request from A, it must keep track of the partially opened connection in a "listen queue" for at least 75 seconds. This is to allow successful connections even with long network delays. The problem with doing this is that many implementations can only keep track of a very limited number of connections (most track only 5 connections by default). A malicious host can exploit the small size of the listen queue by sending multiple SYN requests to a host, but never replying to the synchronise and acknowledge (SYN&ACK) replies. By doing so, the other host's listen queue is quickly filled up, and it will stop accepting new connections, creating a denial-of-service.

- **ISN Attacks** - A sequence number (32-bit), ISN, is used in TCP connections to synchronise and regulate data flow. Normally it is impossible to guess the ISN (5 billion unique numbers), however, if the ISN for a connection is assigned in a predictable way, it becomes relatively easy to guess. This flaw in TCP/IP implementations was recognized as far back as 1985. By first establishing a real connection to the victim, the attacker can determine the current state of the system's

counter, used to generate the ISN. The attacker then knows that the next ISN to be assigned by the victim is quite likely to be the predetermined ISN, plus 64. The attacker has an even higher chance of correctly guessing the ISN if he sends a number of spoofed IP frames, each with a different, but likely, ISN. However, when the host receiving spoofed packets completes its part of the three-way handshake, it will send a SYN&ACK to the spoofed host. This host will reject the SYN&ACK, because it never started a connection -- the host indicates this by sending a reset command (RST), and the attacker's connection will be aborted. To avoid this, the attacker can use the aforementioned SYN attack to swamp the host it is imitating. The SYN&ACK sent by the attacked host will then be ignored, along with any other packets sent while the host is flooded. The attacker then has free reign to finish with his attack. Of course, if the impersonated host happens to be off-line (or was somehow forced off-line), the attacker need not worry about what the victim is sending out.

- **Desynchronisation Attacks** – It is possible for an attacker to desynchronise the TCP connection of the victim from the service host. During the three-way handshake process, after host B sends the SYN&ACK packet to host A, the attacker forges new packets from B (to A) in which the connection is first closed via the RST bit, and then a new three-way handshake is initiated with A; identical to the original, "real" handshake but with different sequence numbers. Host B now ignores messages from A (because A is using the attacker's new sequence numbers), and Host A ignores messages from B (because A is expecting messages with the attacker's sequence numbers). The attacker then replicates new packets, with the correct sequence numbers, whenever A and B try to communicate. In doing so, the attacker may also modify the messages or inject his own. If a RST packet is sent in the middle of a legitimate connection, the connection is closed and the application/user is notified of this. To cause desynchronisation in the middle of a connection, without closing the connection, only the sequence number counters should be altered. The Telnet protocol, in particular, provides an interesting mechanism to do this. By sending enough NOP commands, an attacker can cause the connection to become desynchronised and can then begin replicating new packets, with the correct sequence numbers, as before.

Software to realise these attacks is readily available on the Internet [18]. The most common of these is SYN flooding, which is often used as a prelude to a more complicated type of attack. It is surprising perhaps that the other attacks

have not been used more because of their obvious benefits, in particularly hijacking telnet sessions.

There are however a number of ways to address these attacks and minimize their impact. Authentication is a major problem and will only be addressed with the implementation of IPv6. Replacing the Berkeley system ISN with a pseudo-random number generator would decrease the probability of ISN attacks by several orders of magnitude, however there would always remain the possibility the attacker could guess the right number sequences. TCP wrappers, small programs, which run in place of the service daemons, can perform some useful security functions. These programs can provide extra logging, authentication and even be used as tripwires for detecting unusual or unauthorised actions. However these should be viewed as repairs to a flawed system and therefore of limited value. The majority of these attacks are more adequately dealt with by the protocols defined in IPv6.

4.2.4.2 User Datagram Protocols (UDP)

The User Datagram Protocol (UDP) is used when reliability mechanisms in TCP are not required; UDP is often referred to a best-effort protocol and TCP as a reliable protocol. In contrast to TCP, UDP is a connection-less oriented protocol, relying on the application to divide the data into packets (datagrams) and reassembling them correctly at the other end. Therefore fewer overheads are required and applications are less sensitive to network usage. It currently supports much of bandwidth-intensive, multimedia and multicast applications on the Internet, which uses small datagrams and requires little reassembling overheads.

In order to maximise the benefits of TCP and UDP many programs will use a separate TCP connection as well as a UDP connection. Important status information is sent along the reliable TCP connection, while the main data stream is sent via UDP. It can also prevent some of the attacks on UDP connections, by providing greater control, authentication and integrity of datagrams.

It should be noted that UDP is even more vulnerable to attack then TCP. The attacks described in the previous section can be replicated with UDP, without the complication of SYN & ACK packets. Messages can be altered, delayed, destroyed and replayed in order to attack a communication channel, because of the lack of control, authentication and reliability. The most common attack, analogous to SYN flooding, is *packet storming*, where a UDP socket is flooded with datagrams. Other attacks allow system logs to be modified on vulnerable hosts, to scan hosts for open ports and to create or steal arbitrary packets. These

attacks make UDP channels extremely insecure and vulnerable to unskilled opponents.

4.2.4.3 TCP/IP Suite

Transmission Control Protocol / Internet Protocol (TCP/IP) is the de facto protocol used on the Internet for connectivity and transmission of data across heterogeneous systems. It is an open standard which is available on most Unix systems, VMS, other minicomputer systems, many mainframe and super-computing systems and PC systems. The most common hardware solution is Ethernet, but TCP/IP will also run on Token-Ring, AT&T StarLAN, microwave & spread spectrum systems, LocalTalk (needs a gateway), serial lines (modems, serial connections) and other systems as well.

TCP/IP encompasses a suite of networking protocols, taking its name from two of the fundamental protocols in the collection, TCP and IP. Other core protocols in the suite are UDP and ICMP. These protocols work together to provide a basic networking framework that is used by many different application protocols, each tuned to achieving a particular goal. This suite of protocols is one of the triumphs of the open source movement, because, unlike other standards such as Ethernet, they have been universally developed, with no single author or owner, and are widely and freely available. They were developed as part of ARPANET in the 1970's by the US DoD to be robust and automatically recover from any node failure. This has the advantage that large networks can be constructed with less central management, however it can also mean that network problems can go undiagnosed and uncorrected for a long time. The protocols are now defined in a series of RFCs (Requests for Comments) [19] managed by the Internet Engineering Steering Group (IESG) based on recommendations from the Internet Engineering Task Force (IETF) [20].

Critiques as to the extent of security problems in the current version of TCP/IP can be found here [21] and software to exploit them here [22]. The ability of the open source movement to respond to the changing needs of the Internet is reflected in the development of IPv6 to address the security and address issues of IPv4. The publicly scrutinised development has been useful, however a number of problems remain, as discussed earlier.

It is unlikely that any other de-facto protocol suite will replace TCP/IP within the next decade because it has been so widely implemented and is so flexible to the needs of the Internet. Serious thought should be given to whether such a vulnerable network suite should be used on critical computers, a process often overlooked.

4.2.5 Session and Presentation Layers

The session layer establishes, manages, and terminates communication sessions between presentation layers. Communication sessions consist of service requests and service responses between applications located on different network devices. *Daemons*, programs that listen for incoming requests targeted at logical ports, manage the session layer. They start application layer programs, as required, to deal with the incoming requests. The session layer supports the Remote Procedure Call (RPC) protocol to support client/server interaction.

The presentation layer, or syntax layer, provides a variety of coding and conversion functions that are applied to application layer data. These functions, normally incorporated in the operating system, ensure that information sent from the application layer of one system will be readable by the application layer of another system. Some examples of presentation-layer coding and conversion schemes include: data representation formats (graphics, sound, video), data compression schemes, and data encryption schemes.

Common TCP data presentation protocols are:

- **SNMP** – Simple Network Management Protocol - designed to facilitate the exchange of management information between network devices. By using SNMP to access management information data (such as packets per second and network error rates), network administrators can more easily manage network performance and find and solve network problems. Clearly, access to such a resource must be heavily protected. However it is possible to have a null authentication service; this is a bad idea. Even a "read-only" mode is dangerous; it may expose the target host to netstat -type attacks if the particular Management Information Base (MIB) used includes sequence numbers.

- **FTP** - File Transfer Protocol [23] – FTP is one of the most widely and heavily used Internet applications. FTP can be used to transfer both ASCII and binary files between computers. Separate channels are used for commands and data transfer. The username and password used to access this service are transmitted in plaintext and particularly vulnerable to interception. Anonymous FTP allows external users to retrieve files from a restricted area without prior arrangement or authorisation. The FTP daemon runs with extremely high privilege levels and there have been several bugs in the daemon, which have opened disastrous security holes. Trivial file transfer protocol (TFTP), a variant of FTP, permits file transfers without any attempt at authentication. Thus, any publicly readable file in the entire universe is accessible. It is the responsibility of the implementer and/or the system administrator to make that universe as small as possible.

- **Telnet** - Terminal Emulation Protocol [23] - allows virtual terminal emulation. The user is normally authenticated based on user name and password. Both of these are transmitted in plain text over the network however, and is therefore susceptible to capture. One solution is to use asymmetric encryption for key management of encrypted telnet sessions, for example the SSH protocol [24].

- **SMTP** - Simple Mail Transfer Protocol [23] - provides an electronic mail transport mechanism. The most common implementation of this, the sendmail program, is a security nightmare. Sendmail violates the principle of least privilege as it runs with root privilege. The content of mail messages can also pose dangers. Automatic execution of messages, the ability to mail executable programs and the ability to mail postscript files are all dangers. The extent of sendmail exploits [25] exceeds any other presentation layer protocol. The current favourite is buffer overruns, though it is also worth noting that sendmail can be exploited for reconnaissance attacks using the VRFY function.

- **HTTP** - Hypertext Transfer Protocol [23] - enables services to terminals running WWW clients and browsers. The incidence of attacks on web sites is increasing rapidly [26] and WWW security is a significant cause for concern. The HTML specification allows protocols other than HTTP to be used (e.g., FTP, TELNET, RLOGIN), bypassing the filters normally applied to those protocols by a firewall. This can be rectified by using a HTTP proxy, which filters the relevant protocols as required. Other problems include unexpected input values can cause actions which were not intended, special characters may allow unauthorised access to the host, unexpectedly large input may cause a buffer overflow resulting in inappropriate actions and the potential for data driven attack especially for Trojan horses. Authentication / confidentiality / integrity issues are of particular concern for electronic commerce. The most common web servers have a number of vulnerabilities, apache is relatively responsive and secure [27] especially when compared to Microsoft's IIS [28]. Javascript, ActiveX and HTML overflows are perennial favourites for attacking a victim's web browser and there is a good security case for disabling advanced features in web browsers.

- **RPC Protocols** – Remote Procedure Call Protocols [23] – RPC is layered on top of TCP or UDP. However it is used as a general purpose transport protocol in the same way as TCP and UDP by a variety of application protocols such as Network File System (NFS) and Network Information Service (NIS). RPC services are vulnerable to a number of attacks [29]. In particular one of the most dangerous RPC applications is the Network Information Service (NIS). NIS is used to distribute a variety of important databases, including the password file, the host address table, and

the public and private key databases used for secure RPC. The Network File System (NFS) also exhibits some serious security problems because clients are allowed to read, change or delete files without having to log onto the server or enter a password, NFS has very weak client authentication and if not properly configured NFS can allow any other host to simply mount its file system.

- **POP & IMAP** - Post Office Protocol & Internet Message Access Protocol - these are protocols for management of electronic mail, one of the most valuable services on the Internet. Nevertheless, they are vulnerable to misuse. As normally implemented, the mail server provides no authentication mechanisms. This leaves the door wide open to faked messages. These protocols allow a remote user to retrieve mail stored on a central server machine. Authentication is by means of a single command containing both the user name and the password, sent in clear text and extremely vulnerable to packet sniffing. As an alternative, some sites are adopting "one-time passwords" using a cryptographic key scheme to defeat eavesdroppers.

- **Finger Protocol** - Many systems implement a finger service, which displays useful information about users, such as their full names, phone numbers, office numbers, etc. Unfortunately, such data provides useful grist for the mill of a password cracker. By default such services should be disabled and other methods found for disseminating necessary information in a more secure way.

- **DNS** – Domain Name Service [23] - DNS is a distributed database system used to match host names with IP addresses. An intruder who interferes with the proper operation of the DNS can mount a variety of attacks, including denial of service and password collection. There are a significant number of vulnerabilities, which can be readily exploited. In some implementations, it is possible to mount a sequence number attack against a particular user, similar to the ISN attack described earlier. A combined attack on the domain system and the routing mechanisms can be catastrophic. The intruder can intercept virtually all requests to translate names to IP addresses, and supply the address of a subverted machine instead; this would allow the intruder to spy on all traffic, and build a nice collection of passwords if desired. For this reason, domain servers are high-value targets; a sufficiently determined attacker might find it useful to take over a server by other means, including subverting the used machine, or even physically interfering with its link to the Internet. Even when DNS is functioning correctly, it can be used for some types of spying. Zone transfers (AXFR) can be used to download an entire section of the database; by applying this recursively, a complete map of the name space can be produced. Such a database represents a potential security risk; if, for example, an intruder knows that a particular brand of

host or operating system has a particular vulnerability, that database can be consulted to find all such targets. There is a further discussion of DNS weaknesses in Chapter 8. Altogether, for increased security an address-based authentication should be used. Although still weak, it is far better than unmodified name based authentication. The use of Kerberos can greatly improve security, however is not trivial to implement in a working, distributed environment.

Other common protocols are: DHCP, DNS, NNTP, rlogin, rsh, rexec, and the X Window System providing common connectivity for applications layer programs. They also exhibit a wide variety of exploits [23].

To improve security, only essential daemons should be used on any computer connected to a network. All daemons divulge information about a host, making it more vulnerable to attack. On a server, daemons should be run with the lowest possibly priority and offer generic challenges, preventing identification of the operating system. Protocols should be secured wherever possible using stronger authentication, privacy and integrity controls, for example in the case of using SSH in place of telnet. Users should also be made aware of the security issues involved, and avoid sending their usernames and passwords across a public network in clear text.

4.2.6 Applications Layer

The application layer is the OSI layer closest to the end user. Application-layer functions typically include identifying communication partners, determining resource availability, and synchronizing communication.

Two key types of application-layer implementations are TCP/IP applications and OSI applications. TCP/IP applications are protocols, such as Telnet, File Transfer Protocol (FTP), and Simple Mail Transfer Protocol (SMTP) that exist in the Internet Protocol suite. OSI applications are protocols, such as File Transfer, Access, and Management (FTAM), Virtual Terminal Protocol (VTP), and Common Management Information Protocol (CMIP), which exist in the OSI suite. The security of some of these protocols was discussed in the previous subsection.

The applications layer provides the major of software vulnerabilities for computer systems and the extent of these will be explored more fully in the next chapter. From this section it is obvious the extent to which a network is vulnerable to attack from a wide variety of attacks. In the next section we will look at the origins of these attacks and general steps that can be taken towards improving the security of a network.

4.3 Threat & Security Issues

There is an amazingly lax attitude towards security by the majority of organisations. This is especially surprising given that one third of businesses connected to the Internet in 1995 reported up to $100,000 in financial loss over a two-year period due to malicious acts by computer users outside the firm. A little more than two percent of connected companies reported losses of more than $1 million. The hardware, software and information that constitute computer systems are increasingly mission-critical for these businesses. Protecting them can be as important as protecting other valuable resources, such as money, buildings, or employees.

This lax attitude in part is due to the IT press, which often trumpets crypto-systems as the universal problem-solver. It is important to draw a distinction between data security and service security. In this section we will focus on threats to services and techniques which can be used to protect these services.

There are four areas of concern when a trusted network is attached to an untrusted network, for example a private intranet to the public Internet:

1. that inappropriate material will deliberately, or inadvertently, be passed to and from the untrusted network;

2. that unauthorised users will be able to gain access to the trusted network from the untrusted network

3. that the operations of the trusted network may be disrupted as a result of an attack from the untrusted network.

4. that operation of the trusted network may be affected as a result of misuse within the trusted network

In this section we will consider the steps which can be taken to minimize the risk of these networking issues. We will also consider the physical security of computers and networks, complementing the analyses in the other chapters. A good, complementary analysis of intruders, their techniques and steps which can be taken to protect networked computers can be found amongst the publications of Lance Spitzner [30].

4.3.1 Actors, Their Motivations and Objectives

The identity of the people trying to gain unauthorised access to system resources is an important factor in the types of attacks used against a system. The Internet has brought about a large cultural change in hacking, as unskilled people have

relative easy access now to exploits and programs on the Internet, which they can use for 'fun' without a full understanding the nature of the attack.

The groups and individuals who pose a threat to computers and computer networks can be broken down into a number of major risk groups. In this subsection we will consider some of these groups and their motivations for reaching a particular goal. These groups are:

1) **Common Consumer** – this group corresponds to widely available knowledge and techniques. The majority of people have access to a computer either at home or work and know the rudiments of computer and network security. Their attacks tend to be opportunistic, for example access to a computer console where it has been left unattended, or accessing files on a world-readable share. Their motives can be varied, however they tend to be poorly resourced and can be defeated by good security practices.

2) **Amateur** – enthusiasts with recreational interest in computers and networking pose a significant threat. With access to the Internet and consequently a wide base of information, they can be knowledgeable attackers though may be limited by resources. Tools used by this group are typically standard tools, like generic scripts and network scanners, for probing and exploiting vulnerabilities. Again motivation varies widely from intellectual interest to malicious intent. The large numbers of attacks by this group require organisations to ensure there is proper security at all points of their computer systems and there is monitoring of unusual system activity. Employees are the biggest sub-group, and it should be considered that all employees pose a threat at least at this level.

3) **Restricted Professional** – individuals who make a living from the computing industry pose several threats. The first is that sensitive information regarding a network can be unwittingly or deliberately disseminated to unauthorised entities; the former by a careless employee not shredding sensitive documents, the later by a disgruntled employee. An employee with a grudge could also potentially booby-trap or trojanise software controlling networks for his or her own purposes. Their motivation tends to be for financial reasons. They primarily commit active and passive attacks, which can be detected by implementing good management of resources and codes of practice. At this level there is significant development of new intrusion techniques, which can often lead to penetration without detection, though which should be detected by IDS systems.

4) **Professional Organisation** – multi-national companies often employ or have internal departments responsible for information technology and hence data security. These groups have access to substantial budgets and the latest

equipment and the potential to develop their own attacks. They offer a formidable threat backed with substantial resources generally targeted at other organisations that pose a financial threat. Again the attacks are primarily active in order to penetrate a hostile environment, however their secondary task can also be to monitor and audit internal network usage. Hardware and software attacks are common and can include computer forensics and penetration testing.

5) **Intelligence Agency** – the highest risk is posed by governmental intelligence agencies that have essentially unlimited resources, access to cutting edge research technologies and can operate above the law to some extent. They have access and the resources to employ hardware and network attacks at all levels, including TEMPEST monitoring and backbone data sniffing.

Rogue motivations of individuals and small groups also need to be considered. For example, revenge for a perceived injustice (e.g. trojanised programs), terrorism of individuals or minority groups (e.g. mail-bombing), poverty (e.g. stealing equipment), corruption of employees by moral or financial means (e.g. moles leaking sensitive information) and poor security procedures (e.g. careless disposal of sensitive information) all need to be considered when assessing risks and threats to computers and networking.

The goal of an attack can be one of many objectives: to destroy or disrupt computer services, attack or passively monitor communication channels, or to modify hardware, software and/or users. We will consider the threats and security measures for hardware in the next subsection and networking issues in the subsequent subsection.

4.3.2 Hardware Issues

At close quarters, an attacker can have a much more devastating effect on computing resources, especially if they have been left unattended. Therefore physical computer safety is as at least as important an issue as network security. Normally computers are protected by normal access control mechanisms, such as locks and burglar alarms and these are sufficient to deter most opportunistic attacks. However there is little hardware security when the machine is switched off and removed by the attacker. In that case most of the data and information on the system is at risk. In this subsection we will consider the security of hardware and the ways in which it can be attacked.

4.3.2.1 Non-Invasive Attacks

With access to a computer, but not the means to physically access the components, an attacker can perform several attacks. The most obvious attacks are to destroy or disrupt service, by preventing power to the computer (e.g. jamming the power supply fan), destroying data (e.g. placing a strong magnet near the computer) and preventing network access (e.g. cutting the network cable). These can often be easily fixed and do not pose a significant threat to most organisations.

Of greater threat are passive and active attacks. Passive attacks include opportunistic views of other users' screens and password entry. More complicated attacks, such as cleaning a keyboard before use and then dusting after password entry, and using hidden video cameras to monitor key entry and screens have also been used. Active attacks on a computer include trying to access the system by password entry and opportunistic use of a computer left unattended but unlocked. Active attacks of a computer's network or serial connections can provide access to the system by mimicking other trusted devices, or impersonation of the computer under attack by another machine.

The high-risk factor of being detected during a non-invasive attack, limits attacks to opportunistic occasions and by the lower risk groups. Higher risk groups generally plan attacks and will use invasive techniques if they are going to penetrate the physical security surrounding the computer. If non-invasive monitoring is required by the highest-level threats then TEMPEST schemes will be used.

4.3.2.2 TEMPEST

TEMPEST (Transient Electromagnetic Pulse Emission Standard), the interception and reconstruction of electromagnetic radiation from computer components, is a growing tool of intelligence gathering operations [31]. It can be carried out actively or passively over several hundred metres and is difficult and expensive to fully guard against. Several aspects of it are discussed in different chapters of this book and a growing range of information can be found on the Internet [32].

It is generally broken down into three categories: Tempest which strictly refers to stray RF emissions, Hijack which refers to electromagnetic leakage through components forming a secondary weak signal and Nonstop which refers to secondary RF signals cause by placing machines in close proximity. Large numbers of machines do not provide protection against such effects and only expensive and complete shielding provides a defence. However, to mount such

an attack requires getting sufficiently close (hundred of metres) of the target and is resource intensive and therefore only likely to be carried out by governments and large organisations for the foreseeable future. Simple attacks can however be carried out with off-the-shelf hardware [33].

Protection against TEMPEST can be implemented in a number of ways. Primarily, software can be used to make recreation of the screen contents more difficult. In particular fonts can be changed [34]. Secondly, physical shielding, for example Faraday cages, can be used to insulate computer components [35]. Taken to its extremes, one group has bought a used nuclear bunker for its data haven [36].

4.3.2.3 Invasive Attacks

With physical access inside computer systems, an attacker has widen the options of attack. Only the basic security features of a computer system can protect against full access to information.

Most PCs have the option to set a password within the BIOS (basic input-output operating system), which will prevent system access if the computer is turned on, or rebooted. This is only a deterrent for very unskilled attackers. There are many ways around such passwords, from shorting the BIOS, re-blowing the BIOS chip, using an alternative BIOS or using software to extract the password [37]. With access to all the hardware components of a computer, the ones of most interest to an attacker are the ones which store data, in particular hard discs.

These discs contain sensitive information, including passwords, data files and cryptographic keys and authentication codes. These can be recovered in a number of ways. By using a floppy disc, or other filling system, to reboot the computer system access can be easily obtained to the files. Trojanised programs, viruses and other malicious code could also be installed and then the machine rebooted to appear as normal. The only protection against these attacks is to remove other filling systems and the possibility to boot off them. Tamper detection cases should also be used to detect when they have been opened. Filling systems, or at least sensitive data files, can be encrypted, making data recovery several orders of magnitude more difficult. If the attack is easily able to remove hardware from the premises, then this can be taken away for further examination. In the next subsection we will consider this field of computer forensics in more detail.

4.3.2.4 Computer Forensics

The desire to recover data, both authorised and unauthorised, has spawned the recently field of computer forensics. Data recovery from storage systems, such as tape backups and hard discs are the goal of this field. Previously this was the domain of security agencies because of the technical requirements. However as the tools and training have filtered to the civilian sector, along with the need for evidence preservation and preparedness, the demand for computer forensics has rocketed.

A wide range of tools is now available [38] for forensic usage such as incident response, data elimination, audit tracking, document recovery and ambient data processing. This last category is potentially the most powerful for forensics experts and hackers alike; ambient data is found in everything from swap files, memory, the stacks, file slack and unallocated file space. These areas can contain fragments of processes run on the system (immediately before shutdown), including passwords, networking and security data and applications data such as e-mail, word processing, calculations and network connections.

Protecting against this type of attack is extremely difficult. Encryption is the strongest method, followed by tamper detection systems. Although very desirable, constructing tamper resistant devices is a difficult art. IBM defines three levels of tamper-resistance though no formal standards exist [39]:

1. **Casual Attack** – resistance against attack by unskilled individuals – for example VISA hardware security modules used in banks, which are disabled by micro-switches when opened.

2. **Knowledgeable Attack** – resistance against attack by knowledgeable individuals. An example would be IBM's 47xx cryptographic units, which contain many fine wires that are damaged when the units are opened.

3. **Funded Attacks** – resistance against attacks by well-funded organisations – for example the design of the Clipper chip.

The critical question with regard to resistance is whether an opponent can obtain unsupervised access to the device. If they cannot, then relatively simple measures may suffice, however, as has been well illustrated by Pay-TV even smart cards rated by government signals agencies as 'the most secure processor generally available' are routinely broken because of the technologies available and financial incentive. Normally these attacks against such targets as pre-payment cards, metering of resources such as electricity and gas, remote locking devices for cars, and mobile phones are well funded and use an extensive range

of technologies and have access to multiple copies of the cryptographic equipment.

Attacks range from testing ones, such as single stepping, input voltage varying and using unusual temperature, to destructive ones using high technology like laser drilling and focused ion beams. The EPROM (erasable programmable read-only memory) is the main target of these attacks and these can be accessed and analysed using relatively simple and inexpensive techniques. Reverse engineering integrated circuits is commonly done, even to the level of processor chips; for example an Intel 80386 took two weeks and six chips to determine its structure and the Clipper chip was reverse engineered by at least one US chipmaker shortly after release. More recent tamper proof modules, employing many sophisticated and refined countermeasures, such as the Dallas DS5002FP secure microcontroller, have also been shown to be susceptible to attack. There is strong evidence though that the military have technologies capable of withstanding limited level three attacks, such as the prescribed action systems used to protect nuclear weapons and the seismometers used to detect test-ban treaty violations. Therefore, the current best is that one can impose cost and delay to a capable and motivated opponent. An excellent review of tamper resistance and attack methods has been published by Anderson et al. [40].

There have been a growing number of incidents were sensitive data has been recovered from a computer either intentionally or unintentionally left in the hands of a third party. For example the discovery of child pornography on his computer, led to the imprisonment of Gary Glitter [41], and there have been several high profile cases of senior civil servants and military personnel having their laptops stolen, only to be returned several days later [42].

4.3.3 Network Issues

The threat of a software attack is the greatest single threat to a networked computer. The ability to attack from a remote site with a high degree of anonymity is a powerful defence against detection. In the next chapter we will consider software threats to network services in more detail and in Chapter 2 we considered the attacks that can be carried out on the hardware infrastructure of a network. In this subsection we will consider the threat to network services from a software attack and how the risk of service loss can be minimized.

Attacks can be analysed more fully in terms of a five step plan:

1. The first step of any attack is *outside reconnaissance*, gathering intelligence without giving away intention; for example DNS entries can give useful information such as host operating system.

2. The second step is *insider reconnaissance*. In order to determine whether particular vulnerabilities exist, more invasive techniques are used to scan for information. For example TCP/UDP scans give away which ports are open and which services are running.

3. Until this point nothing illegal has occurred. Once a vulnerability has been identified however, the intruder will attempt to *exploit* it. Several exploits may be attempted, for example first getting access to a user's account and then attempting to get root access.

4. Once unauthorised access has been achieved, an intruder will ensure a proper *foot-hold*. The intruder's main goal is to hide evidence of the attacks (doctoring the audit trail and log files) and make sure they can get back in again. They may install 'toolkits' that give them access, replace existing services with their own Trojan horses that have backdoor passwords, or create their own user accounts. The intruder will then often use the system at a later date as a stepping-stone to other systems, since most networks have fewer defences from inside attacks.

5. With access, an intruder will take advantage of their status for personal *profit*. This can be to attack other machines or to steal confidential material or deface web sites.

In assessing the risk and threat posed by particular attacks it is worth considering the steps an intruder needs to take. This is useful in configuring intrusion detection systems, described later. Common points of weakness can also be identified and measures taken to reduce risk. A useful analysis of security incidents between 1989-1995 can be found here [43]. The National Infrastructure Protection Center [44] and CERT publish regular reports into the impact of intrusions and their analysis of future trends, which is helpful in the planning of longer-term security measures.

We will consider attacks in two broad classes, active and passive attacks. The responses to these threats fall into two main groups: measures that provide protection and those that provide detection. We will consider two of these countermeasures, firewalls and intrusion detection systems.

4.3.3.1 Active Attacks

There is a constant battle between the *whitehats* [45], the network and security administrators, and *blackhats* [46], the hackers and crackers, over network security. Ironically both groups use the same tools and techniques to discover and detect vulnerabilities. Often this distinction can be blurred as hackers and their groups move from one side to the other or sit in the middle, providing both

hacking information and tools to prevent it. Sites like www.rootshell.com act as repositories for software exploits and often the people who use them are referred to as 'script kiddies', but they also act as useful databases for network administrators. Internet mailing lists such as bugtraq [47] and meetings such as DEFCON [48], play a dual role of informing both hackers and network administrators.

In the previous section a large number of active attacks were outlined at all levels of the OSI model. The common types of active attacks are reconnaissance, exploiting weaknesses and denial-of-service attacks. These attacks make use of software bugs, system configuration errors, passive interception, password cracking and design flaws. Reconnaissance in itself is not illegal, however often it is a prelude to an attack and therefore it is important to determine if a network is being probed, so steps can be taken to ensure the risk of intrusion is minimized. Detection systems can also give useful intelligence as to who poses the real threats and current intrusion techniques.

Ironically one of the reasons why the Internet was first developed was to provide resilience against denial-of-service attacks, initially conceived of as being nuclear strikes against the defence computers in ARPANET. Denial of Service (DoS) attacks have become a common feature of the Internet, often used to bring down web sites. The most infamous case of a DoS is the Internet worm in 1988 [49], which paralysed substantial parts of the Internet. More recently SYN flooding has become the most popular method for creating a denial-of-service, because of its simplicity and low overheads.

Active attacks can be combated in a number of ways. Later in this subsection we will consider the two most common: firewalls and intrusion detection systems. Practical steps, such as disabling vulnerable services, maintaining up to date software especially after CERT advisories [50], using TCP wrappers, and installing cryptographic and authentication schemes can improve the security of a system considerably. Third party penetration testing can be used as a proactive measure to identify and correct vulnerabilities before they are maliciously exploited. Plans should also be made for a migration to IPv6 because of the obvious security benefits its provides.

4.3.3.2 Passive Attacks – Network Monitoring

The use of software tools to monitor networks is a double-edged sword. On one hand it can provide a means to analyse use, target resources more effectively and watch for abuse of the system. On the other hand they can be used for packet sniffing and analysis of traffic to identify hosts and vulnerabilities. In Chapter 2

hardware tools were described which can assist in network monitoring and these are complementary to the software tools described in this section.

The provision of integrated packages for network administration is big business now with a rapidly growing number of intranets within organisations and wider use of the Internet. There is also a wide range of public domain tools, in particular for Unix-like operating systems, which can be used, or exploited. These generally do not offer the level of integration, or user-friendly interfacing, but can be as effective in the hands of a knowledgeable user.

The largest challenge to legitimate network monitoring is to provide a high ratio of true:positives (correctly identified attacks) to false:positives (attacks not detected). For example, the US National Computer Security Agency (NCSA) asserts that most attacks against computer systems go undetected and unreported, citing attacks against 9000 Department of Defence computers by the US Defence Information Systems Agency (DISA). These attacks had an 88% success rate and went undetected by more than 95% of the target organisations. Only 5% of the 5% that detected an attack, a mere 22 sites, reacted to it [51]. It is noteworthy that these sites belong to the US Department of Defence and were not commercial sites, which indeed may have even lower levels of security. NCSA also quote the FBI as reporting that in more than 80% of FBI investigated computer crimes, unauthorised access was gained through the Internet.

The range of networked services, and hence the range of types of system abuse, is enormous and growing rapidly with the amount of network traffic and the demands of customers. Therefore it can be difficult to correctly identify attacks amongst the overwhelming major of legitimate traffic, especially of unannounced vulnerabilities. Consequently there is a lot of interest in developing smart, automated systems, based on technologies such as neural networks and intelligence databases, which can be used to give accurate detection of unusual network traffic occurrences. In the next chapter we will consider similar systems that are used to detect system intrusion and abnormalities.

Packet Sniffers

Anyone with access to a LAN can install a packet sniffer, which will harvest all the traffic, or filter traffic looking for particular key words and signatures on a network. Access control by passwords, for services like telnet and ftp, are normally transmitted in the clear and therefore are very susceptible to sniffing. Backbone sniffing has been done for a large number of years by intelligence agencies, and by others in more recent years.

Packet sniffers [52], such as Sniffit [53], and network monitoring tools like Sniffer Pro [54], work by putting the network information controller of the computer into a promiscuous mode; instead of only listening out for packets addressed to that computer, the computer will listen to all packets on the network. By spooling all these to disc it is possible to generate log files of all network traffic, or by using a filter to look for particular traffic. From an administration point of view this is a powerful tool, enabling identification of problems on the network and potential misuse. From an attacker's point of view they can discover the services being run on the network, users habits and intercept passwords and data files. There are now programs, such as AntiSniff [55] which can detect some of the characteristic signs of a packet sniffer, however the most effective way of stopping passive attacks which use packet capture and analysis is to use encrypted data on the software side, and hubs which do not broadcast packets to all computers on the net. Although these are relatively easy solutions the overwhelming majority of networks and users have not taken these precautions.

4.3.3.3 Protection – Firewalls, Security Policies & VPNs

One answer to attacks is to introduce firewalls to protect intranets. It is surprising that only approximately 40% of the fortune 500 companies using the Internet have installed a firewall, given their obvious advantages. As the Internet continues to double annually, it is not surprising that the security auditing business is booming.

A firewall is a set of related programs, which run on a gateway server (single point of entry) using a set of security policies (access control policy ACP), to limit access to and from the intranet. The most basic form of firewall uses simple packet filtering to deny non-standard packets, in particular to prevent packets for and from unusual TCP/UDP ports and IP addresses. Such firewalls are commonly available on many Unix-like systems, however they can be defeated in a number of ways. These firewalls, working at the network layer (3) level, do not attempt to analyse the payload of packets and therefore are limited to blocking scans, illegal port access and avoiding bad addresses. They are susceptible to buffer overruns, IP spoofing and ICMP tunnelling, but have a high data throughput rate and are inexpensive to implement and maintain.

More complex firewalls exist, called *circuit gateways*, working at the applications layer (7), which reassemble and examine all packets. These firewalls can be configured for active filtering and proxy support for e-mail, ftp, http etcetera. The power of filtering data channels is significantly more effective that network layer filtering and prevents a wider range of attacks and exploits. In particular it is useful for preventing tunnel-and-pry attacks where the payload of legitimate

looking packets is a virus or worm. The more detailed logging of application layer firewalls can also form an important part of auditing and forensic investigations. All of these benefits are however achieved at the expensive of throughput, which can significantly decrease if the firewall is faced by a large volume of traffic. This protection ironically also forms its major weakness – denial-of-service attacks. By flooding a firewall with traffic it can cause the software to crash, effectively cutting off network access, because it is usually the only access point.

Firewalls are used for a number of different reasons. Primarily they are aimed at detecting and preventing unauthorised intrusion attempts. They are also increasingly being used to stop inappropriate Internet usage by employees, by blocking services such as ICQ [56]. It should be remembered however that a firewall is only as good as the rules governing it, and does not have intelligent inspection capabilities. Therefore it is relatively easy to attack a machine behind a firewall or for internal users to "tunnel" out. For example http exploits can be used to attack a web server behind a firewall as the traffic rules will be configured to allow packets from and destined for port 80. Another often misunderstood belief is that firewalls can prevent viruses and worms. Generally firewalls are not sufficiently intelligent to detect malicious payloads within legitimate documents such as Word macros and e-mail attachments. This illustrates why an IDS is useful in conjunction with a firewall.

Firewalls are widely available to suit all tastes [57], and used by individuals, many large companies and organisations that can afford to maintain them. There is also evidence that a number of countries, in particular in Asia, that use firewalls to limit foreign influence [58]. Whether firewalls actually provide increased security is a mute point amongst some analysts, as properly configured machines ultimately provide more security, with or without a firewall. However security is not often the focus or background of those making the decisions; a 'solution in a box' is often required by inexperienced network managers, making a firewall an attractive proposal, though in effect this can be false security if not properly maintained.

Firewalls can create a number of serious bottlenecks and problems. Some organisations have a high percentage of internetwork traffic which all has to be processed by the firewall. In addition if a user has any particular application requirements or if new applications are introduced using other network services, the firewall software needs to be updated to take account of all the changes, requiring time consuming, expensive maintenance. The initial design and implementation of a firewall can also be time-consuming [59]. Firewalls can also be counterproductive: the classic analogy is with burglars trying to break into premises guarded by alarm systems; by activating the alarms and assessing response, burglars can deny the premises a normal working pattern as well as

decrease the police's confidence in the alarm system, while increasing their own ability to burgle the premises successfully.

There is also a philosophical question in that most attacks on computer services occur within intranets, rather than originating from external networks. Perhaps the most telling shortcoming is noted by the military who have found the only real method of stopping attacks is to use multi-level secure machines that confirm to security policies such as the Bell-LaPadula model [60] and which have been subjected to independent testing. Whereas the military is primarily interested in eavesdroppers, other sectors, for example the banking and commercial sectors, are more interested in ensuring integrity and are more likely to follow other security models, such as the Clark-Wilson model [61].

Virtual Private Networks (VPNs)

To combat the threat of sending data of public networks, many organisations are turning to virtual private networking (VPN) [62]. Some liberal estimates put the market as large as $10 billion in 2001 [63], indicating the seriousness with which organisations are treating confidentiality and the increasing mobility of employees and diversity of network access points.

VPNs use cryptography to encrypt packets when they are sent over public networks. They can be included in the firewall structure of an organisation to encrypt only packets destined for machines outside the intranet, reducing overheads associated with encryption and decryption. Similarly, routers can also be used for VPN.

Generally four different protocols are used for creating VPNs over the Internet: point-to-point tunnelling protocol (PPTP), layer-2 forwarding (L2F), layer-2 tunnelling protocol (L2TP), and IP security protocol (IPSec). PPTP is the default VPN included in Microsoft products and benefits from working at the datalink layer (2), however it does not include strong encryption or authentication protocols. Like PPTP, L2F uses PPP for authentication, but also includes support for a wider range of authentication and control systems, supports multiple connections through a single tunnel and is not reliant on IP. L2TP is being designed by an IETF working group as the heir apparent to PPTP and L2F, designed to address the shortcomings of these past protocols and become an IETF–approved standard. IPSec, discussed earlier, is perhaps the most important protocol, because of its strong security measures, however it is designed only to handle IP packets and therefore not suitable for NetBEUI, IPX, or AppleTalk.

A wide range of VPN software is available using these protocols [64]. They are attractive for a number of reasons. They offer secure site-to-site and remote access connections through public networks. This is considerably cheaper than renting dedicated, long-distance lines for the same purpose. The security of the transmitted packets prevents opportunistic sniffing for information and reduces the risk of knowledgeable attack by an intruder. The major disadvantage however is that safety is achieved at the cost of performance [65].

The future role of VPNs is slightly uncertain given the development of IPv6, which uses IPSec as its security protocol. The only remaining advantage of some VPNs would be their ability to support other networking protocols, however these form only a small percentage of the market. Therefore it is probable that the impact of VPNs will diminish over the next decade.

4.3.3.4 Detection - Intrusion Detection Systems

The other major form of monitoring unauthorised use of computer resources is to use intrusion detection systems (IDS) [66]. This is one of the fastest growing areas of computer security research, with US military funding now exceeding $500 million. This has been prompted by the realisation that attacks happen, there are ways of monitoring and controlling them, but that currently there are very few ways in which they are logged or audited. The most basic IDS uses simple thresholding to determine unusual behaviour and flag for further attention. Examples are phone calls lasting more than six hours, three or more failed logins and unusual expenditure patterns.

More sophisticated systems use a number of other techniques to detect unauthorised behaviour. Misuse detection systems try to detect the likely behaviour of an attacker. For example, a rapid increase in network resources by a user for file transfer or processing time. Some forms of misuse are easy to detect given a clear model, however attackers have become more subtle and consequently more sensitive anomaly detection is much more difficult. The detection of anomalies is difficult because of the fine line between normal use and unusual usage due to external variables and therefore AI systems are used with varying degrees of success to try and improve detection. Further details of IDS products and methods can be found here [67].

In contrast to, and complementing firewalls, IDS software is designed to detect threats. The two are often used together, with the IDS system placed outside the firewall or inside the firewall on the intranet, to detect all attacks, or just those which penetrate the firewall. An IDS system within a firewall can also be used to detect if the firewall is properly configured and unauthorised activity within

the intranet. This later use is becoming particular popular, given that the statistics suggest the vast majority of attacks are from insiders.

IDS systems also provide auditing of computer networks, an increasingly important issue in the fight against computer crime. A well organised audit trail which can be rapidly and efficiently interrogated for instances of newly discovered vulnerabilities is a boon to all large organisations in assessing risk.

In addition there are a number of tools that can be used to complement an IDS. System integrity verifiers, for example Tripwire, can watch software components to detect changes made by an attacker. This is useful in understanding the techniques used and the programs and logs which can be modified to maintain unauthorised, undetected entry. Automated log file monitors can be used to scan large log files for patterns that suggest an intruder is attacking. These are particularly useful in detecting http server attacks, which may only send one suspicious request every few minutes, amongst other legitimate traffic. The third set of tools, and which cause controversy are deception systems. These fly-traps or honey-pots are designed to lure an intruder to attack a system because of apparent vulnerabilities. The intruder's actions can then be studied in a secure environment without compromising system security. A number of deception papers and software packages are publicly available [68,69].

4.4 The Internet

4.4.1 Introduction

It is important to distinguish between an internet (with a small 'i', a contraction of internetworks) referring to a network of networks and the Internet, which refers to the global internet which has developed out of the original ARPANET. The Internet has essentially evolved as a solution to three key problems: isolated local area networks (LANs), duplication of resources, and a lack of network management. The US military originally funded ARPANET as an experiment in distributed, resilient networking for an effective defence system. It was gradually replaced in the 1980's by a new defence network, the Defence Data Network, and NSFNet funded by the National Science Foundation. This was matched by national networks in other countries, such as JANET (Joint Academic Network) in the UK, connected by international links. In 1995 NSFNet began to be phased out and now the backbone of the internet is run by a consortium of commercial backbone providers.

Currently there are more than 250 million people worldwide who regularly use the Internet, with an approximately growth of 10% per month. More than 65%

of computer sold are for the home market are more than 90% with modems, allowing connection to the Internet through service providers. With more than a third of families in the USA with an Internet connected PC, there is a great demand for information gathering, processing and distribution. As with computer speeds, the demand of large bandwidth applications is also driving forward network technology, with a doubling of network performance every nine months.

The popularity of the Internet has originated from a number of key concepts, most importantly that connection for home users is flat rate in the USA. The commonality, using TCP/IP as the common protocol base and with the majority of systems running a Unix-like operating system is also an important factor in giving seemless interfacing between services. There are now more 56 million hosts on the Internet, with nearly $500 billion dollars of revenue being generated each year from the Internet, indicating its staggering growth [70].

The Internet does however have a number of drawbacks, as indicated in this chapter. Primarily it was not designed originally to grow to this size and its protocols have had to be patched in order to cope with the huge user base. The distributed chaotic nature is frustrating for many and has led to a number of exploits as outlined earlier. Security within protocols was not a major consideration when the original protocols were developed as it was assumed that attackers would rarely have direct access to the network, instead its resilience to physical attack was deemed more important.

Therefore there has been growing interest in upgrading the Internet to meet these problems and to provide the bandwidth and flexibility for new services.

4.4.2 Internet 2 – The Future of the Internet?

Internet 2 is a collaborative effort among a number of US universities, federal R&D agencies (who refer to it as the Next Generation Internet (NGI)), and private sector firms to develop a next generation Internet for research and education, including both enhanced network services as well as the multimedia applications, which will be enabled by those services. The project will be conducted in phases over the next three to five years. In the initial project phase, end-to-end broadband network services will be established among the participating universities. On a parallel basis, teams of university faculty, researchers, technical staff and industry experts will begin designing applications. It is expected that within eighteen months, "beta" versions of a number of applications will be in operation among the Internet2 Project universities.

The project includes hardware, middleware and software development. Hard-

ware development includes Abilene (advanced backbone infrastructure) to provide connects between gigaPoPs (gigabyte bandwidth points-of-presence). The protocols used will be IPv6, QoS and multicast to provide a wider, more secure set of services. Software is being developed for Tele-immersion, virtual laboratories, digital libraries, and distributed instruction amongst other ideas [71].

Will Internet2 be the future of the Internet? It is perhaps best seen as a testbed for software, protocols and hardware before implementing then on a world-wide basis. Many of the concepts will undoubtedly be developed on a world-wide basis, however the large size of the Internet makes fast implementation difficult. Internet2 is touted as a means to test technology and develop new networking services.

Other, smaller scale, projects are going on world-wide, for example SuperJANET4, the replacement for the academic network connecting universities in the UK [72]. Again these are seen as test beds for universities to iron out problems with technology before they are implemented more widely on the Internet.

All of these developments look promising because of the inclusion of stronger, publicly developed protocols, in keeping with the history of the Internet. The use of IPv6 and its associated protocols will remove a number of security issues and the inclusion of a public key infrastructure (PKI) within Internet2 is a sign that security is being taken seriously.

However these developments will not remove application vulnerabilities, which will continue to be the main point of entry into networked computers. The rapid development of network services will prevent stable, effective operating systems from existing for a useful period. Instead, the rapid turnover of software products will lead an increase in vulnerability through poor installation and management and the growth of exploits for each different version.

4.5 Conclusions and Future Prospects

In this chapter we have looked at the security and vulnerabilities of computers and computer networks from a service perspective. In the second section, network design was considered using the OSI model. Design flaws and vulnerabilities were identified in each of the seven layers and steps identified to strength the security at each layer. Of particular concern was the TCP/IP suite, which forms the basis on the Internet. Many of these issues will be addressed by the next version, IPv6, though undoubtedly it will open the possibility for new attacks.

In the third section hardware and network service security issues were considered. Invasive and non-invasive techniques for gathering intelligence from computer hardware were described, complementing the techniques described in Chapter 2. The future role of computer forensics is likely to become increasingly important with society's greater reliance on IT products. Security of network services, from both passive and active attacks was also considered. Specific software products were described and the threat they pose. Two solutions to these threats were addressed in terms of protection and detection; firewalls provide protection against undesirable network traffic, while IDS systems provide detection of unusual behaviour. The increasing awareness of the need for security will cause rapid growth in these sectors as organisations and individuals implement these systems.

The Internet was considered in more detail in the fourth section, reflecting its growing importance in daily life. Of particular interest was the technologies which will change the nature of the Internet over the next decade, and affect security of computers and networks.

References and Further Reading:

Cheswick and Bellovin, *Firewalls and Internet Security – Repelling the Wily Hacker,* Addison-Wesley (1994)

S. Garfinkel and E. Spafford, *Practical Unix and Internet Security,* O'Reilly & Associates (1996)

Simson Garfinkel, Gene Spafford, *Web Security & Commerce* (O'Reilly Nutshell), O'Reilly & Associates (1997)

Deborah Russell, G. T. Gangemi, *Computer Security Basics*, O'Reilly & Associates (1991)

Chris Brenton, *Mastering Network Security*, Sybex Inc. (1998)

Stephen Northcutt, *Network Intrusion Detection: An Analysts' Handbook*, New Riders Publishing (1999)

Elizabeth D. Zwicky, Simon Cooper, D. Brent Chapman, Deborah Russell, *Building Internet Firewalls*, O'Reilly & Associates (2000)

Stuart McClure, Joel Scambray, George Kurtz, *Hacking Exposed: Network Security Secrets and Solutions*, Computing McGraw-Hill (1999)

Maximum Linux Security : A Hacker's Guide to Protecting Your Linux Server and Workstation, Sams (1999)

Bruce Schneier, *Secrets and Lies: Digital Security in a Networked World*, John Wiley & Sons (2000)

[1] Data Communications Cabling FAQ

http://www.cs.ruu.nl/wais/html/na-dir/LANs/cabling-faq.html

[2] Doug Klaiber, *A new switching architecture for a new competitive environment*, CTI Developer (June 1999)

http://www.tmcnet.com/articles/ctimag/0699/0699taqua.htm

[3] arpwatch software package

http://freshmeat.net/search/?q=arpwatch

[4] ARP Exploits of Rootshell

http://rootshell.com/secengine/search.cgi?query=arp

[5] *A non-standard for transmission of IP datagrams over serial lines – SLIP*, rfc1055, Network Working Group

http://www.cis.ohio-state.edu/htbin/rfc/rfc1055.html

[6] Point-to-Point Protocol Frequently Asked Questions

http://www.faqs.org/faqs/by-newsgroup/comp/comp.protocols.ppp.html

[7] AOL Hacking

http://www.zdnet.com/zdnn/special/aolhack.html

http://www.aolsucks.com/

[8] *Extending Ethernets with Hubs*, Quick Reference Guide to the Ethernet System

http://wwwhost.ots.utexas.edu/ethernet/100quickref/ch1qr_16.html#HEADING15

[9] e.g. AntiCode's Router Exploits

http://www.AntiOnline.com/cgi-bin/anticode/anticode.pl?dir=router-exploits

[10] Marcin Dobrucki, *Seminar on Network Security*, Helsinki University of Technology

http://www.tml.hut.fi/Opinnot/Tik-110.501/1999/papers/ipv6/ip6sec.html

[11] Phrack Magazine, vol.8 issue.54

http://www.phrack.com/

[12] Kerberos Primers

http://uk.dir.yahoo.com/Computers_and_Internet/Security_and_Encryption/Kerberos/

[13] SKIP Website

http://skip.incog.com/

[14] Linux Security Administrator's Guide - Exploits

http://www.nic.com/~dave/SecurityAdminGuide/SecurityAdminGuide-11.html

[15] Craig A. Huegen, *The latest in denial of service attacks: Smurfing*

http://www.pentics.net/denial-of-service/white-papers/smurf.cgi

[16] ICMP Exploits

http://rootshell.com/secengine/search.cgi?query=icmp

http://search.atomz.com/search/?sp-q=icmp&sp-a=00051847-sp00000000

[17] ARP Exploits

http://rootshell.com/secengine/search.cgi?query=arp

[18] Packetstorm TCP Exploits

http://packetstorm.securify.com/unix-exploits/tcp-exploits/

[19] RFC Editor Website, Internet Society

http://www.rfc-editor.org/

[20]The Internet Engineering Task Force Website

http://www.ietf.org/

[21] TCP/IP Security Critiques

http://www.ja.net/CERT/Bellovin/TCP-IP_Security_Problems.html

http://www.firstmonday.dk/issues/issue2_5/rowland/

http://www.linuxsecurity.com/resource_files/documentation/tcpip-security.html

[22] TCP/IP Exploits

http://members.tripod.com/html_editor/exploits.htm

http://www.networkice.com/advice/Exploits/Ports/

http://search.cyberarmy.com/Exploits_and_Internet_War/

[23] William Cheswick and Steven Bellovin. *Firewalls and Internet Security*, Addison Wesley, (1994)

[24] SSH Primers

http://uk.dir.yahoo.com/Computers_and_Internet/Communications_and_Networking/Software/Unix_Utilities/Ssh__Secure_Shell_/

[25] sendmail exploits

http://packetstorm.securify.com/exploits/apps/mail/sendmail/

http://search.cyberarmy.com/Exploits_and_Internet_War/Unix/SendMail/

[26] Steven Miller, *Civilizing Cyberspace*, Addison-Wesley (1996)

[27] Apache Vulnerabilities

http://rootshell.com/secengine/search.cgi?query=apache

[28] IIS Exploits

http://rootshell.com/secengine/search.cgi?query=iis

[29] RPC Exploits

http://rootshell.com/secengine/search.cgi?query=rpc

[30] Lance Spitzner, *Technical Whitepapers and Publications*

http://www.enteract.com/~lspitz/pubs.html

[31] P. Wright, *Spycatcher – The Candid Autobiography of a Senior Intelligence Officer*, William Heinemann, Australia (1987)

[32] Joel McNamara, *The Complete Unofficial TEMPEST Information Page*

http://www.eskimo.com/~joelm/tempest.html

[33] W. van Eck, *Electromagnetic radiation from video display units: an eavesdropping risk?*, Computers & Security, vol.4, no. 4, pp. 269-286 (Dec 1985)

[34] Ross Anderson's Soft Tempest Work

http://www.cl.cam.ac.uk/users/rja14/#Tempest

[35] The Complete, Unofficial TEMPEST Information Page

http://www.eskimo.com/~joelm/tempest.html

[36] The Bunker Internet Site

http://www.thebunker.net/

[37] The PC Hacking FAQ

http://www.skyinet.net/~siobee/densold/hacker-faq.htm#biospass

[38] e.g. New Technologies Inc. Forensic and Security Software

http://www.secure-data.com/tools.html

[39] D.G. Abraham, G. M. Dolan, G.P. Double, J.V. Stevens, *Transaction Security System*, IBM Systems Journal, 30:2, 206-229 (1991)

[40] R. Andrews & M. Kuhn, *Tamper Resistance – a Cautionary Note*, The Second USENIX Workshop on Electronic Commerce Proceedings, Oakland, California, pp 1-11 (1996)

[41] *Gary Glitter sentenced to four months*, RTE News, 12th November 1999

http://www.rte.ie/news/1999/1112/glitter.html

[42] *Laptop stolen from British Minister in charge of nuclear secrets*, RTE News, 4th June 2000

http://www.rte.ie/news/2000/0604/laptop.html

[43] John D. Howard, *An Analysis Of Security Incidents On The Internet 1989 – 1995*, Ph.D. Carnegie Mellon University

http://www.cert.org/research/JHThesis/Start.html

[44] National Infrastructure Protection Center Webpages

http://www.nipc.gov/

[45] Whitehats Network Security Resource Website

http://whitehats.com/

[46] Hack News Network Website

http://www.hackernews.com/

[47] Bugtraq Archive on securityfocus.com

http://www.securityfocus.com/frames/?content=/templates/archive.pike%3Flist%3D1

[48] DEFCON Website

http://www.defcon.org/

[49] Charles Schmidt and Tom Darby, *The What, Why, and How of the 1988 Internet Worm*

http://www.software.com.pl/newarchive/misc/Worm/darbyt/pages/worm.html

[50] CERT Coordination Center

http://www.cert.org/

[51] Jorge Cobb and Prathima Agrawal, *Congestion or Corruption? A Strategy for Efficient Wireless {TCP} Sessions*, IEEE Symposium on Computers and Communications, Alexandria, Egypt, 262-268 (1995)"

[52] Packet sniffer packages

http://freshmeat.net/search/?q=sniffer

[53] Sniffit Package Website

http://reptile.rug.ac.be/~coder/sniffit/sniffit.html

[54] Network Associate's Sniffer Pro Package

http://www.axial.co.uk/products/manufacturers/nai/sniffer/

[55] Lopht's AntiSniff Package

http://www.l0pht.com/antisniff/

[56] ICQ Communication Network

http://www.icq.com/

[57] Firewall Primers

http://uk.dir.yahoo.com/Business_and_Economy/Companies/Computers/Business_to_Busines

s/Security_and_Encryption/Software/Firewalls/

[58] Heather McCabe, The Net: Enemy of the State?, Wired News (12th Aug 1999)

http://www.wired.com/news/politics/0,1283,21240,00.html

[59] John Wack and Lisa Carnahan, *Keeping Your Site Comfortably Secure: An Introduction to Internet Firewalls*, NIST Special Publication 800-10

http://csrc.ncsl.nist.gov/nistpubs/800-10/main.html

[60] A Zaslavsky, *Protection & Security in Distributed & Mobile Computing Systems*

http://nemesis.csse.monash.edu.au/~azaslavs/cot5701_link/

[61] D.R. Clark, and D.R. Wilson, *A Comparison of Commercial and Military Computer Security Policies*, IEEE Symposium on Security and Privacy, Oakland, April 1987.

[62] VPN White Papers

http://www.vpnc.org/white-papers.html

[63] Nortel Network's VPN Exhibition

http://www.iec.org/exhibits/nortel_03/

[64] VPN Suppliers

http://uk.dir.yahoo.com/Business_and_Economy/Business_to_Business/Computers/Communications_and_Networking/Virtual_Private_Networks__VPNs_/

[65] PC Magazine Performance Tests (Dec 1999)

http://www.zdnet.com/pcmag/stories/reviews/0,6755,2400750,00.html

[66] Network Intrusion Detection Systems FAQs

http://www.ticm.com/kb/faq/idsfaq.html

[67] IDS Primers

http://www.cerias.purdue.edu/coast/ids/

http://www.networkintrusion.co.uk/

[68] e.g. Deception Toolkit

http://www.all.net/dtk/

[69] Lance Spitzner, *To Build A Honeypot*

http://www.enteract.com/~lspitz/honeypot.html

[70] NUA Internet Surveys

http://www.nua.ie/surveys/

[71] Applications of Internet 2

http://apps.internet2.edu/

[72] SuperJANET4 Website

http://www.superjanet4.net/

Chapter 5 - Software Threats and Vulnerabilities

5.1 Introduction

In 1994 an Ernst and Young / Information Week survey [1] found that 54% of companies reported some type of financial loss in the previous two years as a result of computer problems, some from crashes and internal problems, but an increasing number from malicious damage. Since then, the Internet revolution has increased connectivity and computer based commerce as well as the range of ways in which computers can be attacked by software.

Five years later, a 1999/2000 CSI/FBI survey found 90% of respondents to have detected computer security breaches in the last twelve months, primarily large corporations and government agencies [2]. Seventy percent reported a variety of serious computer security breaches other than the most common ones of computer viruses, laptop theft or employee "net abuse"; for example, theft of proprietary information, financial fraud, system penetration from outsiders, denial of service attacks and sabotage of data or networks.

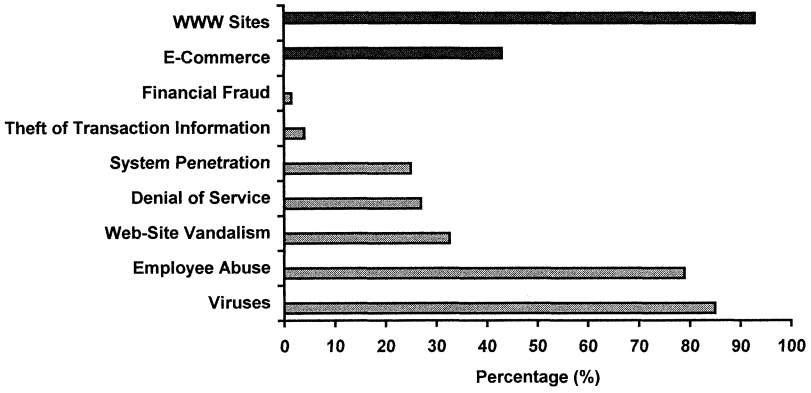

Figure 1 - Software attacks experienced in one year by 643 organisations [2]

Seventy-four percent acknowledged financial losses due to computer breaches. Forty-two percent were willing and/or able to quantify their financial losses. The losses from these 273 respondents totalled $265,589,940 (the average annual total over the last three years was $120,240,180), with the largest growth areas in sabotage of data or networks, theft of proprietary information and financial fraud. What is surprising therefore is the lack of proper auditing of misuse (Figure 2), with more than 80% of organisations not properly quantifying or identifying breaches of security.

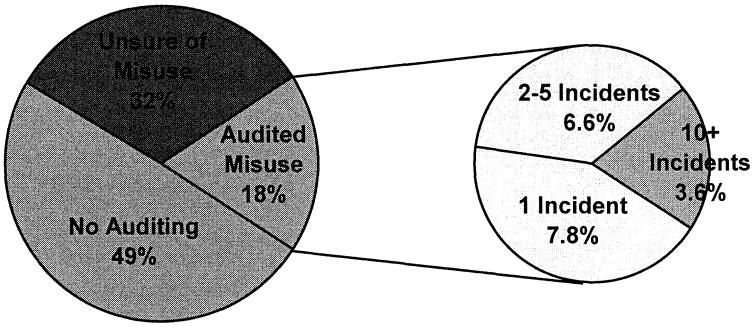

Figure 2 – Auditing of Computer Attacks [2]

Although more than seventy percent of organisations reported unauthorised access by insiders, primarily the abuse of Internet access privileges (for example, downloading pornography or pirated software, or inappropriate use of e-mail systems), there is a steady trend to remote attacks with fifty-nine percent citing the Internet connection as the point of attack.

This chapter will look at the current range of software threats, and the programs available to protect against and track these attacks. The second section will consider the origin of attacks: who, how and why. The goal of an attack is important in deciding the tools used and the probability of detection. The third section describes ten different techniques used to penetrate systems, all widely used. Particular packages will be described and compared with others, highlighting their signatures.

The second half of this chapter looks at software solutions dealing with these threats and their effectiveness in dealing with mutated and entirely new attacks. Five main areas are covered: operating system security, database security, anti-virus software, auditing tools and network security tools. This final area is the major growth area in both attacks and security tools because of the growing reliance on networking but also the most poorly understood. The chapter concludes with some comments on the future of software threats and the steps that can be taken to minimize them.

5.2 Actors, Their Motivations & Tactics

The origin and goal of an attack is important to understand in order to deal effectively with the threat it poses. The anonymity and remoteness of software attacks make them attractive for a wide range of groups and purposes. The

difficulty in preventing, tracking and prosecuting attackers makes them a powerful weapon in the virtual battlefield.

5.2.1 Actors

Software plays a central role in both threatening and protecting data and networks. In the previous chapters we have discussed hardware and data security, indicating that software is playing a central role. Later in the chapter we will consider some of the tools used to carry out attacks and defend against them. The main groups and individuals who use these tools include:

- **Law Enforcement and Intelligence Agencies** – with effectively unlimited resources and background knowledge, national agencies are formidable opponents. Their resources are typically only targeted against criminals, foreign governments and large organisations perceived as a threat, though they are also involved in indiscriminate passive interception for intelligence gathering. They typically use the full spectrum of penetration techniques, with options such as forcing software companies to include backdoors in their products, not normally open to other threat groups.

- **Organisations and Businesses** – commercial advantage can be gained by the "misfortunes" of competitors, which can include data loss, disruption of network services, "leaks" of sensitive material and defacement of websites. Interception of sensitive data is the most valuable resource to be gained from an attack but equally any slip in credibility for the victim can have the same dramatic effects. This is illustrated by banks reluctance to report the extent of computer crime.

- **Extremists** – the freedom and publicity offered by the Internet has attracted the attention of many extremist organisations, including terrorists, religious sects and political movements. Ironically they often attempt to stifle opinions and views contrary to their own, highlighted by the struggle of the Church of Scientology and its opponents [3]. The tools used by these groups tend to be destructive and disruptive and targeted specifically against their opponents. Although attacks by these groups are infrequent at present, there is a high probability they will increase in the longer term, particularly with increasing social divisions.

- **Employees** – the majority of computer criminals are employees. Access to internal information, hardware and internal networks is a valuable resource, permitting misuse, interception (sniffers), disruption (denial-of-service attacks), destruction (logic bombs) and modification

(reprogramming). Employee crime is a growing reality and employers need to tread a fine line between protecting their business and becoming a "big brother".

- **Computer Engineers & Administrators** – engineers and administrators with access to hardware and system software have unique access for committing crime. The motivating for committing crime can be varied, from financial gain, revenge or ideological reasons and is difficult to guard against. The attacks used are typically the implementation of Trojanised programs and backdoors and using network monitoring software to gather unauthorised intelligence on users. Proper procedures for auditing software and external testing can help prevent abuse by this group. However motivated and knowledge opponents with insider access are difficult to detect.

- **Hackers** [10] – attacking from a remote location, with good background knowledge in protocols and software programming, hackers pose a significant threat. Often intrusion by a hacker will go un-noticed because of insufficient intrusion detection software (IDS) and their use of unreported exploits. Typically, their attacks are non-destructive and cause minimal disruption, instead focussing on intelligence gathering in line with the 'hacker ethos'.

- **Cracker** – *cracker* is the term that is generally given to malicious hackers. These individuals are likely to use sophisticated techniques to gain unauthorized access with the specific intent of causing malicious damage (such as defacing a web page), or stealing confidential information for financial gain. Criminal groups, activists and security agencies could employ a cracker for their own purposes. This group is difficult to defend against as their attacks can be unpredictable and acting without any previous history.

- **Script Kiddies** [4] – many script archives exist on the Internet that contain simple scripts, in C, PERL, shell script etc. to exploit common vulnerabilities. In security circles, users of these scripts are often referred to as *script kiddies* because they use these standard scripts on multiple machines without a full understanding of their operation and how they can be tracked.

- **Spammers & Fraudsters** – with the increasing commercialisation of the Internet, there is an increasing group of people wishing to exploit vulnerabilities for their own purposes. *Spam*, the general term for junk mail, makes use of spider programs to collect email addresses, which are then sold on to advertisers who send unsolicited email to the addresses.

Fraud is committed in many different ways on the Internet; fake web sites to gather credit card numbers, selling pirated software and spreading false information are some of the techniques used. The growing use of the Internet, and the future growth of mobile financial services, will only encourage these practices. The use of authentication protocols and firewall software can minimize the impact of these attacks along with legal action against the perpetrators.

These groups exploit a wide range of tactics in order to exploit computer services. For each organisation the individuals and groups posing a threat should be categorised and weighted as part of security policy. The motivation and the tactics used by these groups will be discussed in the next two subsections.

5.2.2 Motivation

The purpose of a software attack is usually to misuse, destroy, disrupt, intercept, disinform or gain information and/or financial advantage; the same goals as discussed in the previous chapters. The motivation behind software attacks is varied but can be classified into several broad categories:

- **Military and Intelligence** – the cloak and dagger of espionage has turned virtual during the Cold War, with valuable information to be gleaned from computer break-ins and traffic analysis. Clifford Stoll, in *The Cuckoo's Egg* describes how a West Germany hacker extracted information from the defence computers of ten nations and then reportedly sold them to the Soviet KGB. *Information warfare* has been used to great effect in recent international conflicts to provide publicity, annoyance and disruption of services [5], a trend that will undoubtedly increase.

- **Business** – the phenomenal growth of the Internet and the strong dependence of business on computers make businesses increasingly the target of both competitors and the curious, although most business crime is still committed by employees. The cost of computer crime to businesses is estimated at £5 billion per annum in the UK alone [6].

- **Financial** – most of our financial transactions are now carried out electronically. Banks are a tempting target for computer criminals and with more e-banking services being offered daily and commercial data being stored and transmitted, this will become a growth area for software attacks.

- **Ideology / Terrorism** - During the 1980's several terrorist organisations attempted the physical destruction of computer resources, for example the Italian Red Brigade. Now the Internet offers them a two-sided sword, offering them publicity and exposure but also a means by which they can strike at their targets. A list of the terrorist groups, freedom fighters and propagandists active is available here [7] with details of some of their activities here [8].

- **Grudge** – Not all attacks seek to gain information; some simply want to wreak damage and destruction. One well-known case of a grudge attack was a logic-bomb planted by a Texas insurance company employee, Donald Gene Burleson, causing large financial losses [9].

- **Recreation** – in some ways computer crime is ideal: it is bigger, faster and more anonymous than traditional crimes. The largest recreational attack is by employees using extensive computer resources for their own purposes, like Internet surfing, personal e-mail etcetera. There is also a large section of computer criminals, traditionally referred to as *hackers* [10], who are not in it for the money, but perpetrate crime as an intellectual challenge, or occasionally with more malicious intent.

One of the major problems with the IT sector is that organisations have not invested sufficiently in staff, training and monitoring software to detect and combat computer crime. In the mushrooming rush for Internet based commerce, many have neglected risk analysis and sound auditing procedures. Policies are typically drawn up in reaction to events instead of proactive planning, in part causing a veil of secrecy over the extent and nature of computer crimes. Legal cases involving computer crime [11] are typically high profile and are considered bad publicity for companies involved and therefore lessons are difficult to learn.

5.2.3 Goals of an Attack

The motivated groups and individuals outlined in the previous subsections have a number of basic goals, which we will consider in this subsection. The vast array of attacks now used can be broken down into groups according to the objective of the attack, the most common of which we will consider now.

5.2.3.1 Destructive Attacks

The most rudimentary goal for an attack is the destruction of data. The benefit margin of launching a destructive attack is limited because the attacker gains

little advantage for the risk involved in carrying out the attack. The classic destructive attack is the use of a virus to penetrate a system and destroy data at a pre-determined time. In this case, the attack is indiscriminate and uncontrolled. More generally, trojanised programs can be triggered internally or externally to cause data loss or social engineering can be used to subvert the data storage cycle, for example by stray magnetic fields.

The motivation for these attacks tends to be emotionally based where the attacker does not seek information, rather revenge on the target. Typically, they are perpetrated by individuals and small groups. The major actors involved in destructive attacks are corrupted or sacked employees, terrorist organisations seeking publicity and crackers for kudos. The threat of such attacks has diminished in the last decade because of the ease with which other more beneficial attacks can be carried out and the smaller margins of loss inflicted on the targets.

Such blatantly hostile attacks are immediately apparent and therefore can generally be dealt with effectively. Software modification and anti-virus software are useful in detecting and preventing attacks, while frequent backups minimize data loss. Good management of security, employees and data handling substantially reduces the threat against all but nation-sponsored attacks.

5.2.3.2 Disruption Attacks

The safeguards against data destruction have prompted another, easier, form of attack where services are disrupted instead, circumventing these some of safeguards. A disruption attack can use a variety of techniques, most commonly software vulnerabilities to prevent normal operation and flooding the network with traffic to cause a denial-of-service.

In contrast to destructive attacks, disruption attacks are generally coordinated against particular targets in real-time. In some cases, the attacks are solely to provide disruption, although a disruption attack can also be used to precede a more dangerous active interception attack. To date most attacks have been carried out by individuals and small groups from a remote location motivated by revenge, ideology, or for recreation. There is strong evidence that a national information warfare attack would involve attempting massive, concerted disruption attacks although only limited incidents have been publicised so far [8].

The most common disruptive attack on the Internet, a *denial-of-service*, can be achieved in three ways: by attempting to flood a network to prevent legitimate traffic, by preventing or disrupting access to a service or user, or by disrupting

the network between the user and the service. There are three common modes of attack:

- consumption of scarce, limited, or non-renewable resources - this includes SYN flooding [12], turning UDP services against each other [13], using up bandwidth with ICMP ECHO packets, and email bombing [14]

- destruction or alteration of configuration information – a poor configured computer is more vulnerable and therefore it is essential that computer are properly configured and their configurations checked regularly, especially after suspicion of unauthorised access

- physical destruction or alteration of network components – physical protection of computers and their peripherals is equally important because of the high incidence of employee computer crime with direct physical access to the computer hardware.

Some denial-of-service attacks can be executed with limited resources against a large, sophisticated site. This type of attack is sometimes called an *asymmetric attack*. For example, an attacker with an old PC and a slow modem may be able to disable much faster and more sophisticated machines or networks. At the beginning of 2000 a number of high profile world wide web (WWW) servers were disabled for several hours by distributed denial-of-service attacks, perpetrated by a 15 year old [15]. This was the first demonstration of a mass attack against a large Internet portal using slave machines, whose security had been breached earlier. Two tools, Trin00 and Tribe Flood Network, are now well circulated on the Internet for carrying out such an attack and are likely to continue disrupting services in weak systems [16].

Illegitimate use of resources may also result in denial of service. For example, an intruder may use an anonymous ftp area as a place to store illegal copies of commercial software, consuming disk space and generating network traffic. The use of hacked computers to store information and pirated software looks set to increase with the blooming warez market and peer-to-peer networking programs like Napster.

Disruptive attacks are difficult to prevent because they use normal services. Programs to detect and block abnormal traffic, such as TCP wrappers, are generally the first line of defence. Development of "smart" intrusion detection systems and network monitoring software to block attacks will also become more common as the demand for protection against disruption increases, particularly for companies operating on the Internet. Management of resources and their allocation can limit the impact of these attacks.

Inevitably, there will be an increase in networked disruption attacks because of their anonymity and effectiveness against large organisations. New exploits with which to disable services will be discovered, aided by the increasing sophistication of services offered. Therefore, a determined and well-resourced attacker intent on disrupting services will remain the most common, and potentially damaging, threat to networked resources.

5.2.3.3 Active Interception

The active penetration of a network system in order to gather intelligence is the most risky but also potentially the most beneficial of all attacks. Unauthorised access to, and in some cases control of, remote systems is a powerful tool for information gathering and further attacks. Active interception requires either penetration of the target system, or *spoofing* a legitimate communication channel. We will now consider both of these attacks.

Continual unauthorised access to data and resources is the ultimate goal of any attack. Access can be gained using a number of exploits, such as software vulnerabilities, and backdoors programs like trojan horses, which will be discussed in the next section. Once access has been achieved, the attacker will cover their tracks and install methods by which unauthorised access can be maintained. Normally this involves the modification of logs to remove details of the penetration and installation of trojanised programs. In addition, software may also be installed to passively monitor network traffic and password files accessed to gain access to other accounts.

In *network spoofing* a system presents itself to the network as though it were a different system (system A impersonates system B by sending B's address instead of its own). The reason for doing this is that systems tend to operate within a group of other *trusted* systems. Trust is imparted in a one-to-one fashion; system A trusts system B (this does not imply that system B trusts system A). Implied with this trust, is that the system administrator of the trusted system is performing his job properly and maintaining an appropriate level of security for his system. Network spoofing occurs in the following manner: if system A trusts system B and system C spoofs system B, then system C can gain otherwise denied access to system A.

One of the best known spoofing attacks occurred on Christmas Day 1994 when Kevin Mitnick [17] gained access to Tsutomu Shimomura's computers through IP-spoofing [18]. He was arrested less than 2 months later after a series of intrusions to recover confidential material from Motorola and 20,000 credit card numbers from Netcom.

Spoofing can also allow email to be forged, access to restricted information and peripherals and spamming. The first line of protection against spoofing is considered to be a firewall that filters out undesired or malformed packets. Spoofing however has also changed tactics and can be used to carry out denial of service attacks and scans of systems without the originator being traced, or so called 'tunnel-and-pry' attacks [19]. More complex solutions exist including using an encrypted tokenised protocol like Kerberos for controlling remote services. There has been no known cases of a successful spoof of a Kerberos system [20], however its complexity makes installation and maintenance difficult. It is problematical to stop spoof attacks, however properly configured daemons and logging can provide adequate defence and auditing, especially in combination with a firewall.

5.2.3.4 Passive Interception

Software wiretapping can be as effective as the physical wiretapping discussed in Chapter 2 and demanding to detect. Normally a network card within a computer will only listen to traffic destined for it, but it is possible to make a card become promiscuous, listening to all network traffic. Therefore all computer attached to the network path over which data is transmitted can potentially intercept the packets.

The software used to force a card into promiscuous mode has many descriptions, the two most common of which are packet sniffers and network monitoring software. These packages are used both by network managers for optimising systems and for attackers to gather information. The current popularity of these tools is because a lot of sensitive information is transmitted in clear text form, for example in the case of telnet sessions, electronic mail interactions and remote shells.

Passive interception can be combated in a number of ways. First hardware, such as packet switching routers can be used, to prevent global traffic passing each network card. Secondly encryption makes the complexity of the attackers job several orders of magnitude more difficult. Thirdly, if access to machines is available, auditing of running processes can detect if a promiscuous program is running. If access to machines is not possible, there are techniques such as causing bursts of spurious network traffic, ignored by normal network cards, to detect any speed changes and unusual information requests from a machine intercepting traffic.

Although these defences are very effective, intelligence can still be gathered by watching the ports and traffic sent, perhaps in the preparation of an active interception or a disruptive attack. The risk posed by promiscuous software

cannot be over emphasised especially on the Internet. Ever threat group has access to, and can use this software to great effect.

5.2.3.5 Modification Attacks

Modification of system settings and software by a user can achieve a number of goals. It can provide destruction, disruption or interception of data channels. It can also be used to maintain unauthorised access to a system or to open covert data channels. Authorised users pose the greatest threat for modification attacks as generally intranet traffic is not monitored as closely as internet traffic.

One particular modification attack commonly used by programmers is a *salami attack*. The classic apocryphal story of a salami attack involves computations of interest on bank accounts. A clever programmer redirected the rounding errors of the interest payments to whole currency units so that it tallied with the bulk interest paid on the bulk balance of all the accounts. Programs which compute amounts of money are subject to these salami attacks were small amounts are shaved off, though can be found through adequate auditing and error correcting and detecting techniques. Small scale attacks persist however because corrupt programmers have many advantages such as access to the large complex programs and the ability to manipulate small amounts over many iterations. Access and knowledge will limit these attacks to small but potentially spectacular cases.

To protect against modification attacks a number of steps can be taken. Auditing of system software, ensuring programs, logs, and configuration files have not been altered can be used to detect penetration by an attacker. Auditing of in-house code is essential and part of any good systems analysis and can be used to identify salami attacks.

In the next two sections we will consider in more detail first the software which is a threat to network security and then the techniques which can be employed to strength the security of a system.

5.3 Penetration Techniques

The traditional threats to a network were considered in Chapter 2, examining susceptibility of networks to eavesdropping and disruption of hardware by skilled actors. The greater flexibility of software and the ability to mount a remote and potentially untraceable attack are increasingly attractive to cyber-criminals and widens the net of attackers to include those with minimal experience. This is compounded by the proliferation of poorly secured network sites offering rich pickings. Therefore, it is highly probable that software attacks will become increasingly prevalent over the next decade.

In order to perform the attacks described in the previous section, a number of techniques can be used which will be outlined in this section. We will consider ten of these techniques in detail:

1) **Scanning for Vulnerabilities** – In order to identify hosts which are vulnerable to attack, a range of tools are available to scan networks

2) **Software Vulnerabilities** – Vulnerabilities in the software running on a computer system can offer access unauthorised access

3) **Viruses** – The oldest technique for penetrating systems is to spread electronic viruses, which are cloaked inside other seemingly innocent programs

4) **Worms** – Worms are the networked extension of viruses, making use of computer interconnection in order to spread

5) **Trojan Horses** – Programs which have a hidden purpose other than the one for which they were designed, can provide a route for unauthorised access

6) **Trap Doors** – Undocumented features of software which weaken its effective, in particular encryption programs, can reduce the time to gather intelligence

7) **Covert Channels** –Where direct access to data is not possible, covert channels, such as uncleared computer memory, can be used

8) **Protection Crackers** – If an actor is in possession of encrypted information or protected programs, they can use cracking programs to bypass protection or encryption

9) **Social Engineering** – An alternative approach to gaining unauthorised access is to social engineer information from faults in operating procedures

10) **Myths & Hoaxes** – Disinformation can also act as a powerful technique against positive protection and procedural measuring being implemented

The most well-known software threat is from computer viruses, which traditionally were spread by floppy disc, but they have now mutated to take advantage of more powerful computer programs and the Internet with the

networking facilities it offers. The proliferation of the first eight attacks is testament to the growth of the Internet and generally poor security employed on networked machines. The last two techniques take advantage of human weaknesses and are used to illustrate that the weakest link of network can lie beyond the hardware and software used. We will now consider these ten techniques in more details and the software used.

5.3.1 Scanning for Vulnerabilities

Many different types of network scanner exist, common to all computer platforms. Some, like the well-known SATAN (Security Administrator's Tool for Analysing Networks) [21], are a dual edge tool designed for probing vulnerabilities to allow administrators to patch them, whilst also offering the same possibility to hackers to probe the network.

Network or port scanners probe remote networks for open services. They are routinely used both by network administrators to determine the effectiveness of firewalls and security and by hackers using them to identify vulnerabilities. Arguably, the most powerful scanner currently available is Nmap [22], however more specialised scanners like Queso for host identification [23] and spidermap [24] for selective scans are available.

In addition, to network scanners, there are two other specialist types of scanners: host scanners, such as COPS [25] and Tiger [26] which scan individual hosts to detect open ports and possible vulnerabilities on a local machine; and intrusion scanners, such as Nessus [27], Saint [28] and Cheops [29], which specifically probe for vulnerabilities on remote machines. Intrusion scanners can also be used to actively attack a system using one of many exploits [30] and therefore pose the most direct threat to secure computer networks.

5.3.2 Software Vulnerabilities

The CERT Co-ordination Center recorded nearly 10,000 incidents of network attacks in 1999, more than 250% up on 1998 [31]. The number of software vulnerabilities it reported also grew from 262 to 417 in 1999. It maintains a substantial archive detailing the attacks [32] and vulnerabilities [33] and posts frequent alerts and advisories to the computer community. However, the logistics of patching software and taking services off-line means that such services have a useful but limited effect on keeping software up-to-date.

Vulnerabilities generally arise from programming short cuts and the modular nature of many operating systems. Network daemons, which listen for incoming

network traffic on a computer, are normally the target, probed with unconventional traffic to discover flaws. The majority of daemons, in particular "sendmail" (which handles electronic mail) and "httpd" (which handles WWW traffic), can be exploited resulting in everything from stopping network services to unauthorised access to the filing system. A discussion of the vulnerabilities in networking protocols can be found in Chapter 4.

Practically, little can be done to remove vulnerabilities unless there is a top down approach to security. With a securely designed operating system operating limited, well-designed daemons, problems can be minimized, however in the real world this is not an option for most organisations. Instead, they rely upon generic, mass-marketed operating systems and software, which ironically are also the target of most exploits. Downtime then becomes an issue in order to patch vulnerabilities and generally, this is not done unless it is a particularly nasty exploit or there is another reason. Often the demands of business outweigh the results of risk analysis.

The breadth and depth of vulnerabilities is staggering [34]. The growth in exploits of WWW servers is particularly interesting however as the use of the Internet grows and with it e-commerce. The relative youth of these daemons and the burgeoning demands placed upon their operation, combined with the spoils for a success exploitation will lead to a growing range of exploits and patching being posted on the Internet. Indeed there is a long list of well-know WWW sites that have been defaced [35], and a well-publicized case last year of hackers gaining access to all the email accounts on the Microsoft Hotmail servers [36]. Vulnerabilities will not go away and are a major headache for all computer administrators; however proper maintenance and risk assessment can minimize the threat.

5.3.3 Password Crackers

Historically, the most common technique used to gain unauthorized system access involved password cracking. Password cracking is a technique used to surreptitiously gain system access by providing a legitimate password, gained from a successful attack on a password hash file. The two major sources of weakness in passwords are easily guessed passwords based on knowledge of the user (e.g. wife's maiden name) and passwords that are susceptible to dictionary attacks (i.e. brute-force guessing of passwords using a dictionary as the source of guesses). Programs such as Crackerjack [37], Jack the Ripper [38] and the distributed password cracker Slurpie [39], can be used for a brute force attack against a password hash file, often accessible by any user on a poorly secured computer system. To some extent protection can be offered by shadowing the

password file [40], however gaining root access through an exploit can still give access to the file and also to any account on the system.

A related issue is the bypassing of software protection measures. There are many programs and databases available on the Internet [41], such as Oscar and Astalavista, which provide lists of serial numbers, key generators and protection removal patches for commercial software. These programs promote software piracy and the trade of *warez* on the Internet, in turn encouraging system penetration to provide dumps for the illegal trade. Interestingly there are few, if any, programs that have not been pirated, despite many varied and complicated protection measures.

5.3.4 Viruses

A computer virus is a program that can "infect" other programs by modifying them in an analogous way to human viruses invading normal cells. The term virus arises because the infected program can be modified to include a copy of the virus program itself, so that the infected program then begins to act as a virus, infecting other programs.

The time to develop a virus can be surprisingly short; a virus of 200 lines of Fortran code plus 50 lines of commands was developed in less than 24 hours with little experience of the machine under attack [42]. This speed of development in combination with their small size and potential dormancy make them difficult to non-destructively detect without prior identification. The ready availability of virus constructors [43] also opens the possibility of people with little experience of programming constructing their own viruses, increasing the mutation rate and number of outbreak centres.

The term "computer virus" was first formally defined by Fred Cohen in 1983 and described main characteristics of replication, a carrier host program, activation by an external action and the replication is limited to the system. This has led to viruses primarily being developed to attack single-user, personal computer platforms, overwhelmingly the PC. Academic research has shown however that viruses are possible for multi-tasking systems, but they have not yet appeared. This point will be discussed later.

The first IBM-PC virus appeared in 1986 [44]; this was the Brain virus. Brain was a bootsector virus and remained resident in the computer memory. In 1987, Brain was followed by Alameda (Yale), Cascade, Jerusalem, Lehigh, and Miami (South African Friday the 13th). These viruses expanded the target executables to include COM and EXE files. Cascade was encrypted to deter disassembly and detection. Variable encryption appeared in 1989 with the 1260 virus. Stealth

viruses, which employ various techniques to avoid detection, also first appeared in 1989, such as Zero Bug, Dark Avenger and Frodo. In 1990, self-modifying viruses, such as Whale were introduced. In 1991 the GP1 virus was unleashed which was "network-sensitive" and attempted to steal Novell NetWare passwords. As illustrated, since their inception viruses have become increasingly complex.

Personal computer viruses exploit the lack of effective access controls in their operating systems. The viruses modify files and even the operating system itself. These are "legal" actions within the context of some operating systems, in particular Microsoft products. While more stringent controls are in place on multi-tasking, multi-user operating systems, configuration errors, and security holes (security bugs) make viruses on these systems more than theoretically possible. It has been suggested that viruses for multi-user systems are too difficult to write. However, Fred Cohen required only "8 hours of expert work" to build a virus that could penetrate a UNIX system [45]. This potential for multi-user system viruses has not been fully realised to date however, leading to a demand for a minimum population and connectivity cycle to sustain and mutate the viruses an operating system can support.

As with human viruses, the spread of a computer virus is dependent on sharing and transitivity. They exploit weaknesses in the operating systems controls and human patterns of use. Destructive viruses are more likely to be detected and eradicated quickly, while an innovative dormant virus can have a larger window to propagate in before it is discovered. To spread effectively the virus also requires a large population of homogeneous systems which an exchange of executable software. This is highlighted by less than 100 known viruses for Apple Macintoshes [46] while there are more than 10,000 for PCs [47]. Other viruses exist [48], though one area where viruses are not yet prevalent is in the WWW scripting languages, though the first signs of Java based activity are already appearing [49].

The primary problem with viruses is that evidence and facts are difficult to uncover. The computer industry, lead by virus fighters, use a two edged sword to control the marketplace: on one hand people are reluctant to report virus problems except on broad generalised terms and often attribute other problems to viruses, on the other ‘facts’ promoted by the anti-virus industry are sensationalised from by their subjective opinion, anecdotes and urban myths. To date there has been no significant independent measure of computer virus problems; indeed, there are no virus metrics in existence, in sharp contrast to the Internet. This is one area in which progress is required in order to assess the true value of anti-virus software and its effectiveness and the real risks and occurrences of networked computing problems.

5.3.4.1 Macro Viruses

Macro viruses, a more recent twist on the virus philosophy, use the features of Macro languages that are built into modern data processing systems (word processors and spreadsheets). To allow the viruses to spread in the system there has to be a built-in macro language that allows the user:

1. to assign a specific macro program(s) to specific file
2. to copy a macro program(s) from one file to other ones
3. to pass the control to some macro program(s) without user's permission (auto and standard macros).

There are three systems that meet these conditions: MS-Word, MS-Excel and AmiPro. These systems contain built-in Basic-like macro languages (Word - Word Basic, Excel - Visual Basic). These features of modern systems were designed to write "document auto-processing systems", but they also allow the viruses to spread, i.e. to infect other files.

Under Microsoft Word, Excel and AmiPro the viruses receive control while opening/closing an infected document, then they hook one or more system events (functions, macros), and infect the files that are accessed with these functions. The macro viruses are "memory resident". They hook the system events and are active not only at the moment of file opening/closing, but during all time when the system is working. The virus can also lie dominant under non-english versions of the carrier programs, but still infect files transferred providing a new twist.

Macro viruses are becoming more prevalent because of the ease with which they are written (constructors are freely available on the Internet [50]) and then can be easily attached to seemingly innocent data files. To date the most well-known macro virus has been the Melissa virus, spread using Microsoft Word and Microsoft Outlook. Reports [51] estimated about 80% of major organisations suffered infection in 1999 as the virus rapidly spread, emailing itself to the first fifty addresses in the user's Microsoft Outlook address book. The alleged creator of Melissa, David L. Smith, was very publicly hunted by the FBI, though many in the computer industry are sceptical about the hype surrounding the case. Other well-known macro-viruses include ExploreZip [52] and WM97/Marker [53].

To date none of the macro viruses in the wild have been as destructive as the machine code viruses described earlier, though potentially they can create as much damage. With the growth of more powerful data processing programs, the

power of macro languages will also expand and the distinction between viruses and worms will become more blurred.

5.3.5 Worms

Worm programs, first described by Shoch and Hupp [54], are network extensions of viruses. The worms use the network management mechanism of a computing system to identify free machines on the network and to pass the worm program on to these machines. Once active, the worm then tries to find other free machines. Some worms are benign, managing resources and sharing within large networks, or providing distributed computing resources for projects like SETI@HOME [55] and distributed.net [56]. Indeed this is how they were first conceived in 1982.

The Christmas Tree Exec was the first malicious worm, attacking IBM mainframes in December 1987. It brought down both the world-wide IBM network and BITNET. Strictly, it was a trojan horse with a replicating mechanism, sending an executable to everyone on a user's address lists. The Internet Worm [57] was a true worm; it attacked Sun and DEC UNIX systems attached to the Internet (it included two sets of binaries, one for each system). It utilized the TCP/IP protocols, common application layer protocols, operating system bugs, and a variety of system administration flaws to propagate. Various problems with worm management resulted in extremely poor system performance and a widespread denial of network service.

The Father Christmas worm was also a true worm with an additional feature of successful system penetration to a specific site. This worm made no attempt at secrecy; it was not encrypted and sent mail to every user on the system. About a month later another worm, apparently a variant of Father Christmas, was released on a private network. This variant searched for accounts with "industry standard" or "easily guessed" passwords.

Worms display the same increasing complexity found in the development of PC viruses. They exploit flaws in the operating system or inadequate system management to replicate with usually results in brief but spectacular outbreaks, shutting down entire networks.

5.3.6 Trojan Horses

In a similar way to the mythological Trojan horse, the computer Trojan horse is a program which performs a hidden function in addition to its stated, obvious function. Typically, Trojan horses are offered to other users who run them and

provide unauthorised access to their computer accounts or files without them being aware of the breech because the software functioned normally. Unlike viruses, they do not replicate independently of the host program, though some mutate into viruses once activated.

By viewing the source code of the program it is however easy to spot the inclusion of a Trojan horse or by using system monitoring software watching for unauthorised commands. Some operating systems, particular Microsoft Windows 95/98, are very susceptible to these programs however because source code is rarely distributed with the compiled binaries and it is difficult to track unusual usage. Complex Trojan horses have also been written which can lie dormant and triggered on demand, with the relevant instructions scattered through the overt program or encrypted.

Macro and fake installers are the growing trend for Trojans on the Internet [58]. For example Trojan versions of Internet Explorer for the PC and TCP Wrappers and util-linux for Linux have recently been circulated [59]. Users are fooled into thinking they are installing an authentic version of the software because in all outward ways they are installing the software as normal.

Other programs start as a virus/worm launcher inside a Trojan, for example ExploreZip, or as a hybrid of the two [60]. The most well-known Trojan and back-door to date has been the Back Orifice program, touted as a remote administration tool by its creators, the Cult of the Dead Cow [61]. It has been superseded by the smaller and more flexible BO2K [62]. As with recent macro viruses like Melissa, the hype surrounding these programs is much greater than the extent of penetration actually using them. Its potential however is staggering in comparison to viruses, essentially providing full remote control of a computer and its facilities. The publicity [63] generated by its release and the rapid response of computer community quickly identified how to identify it and deal with it effectively. More than 100,000 copies of the program have been downloaded with an easy-to-use interface making it essentially child's play to use, however it does not present a significant risk to any system where anti-virus or network monitoring software is installed and regularly updated [64].

5.3.7 Trapdoors & Backdoors

A trapdoor is a secret, undocumented entry point into a module or feature of a software program; a benign trapdoor is often referred to as an 'Easter Egg' [65]. The trapdoor is typically inserted during code development to enable fast module testing or to allow for hooking into other software modules. They are also useful for subsequent debugging and access of authorised users and for auditing systems to check data flow through the system. However, trapdoors can

also be introduced by poor error checking, for example failing to catch unacceptable input.

Historically stronger nations have normally sold cryptographic hardware to less developed nations either without informing them that the hardware was insecure. For example, the British sold Enigma-like machines to developing countries, and the products sold by Crypto AG were weakened at the request of the NSA [66]. One of the best-documented recent cases of a trapdoor is in the international version of the popular Lotus Notes [67]. The US software manufacturer installed a trapdoor to allow the National Security Agency to intercept both e-mail and conference messages from the software. This caused shock waves in many European countries, in particular Sweden where approximately 500,000 people including MPs and government officials used it, who remain suspicious of US manufactured software.

The second well-publicised case of a potential backdoor into secure communications is in the encryption library shipped as standard by Microsoft as part of Windows 95/98/2K/NT [68]. The file, ADVAPI32.DLL, contains two keys for verifying digital signatures of service providers, one named _NSAKEY, which Microsoft claim is merely an unfortunate name for a backup key and not in the possession of the National Security Agency [69]. Speculation has also surfaced about a third key present in Windows 2000 [70] and Netscape Communicator/Navigator software, further denting the reputation of US-based software companies abroad.

Other programs like Back Orifice, described in the previous subsection, and Netbus [71] are often considered as backdoor programs, offering access to a system while potentially offering other services. As networking continues to grow and users become less computer literate, the range of backdoors will grow. Backdoors in operating systems and crucial programs are more of a serious issue, but are unlikely to be addressed in the near future while the US monopoly on software holds and the security agencies are keen to gather indiscriminate information.

5.3.8 Covert Channels

Programmers working on sensitive data projects generally do not have access to the data directly, instead working with generic data. A corrupt programmer can introduce a covert channel, a hidden means for the program to communicate information, by manipulating the output to contain sensitive data. The covert channel for example can be through the least significant bit of numbers the programmer has access to, or the number of spaces or lines used in a print out. Because of the need for secrecy and to prevent detection, the amount of

information which can be transferred in this way is limited to tens of bits, however this can provide highly valuable information in an almost undetectable way.

A good example of a covert channel is described by Neal Stephenson in the classic techno-thriller *Cryptonomicon* [72]. One of the main characters, afraid of TEMPEST monitoring, uses the lights on his keyboard to display a Morse code representation of a map location. Covert channels are exceedingly difficult to account for and generally are not as significant a threat as other penetration techniques, except in the case of national security.

5.3.9 "Social engineering"

In addition to the software threats discussed so far, there are also additional threats posed by human weaknesses. People have been known to call a system operator, pretending to be some authority figure, and demand that a password be changed to allow them access or sending a forged email to get users to execute a Trojan or virus attachment. This kind of social engineering can be very effective for gaining intelligence in much the same way as dumpster diving for sensitive printed information which has been inadvertently disposed of.

There is limited protection against such attacks. Many of the larger email providers and banks on the Internet have succumbed to these relatively low-tech attacks, which can offer a very effective backdoor against strong protection from direct Internet attacks [73]. Clear codes of practice and a knowledgeable user base are important in preventing engineering of this sort and restricting access to only those that require it. Inevitably, users are the weakest point and the most vulnerable in any networked system and therefore need to be informed and protected from such risks, while steps taken to ensure any unintentional breaches are dealt with in a suitable manner.

5.3.10 Myths, Hoaxes, Spam and Fraud

Perhaps as damaging as social engineering in a more subtle way is hoaxing. Hoaxes can tie up computing resources as people send details of these fakes around the Internet, forming a virus in itself, while causing undue panic to less experienced users. Stories of these incidents are spread widely around the Internet, forming urban legends [74] amongst casual and inexperienced computer users.

There are many forms of viruses reported, which are in fact hoaxes. It is not always easy to distinguish which are real and which are hoaxes, because some

could theoretically could be used, however there are as many that are implausible from a technical perspective. A common example is that no known viruses have been found in standard interchange formats (e.g. WAV/AU, AVI/MOV, GIF/JPG).

The mass media is as guilty as Internet users in passing on alerts without any basis, surprising considering their resources at checking a story's validity. The history of hype started to some extent as viruses began to be spread on the Internet and virus scanners became widely available. Indeed, the founder of McAfee Associates, one of the main players in virus protection market, became the first doom and gloom merchant to boast his own financial ends [75].

A number of classic instances of viruses, like the ExploreZip, Remote Explorer, Hare and Michelangelo have all been over hyped and/or distorted by the media, including the technical computer press, turning out to be rather damp squids [76]. A similar effect has been observed with Y2K problems at the beginning of 2000. The recriminations have already begun as to extent of the hype and expenditure employed, given the few bugs that actually arose [77]. The primary beneficiaries of the Y2K hysteria have been the computer industry, to whom the whole process has been worth in excess of $100 billion, and who will continue to reap the rewards along with the lawyers. One of the few objective estimates reported approximately 5% of companies were affected in some way, often minor problems relating to dated computer equipment, with this figure falling to 1% after the first week of employees being back at work [78].

Government agencies and large corporations are also far from immune to hoaxes. For example, the FBI published a Law and Enforcement Bulletin, citing at least five hoax viruses (with names such as Gingrich, Clipper, Lecture and Clinton), the US Joint Chiefs of Staff issued a priority message to all military offices around the world warning of the hoax Wazzu virus, Penguin Press became embroiled in a fake virus alert which was used as a market ploy to promote a book and a host of incidents caused or engineered by the major anti-virus companies [79].

The commercial value of the Internet has also developed its own form of junk mailing, *spamming*. By using automated spiders to collect email addresses from the world-wide web, retailers sell lists of millions of email addresses to advertisers. These advertisers then use mailing programs to send out mass mailings to all these addresses, providing a significant drain on resources, none of which is paid for by the advertiser. Unsolicited email is a significant problem, in particular when fraud is involved. Inexperienced users can make life significantly more difficult for themselves both by actively replying affirmatively or negatively to spam.

Tools and techniques for these attacks are readily available on the Internet, fuelling the underground hacker culture. The power of the Internet is evident in the spread of security information and the constant battle in keeping systems updated against software threats. The anonymity and remoteness of attacks combined with the uncertainty of the law and the notoriety and standing, both socially and financially, of successful hackers fuels the demand. Potentially many of these tools can be used as both a means of security a system and of penetrating it; therefore it is difficult to imagine any sensible way of preventing their distribution.

5.4 Detection and Security

In this section, we will consider a range of software tools and techniques, which can be used to improve the security of a networked computer. Five major areas in which security can be improved will be dealt with in detail: operating systems, databases, auditing software, protection against malicious programs and network security. In the following five subsections we will consider each of these areas and the protection which can be afforded against the attacks described in the previous section.

Although there are many security tools and products discussed in this section, good management practices should form the first line of defence against the threat of attack. Software tools are only as effective as the people utilising them; poorly installed software can be more damaging than none at all because of a false sense of security. Solutions should also match the risks involved. For example, critical infrastructure machines should run minimal services in as secure an environment as possible, whereas similar security would not be appropriate in an Internet café.

5.4.1 Operating System Security

Operating systems (OSes) are the primary providers of security in computing systems, particularly for legitimate users. Operating system security can be divided into four major areas, or services, offered to a user:

1. **Memory Protection** – each user on a computer system has memory allocated for their use, which is managed by the operating system. A multi-user system may have users of many different levels using the same physical memory and therefore it is important that one user cannot access memory allocated to another, potentially higher level, user. Memory protection can also control a user's own access to programs, using

differential security like read, write and execute flags, to avoid corruption of programs and data.

2. **File Protection** – in addition, the operating system must protect user and system files from access by unauthorised users. Similarly, input/output devices, such as modems, other network cards and computer-controlled hardware must be protected.

3. **General Object Protection** – General objects, such as mechanisms to provide concurrency and allow synchronisation must be provided to users of a network. However, use of these objects must be controlled so as not to have a negative effect on other users. For example it is important to limit network bandwidth through software so one user cannot monopolise a data connection.

4. **Access Authentication** – access to the operating system must only be given to legitimate users who have identified themselves. The most common authentication mechanism is password comparison with a stored hash function. This is analogous to a castle wall that can be fortified in many different ways and can successfully prevent attack from outside, however has limited protection against internal deception and illegal activity.

Many models exist to describe security properties of computing systems and users, with access control being the basis of most of these models. These generic models can then be adapted to a specific access policy defined by system managers. Models exist from simple binary options on whether an object is sensitive or not and whether a user is authorised or not, through to the complex Bell-LaPadula [80] and Biba Models [81]. These two models form the basis of the U.S. Department of Defense Trusted Computer System Evaluation Standard (Orange Book), described in more detail later. Other more abstract models, such as the Graham-Denning and Take-Grant systems, deal with the limits and properties of the security implemented and are employed in risk-analysis of sensitive computer networks.

Security functions pervade the design and structure of an operating system at every level and therefore need to be considered at every level of design and tested to ensure security is enforced or provided. Ideally, the security should be of open design and easy to use, while providing least privilege and least common mechanisms with economy and multi-authorisation schemes. Separation or isolation of processes can be achieved in a number of ways, including physical separation of hardware facilities, logical separation using a software monitor program, temporal separation of when processing occurs (e.g.

sensitive data in the afternoon, and general data jobs scheduled for the morning) and cryptographic separation so unauthorised users cannot access data in a readable form.

The security of an operating system is limited primarily by I/O processing and often flaws are introduced in the rigorous, intensive coding for I/O channels. Short cuts used for speed can weaken other protection features and channels can bypass some normal security features such as page or segment translation. The generality of off-the-self operating systems, ambiguity of access policies for users and incomplete mediation between different components of the systems are the other main vulnerability points of operating systems.

Validation and more formal verification of secure operating systems can be carried out in a number of ways. Verification is a time-consuming and complex activity, requiring analysis of the logic used at all points in the operating system; this is not always possible for an operating system not designed with logic in mind. *Verification* uses less rigorous techniques, such as requirement checking, design and code reviews and module and system testing to check, correct and confirm. Penetration testing can then be carried out against the operating system by a team of skilled independent experts and flaws identified. Although empirical in comparison to formal verification, validation can offer a sufficient degree of confidence for all but the most secure systems.

The U.S. Department of Defense has identified six requirements for secure computers systems, which form the assurance classification levels for the Orange Book:

1. **Security Policy** – there must be an explicit and well-defined security policy enforced by the system.
2. **Identification** – every object must be uniquely and convincingly identified.
3. **Marking** – every object must be marked with a security label
4. **Accountability** – the system must maintain complete, secure records of actions that affect security.
5. **Assurance** – the security mechanisms of the system must be evaluated for their effectiveness.
6. **Continuous Protection** – the mechanisms which implement security must be protected against unauthorised change.

Most common operating systems fall in the Class D (minimal protection – e.g. MS-DOS), Class C1 (discretionary security protection – e.g. Windows 95/98) and Class C2 (controlled access protection - e.g. VAX/VMS, UNIX). Interestingly Windows NT was classified as C2 TCB (trusted computer base) on a

stand-alone basis (i.e. no longer valid when connected to a network) in a bid to secure contracts from the US military. A useful description of the steps taken to reach class C2 is given in the case of Novell's bid for C2 status with version 4.22 of its networking software [82]. The three subdivisions of Class B (labelled security protection, structured protection & security domains) offer increasing validation with Class A1 representing a formally verified operating system, of which only there is only a handful, including the Honeywell/Bull SCOMP system.

In addition to the Orange Book in the Rainbow Series [83], the Red Book (Trusted Network Interpretation), Yellow Book (Methodology for Security Risk Assessment and Lavender Book (Database Security Evaluation) are relevant to system security. In particular, the Yellow Book provides quantifiable risk analysis techniques, useful for determining the level of security required. Critical services should only be run on Class A or B hosts with limited user access and service provision depending on resources. Important resources should operate in a Class B or C system, for example a firewall or DNS server, which are important for operation and / or security. User workstations should at least be Class C1 if not C2, especially when access to more sensitive network resources is possible. The difference between the maximum and minimum security level of information stored versus the difference between the maximum and minimum security of users is a useful indicator in system design. The larger these two numbers and the larger the difference between them, then the more secure a system needs to be.

It is worth noting that these levels of security can be implemented both in software and hardware. In the last decade there has been many widely reported news stories concerning flaws in Intel processors, for example the introduction of personal serial numbers in Pentium III chips [84]. The flexibility of hardware is also its weakness, introducing the possibility of corruption during an attack. Microcode can be uploaded to the processor, or EEPROMs (e.g. BIOS) rewritten in software, causing hardware to perform in unexpected ways [85]. Viruses already exist which can trash BIOSes [86], but of wider concern are the exact details of the processes within a processor and the function of updates; do backdoors exist within processors for intelligence agency use? A more detailed discussion of the security and risk in operating systems can be found here [87].

None of the commercially available hardware or operating systems used currently is specifically designed from a security perspective. Products such as the Viper processor developed by the UK's RSRE [88] and the SCOMP operating system [89] are verifiably secure, however all others carry a degree or risk and this has to be quantified in a risk assessment scheme.

5.4.2 Database Security

The majority of modern organisations now operate substantial databases, storing information on everything from clients to products and salary details. Security of these databases is an area of increasing interest in computer security because of the boom in e-commerce, making the information contained in the databases extremely valuable and more critical to business confidence. Databases, the most common of which is the relational database, are controlled by a front-end or a database management system (DBMS). This front-end controls shared access to a centralised set of data for multiple users with minimal redundancy while preserving consistency and integrity.

The requirements for databases are similar to those described for operating systems: to address the basic problems of authentication, access control, reliability and exclude spurious data. Integrity and reliability are supported by protection features such as audit trails of changes to the database, phased updating, consistency checking and correcting and the use of constraints.

Securing sensitive data is however more problematic due to the range and nature of possible data which can be stored. Within a database, a field or record may contain sensitive data, or the uniqueness of a record requires it to be treated more sensitively, for example in the case of medical databases. Data can be disclosed in a number of ways, often inferred. If the exact data cannot be accessed, then bounds can be placed on the content, for example by simply looking at the length of the entry, or comparing it to other database entries. If an attacker can make searches but cannot access the data directly, then they can use logical probability attacks to infer values by querying the database correctly, or by using external data to clarify generalised results. Unfortunately, there are no perfect solutions to these attacks. However they can be minimized by suppressing obviously sensitive information, tracking what the user has requested, and disguising the data where necessary.

For large multi-user, multi-purpose databases, differential security adds further complexity. This is highlighted because to date no trusted DBMS have been developed however there are guidelines and proposed standards. These standards include the encryption of data, portioning of data, integrity and sensitivity locks, with query filtering and efficient analysis.

There is now an annual conference dedicated to database security issues [90], concentrating on the access to databases from web servers on the Internet. This is a growing area of research [91], as more services become available through the Internet, for example medical information services [92], which use data warehousing and data mining.

There are many key issues of knowledge discovery versus personal privacy [93] in large databases. These issues cause concern for the general public with over 70% of ordinary US citizens not willing to shop on-line for fear of identity theft and wary of false credit checking [94]. On the other hand, such databases have been useful for detecting fraud [95] and network intrusion [96]. As Internet services begin to play a greater role in daily life, the issues of security and privacy within databases will become more important and of more public concern.

5.4.3 Anti-Virus Security

Currently the most common security programs are the ones that detect and prevent computer viruses. Primarily this is for historic and user-base reasons, with computer viruses common on the dominant PC platform before the Internet was realised at a home user level. With the blossoming of virus-like programs, like Trojan horses and worms, the software has been adapted to deal with these type infections. With the growing connectivity to the Internet of home-users, it is generally believed that they will eventually morph into all-round security programs protection against network attacks as well as infection.

There are three classes of anti-virus products currently available: detection tools, identification tools, and removal tools. Scanners, an example of both a detection and identification tool, search for "signature strings" or use algorithmic detection methods to identify known viruses. Disinfectors, removal tools, rely on substantial information regarding the size of a virus and the type of modifications to restore the infected file's contents. Vulnerability monitors, detection tools, attempt to prevent modification or access to particularly sensitive parts of the system, may block a virus from hooking into sensitive software interrupts. This requires a lot of information about "normal" system use, since personal computer viruses do not usually circumvent any security features. This type of software also requires decisions from the user.

Modification detection, also a detection tool, is a very general method, and requires no information about the virus to detect its presence. Modification detection programs, which are usually checksum based, are used to detect virus infection or Trojan horses. This process begins with the creation of a baseline, where checksums for clean executables are computed and saved. Each following iteration consists of checksum computation and comparison with the stored value. It should be noted that simple checksums are easy to defeat; cyclical redundancy checks (CRC) are better, but can still be defeated; cryptographic checksums provide the highest level of security.

The major players in the anti-virus software market are Symantec [97], McAfee [98], Data Fellows [99] and Dr.Solomon's [100] for PCs, similar products for MacOS and McAfee for Unix-like operating systems [101]. Currently there is little evidence of viruses for Unix-like operating systems, however their dominance of the Internet server market and the potential value of penetration is likely to cause a surge of underground interest, while for the PC platform network worms are likely to be an increasing threat, highlighted for example by the recent 'love bug' macro virus [102].

The lack of hard quantitative evidence and the growing false and exaggerated reporting by the media of computer security make metrics of anti-virus security difficult. There is however widespread acceptance of their benefit and they are widely implemented, often as the only security defence. With a twenty year history, this is a relatively mature field of computer security and unlikely to change drastically in the next decade, instead reacting to new forms of attack and incorporating other security programs to offer a more holistic service.

5.4.4 Auditing Software

Auditing software, for checking computer systems, is beginning to be more widely used for generic checking of common security loopholes. Tracking of abuse is also important in determining future security requirements and for law enforcement, as well as providing more quantitative data to work with.

The main division of auditing software is between local and remote analysis, with some packages offering both services. Local auditing software can be split into three basic sub-types: configuration checkers, file-integrity checking, and logging management. Configuration checking software, like COPS and Tiger mentioned earlier, check.pl [103] and titan [104], examine the configuration of the local machine: checking password security, file and remote access tables, port configurations, shares, services, access to hardware etc.

Integrity-based packages, which check for changes to critical files and parts of the filing system / operating system, such as Tripwire [105], L5 [106], ViperDB [107], and audit [108], also can provide detection and protection against Trojans, file changes and changes in permission levels and add an extra layer to security. Variations are also available, for example installwatch [109], examine changes made during installing new software on a machine, and cpm [110], which examines the status of the network card.

Many computers generate substantial log files, detailing everything from system information to critical errors, which can be filtered by logging management software for the telltale patterns of attack. Some loggers look at specific log

files, such as flog [111] for analysing ftp logs, checksyslog [112] for analysis the syslog files, and others such as sportal [113] which watch hot files and send alerts when particular words appear in those files. Replacements for the standard logging daemons are also available with added features such as cryptographic storage [114] and direct filtering of the messages to give immediate detection [115].

Remote auditing software is similar in nature to IDS and firewall tools and therefore many of them can serve a dual purpose. Security scanners [116], such as Nessus, SATAN, SAINT, SARA and ISS are common programs used on Unix-like platforms. Network traffic monitoring software, for example based on the sniffer programs described earlier, also form an important auditing tool. Commonly available traffic monitors include Argus [117], IP flowmeter [118] which can also be used for transaction billing, Arpwatch [119] which monitors IP/Ethernet address and flags any mismatches, and packet logging programs such as iplog [120] which logs TCP, UDP and ICMP traffic and detect scans and protect against floods and smurf attacks.

There are many other more specific auditing tools available: for example, firewalk [121] can be used to audit filter rules for a firewall or other packet filtering devices, Buffer Syringe [122] which checks for buffer overflow exploits in server daemons, and zodiac [123] which can be used to test and monitor DNS servers against spoofing, denial of service attacks and floods.

In addition to the software described previously, another growing subset of auditing tools is becoming available, particularly for Unix-like operating systems. Source code analysers can scan through uncompiled source code and determine poor programming, which may lead to exploits. Currently such programs are used in the black hat community for discovering exploits, however they are becoming commercial available, such as SLINT [124] and ITS4 [125]. Looking for generically weak structures, they can look for bad memory allocation, buffer overflows, improper allocations and other potential problems, including Trojan code.

Auditing software provides a crucially important second row defence against attack. It can be used proactively to detect vulnerabilities, actively to detect intrusion attempts and reactively to respond swiftly and decisively to any successful penetrations. Although not as glamorous as the front-end security products like firewalls and anti-virus software, auditing tools will become more widely available over the next decade as organisations realise the added value they give.

5.4.5 Network Security

With the increasing connectivity of computers, both at home and in the work place, the major growth area in computer security programs are packages that protect against disruptive network traffic. Protecting a system against network attacks requires a combination of basic system security and good network security. There are a variety of procedures and tools that can be applied to protect a system against remote attacks; five of the main groupings will be described in the following subsections, which give a flavour of the current trends. Good overviews of network security can be obtained from several sites on the Internet, with detailed descriptions of tools and exploits [126].

5.4.5.1 System Management

In basic system security, the most important means of defence against network attacks is the identification and authentication controls, which are usually integrated into the system. If poorly managed, these controls become a vulnerability, which is easily exploited. Resources are available to system managers to keep them abreast of security bugs and bugfixes, such as the CERT computer security advisories [127], to help prevent vulnerabilities being exploited. Many of the network security programs monitor the activity of these two controls both locally and on local networks. The difficultly however is achieving a high true-positive response compared to false alerts.

In general, services and protocols should be disabled that are not required, password and access control mechanisms checked for weaknesses, software regularly updated and basic security programs installed, such as anti-virus suites. Regular checks should be made on all machines, as an attack will often go for the weakest host to get a foothold before attacking a more secure host. Specific details of these steps and their advantages are discussed more thoroughly elsewhere in this, and other chapters.

5.4.5.2 Network Monitoring

The outermost ring of defence on a network, equivalent to watching the surrounding countryside from a castle, is network monitoring. This incorporates both passive and active monitoring of the network for any suspicious activity.

Passive monitoring is normally described as sniffing and is a powerful tool for both authorised and unauthorised users. Normally a network card will only process information destined for its address, however it is possible to put most network cards into a promiscuous mode where they will process all network packets, effectively eavesdropping to all information on the network. A protocol

analyser, or sniffer, has the capability of capturing every packet on a network and of decoding all seven layers of the OSI protocol model, making it very powerful. These programs are a very significant threat because they are very difficult to detect due to their passive nature.

In authorised hands, they can be used to monitor and audit traffic flow and detect/prevent unauthorised packets, both in real-time and in hindsight, with the potential for feedback to prevent bottle-necking. Sniffers are also used for diagnostics and debugging of network, analysing protocols and generating test traffic.

In unauthorised hands, sniffers can be used to intercept everything from e-mail, to WWW traffic and most worryingly terminal sessions on remote machines. As most of this traffic is in clear text, logins, passwords and personal details such as credit cards, can easily be extracted from the traffic for future use.

Sniffers for monitoring networks are readily available both commercially and as freeware. For example, Sniffer Pro 98 LAN for Windows 2000/NT [128] can cope with up to 100Mbps and the Distributed Sniffer System [129] can manage multiple subnets. Freeware sniffers, like Sniffit [130] and Snort [131], have less front-side features, though are generally more versatile and offer the possibility of greater customisation.

Recently there has been a growing interest in being able to determine whether a packet sniffer is running on a local network. The first sniffer-detection program released, causing real interest, was AntiSniff [132] by LOPHT Heavy Industries, a hacker group that recently joined forces with @Stake to provide computer security services. Anti-sniff is a proactive program using a number of techniques to determine whether a sniffing program is running; sniffers normally run DNS checks to resolve the names of where packets are going/coming from on the network and by introducing false ones and watch requests to DNS servers it is possible to detect a sniffer. Similarly, the load of a machine running a sniffer is affected by the amount of traffic. By flooding the network with useless packets that other machines will ignore and watching the ping response of suspect machines, it is also possible to infer whether the suspicious computer is running in promiscuous mode. Since its release, AntiSniff has been joined by other sniffer detection programs including sentinel [133] and bogon [134]. Although such programs cannot detect a sniffer looking for particular data from a known source, it can detect indiscriminate sniffers.

Many active network-monitoring tools exist. These come in different flavours according to application. One subset actively monitors the status of other machines on the network to detect any changes (for example Big Brother [135]). Others can scan for services and vulnerabilities: XpsScanner [136] scans for

CGI vulnerabilities, NSS2000 [137] and VeteScan [138] scan multiplatforms for hundreds of remote vulnerabilities, Nmap (described earlier) and Cheops the network "swiss army knife" [139]. Several lists of network scanning software are readily available on the Internet [140]. The vulnerability scanners are increasingly popular and provide a useful proactive technique for detecting possible intrusion points.

5.4.5.3 Firewalls

A firewall [141] is one of a number of ways of protecting a trusted network from another untrusted network. The actual mechanism whereby this is accomplished varies widely, but in principle, the firewall can be thought of as a pair of mechanisms: one which exists to block traffic, and the other which exists to permit traffic. Some firewalls place a greater emphasis on blocking traffic, while others emphasize permitting traffic. Firewall systems are found in two forms: simple or intelligent. An intelligent firewall filters all connections between hosts on the organizational network and the world-at-large. A simple firewall disallows all connections with the outside world, essentially splitting the network into two different networks. A more detailed discussion of firewall operation can be found in Chapter 4.

The policies which govern the flow of information through a firewall must be realistic and reflect security policies elsewhere. The three important criteria in determining firewall policies are: the level of security paranoia required, the level of monitoring, redundancy and control required, and finally the cost. The resources required to properly maintain a firewall can also play an important role.

A proxy service can also be used to act as a firewall. Generally users have direct access to the Internet, with requests going directly to remote machines. However as commercial and security pressures grow, many organisations are beginning to use proxy programs for WWW and ftp services. These programs act as transparent intermediaries, caching commonly accessed web pages, but more importantly filtering requests and responses if deemed inappropriate by the proxy software rules.

A properly configured router can act sufficiently well as a firewall for low-end security networks, with no additional costs though may require support time to upgrade the rules periodically. At the other extreme, a complete commercial firewall product may cost $100,000 upwards, not including maintenance costs. Many commercial packages are available [142], including IBM SecureWay [143], AltaVista Firewall [144], McAfee Personal Firewall [145], CiscoWorks [146] and Netra Server [147] to name but a few. Similarly, there is a wide range

of proxy software available [148], including Netscape Proxy Server [149] and Squid [150]. In addition, there are a wide range of tools for testing and monitoring firewalls [151]. Interestingly firewalls are increasingly being used as "ambassadors" for an organisation, hosting all its external services, while protecting internal systems from outside demands, a trend likely to continue as more businesses become concerned about the security of their computer systems and the productivity of employees.

5.4.5.4 Wrappers

Another type of network security tool is the wrapper program. Wrapper programs can be used to "filter" network connections, rejecting or allowing certain types of connections (or connections from a pre-determined set of systems). This can prevent worm infections by "untrusted" systems. Overlaps in trust may still allow infection to occur (A trusts B but not C; B trusts C; C infects B which infects A) but the rate of propagation will be limited.

Many operating systems now ship with wrappers included, though it is left to the system administrator to decide how these wrappers are configured. The freeware TCP_Wrappers suite [152] is the most common, filtering SYSTAT, FINGER, FTP, TELNET, RLOGIN, RSH, EXEC, TFTP, TALK, and other TCP network protocols and logging any unusual behaviour. They cannot however protect Portmap and RPC requests and UDP daemons effectively, though these services can be protected in other ways.

5.4.5.5 Intrusion Detection Software (IDS)

The division between auditing and intrusion detection is generally small, with many programs performing both tasks. As with auditing, there are both IDS programs to detect intrusion on a local machine and remote network monitoring of possible intrusions. Local tools include check-ps [153] that detects rogue programs running, bgcheck [154] for checking and limiting background processes.

Common network tools for IDS include the SHADOW TIDIS package [155] developed by the SANS Institute, Network Flight Recorder [156] an off-shoot of the same project, the Abacus Project's PortSentry and HostSentry [157], and Big Brother for monitoring multiple machines [135], with lists of others available on the Internet [158]. More recently, a range of IDS programs using a non-promiscuous mode and incorporating firewalls have become available aimed specifically at personal computers, including BlackICE Defender [159], CyberCop [160] and RealSecure Micro-Agent [161]. This appears to be a

growing market with the number of home users connecting to the Internet and using networked services.

These programs detect intrusion either by anomaly detection (unusual network or computer usage statistics) or more commonly by signature recognition (patterns of attacks). The former is better in that no knowledge of the mechanics of the attack is required, however it is difficult to obtain good true:positive statistics. Signature matching occurs at both the protocol and application layers and by combining this with logging all the events in these layers, a database of normal use can be developed. Using both detection methods providers stronger approach, similar to that used in virus scanning. Indeed the similarity will probably help unite both fields in providing all-round computer security products. A more technical discussion of IDS systems can be found in Chapter 4.

One area of debate within the IDS community is the value of honey pots [162]. These are machines or ports on machines, which are used to lure hackers by seeming insecure or of vital importance. Several programs act as honey pots for various popular backdoor programs, such as GBS for Netbus connections [163] and Back Officer [164] for Bark Orifice and decfingerd [165] for replacing the normal finger daemon, to log scans for these programs. A more encompassing Deception Toolkit is also available [166], which can be used to replace all known port services with deceptive responses. These tools can be useful and have provided insight into the hacker community, however many of them do not appreciate being duped.

5.5 Conclusions and Future Prospects

In this chapter, we have considered the threats and risks posed by remote attacks conducted by software. In the second section, we considered the motivations and goals of an attacker and the threat they would typically pose to an organisation. A proper risk assessment is essential for any organisation to understand threats and target resources accordingly, to minimize disruption.

The third section described ten techniques an intruder could use to gain unauthorised access to a remote computer. Eight of these techniques made use of vulnerabilities in software and protocols, while the last two made use of human vulnerabilities. The wide spread and successful use of these techniques across the world is an indication of generally poor security policies implemented by organisations, in contrast to their expenditure and reliance on computer resources.

Responses to these penetration techniques where discussed in the fourth section. Five major areas were security can be improved were identified, with examples of products and procedures which can be used. Management of hardware and personnel is as important as utilising detection and prevention software in minimizing the threat posed by attacks. Firewalls and IDS systems are both important tools in the fight against remote attackers, however before they can be properly effective education of all computer users and administrators must be the highest priority.

References and Further Reading:

Joel Scambray, Stuart McClure, George Kurtz, *Hacking Exposed: Network Security Secrets and Solutions*, McGraw-Hill Professional Publishing (2000)

Laura E. Quarantiello, Cyber Crime: How to Protect Yourself from Computer Criminals, Tiare Publications (1996)

Stephen Northcutt, *Network Intrusion Detection: An Analysts' Handbook*, New Riders Publishing (1999)

Keith Brown, *Programming Windows Security* (The DevelopMentor Series), Addison-Wesley Pub Co. (2000)

Scott Mann, Ellen L. Mitchell, *Linux System Security: The Administrator's Guide to Open Source Security Tools*, Prentice Hall (1999)

Maximum Linux Security: A Hacker's Guide to Protecting Your Linux Server and Workstation, Sams (1999)

Robert L. Ziegler, *Linux Firewalls*, New Riders Publishing (1999)

Gerald Kovacich, *The Information Systems Security Officer's Guide: Establishing and Managing an Information Protection Program*, Butterworth-Heinemann (1998)

Arthur E. Hutt (Editor), Seymour Bosworth (Editor), Douglas B. Hoyt (Editor), *Computer Security Handbook*, John Wiley & Sons (1995)

Deborah Russell, G. T. Gangemi, *Computer Security Basics*, O'Reilly & Associates (1991)

[1] Toronto Financial Post, 15th December 1994

[2] *Ninety percent of survey respondents detect cyber attacks, 273 organizations report $265,589,940 in financial losses*, Computer Security Institute Press Release (22nd March 2000)

http://www.gocsi.com/prelea_000321.htm

[3] Ron Newman, *The Church of Scientology vs The Net*

http://www2.thecia.net/users/rnewman/scientology/home.html

[4] Internet Security Tools

http://kepler.informatik.uni-oldenburg.de/lehre/SicherheitVS/Denial/security-tools.htm

[5] Military and C4I Homepage, Infowar.com

http://www.infowar.com/mil_c4i/mil_c4i.shtml

[6] Gulshan Rai, R.K. Dubash, A.K. Chakravarti, *Computer Related Crimes*, Government of India

http://www.mit.gov.in/crime.htm

[7] Bob Cromwell, *Terrorists, Freedom Fighters, Crusaders, Propagandists, and Military*

Professionals on the Net

http://RVL4.ecn.purdue.edu/~cromwell/lt/netusers.html

[8] Class III Infowar, Infowar.com

http://www.infowar.com/class_3/class_3.shtml

[9] Texas vs Burleson, No. 2-88-301-CR, Court of Appeals of Texas, Second District, Fort Worth

http://rampages.onramp.net/~dgmccown/c-txblsn.htm

[10] Hacking and Computer Knowledge Website

http://www.hack.org/

[11] *Computer Crimes*, Davis McCown, Attorney at Law, Hurst, TX

http://www.davismccownlaw.com/articles/ix-crime.htm

[12] CERT® Advisory CA-1996-21 TCP SYN Flooding and IP Spoofing Attacks

http://www.cert.org/advisories/CA-1996-21.html

[13] CERT® Advisory CA-1996-01 UDP Port Denial-of-Service Attack

http://www.cert.org/advisories/CA-1996-01.html

[14] *Email Bombing and Spamming*, CERT Coordination Center

http://www.cert.org/tech_tips/email_bombing_spamming.html

[15] Robert Melnbardis, *Canada arrests 15-year-old "Mafiaboy" hacker*, Yahoo News (19th April 2000)

http://uk.news.yahoo.com/000419/91/a46ir.html

[16] *Denial of Service Attacks*, CERT Coordination Center

http://www.cert.org/tech_tips/denial_of_service.html

[17] Free Kevin Mitnick – The Official Kevin Mitnick Site

http://www.kevinmitnick.com/

[18] Tsutomu Shimomura, John Markoff, *Takedown: The Pursuit and Capture of Kevin Mitnick, America's Most Wanted Computer Outlaw-By the Man Who Did It*, Hyperion Books (1996)

http://www.takedown.com/

[19] Will Knight, *Method could be taken up by evil hackers, warns telco monolith*, ZDNet News (5th Nov 1999)

http://www.zdnet.co.uk/news/1999/44/ns-11244.html

[20] Kerberos Authentication System

http://uk.dir.yahoo.com/Computers_and_Internet/Security_and_Encryption/Kerberos/

[21] Muffy Barkocy, *Security Administrator's Tool for Analyzing Networks (SATAN)*

http://www.fish.com/~zen/satan/satan.html

[22] Nmap – Stealth Port Scanner for Network Security Auditing

http://www.insecure.org/nmap/index.html
[23] Portscanner Homepage
http://www.ameth.org/~veilleux/portscan.html
[24] Spidermap Package
http://www.secureaustin.com/spidermap/
[25] COPS (Computer Oracle and Password System) Package
http://packetstorm.securify.com/UNIX/audit/cops/
[26] TAAMU Security Tools - Tiger Package
http://net.tamu.edu/network/tools/tiger.html
[27] Nessus Package
http://www.nessus.org/
[28] Security Administrator's Integrated Network Tool (SAINT) Package
http://www.wwdsi.com/saint/
[29] Cheops Network User Interface Package
http://www.marko.net/cheops/
[30] Rootshell Homepage
http://www.rootshell.com/
[31] CERT/CC Statistics 1988-2000
http://www.cert.org/stats/cert_stats.html
[32] CERT/CC Incident Notes
http://www.cert.org/incident_notes/
[33] CERT/CC Vulnerability Notes
http://www.cert.org/vul_notes/
[34] Matt Bishop, UC Davis Vulnerabilities Project
http://seclab.cs.ucdavis.edu/projects/vulnerabilities/
[35] Archives of Hacked Websites
http://www.onething.com/archive/
http://www.attrition.org/mirror/
[36] Terho Uimonen, *Who's Reading Your Hotmail?*, IDG News Service (30th Aug 1999)
http://www.pcworld.com/pcwtoday/article/0,1510,12549,00.html
[37] Crackerjack Package
http://tms.netrom.com/~cassidy/utils/jack14.zip
[38] Jack the Ripper Package
http://tms.netrom.com/~cassidy/utils/13a3dos.zip
[39] Slurpie Package

http://www.jps.net/coati/archives/slurpie.html

[40] Shadow Password File Utilities, freshmeat.org

http://freshmeat.net/appindex/1998/02/15/887562187.html

[41] Oscar Serials Package

http://www.gulli.com/oscar.shtml

[42] Fred Cohen, *Computer Viruses - Theory and Experiments* (1984)

http://www.all.net/books/virus/part5.html

[43] AVPVE: Virus Constructors

http://www.metro.ch/avpve/constr.stm

[44] Peter Denning, *Computers Under Attack: Intruders, Worms, and Viruses*, ACM Press, (1990)

[45] Lance Hoffman, *Rogue Programs: Viruses, Worms, and Trojan Horses*, Van Norstrand Reinhold, (1990)

[46] Apple Macintosh Virus Information

http://macos.about.com/compute/macos/cs/virusesvirusutil/index.htm

http://madokan.fortunecity.com/mvic/virusmac.html

[47] F-Secure Virus Information Center, DataFellows

http://www.datafellows.com/virus-info/

[48] Sophos Virus Info

http://www.sophos.com/virusinfo/

[49] Java Viruses Primer

http://www.sophos.com/virusinfo/scares/javaviruses.html

http://www.knowledgestor.com/info/com.g2news_csn_146_14.html

[50] Macro Viruses

http://www.metro.ch/avpve/macro.stm#constructors

[51] Kathleen Ohlson, Melissa: The day after, Computerworld News (30th Mar 1999)

http://www.computerworld.com/cwi/story/0,1199,NAV47_STO27605,00.html

[52] W32/Exlorezip Macro Virus, Sophos Virus Info

http://www.sophos.com/virusinfo/analyses/explorezip.html

[53] WM97/Marker Macro Virus, Sophos Virus Info

http://www.sophos.com/virusinfo/analyses/wm97marker.html

[54] J. Shock and J. Hupp, *The Worm Programs – Early Experience with a Distributed Computer*, Comm. ACM, 25:3, 172 (1982)

[55] SETI@HOME Homepage

http://setiathome.ssl.berkeley.edu/

[56] Distributed.net Homepage

http://www.distributed.net/

[57] Eugene Spafford, *The internet worm program: An analysis*, Computer Communication Review, 19(1) (January 1989)

http://www.software.com.pl/newarchive/misc/Worm/darbyt/pages/worm.html

[58] Trojan Packages

http://www.avx.ro/avxvb.php?srcname=&sortby=1&srctype=3

[59] CERT Advisory CA-1999-02 Trojan Horses, CERT Coordination Center

http://www.cert.org/advisories/CA-1999-02.html

[60] CERT® Advisory CA-1999-06 ExploreZip Trojan Horse Program, CERT Coordination Center

http://www.cert.org/advisories/CA-1999-06.html

[61] Back Orifice, Cult of the Dead Cow's Tools

http://www.cultdeadcow.com/tools/

[62] Back Orifice 2000 Website

http://www.bo2k.com/

[63] Ira Sager, *Now Any Hack Can Be A Hacker*, Business Week (31st Aug 1998)

http://www.businessweek.com/1998/35/b3593100.htm

[64] CERT Vulnerability Note VN-98.07 (Back Orifice), CERT Coordination Center

http://www.cert.org/vul_notes/VN-98.07.backorifice.html

[65] Easter Eggs Primer

http://uk.dir.yahoo.com/Computers_and_Internet/Information_and_Documentation/Easter_Eggs/

[66] Wayne Madsen, *Crypto AG: The NSA's Trojan Whore?*, Covert Action Quarterly 63

http://mediafilter.autono.net/CAQ/caq63/caq63madsen.html

[67] Duncan Campbell, *Only NSA can listen, so that's OK*, Telepolis News (1st June 1999)

http://www.heise.de/tp/english/inhalt/te/2898/1.html

[68] Microsoft and the NSAKEY Frequently Asked Questions

http://www.cryptonym.com/hottopics/msft-nsa/faq.html

[69] *There is no "Back Door" in Windows*, Microsoft TechNet Security (3rd Sept 1999)

http://www.microsoft.com/TechNet/security/backdoor.asp

[70] *Security Expert Says Microsoft Placed NSA Backdoor In Windows*, HackWatch News (3rd Sept 1999)

http://hackwatch.ie/~kooltek/nsabackdoor.html

[71] NetBus Package

http://www.netbus.org/

[72] Neal Stephenson, *Cryptonomicon*, Harperperennial Library (2000)

http://www.cryptonomicon.com/

[73] *Low-tech hacking a big problem*, ZDNet UK News (3rd Aug 1998)

http://www.zdnet.co.uk/news/1998/30/ns-5153.html

[74] Urban Legends Primer

http://uk.dir.yahoo.com/Society_and_Culture/Mythology_and_Folklore/Urban_Legends/

[75] Rob Rosenberger, *Bizarre quotes from the experts*

http://kumite.com/myths/opinion/wildquot.htm#mcafee

[76] Truth About Computer Virus Myths & Hoaxes

http://www.Vmyths.com/

[77] Ben Berkowitz, Y2K Media Watch: Y2K Post-Mortem, USC Annenberg School for Communication (6th Jan 2000)

http://ojr.usc.edu/content/y2k.cfm?request=310

[78] ZDNet UK News Y2K News Special

http://www.zdnet.co.uk/news/specials/1999/03/y2k/

[79] Rob Rosenberger, *John McAfee Awards for Computer Virus Hysteria 1997*

http://kumite.com/myths/cvha/1997/

[80] D.Bell, L.LaPadula, *Secure Computer Systems: Mathematical Foundations and Model.* MITRE Report MTR 2547, 2 (1973)

[81] K.Biba, *Integrity Considerations for Secure Computer Systems*, US Air Force Electronic Systems Division (1977)

[82] Linda Boyer, *NetWare 4: The Climb to C2*, NetWare Connection, Nov./Dec. 1995, pp. 6-14

http://www.nwconnection.com/nov-dec.95/nw4clin5/

[83] Rainbow Series, NCSC

http://www.radium.ncsc.mil/tpep/library/rainbow/

[84] Intel Pentium III processor security cracked, securiteam.com

http://www.securiteam.com/securitynews/Intel_Pentium_III_processor_security_cracked.html

[85] Upgrade Your Pentium's Microcode, Slashdot

http://slashdot.org/articles/00/10/27/126258.shtml

[86] Global CIH Virus Information Center, F-Secure Website

http://www.F-Secure.com/cih/

[87] Charles Pfleeger, *Security in Computing*, Prentice Hall International Editions (1989)

[88] The Risk Digest, Vol. 9, Issue 1, ACM Committee on Computers and Public Policy (July 1989)

http://catless.ncl.ac.uk/Risks/9.01.html#subj2

[89] Products evaluated by NSA's Trusted Product and Network Security Evaluations Division
http://www.cs.uah.edu/~thinke/epl.html
[90] Fourteenth Annual IFIP WG 11.3 Working Conference on Database Security
http://www.cs.vu.nl/ifip-2000/
[91] Rick Noel, *Security Issues related to Database Access from the Web*
http://www.cs.rpi.edu/~noel/Security/index.html
[92] Security of Medical Information Services, Ross Anderson's Homepage
http://www.cl.cam.ac.uk/users/rja14/#Med
[93] *Knowledge Discovery in Databases vs. Personal Privacy*, IEEE Expert Symposium (Apr 1995)
http://www.kdnuggets.com/gpspubs/ieee-expert-9504-priv.html
[94] Sal Stolfo, *Comments on Web and Privacy*, Kdnuggets mailing list (21st Mar 2000)
http://www.kdnuggets.com/news/2000/n07/i3.html
[95] Fraud Detection Solutions, Kdnuggests.com
http://www.kdnuggets.com/solutions/fraud-detection.html
[96] Wenke Lee, *Data Mining Approaches for Intrusion Detection*
http://www.cs.columbia.edu/~sal/JAM/PROJECT/ID-SLIDES/index.htm
[97] Symantec AntiVirus Research Center
http://www.symantec.com/avcenter/
[98] McAfee Corporation Homepage
http://www.mcafee.com/
[99] F-Secure Corporation Homepage
http://www.DataFellows.com/
[100] Dr.Solomon's Homepage
http://www.drsolomon.com/home/home.cfm
[101] Virus Checkers
http://www.royaldataservices.com/virusprog.htm
[102] Love Bug Primers
http://uk.search.yahoo.com/search/ukie?p=love+bug+virus&y=y
[103] check.pl package
http://packetstorm.securify.com/UNIX/audit/check.pl
[104] Titan Package
http://www.trouble.org/titan/
[105] Tripwire Data Security Software

http://www.tripwiresecurity.com
[106] L5 Package
http://packetstorm.securify.com/UNIX/audit/L5.tgz
[107] ViperDB Package
http://packetstorm.securify.com/UNIX/IDS/ViperDB-0.7.tar.gz
[108] Audit Package
http://members.home.net/jefftranter/audit.html
[109] Installwatch Package
http://linuxberg.iol.it/conhtml/preview/8335.html
[110] CPM Package
http://packetstorm.securify.com/UNIX/IDS/cpm/
[111] Flog Package
http://linuxberg.iol.it/conhtml/adnload/8239_1481.html
[112] James W. Abendschan, *Unix and Network Security Resources*
http://www.jammed.com/%7Ejwa/Security/
[113] Sportal Package
http://sportal.sourceforge.net/
[114] Cryptographic Syslog Package
http://packetstorm.securify.com/UNIX/loggers/ssyslog.1.21.tar.gz
[115]syslog-ng Package
http://www.balabit.hu/products/syslog-ng.html
[116] Unix Security tools Selection, Packetstorm
http://packetstorm.securify.com/UNIX/audit/
[117] Argus Package
ftp://ftp.sei.cmu.edu/pub/argus/
[118] IP Flow Meter Package
http://www.via.ecp.fr/~tibob/ipfm/
http://www.ipmeter.com/
[119] Arpwatch Package
ftp://ftp.ee.lbl.gov/
[120] iplog Package
http://ojnk.sourceforge.net/
[121] Firewalk Package
http://www.packetfactory.net/firewalk/
[122] Buffer Syringe Package

http://packetstorm.securify.com/UNIX/audit/bsyrin1.zip
[123] Zodiac Package – DNS Protocol Monitoring and Spoofing
http://www.packetfactory.net/Projects/Zodiac/
[124] Slint Packages – Source Code Security Analyzer
http://www.l0pht.com/slint.html
[125] ITS4 Package – Software Security Tool
http://www.rstcorp.com/its4/
[126] Packetstorm Website
http://packetstorm.securify.com/
[127] CERT Coordination Center
http://www.cert.org/
[128] Sniffer Pro High Speed – Sniffer Technologies
http://www.sniffer.com/asp_set/products/tnv/snifferprohighsp_intro.asp
[129] Distributed Analysis Suite, Network Associates
http://www.nai.com/international/uk/asp_set/products/tnv/das.asp
[130] Sniffit Package
http://reptile.rug.ac.be/~coder/sniffit/sniffit.html
[131] Snort Package
http://snort.whitehats.com/
[132] AntiSniff Package
http://www.l0pht.com/antisniff/
[133] Sentinel Package – Remote Promiscuous Detection
http://www.subterrain.net/projects/sentinel
[134] Bogon Package
http://packetstorm.securify.com/UNIX/IDS/bogon.c
[135] Big Brother Package – System and Network Monitor
http://bb4.com/
[136] XpsScanner Package – CGI Vulnerability Scanner
http://packetstorm.securify.com/UNIX/scanners/HTTP-XpsScanner.tgz
[137] Narr0w Security Scanner 2000 – Remote Vulnerability Scanner
http://www.zone.ee/unix/
[138] VetesCan Package – Vulnerability Scanner
http://www.self-evident.com/sploits.html
[139] Cheops Package – Network User Interface
http://www.marko.net/cheops

[140] Network Scanners, Packetstorm

http://packetstorm.securify.com/UNIX/scanners/

[141] *Internet Firewalls*, COAST

http://www.cerias.purdue.edu/coast/firewalls/

[142] Firewall Primers

http://ipw.internet.com/protection/firewalls/

http://uk.dir.yahoo.com/Business_and_Economy/Business_to_Business/Computers/Security_and_Encryption/Software/Firewalls/

[143] Secureway Firewall Package, IBM

http://www-4.ibm.com/software/security/firewall/

[144] Altavista Firewall Package

http://altavista.software.digital.com/

[145] McAfee Personal Firewall

http://www.mcafee.com/myapps/firewall/ov_firewall.asp?

[146] Cisco Works 2000

http://www.cisco.com/warp/public/cc/pd/wr2k/index.shtml

[147] Sun Netra Server

http://www.sun.com/960325/cover/press/netra.pr.html

[148] Web Proxy Primer

http://uk.dir.yahoo.com/Computers_and_Internet/Software/Internet/World_Wide_Web/Servers/Proxies/

[149] Netscape Proxy Server

http://home.netscape.com/proxy/v3.5/index.html

[150] Squid Package – Web Proxy Cache

http://www.squid-cache.org/

[151] Internet Firewall – Resources, COAST

http://www.cerias.purdue.edu/coast/firewalls/fw-body.html#firewalls

[152] Wietse's tools and papers – TCP Wrapper

ftp://ftp.porcupine.org/pub/security/index.html

[153] Check-ps Package – Reports or kills hidden processes

http://freshmeat.net/search/?q=check-ps

[154] bgcheck Package – Process Monitor

http://freshmeat.net/search/?q=bgcheck

[155] Cooperative Intrusion Detection Evaluation and Response (CIDER) Project

http://www.nswc.navy.mil/ISSEC/CID/

[156] Network Flight Recorder Inc.
http://www.nfr.net/index.html
[157]Abacus Project, Psionic Software
http://www.psionic.com/abacus/
[158] Network Intrusion Detection Systems FAQ
http://www.ticm.com/kb/faq/idsfaq.html#4
[159] Network ICE Homepage
http://www.networkice.com/
[160] Network Associates Homepage
http://www.nai.com/
[161] Internet Security Systems Homepage
http://www.iss.net/
[162] Lance Spitzner, *To Build a Honeypot*
http://www.enteract.com/~lspitz/honeypot.html
[163] Grazer1's Bait System
http://packetstorm.securify.com/UNIX/IDS/Gbs.c
[164] BackOfficer Friendly Package, Network Flight Recorder Inc.
http://www.nfr.net/products/bof/
[165] Deception Finger Daemon Package
http://wwwinfo.cern.ch/dis/security/general/tools/honey.html
[166] The Deception Toolkit Home Page, Fred Cohen and Associates
http://all.net/dtk/dtk.html

Chapter 6 - Electronic/Cyberpayment Technologies

6.1 Introduction

Electronic payment technology is less than thirty years old and in its infancy when compared to the use of coins and bank notes. Now more than 30,000 financial institutions worldwide issue cards, increasingly multi-functional, to their clients that are used to purchase more than $2 trillions in products and services worldwide. These technologies have been successful in reaching mainstream usage for several reasons: Firstly, they offer convenience to the consumer; instead of having to visit a financial establishment during working hours, the consumer is able to make transactions at will and at any time of day. Payments can also be made without cash having to change hands, increasing security against loss and theft for both the consumer and businesses. There are also advantages to the financial institutions; although there was an initial investment in implementing the technology, once it became widely used, it allowed the institutions to make cost savings by reducing the number of branches open to the public. In sum, all parties have benefited and the amount of cash in circulation has decreased.

With less cash in circulation, crimes such as theft and robbery are less attractive than they used to be because the proceeds are likely to be lower. At the same time the risk of getting caught has increased due to modern security systems and faster law enforcement response times. Nevertheless, crime has evolved to embrace new technology. As new types of electronic payment technologies have been introduced, new forms of crime have followed soon after. Examples have included using stolen credit cards, mugging people who have just visited an automatic teller machine (ATM), setting up direct debit transactions for small amounts from other people's bank accounts, and in extreme cases stealing an ATM. In the last example, an adaptation of "smash 'n' grab" has been perpetrated, where a JCB digger has been used to wrench an ATM from a bank wall and dump it in the back of a getaway truck.

Threats and problems associated with existing payment technologies are generally well understood. Unfortunately, this cannot be said for new emerging field of E-Commerce and "Cyberpayment". The phenomenal growth of the Internet and private computer network technology has put large demands on being able to conduct financial transactions using networked computers, which has become popularly known as "Cyberpayment" systems. At present Cyberpayment systems are being introduced rapidly and offering the public a much broader spectrum of services than has ever been available before. Many financial institutions are offering their own proprietary Internet based

applications that allow customers to make financial transactions or trade on the stock market directly. At the same time large retailers and wholesalers are starting to do business with each other and consumers directly using Cyber-payment technology. These types of transactions have been introduced to the public and labelled as "business-to-consumer" or "business-to-business" E-Commerce.

A lot of effort is being put into understanding the risks and threats associated with this emerging computer-based technology. To some extent this effort is being hampered because of the large number of different technological standards being introduced and used with no universal standard or accrediting system. Cyberpayment technologies and services will become increasingly popular because of the convenience they offer to both consumers and financial institutions, resulting in increased user dependence and a further reduction in the amount of cash in circulation. Crime will also adapt to take advantage of the new technology. Indeed, there may be unprecedented opportunities for using it for purposes such as money laundering, extortion and electronic theft. The potential for money laundering is especially worrying, because in some cases, the new technology can circumvent and undermine traditional investigative and auditing measures of detection. The range of different products available complicates matters further as regulatory bodies and law enforcement personnel have to come to terms with a vast array of rapidly changing technologies. As this technology becomes an everyday part of a nation's infrastructure and user dependency increases, it is likely to come under scrutiny from other nations and politically motivated entities (such as terrorists) in times of warfare or political conflict.

This chapter has been broken up into five distinct sections. We start in section two by considering some of the different types of electronic payment systems that are currently available. Cyberpayment technologies are presented along with some of the most common technology that is already well established, including comments on their strengths and weaknesses. In section three, we examine some of the properties that should ideally be incorporated to secure financial data transmission in emerging payment technologies. It is worth noting that these properties can be implemented through the use of existing technology to minimize the highlighted risks. Section four deals with the threats and problems that face both existing and emerging technology, and special attention is given to the use of Cyberpayment technology in money laundering. We conclude in section five by considering the future of electronic payment technologies and what can be done to address some of the issues for concern raised in the body of this chapter.

6.2 Types

There are many different types of electronic payment products available. These range from small value transfer systems that are typically used by businesses and consumers up to the large value interbank fund transfer systems on which today's international money and capital markets are based. Currently there are two main ways of representing funds in electronic payment systems:

- **Balanced based systems**: where the account balance is stored and updated after every transaction; and
- **Note based systems**: where electronic notes with a fixed value and unique serial number are transmitted electronically from one system to another.

The two types of technology in widespread use are hardware-based systems and card based systems. An example of the latter would be a bankcard consisting of a plastic card with a magnetic stripe on its underside. Hardware systems on the other hand function by using software that is installed on a computer connected to a computer network.

In the subsections that follow we will examine some of the main electronic payment systems more closely. Well-established technologies such as bank-cards, credit cards, and telephone banking are presented along with new emerging technology such as smartcards and Cyberpayment systems including electronic fund transfer and computer network based E-Commerce.

6.2.1 Bankcards, Credit Cards and Smart Cards

The introduction of ATMs, bankcards and credit cards has resulted in considerable benefits for consumers and financial institutions. This technology has revolutionised the banking industry by offering more convenience, flexibility and safety to consumers while reducing the risks and overheads to the financial institutions. Both bankcards and credit cards are made from a piece of plastic with a magnetic stripe on the underside. In some cases a computer microchip is also embedded on the front of the card to provide added security and additional functionality. This type of card is now generally referred to as a "smart card" because of its multi-functional capabilities.

To the user, magnetic technology appears to be reasonably standardised. For example, one can go overseas on holiday and still withdraw cash from a foreign bank's ATM[1]. While different cards look alike and are used in a similar manner, the underlying technology can vary widely. Cards always remain

property of the issuing institution and the type of information stored on the magnetic stripe depends on how the technology has been implemented. Depending on the implementation, there may also be an issue over personal privacy if information about the user is passed on through use of the card without the user being aware. On the other hand, such information can be used to provide additional security features to help prevent fraud.

Bankcards

Using a bankcard in an ATM to obtain cash is a trivial process; however, the data processing and transactions that go on behind the scenes are quite complex. Although bankcards based on smart card technology are becoming popular in some countries, the magnetic stripe card is still the most widely used because of its cost effectiveness.

When a bankcard is issued, a PIN number is sent separately to the user so that the card can be used in an ATM. PIN numbers are usually 4 digit numeric values and authentication relies on the fact that only the legitimate cardholder knows the number. To ensure that the PIN number remains secret, cryptographic techniques are used to stop the PIN number from being reverse-engineered from any information stored on the magnetic stripe. Encryption is also used so that the PIN does not have to be stored or transmitted in clear text when the user makes transactions. The explanation that follows assumes a basic knowledge of cryptographic techniques, which was presented in the encryption chapter.

When a PIN number is created or chosen, it is mixed together with additional information to create a block of data. The block of data is then encrypted with a one-way algorithm to produce a cipher text. Digits are then taken from the cipher text to form a PIN offset value. The PIN offset is finally stored in a database, or in rare cases on the card's magnetic stripe. When a user enters their PIN number, the PIN offset is recalculated and compared with the stored value to verify that the correct number has been entered. The PIN offset is usually stored in a database because this method allows users to change their PIN numbers without having to replace the bankcard.

ATM devices have a hardware based security module for storing the PIN number entered by a user. Once a user has entered their bankcard and PIN number, a message is constructed and encrypted by the ATM terminal's working key. The encrypted message is the forwarded for processing and PIN verification. The use of encryption in this step prevents the PIN number from being intercepted electronically as it is not transmitted in clear text form. In 1996 a gang in the UK allegedly planned to steal £800m by tapping the data telecommunication lines between ATMs and banks to gather information and

produce counterfeit cards[2]. It is not known if transaction encryption was used is this case. However, if it were being used, the intercepted information would have been of little value to the gang. Also note that the ATM working keys can be changed at regular intervals if necessary. Message Authentication Codes (MACs) can also be used to guarantee the integrity of ATM transaction messages with little processing overhead. More information about the transaction message numbers used by ATMs and their meaning can be found in the ISO[3] standard ISO8583. This standard also documents the significance of the order of binary header bits used in the message protocol.

After magnetic stripe cards became widely used, more people began to understand how they worked. It was not long before the problem of card counterfeiting was discovered. These attacks involved creating fake cards based on information found in discarded receipts and by observing users as they entered their PIN number. With this information, only a home PC and magnetic card writer was required to construct fake cards. Once a card had been made it could then be used with the observed PIN number to withdraw cash. An example of this type of attack allegedly occurred in the UK, where a gang was reported to have used telephoto lenses to film bankcard and PIN numbers being input to an Abbey National ATM[4]. The gang then apparently matched these details up and manufactured fake bankcards allowing them to steal £130,000.

To address card counterfeiting and the type of attacks outlined above, Card Verification Values (CVV) were introduced. A CVV is similar to a PIN offset in that it is constructed cryptographically to form a non-derivable sequence of digits or checksum. It is formed by encrypting data, such as the user's account number, with a one-way algorithm. A selection of digits are then taken from the resulting cipher text and written on to the bankcard's magnetic stripe. The CVV value is then validated in addition to the PIN number when the cardholder makes a transaction. Because the CVV value is only stored on the magnetic stripe, counterfeit cards can no longer be created by simply observing users.

Although the addition of a CVV to bankcards has reduced counterfeiting, it has not stopped the problem entirely. If the cardholder is observed entering their PIN number, and if their card can be obtained for a short period of time, a counterfeit copy can still be made. This type of attack involves using a magnetic stripe reader/writer attached to a computer to copy the information contained on the card's magnetic stripe. The card can then be used in ATMs in conjunction with the observed PIN number. Obviously this type of attack can only be successful if the cardholder is unaware that the card has been outside of their possession. Another method of obtaining the information needed could be to trick users into inserting their card and PIN number into a fake ATM.

Most bankcards also support the purchase of goods and service through Electronic Point Of Sale (EPOS)[5] transactions. EPOS terminals differ from ATMs in the way that they authenticate the cardholder. In some implementations the user is required to enter their PIN number into the terminal, while in others only the user's signature is required. In the latter case, there is clearly potential for abuse if the user's bankcard magnetic stripe can be copied. Once a counterfeit card had been manufactured, it could be used to purchase goods by simply supplying a signature. There are also problems with stolen cards because the user's signature can be found on the reverse of the card and faked when making payments. This problem has been addressed differently in different implementations. One approach is to require the user to provide supporting documentation that proves their identity. Another alternative is to add a photograph of the user to the reverse of the card – this is still uncommon and many users are not happy to provide a photograph to their issuing institution. User transaction profiling can also be used to model a cardholder's typical spending activity. This can be used to raise a computer-generated alarm at the issuing institution if the user's spending habits deviate enough from the profile on record. When this happens, the issuing institution can ask to speak to the cardholder before approving a transaction.

Credit Cards

In contrast to bankcards that generally operate in a balanced-based fashion, credit cards are a note-based system and generally open to greater abuse.

Credit cards can be used with a PIN number in ATMs in the same manner as bankcards for obtaining cash advances. As such, when used in this manner they are subject to the same PIN and CVV checks as bankcards. However, when credit cards are used for purchasing goods they are more susceptible to fraud than bankcards because usually only the user's signature is required to complete the transaction. It is noted that users may also be required to provide proof of identification in some countries. Fraudulent use of credit cards is harder to detect, because of the length of time needed for the cardholder to discover any unauthorised transactions. To counter this many credit card companies now operate intelligent systems to track transactions and look for unusual card usage, alerting customers if necessary.

Other typical credit card uses include telephone, postal and Internet purchases. In these cases no authentication is required; however, the invoice address is required as an additional, but generally ineffective, security measure. In general credit cards offer a relatively risky alternative to cash and cheques; however, their convenience, ease of use, and guarantees offered by the credit card companies have ensured that use has increased rapidly worldwide.

Smart cards

Many banks and credit card companies are now starting to issue smart cards. Although this costs more than issuing a normal magnetic stripe card, there are a number of advantages such as fraud reduction, improved speed and efficiencies and opportunities for access to new business channels.

The embedded microchip on a smart card is capable of processing information and interacting in real time with the application it is being used with. Various industries have found different ways of utilising this technology and in some cases cards are even produced that have multiple applications working on the same card. For example, a financial institution might have a partnership with a mass transit system so that the smart card can also be used to pay for public transport tickets.

Smart cards are manufactured with special semiconductor features designed to prevent them from being reverse engineered and to prohibit the unauthorised access or manipulation[6] of data on the chip. There are also possibilities for improved user authentication methods. This can range from the use of a simple PIN number up to the use of sophisticated biometric checks such as fingerprint or eye retina scanning. The type of authentication implemented will generally depend on the security level that is required by the application with which the smart card is being used.

Fingerprint authentication technology is advanced and cheap enough that it could soon become widely used as a primary smart card authentication method. Fingerprint Identification Units (FIUs) are relatively small and could easily be incorporated into ATMs and EPOS terminals. Fingerprint authentication would require an encrypted template of the user's fingerprint to be stored on the smart card. To use the card, the user would simply insert the smart card in to a terminal. The encrypted fingerprint template is then downloaded from the smart card and sent to the FIU. The user then places their finger on the FIU, which then checks the user's fingerprint against the known template and sends the result back to the smart card terminal. Encrypted transactions can then be carried out from the terminal if authentication was successful.

Smart cards can also be used as an alternative to carrying cash. Some financial institutions have introduced this technology in an attempt to provide an alternative for cash in situations were loose change is often required – for example, at motorway or bridge tollbooths. In this application, the user can upload financial value on to the smart card electronically. The card can then be used for transactions and the appropriate amount of financial value is deducted from the smart card. There have been trials of this type of scheme; however, it is not yet in widespread use. The success of such applications will depend

largely on user acceptance, and it still remains to be seen if this will be considered a viable alternative to cash.

6.2.2 Telephone Banking

Many financial institutions now offer telephone banking to their customers. These systems can be considered as a user-friendly front end to the bank's computer system. A digital telephone exchange (such as a PABX) is used to answer incoming calls and forward users to a computerised voice-prompted menu system. At this level users authenticate themselves by keying account and PIN details on the telephone keypad. Once users have been authenticated, they are prompted to use the telephone keypad to make transactions and choose from a selection of options. In general, these systems are proprietary and the security features will vary from bank to bank. From a user's perspective security is dependent on the access PIN number remaining secret. Tapping the telephone line can compromise PIN number security, as can pressing the telephone redial key after a call has been made. Although telephone banking has some risks, so far it has proved to be very successful and few incidents have been reported.

We have already seen how Card Verification Values (CVVs) can be used by bankcards to help ensure that the card is not counterfeit. A further development of CVV has been developed for telephone authorisations and has been called CVV2. It works in a similar manner to CVV where some predefined static variables relating to the user are encrypted with a one-way algorithm. Digits from the resulting cipher text are then printed on the back of the user's bankcard. This measure can provide call centre staff with a means of checking that the caller is in physical possession of the card.

6.2.3 Internet/Network Based E-Commerce

E-Commerce is an emerging technology that frequently receives media attention. It is on the verge of becoming widely used by the public, with the increasing use of the Internet and the rapid increase of the range of products and services that can be purchased. E-Commerce can be categorised into two types of models:

Business-to-consumer – where, for example, consumers might purchase goods over the Internet; and

Business-to-business - where a business might have an electronic procurement system for its sub-contactors to make bids for competitive tenders.

The most popular and widely used variant is the business-to-consumer type. The main reason for this is the growth of the Internet and the convenience that it offers to individuals. Many businesses have established a presence on the Internet and some are entirely Internet based. The range of merchants offering products for sale over the Internet is enormous and still growing rapidly. With this growth, and partly due to the fact that this new technology lacks sufficient regulation, there have been reports in the media of dishonest merchants failing to deliver the ordered goods. In some cases merchants have been known to disclose credit card details or even charge for goods that were never ordered, though equally there are any increasing number of people who now regularly shop on the Internet.

In the business-to-consumer E-Commerce model, merchants establish their presence by creating a World Wide Web (WWW) site on the Internet. Consumers can then browse the merchant's products by visiting the relevant WWW site. Once goods for purchase have been selected, the consumer is then asked to transmit their credit card details and mailing address to the merchant. The technical processes behind these steps vary from merchant to merchant, but more reputable merchants generally use a secure means of transmitting these details, typically using a technology such as SSL (Secure Socket Layer), which encrypts the customer's details as they are transmitted over the Internet (any details transmitted over the Internet in plain text are susceptible to interception). In cases where SSL is not used one would expect a reputable merchant to invoice the customer after the order has been made in a more secure manner. A good example of a well-known merchant that uses SSL technology is the on-line bookseller Amazon.com[7]. Here, when payment details are requested a small padlock can be seen at the bottom of the web browser indicating that SSL is being used. The prefix of the hyperlink then changes from *http* to *https*.

The security of such transactions relies on the level of encryption used and this is discussed further in Chapter 3. In cryptographic terms 40-bit encryption, such as the exportable version of DES that is used by many companies, is considered weak and susceptible to decryption revealing banking details. However, as with most systems, the storage of this information is inevitably the weakest link, and there have been several well-publicised cases[8] where credit card details stored on a web server have been stolen using exploits of server software.

There are alternatives to using a credit card for Internet based payments. A new type of payment service provider has emerged to offer electronic cash or "e-cash" payments over the Internet. The idea behind this type of product is to offer added convenience to regular Internet shoppers. Choosing products from Internet shops is relatively straightforward; however, submitting credit card details and invoice addresses can be relatively time consuming. E-cash products[9] offer users the convenience of a simple point and click payment

method. When an e-cash account is opened the user has to supply details and deposit money with their e-cash service provider. Once the account has been opened products can be purchased quickly and without the need to disclose personal information. E-cash technology is emerging and competing products may work differently and offer different features. The future success of this type of product is likely to depend on how popular it becomes with consumers.

While on the subject of business-to-consumer E-Commerce, it is worth mentioning on-line banking over the Internet. Many banks now offer this service and in some respects it can be considered a user-friendly front-end product that interfaces into the bank's electronic fund transfer system. Internet banking systems are usually proprietary to individual banks, where the design and implementation is typically out-sourced to a software developer. Almost all products use SSL technology to keep customers' data secure from interception; however, the strength of encryption the SSL uses varies in different countries and depends on local encryption regulations. Use of Internet banking is much the same as visiting an on-line merchant, but a username and password is required before an individual's account details can be viewed.

Although on-line merchants may be using SSL technology, there is still no guarantee that the merchant is reputable and that they will send any goods that have been ordered. To address this issue, there are now companies that offer an independent verification service of the business practices of on-line merchants. The process involves a full audit of the merchant's complete business process, right down to how customer details (such as credit card numbers) are stored in databases. Once the audit has been completed the merchant is issued with a digitally signed seal that can be displayed on its on-line web pages. The reasoning behind this idea is that consumers can have extra confidence in the merchant because the seal of approval has been awarded. They can also click on the seal and find out on what criteria the merchant was audited. Examination of the digital signature can also prove a reasonable assurance that the seal is authentic. Examples of seals currently found on web sites include those issued by Verisign[10]and Webtrust[11]. A more detailed discussion of digital signatures is given in Chapter 3. Another competing service is being introduced by insurance companies, which should also increase consumer confidence. In this case the insurance company provides a seal that guarantees to protect the consumer. In the long run this type of assurance may become more widely used, as it will probably become cheaper for the merchant to obtain and offer more protection to the consumer.

Business-to-business E-Commerce is different in that transactions can occur bi-directionally between businesses. It is conceivable that a business orders a service from a sub-contractor electronically, and in turn the subcontractor has to order components from its client to complete the work. Transactions can also

occur over private networks as well as the Internet and may use other technology instead of WWW pages and browsers. In business-to-consumer E-Commerce SSL technology authenticates the merchant to the consumer. However, when we move to the realm of business-to-business transactions, both parties need to be able to authenticate each other. This also means that organisations wishing to take part in E-Commerce with each other need to be using the same technological standards. In many countries there are no guide-lines defined for developing technological protocols for business-to-business E-Commerce. The danger here is that many organisations are developing different standards for E-Commerce, and it will mean that they can only do business with organisations using compatible standards. This is similar to the situation between some railway companies in the last century that used different loading gauges. In some countries this issue is being addressed by developing a Public Key Infrastructure or PKI that provides a framework for developing a national standard. This subject is discussed in further detail later in this chapter.

In some instances E-Commerce products do not use SSL technology. The most common alternative is to use a Virtual Private Network (VPN). A VPN can be established over private data communication lines or over the Internet and its privacy relies on the fact that a private encrypted data link is established. VPN solutions mean that a dedicate purpose software has to be installed on all computers that will be used for E-Commerce transactions. As a result VPNs are used more often in business-to-business E-Commerce environments where high volumes of transactions are likely to occur between the same parties. A VPN solution might also be chosen in applications where SSL security is considered inappropriate for the value of transactions being transmitted.

E-Commerce in either business-to-consumer or business-to-business form offers gains in competitiveness and efficiency. The main disadvantage at present is the lack of awareness to security issues and the vast array of different technological standards being used. While business-to-consumer E-Commerce is becoming well established, it will still take some time before all of the public have access to, or own the computer-based technology, needed to use it. Generally, transactions are of low value, and if the user is careful the risks are low. It will probably take a little longer for business-to-business E-Commerce to reach its full potential until some outstanding issues have been settled. These issues typically include strong encryption standards as well as the implementation of a Public Key Infrastructure.

6.2.4 Electronic Fund Transfer Systems

Although they have already been available for some time, Electronic Fund Transfer systems are now popularly labelled Cyberpayment systems because of

the computer network technology they utilise. Due to global Internet connectivity and the falling costs of dedicated data lines, there has been a lot of growth in the Cyberpayment arena. Cyberpayment products are similar to traditional wire transfer systems but typically provide additional features such as anonymity, which is normally associated with cash transactions. The growth in demand for this technology has resulted in a number of different systems being developed. In turn this has offered consumers a wide range of different features to choose from.

New features such as anonymity coupled with the variety of technological standards currently available poses new challenges to regulatory bodies and law enforcement personnel. In a Rand Corporation report on Cyberpayments and Money Laundering[12] a number of features were singled out for law enforcement agents to pay special attention to. These can be summarised as follows:

Disintermediation – the transmission of funds from sender to receiver without passing through a third party, which is subject to inspection by a regulatory authority.

Variety of Service providers – where services are not provided by a bank, the organizations in question might not be subject to the same legal requirements.

Peer-to-Peer Transfers – funds are transmitted directly from the sender to receiver making fraudulent activity difficult to trace.

Anonymity – where the source of the funds and the identity of the sending and receiving parties cannot easily be determined.

While Cyberpayment systems offer greater convenience to users, some of the new properties can provide unprecedented opportunities for illicit activity such as money laundering. Of the properties listed above, peer-to-peer transfers and anonymity are likely to pose the greatest problems for law enforcement.

6.3 Ideal Security Properties

Although the variety of Cyberpayment systems differ in their features, they all have to use a secure means to transmit and store financial and personal data. In this section we consider the ideal properties that will allow users to conduct transactions in a secure and risk free manner. Specific application level security and audit features have been deliberately excluded here because of differences in the variety of products available; however, a broad outline is given as to how the information should be protected after being received.

In the next subsections we present the following properties, which assist in making an ideal security system:

- confidentiality,
- integrity,
- authentication,
- non-repudiation, and
- high availability.

If implemented properly, these properties provide a strong measure of protection for Cyberpayment system users from crime, such as electronic theft and extortion. The rapid growth and constantly changing world of E-Commerce and computer systems, however, cannot guard against flaws and technological improvement, so what may be secure today may be weak tomorrow, and new software exploits of holes can be found which can allow unauthorised access even in seemingly secure systems.

6.3.1 Confidentiality

When a payment is transmitted electronically a certain amount of sensitive or private information is usually included in the transaction. A typical example here is a credit card purchase over the Internet where order details, credit card number and cardholder's name and address are transmitted. This information should be kept confidential and should not be disclosed to any third parties during transmission or storage. Confidentiality during transmission can be assured in two ways: Firstly, all locations along the payment's transmission path can be physically secured to prevent interception. In reality this is impossible to achieve. The second, and more realistic alternative is to make use of cryptographic techniques to encrypt payment transactions. Thus, if a payment transaction were to be intercepted it could not be easily understood or decoded. Cryptography can also be used for storage, and safeguards can be built into the database system to prevent illegal access through exploits in the software. Disclosure of large numbers of credit card details can be disastrous to both the credit card company and vendor, and as the losses through this form of fraud increase, credit card companies will start to perform more stringent checks on the companies using their services.

6.3.2 Integrity

When an electronic payment transmission reaches its destination the receiver should take *due care* to ensure that the details received are genuine, i.e., that

they have not been altered or subverted in any way. While encryption can be employed to help ensure confidentiality, it is not always used. This is particularly the case when confidentiality is a low priority. Even in cases where encryption is used, transactions can be deliberately intercepted and altered to cause a breakdown in service. A received transaction could also contain unintentional errors caused by a technical fault on the transmission path used. To prove that a transaction has kept its integrity we need a means of comparing received payment details with those that were used prior to transmission. While this could be achieved in a number of ways, the most practical method is to use digital signatures.

When a digital signature is used the originator of a transaction applies a one-way cryptographic hash algorithm to the transaction data to create a unique checksum. This unique checksum is then asymmetrically encrypted with the sender's private cryptographic key and added to the bottom of the transaction data to form a digital signature. When the transaction is received, the receiver can then use the sender's asymmetric public key to decrypt the original checksum. The receiver then applies the same one-way cryptographic hash algorithm to the transaction data that has been received to generate a fresh checksum. If the freshly generated checksum is the same as the one that has been decrypted from the digital signature, then the transaction data has not been altered and its integrity has been preserved.

6.3.3 Authentication

There are some products available that offer complete anonymity, making no provision for any kind of transaction records to be kept. The trade-off here is that complete anonymity provides no mechanism to facilitate dispute resolution when errors (deliberate or unintentional) occur. In addition, products offering complete anonymity are more likely to be misused for illegal purposes such as money laundering; however, if electronic money is traceable in a way paper money is not, there are also issues of personal freedom, which will alarm many people.

Authentication mechanisms provide a means of proving that the sender and receiver of an electronic payment are in fact who they electronically claim to be. For example, the sender of an electronic payment needs to feel reasonably confident that they are transmitting the payment to the real recipient. One way of doing this might be for the sender and recipient to meet in person, exchange asymmetric public encryption keys, and provide proof of identification, such as a driver's licence, birth certificate or passport. In practice, when payments are transmitted to many different recipients a more convenient method is required.

Again, cryptographic techniques can come to our aid here with digital certificates.

Digital Certificates can be issued by a third-party known as a Certification Authority or CA. Depending on national variations, the CA may be either an independent company or a governmental organization. A Digital Certificate contains information about the individual or organisation that owns it such as the owner's name, their public asymmetric encryption key, under what circumstances the certificate was issued and an expiration date. Other information may also be included at the discretion of the Certification Authority. Essentially, a Certification Authority acts as a trusted third party that is trusted to verify the identity of an individual or organisation using a rigorous checklist. Once identity has been confirmed, the Certification Authority issues the owner a Digital Certificate that has been digitally signed by the Certification Authority. Because a Certification Authority is trusted, it provides a means of authentication for parties previously unknown to each other. Public asymmetric encryption keys can be extracted from a Digital Certificate, and the Certificate itself can be validated with a simple query to the issuing Certification Authority. There are, however, issues about CAs and their role as confidence agencies that are discussed further in Chapter 3.

Asymmetric encryption, Digital Signatures, Digital Certificates and Certification Authorities together can be used to form a Public Key Infrastructure (PKI). Such an infrastructure if implemented properly can be used in electronic payment technologies to provide confidentiality, transaction integrity and user authentication. A more thorough treatment of Public Key Infrastructure and cryptographic techniques can be found in chapter 3.

6.3.4 Non-repudiation

Non-repudiation is the property where the sender and recipient of an electronic payment transaction cannot deny that a transaction took place. Some electronic payment technologies allow for anonymity and do not authenticate the sender or recipient of a transaction. In such cases the parties involved in an electronic payment transaction can easily deny that it took place, but at the same time there is no mechanism provided for them to dispute any errors. Public Key Infrastructure, if implemented properly, provides the mechanisms for transaction authentication and verification of transaction integrity (confidentiality can be assured by optionally encrypting the transaction data). To ensure that transactions cannot be repudiated, the parties involved must be able to authenticate each other and prove that the transaction integrity has been maintained. It follows that if Public Key Infrastructure is used, non-repudiation can only be assumed if appropriate measures are taken to secure the private encryption keys.

6.3.5 High Availability

Finally, the most often overlooked ideal property for any electronic payment system is availability – i.e. it should be available for use at all times and not experience technical difficulties or need to be closed down for routine maintenance.

There are other issues that can also affect availability. Often, after a period of time, demand for a service can grow, and subsequently the infrastructure needs to be updated to cater for larger transaction volumes. If no action were taken, the service would become slower as demand increased and eventually grind to a halt. The point here is that not all systems are scalable, and sometimes this means that a new system has to be designed and implemented. This can mean that significant disruptions may occur as the old system is phased out and users are transferred to the new one. Well-designed systems are built to be scalable and to cater for future capacity increases. In addition, business continuity and disaster recovery plans are usually made to address any unforeseen circumstances that could possibly arise. Typically financial institutions have a geographically separate backup site that electronic payment traffic can be redirected to in the event of a problem occurring at the main site. Business continuity and disaster recovery planning is discussed in more detail in chapter 7.

6.4 Threats and Problems

The successful operation of electronic payment technology can face problems and be threatened from a number of different sources. In this section an overview is given of some of the most common threats and problems including interception, hacking, denial of service, adverse publicity, lack of standardisation, and cases of fraud such as money laundering. A final subsection has been included to consider the identity and motivation of perpetrators that might exploit any of the threats or problems that are presented.

6.4.1 Interception

There are two different types of interception that deserve mention here. The first involves physical theft – i.e. stealing bankcards, credit cards and computer printouts of transaction records. The second type is electronic network interception, which is quite different and likely to be perpetrated by a completely different set of people.

The problem of magnetic card theft and counterfeiting is generally well understood and has already been presented in section 2.1. The impact of theft of

computer transaction records depends on the type of payment system the record was produced by. Generally, transaction records are printed in a physically secure environment or specific details are masked to prevent fraud; however, this is not always the case with modern E-Commerce systems. Some of these systems can even produce a full list of all user credit card details. If a list of this type was stolen, the credit card details contained in it could easily be sold for fraudulent activity.

With electronic fund transfer, encryption is often used to ensure data transaction confidentiality. Unfortunately, some implementations of encryption don't work very well when technical problems are experienced. An example here would be the transmission of encrypted data over faulty data lines that might only be capable of handling half of their normal bandwidth. When this occurs, some electronic payment systems stop using encryption so that the transaction data can still be transmitted. This can present an opportunity to deliberately disrupt encrypted payment transactions so that they cannot reach the intended destination. The reasoning behind this is that the electronic payment product will eventually switch over to using unencrypted transactions, and when it does the thief will be ready to steal the transaction details by using tools such as packet sniffers.

With Internet based payments that use SSL technology, a random session key is used to encrypt the transaction data. Random session keys need to be truly random. This may sound obvious, but machines such as computers are not particularly good at generating genuinely random data, which in turn is used for generating random session keys. Most pseudo random data generated by a computer follows a predictable sequence and may be successfully guessed. To generate truly random data extra measures need to be taken, such as recording the timing interval between keys being pressed on the keyboard. If a Cyberpayment system uses an implementation of SSL that doesn't generate random enough session keys, it would not take a motivated hacker long to guess the random session keys. This would then mean that any encrypted transaction data (such as credit card numbers) could be intercepted, decoded and stolen.

Consider the following example: Suppose a stockbroking company developed an on-line trading service that used a generic version of SSL with predictable session keys allowing usernames and passwords to be intercepted. If enough usernames and passwords were collected a large database of username/password combinations could be compiled. These usernames and passwords could then be used simultaneously at a later date to electronically purchase one particular stock in an attempt to influence the stock market. Thieves would not be able to obtain the stocks that they purchased using the stolen account details, but they could benefit from the outcome.

6.4.2 Hacking

In the area of electronic payment technology, Internet based E-Commerce is the most obvious target for hackers. Because the E-Commerce systems are connected to the Internet, it is relatively straightforward for anyone else connected to the Internet to use hacking tools to attack the system. This problem is further complicated as most E-Commerce implementations use standard commercially available hardware and software for which vulnerabilities are well known and understood. At the same time, hacking tools are becoming increasingly automated and require little technical knowledge to use. The size of the Internet and the vast number of insecure computers connected to it offer unprecedented opportunities to conduct untraceable-layered attacks.

The most likely goal for a hacker would be to obtain access to the system's payment detail database and retrieve details such as credit card numbers. The likelihood of this occurring depends largely on the strength of the security and control measures in place at the organisation concerned. This subject is considered in greater detail along with a description of ethical hacking in chapter 7. It is important to realise that most security breaches occur when hackers use known security exploits that have been published on Internet IT security Web pages. It therefore follows that security measures must be flexible and updated constantly to reduce the chances of a hacker breaking in to the system.

While we are considering computer based hacking techniques, it is important to consider some of the impacts that the Y2K problem has had on computer-based systems. A large number of computers and software products have recently been updated to become year 2000 compliant. The sheer number of different systems in existence has meant that there has been a shortage of skilled programming resources, and as a result some organisations have hired contractors from overseas. The disadvantage is that it is a lot harder to carry out background checks on sub-contractors from other countries. Untrustworthy sub-contractors that are updating high profile systems, such as Cyberpayment systems, could easily build a trojan horse or back door in to the system that could be exploited at a later time. The full impact of Y2K associated trojan-horses has yet to be seen, and an extortionist or saboteur could wait several years before using a back door to the system.

6.4.3 Denial of Service

Denial of service can occur intentionally or unintentionally. A denial of service situation occurs when normal operations have to be suspended and no backup procedures can be used. If adequate IT controls are in place, intentional denial of service is the most likely to occur.

Deliberate denial of service could be carried out in a number of ways from deliberately flooding an organisation with electronic data transmissions to physically destroying the building where the payment systems are housed. Electronic denial of service is a significant problem and is almost impossible to prevent. Recent attacks have been called distributed denial of service attacks because hackers have used a large number of different computers on the Internet to attack a target at the same time. Such attacks are structured so that many machines flood the target with legitimate connections simultaneously. The result is that all of the network bandwidth gets used up and the service grinds to a halt, refusing any more connections.

Physical destruction of payment equipment, transaction clearing centres or data communication lines will clearly cause a denial of service. Terrorists might use high explosives to reach this goal, and to deliberately cause as much financial damage as possible. Such attacks might even go as far as to target both a main and backup clearing centre simultaneously. In addition, for a terrorist organisation, the use of explosives is much less technically complex than trying to attack a system electronically. The impact of the damage can also be far greater because all of the equipment and the building will need to be replaced. Note though, that most large institutions have several backups stored in several locations to prevent data loss and most standard equipment can be easily replaced.

Denial of service can also arise if a Cyberpayment system is overwhelmed by too many legitimate simultaneous payment transactions. This typically happens when a system becomes so popular that more users are subscribing to it than the technology was designed to cope with. The result is a service that is so slow it becomes unusable. Capacity planning can be used to anticipate future growth; however, some Cyberpayment systems are based on technology that is not scaleable. In these cases a complete, and expensive, technological system redesign is required.

Typically, organizations using electronic payment services lease data lines from telecommunications companies to connect to the system. Data lines are typically quoted as having a certain bandwidth, but the guaranteed bandwidth is often less. Generally, higher data bandwidth than is guaranteed can be obtained, but it is dangerous to assume this one hundred percent of the time. It therefore follows that if the guaranteed bandwidth is lower than the bandwidth needed by an electronic payment system at its busiest time, a denial of service can occur. Another issue is the availability that telecommunications companies offer. For example, 99 percent guaranteed availability over a year could allow for three consecutive days without service. Telecommunication companies have also been known to oversubscribe their data lines, which means in periods where all

customers experience heavy data traffic, the guaranteed bandwidth cannot be provided as promised.

6.4.4 Adverse Publicity

Adverse publicity can damage an organisation's reputation and cause problems that might have previously been unforseen. One of the most commonly reported breaches of security comes in the form of WWW sites that have been hacked or defaced. Once a WWW site has been hacked or defaced, the whole world can view the damage immediately. This also means that such intrusions are frequently reported in the media, especially when a high profile organisation is involved. Generally, an organisation would be very lucky to escape without any media attention if their web site was hacked since the perpetrators of such crimes usually announce their work to the Internet community. Defacing or hacking a web site may not have direct consequences to the organisation concerned (unless credit card details were stolen from the organization's database), but the indirect consequences may be enormous. For example, if a financial institution's Web site was defaced, potential investors might come to the conclusion that the organisation's general security measures are inadequate and invest their money elsewhere.

Other types of theft involving electronic money products are not reported as frequently as one might expect. When considering prosecuting a perpetrator it is important to remember that the media attention the case would receive will be far more financially damaging to the organisation than writing off the amount that was stolen.

6.4.5 Lack of Standardisation

The lack of standardisation in electronic payment and Cyberpayment technology is mostly due to the fact that different systems have implemented at different times because of the rapidly changing demands and availability of software and protection schemes. In addition, many systems are based on proprietary designs, and the companies that developed the technology had conflicting ideas about what would become the most widely used standard.

Publicly scrutinised standardisation will become increasingly important in the future because of the advent of E-Commerce. Organisations using different standards will not be able to make transactions with each other, and consequently will be restricted to trading with organisations using the same type of technology. The whole point of having a single standard is that public computer networks such as the Internet can be used to increase competitiveness. Some

countries are now in the process of developing an open Public Key Infrastructure that can be used to create a national standard for E-Commerce. That said, the process is being hampered in some countries over a debate on what constitutes an acceptable strong encryption standard and law enforcement access to data. Once these debates have been settled in individual countries, it is likely that different countries will have set different standards. Basically, this will mean that international E-Commerce transactions will have to support several different standards.

Earlier in this chapter we presented bankcard technology and saw that even this technology is not totally standardised. One of the main differences comes in the bankcard itself. For example, some cards have smart chips built in and others do not. Bankcards containing the smart chips allow extra security protection and make the cards far more difficult to copy. Ideally all bankcards should be using smart chips, as they are far superior to the older cards that only have a magnetic stripe. Generally, cards using smart chips have only been implemented by a small amount of financial institutions, and often this has been done specifically reduce card related fraud. The likely reason for smart chips not being widely used is that other technology, although less secure, is cheaper. In addition, there would be large costs involved to the financial organisations concerned in upgrading their technology.

6.4.6 Money Laundering

Everyone, including criminals, likes using cash because of the anonymity that it offers. The process of money laundering allows criminals to legitimise and hide the origin of "dirty" money so that it can re-enter the mainstream economy. Traditionally, the money laundering process follows three distinct stages, which have been described by law enforcement personnel as follows:

Placement – where the cash is physically deposited at a financial institution,

Layering – where layers of financial transactions obscure the source of the funds,

Integration – when the funds are integrated with legitimate funds and re-enter the economy.

More details of these stages and the traditional money laundering process can be found in the Rand Corporation's report "Cyberpayments and Money Laundering"[13].

It is important to ensure that the new technology does not hinder a law enforcement agency's ability to detect and prevent fund transfers that are linked

to criminal activity. This problem becomes more complicated if different countries place different emphases on achieving these objectives, and set about doing them in different ways. For example, differences may occur due to regulatory traditions, statutory mandates, trans-national policies, or any other relevant factors. We have also seen that electronic money transfers may be vulnerable to manipulation or interception when being transmitted over computer networks; this can introduce other concerns that do not exist in traditional payment technologies, such as transaction records being inadequately detailed to allow prompt resolution of disputes and errors. At the same time there are few examples of codes of practice and self-regulation in terms of disclosure and fair practice.

Electronic fund transfer can clearly pose new challenges and threats to law enforcement, yet at the same time it can also bring new benefits that were not available with traditional payment methods. In the case of the former, payment systems may be exploited for criminal activity such as money laundering or tax evasion. They also pose the threat of being more directly exploited by methods such as fraud, counterfeiting and system disruption. For the latter, depending on the type of system being used, the fact that some degree of transaction records can be generated compensates for the convenience of electronic systems compared with more traditional methods.

Earlier in this chapter we considered some of the different types of electronic money products that are currently in use. The different features that these products offer allow trade-offs to be made depending on customer requirements. We will now examine some of these features more closely and consider the implications for law enforcement.

There are three main categories of transaction record keeping:

- total anonymity where no records are kept,
- partial record keeping, and
- full record keeping.

Full record keeping can assist in error resolution, operational failures and in protecting customers from being vulnerable to attack. Clearly there is a trade-off between anonymity and transaction verification. In terms of transaction record keeping, systems where full detailed records are kept are likely to use a centralised database for storing transaction records. Such a database allows financial institutions to fulfil their traditional role of acting as an intermediary to help law enforcement officials prevent and detect the illegal movement of funds. Law enforcement relies on financial institutions to a great extent to identify suspicious transactions, maintain records and to identify the customer.

On the other hand systems offering more anonymity and less detailed record keeping can still cater for error and dispute resolution, although a bit more time and effort may be involved. In such cases the customer identification may remain confidential from third parties by using random generated codes, but a database of records is still maintained allowing suspicious activities to be identified. Law enforcement agencies can then use compulsory measures to reveal customer identities and at the same time disputes can be resolved. In the last case, where total anonymity is guaranteed and no records are kept, there is clearly greater scope for misuse and money laundering. The trade-off here is that there are far more operational risks and it may be the case that funds can be lost if system errors occur or if transactions are intercepted. Dispute resolution is also not possible with these types of system.

In the three types of systems that have been outlined, the ones that pose the largest threats are those that offer total anonymity. In 1990 the FATF[14] issued forty recommendations to combat money laundering. In 1996 the FATF's reflection on previous experience led to the adoption of a new recommendation that dealt with new technological developments. The recommendation stated, "Countries should pay special attention to money laundering threats inherent in new or developing technologies that might favour anonymity, and take measures, if needed, to prevent their use in money laundering schemes."

As different countries debate whether or not to apply anti-money laundering laws such as customer identification, and transaction reporting to electronic money products, there are some important things that should be considered. Recording all electronic money transactions would generate massive amounts of data and would add extra costs to electronic payment technologies that do not apply to other payment techniques. Having said this, it is likely that organizations using electronic payment technologies will keep records for their own purposes, such as dispute resolution and fraud detection. These records could be very valuable for fighting crimes such as money laundering. Recently a Group of Ten report of the working party on electronic money[15] made a survey of policies towards electronic money in the G-10 countries, and detailed differences in disclosure requirements, privacy, disputes, guarantees, anti-money laundering measures, licensing, and auditing.

In all forms of electronic payment technologies that use computer networks there are technical risks that need to be addressed. This is especially significant today as some products are now based on the use of open networks, such as the Internet. In order to know that an electronic money system is working properly and generating reliable records, it is necessary to implement ongoing risk monitoring. There are two parts to this process: system testing and auditing. System testing is necessary to identify unusual activity patterns such as attacks and to predict major system problems. Another important aspect of system

testing is penetration testing, which allows identification and isolation of flaws in the system before they can be exploited. Auditing of electronic money systems offers an independent control mechanism for detecting problems and minimising risks. It is also worth considering that as new technologies are introduced, it is important to have the source code for the programs independently reviewed to ensure that there is no extra hidden functionality in the software that might, for example, allow certain types of transactions to go unreported. More examples of possible risks and risk management procedures can be found in a report by the Basle Committee on Banking Supervision[16].

6.4.7 Perpetrators

The skills of an individual using stolen credit cards are clearly going to be different from those of an attacker of a Cyberpayment system. Similarly, while a computer attacker might act alone, a large gang is likely to be involved in the process of laundering the proceeds from organised crime. In this section we consider the skills and knowledge that are needed to exploit some of the threats and problems that have been described elsewhere in this chapter.

Attacks against Cyberpayment systems (Cybercrime) are not easy to commit. Theft or interception of electronic funds requires a detailed knowledge of the intricacies of the system being targeted. This type of information is not generally available for obvious reasons (and because most technology is proprietary), and so it follows that the criminal must either have existing experience of the target system or be receiving insider information from a corrupt employee. Sabotage and extortion are easier to achieve; however, some insider information is still required. While a terrorist organization might only want to know the location of the transaction clearing houses so that they can plant explosives, an extortionist would typically require more detailed technical information so that they could cause a temporary electronic denial of service.

Cyberpayment products are most likely to come under the scrutiny of organised crime groups for purposes such as money laundering. In this case insider information is still needed, but the information is not of the same technical nature of that required by thieves or extortionists. Typically, assistance will be sought from corrupt employees to reveal transaction reporting limits and aggregation periods. Other useful information might include understanding how and under what circumstances audit trails are generated. In some cases a corrupt employee might even be asked to abuse their position and remove transaction records pertaining to illicit transactions.

Hackers differ from traditional criminals in that they generally don't receive any insider information before breaking into or disrupting Cyberpayment systems.

Often they act alone and break into systems as a challenge rather than for financial gain. Nevertheless, there are hackers who are more than happy to help other criminals in committing crimes against Cyberpayment systems.

Finally, disgruntled employees are more than capable of sabotaging a system for a variety of different motives. This is especially worrying because their intricate knowledge of the system concerned may make it easy for them to hide traces of their activity. Typically a disgruntled employee might sabotage a system for a relatively petty motive, but not always necessarily for financial gain. There have been several cases of this type and these are discussed in more detail in chapter 7.

6.5 Conclusion

The rapid growth of Internet technology and Cyberpayment systems has started a chain reaction and many organizations are implementing this new technology through a fear of being left behind, without understanding the full implications and dangers. The range of different technological standards currently available has been driven by market competition to devise an industry standard Cyberpayment system. Ironically, most of these products are proprietary, and have had the opposite effect resulting in a lack of standardisation.

Non-standardisation and rapid development of products means that some systems are more secure than others. In some cases trade-offs have been made between security and functionality. In general, the risks and problems associated with well-established payment systems are well understood; however, this cannot be said for Cyberpayment systems because of the variety of products available. In section three we showed that existing technology could make provision for the secure transmission of financial data. In other words, if due care is taken, threats such as electronic theft, denial of service and extortion can be minimised. PKI has been widely accepted as a way forward for securing financial data transactions. Although PKI generally refers to a public key infrastructure, the same technology could just as easily be used to create a private key infrastructure for use within private organizations.

Cyberpayment application properties are a separate issue. A system might use state-of-the-art technology to secure transmissions and authenticate users while offering peer-to-peer transactions and failing to maintaining transaction records. Systems offering peer-to-peer transactions and anonymity will clearly pose the greatest threat to regulatory bodies and law enforcement personnel because these properties make it almost impossible to trace illicit activity. It should also be noted that systems offering total anonymity have no mechanism for dispute resolution due to the lack of transaction records.

Another issue that is likely to receive more attention in the future is privacy. While anonymity in payment systems can pose challenges for law enforcement personnel, tracking money and profiling user's spending habits could lead to a breach of privacy. There may be a real threat in the future that unauthorised organisations could indiscriminately gather such information and use it for nefarious purposes.

Although non-standardisation has been defined as a problem, it does offer one advantage. In times of warfare or political conflict it is increasing likely that Cyberpayment systems will be targeted by foreign governments and politically motivated entities. As user dependency on these products increases, non-standardisation will make it a lot harder for a nation's entire electronic payment infrastructure to be targeted at once.

The future for Cyberpayment systems looks positive, and it is expected that Cyberpayment systems in conjunction with smart card technology will replace the need for cash payments in the future. This technology is advancing rapidly and it will offer new opportunities for illicit activity such as money laundering. For these reasons, it is vitally important that law enforcement personnel can come to terms with this new technology in a timely manner.

1 This assumes that the foreign bank is connected to an international network service such as Cirrus or equivalent.

2 Convict and vicar foiled £800m cash machine sting, John Steele - Crime Correspondent, UK News Section, The Electronic Telegraph, Thursday 17th December 1996, Issue no 573.

3 The International Standards Organisation (ISO) can be found at http://www.iso.org.

4 Gang stole £130,000 in hole-in-wall cards scam, UK News Section – The Electronic Telegraph, Friday 13th September 1996, Issue no 478.

5 Note that EPOS can also be referred to as EFTPOS. Typical examples include the United Kingdom's SWITCH service or the European MAESTRO service.

6 Data can be irreversibly burnt into the chip by special circuits.

7 http://www.amazon.com.

8 See for example - Credit card web: teenage pair arrested, The Australian Financial Review, Page 29 Monday 27th March 2000.

9 Two examples of e-cash products are "eCash" from eCash Technologies Inc (http://www.digicash.com) and Visa's "electronic wallet" product (http://www.visa.com/pd/ewallet/main.html).

10 More information about certification products offered by Verisign can be found by visiting Verisign's website at http://www.verisign.com/server/index.html

11 Details of the Webtrust certification service can be found by visiting http://www.cpawebtrust.org

12 Cyberpayments and Money Laundering: Problems and Promise, Roger C Molander, David A Mussington, Peter A Wilson, MR-965-OSTP/FINCEN, The Rand Corporation 1998.

13 Cyberpayments and Money Laundering: Problems and Promise, Roger C Molander, David A Mussington, Peter A Wilson, MR-965-OSTP/FINCEN, The Rand Corporation 1998.

14 The FATF (Financial Action Task Force) is an inter-governmental body developed by the G-7 countries in 1989 designed to combat money laundering of criminal proceeds and hiding its original origin.

15 Electronic Money: Consumer protection, law enforcement, supervisory and cross border issues, Group of Ten, April 1997.

16 Risk Management For Electronic Banking And Electronic Money Activities, Basle Committee on Banking Supervision, Basle, March 1998.

Chapter 7 - Managing IT Risks, Threats and Problems

7.1 Introduction

Information is one of the most valuable assets owned by any organization. Accordingly, any information generated or stored in a computer system (or network) needs to be protected with the same degree of security as any other type of media or physical asset. Information managed properly helps an organization run efficiently and effectively. Loss, corruption, or unintentional disclosure of information promise grave consequences for the organization concerned. Examples can include needless financial loss, litigation, regulatory fines and adverse publicity.

Information technology has revolutionized the way organizations conduct their business. Increased storage capabilities in conjunction with lower hardware costs means that more and more information is being stored digitally. Since IBM introduced the first commercially available disk drive in 1957, improvements in technology and miniaturization have led to the density of data stored on today's hard disks increasing by a factor of 1.3 million[1]. Simultaneously, computer network technology has evolved. Over the last ten years the growth of the Internet has seen many organizations becoming part of a global network, providing them with unprecedented opportunities to conduct business electronically.

Improvements in IT have delivered convenience and productivity savings, yet new problems, risks and threats have arisen around them. Latent risks have now the potential of active threats, which may vary according to individual organizations. The perception of nascent threat to an organization's security relies in no small measure on the level of enlightenment possessed by the risk evaluators. Risks and threats will also differ from one organization to another according to the IT system it uses; and the function and value of the information stored in it. For example, the threats and problems faced by an on-line bookseller will be quite different from those faced by an academic institution. That said, most organizations share a common problem in that they seldom fully understand the trade-off between convenience and IT security. In many cases this is due to the novelty of the technology and because IT departments are struggling to implement new features in a timely fashion while coping with a lack of skilled resources. Another problem common to many organizations is their blanket acceptance of IT services — the very fact that the organization cannot survive without them — without any evaluation of the potential for such IT services to be closely responsive to the needs of management. The danger

here is that IT staff are employed for their technical skills rather than their ability to focus on business needs or organizational goals.

This chapter considers how to manage IT related information risks and presents ways to ensure that IT is led and dictated to by an organization's strategic goals and business plans rather than technological developments. Clearly, then, before any risks can be managed, they need to be identified and understood. In turn, and in order to identify risks, we need to be able to understand the potential threats and problems faced by an organization. We may safely conclude that many risks are ever with us. They are recognizable by common sense and have the potential of being managed by more direct responses. Unmanaged risks are the germ of threats. While these will be different for most organizations, some of the most common types are discussed in section 7.2, including information loss, interception, misuse, system failure, unauthorized disclosure, and legal/regulatory action. Once the potential threats and problems have been understood for a given organization, risks can be identified and ranked according to their potential severity. In reaching this assessment, consideration must be given to the potential for the risk, threat or problem materializing and the impact it would have on the organization's day to day operations. This can be best explained by considering the graph shown in figure 7.1 overleaf.

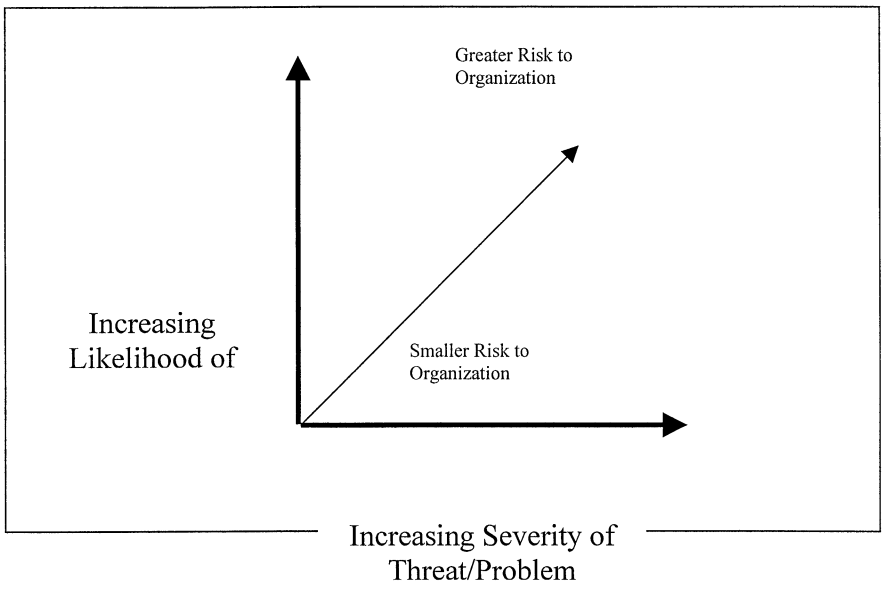

Figure 7.1: Determining the Severity of Risks

Here it can be seen that the risk increases with both the likelihood of occurrence and the severity of impact to the organization. This process of risk assessment can then be carried out for all threats and problems that have been identified, providing the organization with an overall assessment. Do note that this measure is done independently of any preventative measures, which the organization might already have in place; rather, it merely identifies threats and problems faced by the organization and assigns a risk rating.

Section 7.3 examines the controls and procedures which can be put into place to manage threats and problems identified as posing significant risk or exposure to the organization. Here it is important to remember that it is not always practical or financially viable to manage all threats and problems. In many cases threats and problems that pose low risks are accepted because the likelihood of their occurrence and the potential impact are low, hence not justifying additional expense. The controls and procedures covered in section 7.3 include physical security, environmental controls, logical security (including monitoring and escalation procedures), managing changes and documentation, disaster recovery planning and continuity planning. Consideration is also given to the ways that these procedures and controls can be audited. In addition, the relative advantages and disadvantages of using external audit/consulting services are discussed along with ways to obtain added value when these services are used.

Once an organization has identified its potential threats and problems and has categorized them by risk, the controls that are already in place can be examined

to make sure they address all areas of significant risk exposure. For most organizations IT services are changing so rapidly that there will always be some areas of exposure that are not covered by existing controls and procedures. At this point missing controls and procedures can be identified, thus enabling the organization to make plans for improvement. Perhaps the most important point to realize is that managing IT risks requires the process outlined here to be iterative and repeated at periodic intervals.

A method of managing IT threats and problems has been outlined here. Sections 7.2 and 7.3 concentrate on examining potential threats/problems and controls/ procedures respectively. In particular, because this chapter is primarily concerned with IT risk management, logical security issues (such as penetration testing, user access levels and external connectivity) must receive their due attention and are the feature of section 7.3.3. Finally, a conclusion that examines future issues and concerns is presented in section 7.4.

7.2 Understanding Potential Threats and Problems

Understanding potential threats and problems is the first step in determining the various types and magnitude of risks that an organization faces. While consultants can be hired to help identify these, they can only act as facilitators because nobody understands the risks and problems faced better than the organization itself. In cases where external consultants are used, the quality of the results depends on them asking the right people the right questions, the quality of the methodology they use, and their capability to understand and interpret the information gathered from the organization's personnel.

To determine the magnitude of risk for a given exposure the organization must consider its likelihood of it occurring and the severity of its impact. Again, the most qualified individuals to determine these factors are usually members of the organization in question. One of the best ways of determining the likelihood and potential impact of threats and problems faced is to hold a workshop with members of the organization's senior management and members of its IT services department. This type of workshop, when properly facilitated, brings together individuals who normally view the organization from a different vantage point, yielding more accurate risk assessments.

Potential threats and problems that are faced will seldom be the same for any two organizations. In the sections that follow we present some of the most common threats and problems (loss of data, interception, misuse, system failure, unauthorized disclosure, and legal/regulatory action) and likely causes of those problems. Finally, consideration is given to the types of actors associated with different threats and problems.

7.2.1 Loss of Data & Unauthorized Disclosure

Many people associate loss of data with hardware and software problems, and there are few computer users who have not experienced partial or full data loss after a computer crash. Individuals can normally tolerate the loss, but many organizations are increasingly dependent both on the high availability of their data and its integrity; the consequences of its loss can be devastating. Awareness of the cost of data loss has led to most organizations implementing a comprehensive backup strategy, where system backups are made on a regular basis[2]; however, there are a number of other problems that are commonly overlooked, including:

- **Unintentional deletion of data before it has been backed up for the first time.** For example, if data backups are made at the end of every day, the accidental deletion of data created earlier that day would result in data loss. While small losses might be an acceptable risk to some organizations, it is not acceptable to all — for example those conducting financial trading. In such cases the unintentional deletion of data could occur due to a system crash, but it could also occur if write access is granted to inappropriate users who do not understand how to use the system.

- **Inadequate staff training and poorly documented operating procedures.** Although most organizations invest in backup equipment, staff training and documented operating procedures are often overlooked. For example, a high staff turnover could result in new members of staff being unsure to what extent data should be backed up or how to perform backup duties. In some cases, if backup duties are not explicitly written into job specifications, one may find that no one is backing up data – this can go undetected for some time until a backup tape is required to restore a corrupted system. Backup media also needs to be tested to ensure that a restore operation can be performed. Sometimes problems occur when write heads on backup drives fail, resulting in the creation of blank backup media.

- **The deliberate (malicious) or accidental alteration of data.** If a small amount of data is altered and there is no way of identifying the specific data that was changed, the end result is that the integrity of all data has to be questioned. Alteration of data can effectively mean that all data created since the last unaffected backup was made has been lost.

- **Physical destruction of property.** Sometimes the unexpected happens and fire or water (in some locations factors such as terrorism and the use of high explosives also need to be taken into account) destroys buildings and computer systems. When data backups are made they are often stored on-site with the computer equipment, and are also destroyed in the event of a disaster.

Unauthorized disclosure of information contained in IT systems is more likely to be deliberate than accidental. Accidental disclosure occurs most frequently when the wrong data is transmitted electronically to a third party, or when data is transmitted to the wrong destination[3]. In most cases accidental disclosure usually has a low impact on the organization concerned.

The impact of deliberate disclosure of information can be very serious, especially if the information is made available to a rival organization. Problems occur when confidential data is not categorized or logical access not restricted to appropriate user groups. Many organizations employ contractors who may have

unrestricted access to confidential and proprietary data, and who are not subject to the same confidentiality agreements as its employees. In addition, many organizations informally encourage knowledge sharing from individuals who have been recruited from rival organizations.

7.2.2 Interception

It is important to realize that interception can occur physically as well as logically. Unauthorized physical interception, i.e. theft, can occur if sensitive data printouts are not stored in a secure location where access is restricted to appropriate personnel. Organizations must also consider the location of printers when sensitive data is printed. Network topology diagrams and hardcopies of network infrastructure configuration files are another cause for concern. A network diagram provides detailed topology information (often including IP addresses) that would be very helpful to an intruder attempting to break into the organization and gain unauthorized logical access. Hardcopies of network infrastructure configuration files usually contain sensitive information that can be used in planning a logical attack. One example would be a router configuration file, which would typically show packet filtering rule bases and encrypted passwords that can be used to gain access to the device. It often assumed that passwords shown in an encrypted form in configuration files present no threat. Unfortunately, this is far from true, and many older routers utilize weak encryption algorithms that allow passwords to be decrypted[4] with a modern PC in a matter of seconds. Controls and procedures to mitigate physical interception are considered in section 7.3.2.

Logical interception usually requires some degree of technical skill in conjunction with knowledge of the targeted organization's network topology. Note that interception can occur at any point along the path over which data is transmitted; hence, it can occur outside as well as inside organizations. Software based packet sniffing programs have already been discussed in chapter 5. These are used to perform the computer network version of a telephone wiretap. While logical interception may pose technical challenges, successful attempts can yield extremely valuable data, justifying any additional efforts at security. Examples include data transmitted over a network in clear text, such as valid username and password combinations, e-mail, and in some cases credit card details. A packet sniffer can also be used with TCP/IP hijacking software to hijack an existing data connection without having to supply a username/password combination.

7.2.3 Misuse (Unauthorized Access & Inappropriate Usage)

Two types of misuse can occur: The first involves using IT services in a normal manner for an unauthorized purpose. The second, and more serious, involves gaining unauthorized logical access to machines within the organization for nefarious purposes.

Normal usage of IT services for unauthorized purposes can include:

- **Viewing illegal or inappropriate material on the Internet.** When Internet access is given to employees, there is always a temptation to use it for personal reasons. Internet 'surfing' at work can have an impact on the organization's productivity, and may have legal implications, considered in section 7.2.5

- **Using the organization's email account to send content to the Internet that doesn't reflect the organizations values or opinions.** Many employees use their organizational e-mail account for personal use, such as e-mailing friends working for other organizations and posting comments and opinions to newsgroups. This practice can lead to problems if employees are unaware that e-mail content may be admissible in a court of law.

- **Third party usage of network services.** Web server software often comes with an anonymous FTP server that is enabled by default. In cases where these servers are visible from the Internet, they can be used as proxies for the dissemination of illegal material. For example, an internet user could upload pirate software to the FTP server and then post a message to a newsgroup informing other users where to download the software, thus hiding the true source of the software.

- **The Introduction of a virus.** In most cases this is unintentional, and the introduction is usually caused by using floppy disks contaminated by a third party, or by opening e-mail attachments from unsolicited sources. One recent example is the "Love bug" virus[5], which is transmitted via e-mail attachments and has caused problems for organizations worldwide[6]. Virus outbreaks can be devastating and cause excessive amounts of revenue to be lost. While the exact yearly global cost of virus outbreaks is not known, it has been estimated to be in the order of 1.6 trillion US dollars[7].

Unauthorized logical access can be gained through the usage of hacking techniques perpetrated by external third parties or disgruntled/bored employees. Threats and problems include:

- **Unauthorized access not being detected.** Unauthorized access is a major problem; however, the situation can be far worse if it is not detected in

a timely fashion. If a smart intruder gains access to an organization's network, they are more likely to monitor data traffic than cause damage that will be noticed straight away, drawing attention to their activities. Intrusions can be missed if inadequate monitoring and logging tools are in place. Even when such measures are in place, there is no guarantee that log files are reviewed[8] or that monitoring tools are configured properly. Consideration also has to be given to staff training, teaching them how to recognize an intrusion. For details of controls pertaining to logical security issues, see section 7.3.3

- **Damage to reputation.** A breach of logical security can draw adverse publicity to an organization. Cases include hacked web sites, loss of credit card information, and the enumeration of sensitive data. Consider the following case of a hacked website: If a high profile organization publishes an informational web page that becomes defaced during a logical attack, it is bound to receive a lot of media attention. While the information contained on the breached system might have no value, the publicity is likely to cause a public perception that the organization is not logically secure, meaning they will be wary of performing transactions with that organization. More details on defaced websites can be found in chapter 5.

7.2.4 System failure

The impact of a system failure is usually measured by the organization's dependence on its IT services and how long, if at all, it can function without them. Organizations that are highly dependent on their IT services often adopt a redundancy strategy so the impact of failure can be minimized. The strategy can include having two systems operating simultaneously, or having another system ready that can be used as a backup on short notice. Any of the following reasons may cause system failure:

- **Hardware failure.** Sometimes a whole IT system can become unavailable due to the failure of a single piece of hardware, often caused by the failure of an application server or a piece of network infrastructure. While many organizations address this issue by having multiple/backup application servers for critical systems, network infrastructure is often overlooked. Infrastructure devices such as routers are often single points of failure (i.e., all data traffic between two points has to pass through one device), and, because of their high cost, spares are not kept on site. In such cases, a considerable service outage could be experienced if a critical router were to burn out with no disaster recovery facilities being available.

- **Lack of change control.** Administration of a live production system requires good change control procedures to be in place to prevent system outages. Changes made to a production environment without being tested can result in software crashes, followed by time consuming restore processes needed to bring the system back to its original state. Other problems can occur if changes are inadequately documented, resulting in different system components being configured to incompatible specifications. For more details on change control see section 7.3.4.

- **Bugs and glitches.** Typically software contains several bugs when it is first developed. Bugs are often discovered when software is in use in a production environment, and can lead to unstable systems, errors, data corruption, and crashes. Glitches occur because of the limitations of hardware and software. One of the most well known examples is the Y2K problem, which was caused by using two digits to store the current year. Glitches can always occur if consideration hasn't been given to the all problems that might arise in the intended lifetime of a system. The next well known glitch will occur due to problems with the year 2038. For more details see section 7.3.5.

- **Denial of service attacks.** It is often easier to disrupt the operation of an IT system than it is to attempt to gain unauthorized logical access. A number of software-based techniques can be used to create a network-based denial of service attack. A denial of service occurs when enough data traffic is generated to fill all of the available network bandwidth that the system uses (for more details, see chapter 4). The net effect is there is no bandwidth available for legitimate services because all of the network bandwidth has been flooded. Organizations with inadequate network security can also be used as a proxy site to conduct distributed denial of service attacks against other organizations. More details on the latest denial of service issues can be found by visiting the SANS[9] Institute's web site.

- **Capacity and expansion problems.** Capacity problems are not dissimilar from denial of service attacks. Systems are often designed with a certain usage capacity in mind and no provision is made for future expansion. With the growth of the Internet many organizations have found their systems overwhelmed by data traffic demand, resulting in slower and slower response times. In some cases where the system was not designed with scalability in mind, the only option is to implement a new system. Slow response times can also be attributed to network bottlenecks; however, purchasing more bandwidth can easily rectify this. For more details on capacity planning issues, see section 7.3.5.

- **Theft and damage of equipment.** If components belonging to critical systems are stolen or damaged (i.e. by fire or flooding) there will clearly be a system failure. When physical security over an organization's premises is poor, there is a temptation for thieves to gain access and steal equipment. In some cases production systems are found on open floors with no security measures to protect them. Damage to system components can also occur if they are not housed in a suitable room with good environmental controls. Examples of potential problems include overheating, power surges, and staff unplugging the power by mistake. Physical security and environmental controls are discussed in section 7.3.2.

7.2.5 Legal/Regulatory Action

In addition to regulatory action, organizations can face legal action from employees as well as external third parties. Some common examples are summarized below:

- **Offended Employees.** Individuals in many countries have the right not to be subjected to abuse such as sexual or racial harassment in the work place. In some organizations, where acceptable usage policies are not enforced, inappropriate material such as jokes and pictures are disseminated by e-mail with no regard to the sensitivities of the recipients. Failure or unwillingness to prevent this activity can lead to employees taking legal action against the organization concerned. Also consider that third parties that may be offended – i.e., members of an IT helpdesk might be asked to repair an employee's computer and find hard core pornography stored on the hard disk.

- **Inappropriate Usage.** Although we have covered one aspect of inappropriate usage above, there are other problems such as abuse of privileged information that can lead to regulatory action as well as legal action. For example, how much due care does a financial auditing organization need to take to ensure that its employees are not conducting day trading over the Internet based on privileged information? When detailed content monitoring of Internet traffic is conducted, an organization will be not be able to deny knowledge of its employees activities.

- **Copyright Infringement.** Software licenses are expensive; however, fines for using unlicensed software are far greater, and can lead to adverse publicity. In large organizations care has to be taken to ensure that all installations of software are licensed and that the licenses can be shown upon demand. Software licensing can even cause a headache for the most honest organizations. For example, disgruntled employees have been known to

report their employer to a local software protection agency, even though no unlicensed software was actually in use, with the objective of causing maximum disruption to the organization concerned.

- **Inadequate record keeping and obstructing the course of justice.** Some organizations, depending on their nature, are legally required to keep records for a set amount of time. Failure to ensure that backups of critical data are made can result in the loss of vital records and could obstruct the course of justice if the records were required for legal proceedings.

7.2.6 Actors

Incidents can be separated in to two categories: those that occur unintentionally or by accident; and, those that are deliberate and clearly premeditated. Unintentional or accidental incidents are generally well understood and generally require no further explanation; however, this is not always true for incidents that were deliberate or premeditated. Many organizations discover that their logical security has been breached, but are never able to identify the intruder, their motives, or if they were acting alone. There is no easy solution to this problem, so in this section we will discuss the type of techniques that are likely to be used by different type of actors:

- **Hackers**[10] are most likely to break into a system for recreational purposes or as a challenge. In general they act alone or in small groups and are not likely to cause malicious damage.

- **Script Kiddie** is a term used to refer to juvenile or inexperienced hackers. Often they download automated hacking tools and run them against any target hoping to be able gain unauthorized access by chance. They can often be identified through network traffic analysis – for example, they might use a Unix hacking tools against a Windows NT machine. When they do gain unauthorized access, they are most likely to cause some damage and draw attention to their intrusion.

- **Cracker** is the term that is generally given to malicious hackers. These individuals are likely to use sophisticated techniques to gain unauthorized access with the specific intent of causing malicious damage (such as defacing a web page), or stealing confidential information. Criminal groups and activists could employ a cracker for other nefarious purposes.

- **State sponsored hackers** are likely to have better tools at their disposal than any of the actors listed above. In addition, the tools used are likely to have been thoroughly tested so that their impact on state of the art intrusion detection systems will be known. This type of actor is more likely

to subtly establish a backdoor into a system, and use it for long term monitoring or espionage. Likely targets will also include other states' national infrastructures. For more details on this subject, see chapter 8.

- **Fired or disgruntled employees.** It is estimated that approximately 70 to 80 percent of security breaches are internal[11]. Given these statistics, disgruntled or fired employees are a cause for concern. Problems can arise when remote network access facilities are not terminated for users at the same time as their employment. Individuals/employees with network access and malicious intentions can wreak havoc in a number of ways. These include encrypting their data so that no one else can decode it, deleting data, running intrusive hacking and denial of service tools against critical infrastructure, and placing backdoors to the organization's systems for use at a later time.

- **Dishonest system administrators.** Abuse of trust by a system administrator can be very difficult to discover, especially in small organizations where small numbers of staff make segregation of duties impossible. System administrators have privileged access to IT systems. Accordingly, they can delete, view or alter data easily and without permission. Moreover, they can also cover their tracks by deleting relevant entries in log files so that no audit trail of their activities is recorded. For more details on audit trails and how they can be used in investigative techniques for law enforcement issues, see chapter 9.

7.3 Controls and Procedures to Manage Threats / Problems

After the organization has identified threats and problems that pose unacceptable risks, the next step in the risk management process is to ensure appropriate control measures exist to mitigate the identified issues. Since control measures vary with the organization, this section covers typical IT controls and security issues an organization should ideally have in place. These include physical security and environmental controls, security policies, logical security, documentation, change management, and continuity/disaster recovery planning. In the logical security section we also discuss penetration testing services. Given recent increases of logical security breaches, this service deserves special attention as it can often provide organizations with a reasonable assurance that their logical security is adequate and IT controls are working properly.

7.3.1 Security Policies

Organizational Security and End User Computing Policies

An organizational security policy should be sponsored by senior management and provide a baseline for enforcing security procedures and controls across the organization. Ideally, a security officer should be identified and responsibility assigned to conduct monitoring and enforcement tasks. The security policy should reflect the organization's culture and values and take the following items into account:

- Acceptable usage policies;
- The introduction of unauthorized/unlicensed software;
- Logical security (see operating system security standards - section 7.3.3.2);
- Data ownership;
- Segregation of duties;
- Monitoring and escalation procedures; and
- Anti virus controls.

Once an organization's security policy has been created it should be distributed to staff and updated on a regular basis. In order to enforce the policy and allow escalation procedures, such as disciplinary action, it is advisable to get employees to sign an end user computing agreement (EUC) that explains the organization's security policy. By signing an EUC, an employee states that they are aware of the security policy and that they agree to abide by its terms and conditions. Breaches of the security policy can then be followed up formally.

7.3.2 Physical Security and Environmental Controls

Good physical security controls can prevent theft of equipment and restrict access to systems containing sensitive data. An initial barrier should come up when attempting to gain access to the organization's building. For example, many organizations have started using magnetic swipe cards and, in some instances, biometric user authentication devices. Once in the building, access to critical servers, network infrastructure equipment, sensitive IT documentation, and magnetic storage media should be even further restricted.

It is accepted best practice to house critical systems and servers in a dedicated computer room with good environmental controls. Most organizations do this, allowing only key IT staff physical access to the room. Appropriate environmental controls include air conditioning, uninterruptible power supplies and fire

detection/extinguishing equipment. Where water is used to extinguish fires, ensure electrical power is cut before the water is released. Magnetic storage media (such as backup tapes) should be secured in fire/waterproof safe. Also, consider the physical security of any data storage media stored off-site.

Network infrastructure is often more difficult to secure because it is, by definition, distributed. Devices such as routers need to be physically secured to prevent an intruder from gaining console access, from which the devices can be reset. Unused network sockets can also pose a problem due to the ease of packet sniffing. When sensitive data traffic is not encrypted, one can physically disconnect unused network sockets from the network to lower the chances of packet sniffing from an on-site intruder.

7.3.3 Logical Security

7.3.3.1 User Access Levels and Privileged Accounts

Data needs to be classified by confidentiality levels and access to it should be restricted to relevant user groups. This can be achieved by assigning a unique username/password combination to each user, and profiling users into groups with different data access levels appropriate to their positions. For example, some users may only require read access to the data, while others may need read/write access. In the latter case, write access can be monitored by reviewing system log files. Special attention needs to be given to privileged or administrative accounts because they usually have unrestricted access to all data contained in the system. In these cases it is essential that the allocation of privileged accounts is documented and monitored, and that passwords are supplied to senior management for backup purposes.

When users choose passwords they need to be made aware that trivial passwords can be guessed easily and that passwords must be changed on a regular basis. Password policies can be enforced by some operating systems and accounts can be locked out if an invalid password is given more than three times. Administrative staff can also use password-cracking tools to discover weak or easily guessable passwords. Finally, user accounts should be documented and the list should be compared with a human resources list to check that all accounts belong to live employees.

7.3.3.2 Operating System Security Standards

Many operating systems come with insecure default settings that are not appropriate for all organizations. The same can be true for programmable hard-

ware/firmware, such as routers, where older devices come pre-configured with default passwords. An organization can develop operating system security standards such that hosts can be hardened, from a security perspective, to reflect the organization's overall security policy. Obviously, separate security standards will be required for each type of operating system in use.

To create an operating system security standard, many different areas need to be taken in to account, including:

- **The business process needs of the organization**. For any security standard to be effective it must accommodate the needs of the organizational processes it is designed to protect. Clearly, there's not much point in having a security standard that is so strict the organization can't carry out its work.

- **Logical access permission to files and directories**. Logical access to files and directories needs to be categorized by user needs and data confidentiality. Once user needs and data confidentiality has been assessed, logical user domains can be used to restrict data access as deemed appropriate.

- **Network access and the provision for warning banners**. Many operating systems come with a variety of network services enabled. Common examples include Telnet, FTP, SMTP, and SNMP daemons. In some operating systems these services come with default passwords that can be used to gain network access to the machine. Ideally, all unnecessary services should be disabled and all default passwords should be removed from the remaining services. Warning banners should also be added to these services, warning any network user that unauthorised use is prohibited.

- **Operating system patches**. Vendors release operating system patches periodically to fix known vulnerabilities, and it is important to keep up to date with the latest patch level so that vulnerabilities are fixed before anyone has the chance to exploit them.

- **Subscribing to security alert services**. Subscribing to a security alert service allows an organization to be kept aware of any relevant operating system patches or vulnerabilities as soon as they become available.

- **Monitoring log files for inappropriate activity and system errors**. Monitoring log files allows an administrator to become aware of, and fix, any system stability issues before problems arise. Log files also reveal if any employee or intruder has tried to compromise operating system security, or access data that they do not have permission to view.

- **Physical security.** As with key network infrastructure devices, some servers need to have good physical security. Console access to critical devices should be restricted as tightly as possible to ensure the data inside cannot be physically or logically obtained or damaged.

- **The practicality of enforcing the policy**. Operating system security policies needs to be practical and manageable. Some of the issues that might arise are: how one reviews all log files in a very large environment with limited IT staff, and how one reviews logical access to files and directories when there are several thousand on each machine. Technology has come to the aid of these problems in the form of computer assisted audit tools, and there are now software programs designed especially for enforcing operating system security standards[12]. The organization needs to define a security baseline; then the tools can be run on machines within the environment to see if they are in compliance. Similar tools are now in use in large organizations to monitor key log files and e-mail the system administrator if anything out of the ordinary happens.

7.3.3.3 Penetration Testing

Penetration testing is a form ethical hacking, where an organization specifically requests and authorizes known individuals to attempt to gain unauthorized logical access in to their IT infrastructure. By conducting a penetration test, an organization can simulate what would happen if a real hacker tried to attack their environment in a controlled manner. Penetration testing is becoming increasingly popular for a number of reasons.

- Firstly, any weaknesses and vulnerabilities found can be fixed immediately before anyone else has the chance to exploit them and cause real damage.

- Secondly, the testing can independently confirm the adequacy of existing controls in place and reveal any additional vulnerability not previously identified.

- The number of reported computer security breaches is increasing, and many organizations conduct testing to minimize the chances of being attacked and suffering from adverse publicity.

- Some organizations may suspect that there are problems in their IT department – i.e. that their staff is not control conscious or fully aware of the importance of current IT security issues. In these instances the results from a successful penetration test can be used to raise awareness amongst key IT staff.

- Due to previous problems and audits, many organizations are implementing sophisticated monitoring and intrusion detection systems[13] (IDS). Once these systems have been put in place their effectiveness can be put to the test through penetration testing. Staff can also learn how to recognize and react to an attack.

Before intrusive penetration testing can be conducted, relevant background information such as domain names and IP addresses of target systems needs to be gathered. This information can be provided by the organization, or a penetration testing team can attempt to discover it independently using non-intrusive testing techniques. Non-intrusive testing techniques involve IP address discovery through the use of domain service lookup tools, network topology mapping through the use of ping and traceroute services, and the gathering of any other relevant information that might be available. In the latter case, examples include searching technical news groups on the Internet to see if system administrators have posted technical problems using their organizational e-mail address. This can often yield firewall type and version information, making conducting intrusive testing easier.

If an external party is used to conduct intrusive penetration testing, it is most likely that they will ask the organization to sign a formal letter of authorization, which will indemnify the external party form any damage caused during testing. Some intrusive techniques can have an adverse affect on network performance, and some tests can cause machines to crash. Accordingly, penetration testing is often conducted outside normal working hours to minimize the potential impact to the organization. Unauthorized third parties may also be attacking the organization at the same time as the penetration testing is being conducted. Clearly if malicious damage is caused at the same time as the penetration testing is carried out, it is likely the penetration testing team will be blamed. For this reason it is vital that all penetration testing teams obtain a signed letter of authorization before testing is conducted, and that they keep detailed technical log files of their activities presentable for forensic examination.

External penetration testing teams can be expensive, and it is important to get added value wherever possible. One approach is to get employees to shadow the testing team. Shadowing can raise awareness of security issues and has the added benefit that some degree of skills transfer can occur. Detailed technical log files of all testing activity can also be requested. This should ensure that external parties are thorough in their testing and provides proof that work has, in fact, been carried out. In some cases Internet penetration tests can reveal no technical findings, and the technical log files will be the only item deliverable to the organization. Some high profile organizations have penetration tests conducted on a quarterly basis. In these instances engaging two external parties

to conduct the first test and then awarding the on-going work to the party providing the better quality service can lead to added value.

Penetration testing is most effective when a variety of tools, along with manual automated techniques are used. The use of such tools and techniques has already been discussed at length in Chapter 3 and will not be discussed any further here. The table below shows the purpose behind some well known tools used in penetration testing. Details of these and other tools, along with usage instructions, can be found in a book entitled Hacking Exposed[14]. This book also has a companion web site[15], where the tools detailed in the book and the table below can be downloaded.

Purpose	Tool
Gathering background information	Web browsers
Non-intrusive testing	Sam Spade
Operating System fingerprinting	NMAP
Port Scanning	NMAP, Pinger
Firewall/router rule base testing	Hping2
OS vulnerability scanner (Commercial)	Internet Scanner (ISS)
OS vulnerability scanner (Linux based)	Nessus
Packet sniffing	Sniffit
Good multipurpose tool	netcat

7.3.3.4 External Connectivity

An organization should be logically secured against unauthorized intrusion from external connections. Common external connections include Internet connectivity and dial on demand ISDN lines for hardware vendor support. In these cases preventative measures are usually taken by employing a firewall to restrict logical access to specified devices. When firewalls are in place, the organization needs to assign responsibility for monitoring log files, and provide staff with escalation procedures to follow in the event of an unauthorized intrusion.

Some users have desktop modems (i.e. in a laptop) and it can be difficult to stop analogue telephone lines being used to establish dial-up connections. One approach is to document all instances of analogue lines and replace them with digital lines wherever possible. Also consider the security over any remote access servers the organization might have. Note that although some

organizations don't provide their staff with remote access, IT managers often permit remote access to support staff that are on call.

For organizations that are highly dependent on their IT, potential points of single failure need to be identified and consideration should be given to measures that can be used to prevent a denial of service attack. Similarly, organizations owning highly sensitive data should consider implementing an intrusion detection system as an additional security measure.

7.3.3.5 Monitoring and Escalation Procedures

System monitoring can be enabled on most operating systems and network infrastructure, and it should always be enabled unless it has an impact on system performance. System log files should be produced and reviewed on a daily basis for critical devices. Any suspicious activity should be escalated to the organization's security officer for follow-up. Key staff should also be aware of the organization's incident response procedure in the event of an intrusion being detected.

Log files can also provide useful information for capacity and acceptable usage monitoring. In the former case, statistics such as CPU usage, network performance and available disk space can be used to predict system bottlenecks well in advance. In the latter case, sophisticated monitoring software[16] is now available that allows the automated monitoring of Web browsing usage and e-mail content[17]. This type of tool can be used to set a baseline that reflects the organization's acceptable usage policy. Exception reports can then be generated for users that breach the policy.

Finally, we come to virus monitoring and detection. With the recent spate of virus outbreaks it is clear that no organization is safe from a virus attack. Even when up to date anti-virus software has been installed, it is ineffective against new previously undiscovered virii. With a new virus nothing can be done to prevent infection; however, early detection is imperative. When a new virus is detected (often through unexplained performance statistics) the amount of damage can be limited if a containment plan can be put into effect quickly. This can involve isolating infected machines and disconnecting network segments.

7.3.4 Documentation and Managing Changes

Good system and network documentation along with detailed operating procedures facilitate smooth system running and provide information, which can minimize the impact of any changes. Organizational charts detailing job

descriptions and segregation of duties should exist to ensure the relevant parties are informed when changes are due to be made.

A formal change control procedure should be followed when changes to a system are made. Before the change is made, a change request form should be completed detailing the nature of the change and the impact that it will have on the organization. Once the form has been filled out, it should be approved by key IT management staff and then stored in a central location, thus allowing a detailed inventory of changes. When approval has been granted and the change has been made, any relevant documentation should be updated to reflect the changes.

An effective change control procedure will also need to recognize the difference between test and production environments: Where test environments exist, changes should always be made to the test environment first.

7.3.5 Continuity and Disaster Recovery Planning

Planning for the future is especially important where IT systems are concerned. General improvements in technology and fabrication processes have allowed processing power to approximately double every year, while costs have fallen. Application software is also constantly being improved to take advantage of new technology, offering increased functionality and ease of use. Many organizations address these issues by staggering their hardware purchases so that a combination of new and old technology is always in use. In general, hardware components are usually kept for three years before being replaced.

Continuity planning involves more than just planning for the future. Problems arise in the day to day running of systems, and organizations need to have procedures in place to ensure continuity in the event of a disaster.

Examples of unexpected problems include bugs, glitches, and capacity problems. Capacity monitoring (see section 7.3.3.5) should provide an early warning when a system is close to using up all of its processing power, network bandwidth, or data storage facilities. This warning can then justify additional expense for planning to expand or replace the system. As we hinted in section 7.2.4, one of the next glitches will occur in the year 2038. This problem, like the year 2000 problem, is another time related glitch. It is caused because the C/C++ programming language uses a signed integer (32 bits) to store the current value of the time elapsed since the beginning of 1970. This integer value can only hold enough seconds to take us to the year 2038, after which the date will either revert to 1970 or 1901. Although 2038 may seem like a long time away,

some organizations (especially financial) may be affected by this problem in the nearer future and have already started planning necessary upgrades.

In the event of a disaster, an organization may lose or have to vacate its premises. When IT services are critical to an organizations survival, it is prudent to have a second production system located at another site. This could either be a warm backup site (where the data from the production site is mirrored, allowing the site to become operational quickly), or a second production site that shares the system load. In addition to a backup system, detailed disaster recovery plans and procedures should be created. Disaster recovery plans should identify critical systems and detail how they can be recovered and operated at the backup site. Plans should also be tested to ensure they work and to ensure employees are capable of performing the restore operations.

Data backup procedures are a critical part of any disaster recovery plan and can be used to backup anything from an individual machine to a whole production environment. Organizations should formally document required backup procedures and assign responsibility to relevant members of staff to ensure the process is carried out. The type and frequency of backups will vary from case to case; however, it is vitally important that backups are stored off-site as well as on-site, thus catering for a disaster that could destroy the main production site. It is also critical to ensure all backups are tested to ensure archived data can actually be restored from the backup media.

7.4 Conclusion

Increased computing power in conjunction with lower costs means more data is being stored in IT systems. At the same time, global connectivity and Internet growth are offering unprecedented cost savings and productivity opportunities to organizations, especially in the e-commerce arena. While these aspects of information technology are positive, the bad news is that the increasing frequency of logical security breaches means no organization is safe.

No IT system will ever be 100 percent secure. It would be unrealistic to hope for that. We have, however, in this chapter, shown that risks can be *managed* if a responsible approach is taken and auditing is performed on a periodic basis. In section 7.2 we saw how threats/problems can be ranked by risk. Controls and measures to mitigate risks were then discussed in section 7.3. By using the methodology outlined in the introduction, organizations can identify missing controls and develop a plan for improvement.

Information risk management will mean different things for different organizations and some gray areas will always exist. One such area is defining what is

considered to be an acceptable risk. This subject was not discussed in this chapter because it will clearly be different for each organization; however, an acceptable risk threshold is likely to be defined by a trade-off between the severity of the risk and the investment required to mitigate it.

Ultimately, successful information risk management will be dependent on the quality of the organization's threat/problem risk assessment. While the assessment can be carried out directly by the organization, there are clearly advantages in having external consultants help facilitate this process — i.e. advantages such as knowledge of best practices in other organizations, and resources with wider technical knowledge than in-house staff.

Information risk management will be playing an increasingly important role in the future.

1 Avoiding A Data Crunch, Jon William Toigo, Special Industry Report – Scientific American, May 2000.

2 For more details and guidelines of backup procedures see section 3.5 (Disaster Recovery and Business Continuity Planning).

3 In one case a small bulletin board Service (BBS) in Viginia (USA) was mistakenly sent e-mail over a 9-month period that was destined for a Spanish bank. For more details see: Private bank e-mail goes awry, Bob Sullivan, MSNBC.com (Technology), 6 July 2000.

4 For example, weakly encrypted passwords on Cisco routers can be decrypted with Solar Winds' Cisco Password Decryptor tool. Note that this tool is part of a suite of network administration tools that can be run on windows NT. For more details visit http://www.solarwinds.net

5 See Experts Estimate Damages in the Billions for Bug, Paul Festa and Joe Wilcox, CNET Enterprise Computing News, May 5th 2000.

6 This virus is especially virulent, and was spread around the world by infected international organizations in only a few hours.

7 Study Finds Computer Viruses and Hacking Take $1.6 Trillion Toll on Worldwide Economy, Excite.com (news), July 7, 2000.

8 This can often be caused due to the lack of a detailed job description that documents functions and responsibilities. For more details see section 3.2.

9 The SANS Institute is a cooperative research and education organization for system administrators, security professionals, and network administrators. The web site can be found at http://www.sans.org.

10 In this case we refer to the traditional relatively harmless type of hacker as opposed the malicious variety which are also known as crackers.

11 Threat to Corporate Computers is Often Enemy Within, Peter H. Lewis, The New York Times on the Web, March 2nd 1998.

12 Examples of such tools include Enterprise Security Manager (ESM) from Axent Technologies, and Bindview – http://www.bindview.com.

13 One well-known IDS system is Real Secure from Internet Security Systems (ISS). Their website also contains useful technical reference papers that can provide a batter understanding of intrusion detection systems (http://www.iss.net).

14 Hacking Exposed: Network Security Secrets & Solutions, McClure, Scambray &Kurtz, Osbourne/McGraw-Hill, 1999.

15 See http://www.hackingexposed.com.

16 See for example, Internet Monitoring and Management Tools from Webspy (http://webspy.com)

17 A number of US based companies and organizations fired employees last year for inappropriate e-mail and WWW usage. See for example: More employers taking advantage of new cyber-surveillance software, Wolf Blitzer, CNN.com (US news), July 10, 2000.

Chapter 8 - Infrastructure's Dependence and Interdependence on Technology

8.1 Introduction

National infrastructures have traditionally been exposed to risks and interdependencies that are now relatively well understood by governments and military organizations. This is now changing due to the technological advances, standardization and globalization processes that we have seen over the last few years. Critical utility and information infrastructures such as energy supply systems, telecommunication networks and financial networks are now operated and controlled by computer networks. Originally, private and proprietary computer networks were used, but now the evolution of the Internet has led to the development of standardized networking that uses cheap commercial off-the-shelf (COTS) products, which have known weaknesses and vulnerabilities.

Throughout this chapter single and critical points of failure are a recurring theme. Not only do they exist within and between different types of infrastructure, but they also exist in the computer networks that are used to operate and control the same infrastructure. For this reason a lot of attention needs to be given to the different types of vulnerabilities that exist in computer networks such as the Internet. While critical energy supply systems may rely on computer networks for their control, it is also important to remember that a computer network and its underlying infrastructure is also dependent on electricity for its operation. Moreover, computer networks generally rely on the usage of high-speed data backbones that are supplied by large telecommunication companies. As different types of infrastructure are examined more closely, it quickly becomes apparent that they are heavily dependent on each other. Effectively, this means that attacking one type of infrastructure will have a knock-on effect for another. For example, a sustained loss of power to a computer network controlling a gas pipeline would probably result in the pipeline being shut down before all reserve power was exhausted. This same pipeline could be used to supply gas-powered power stations, which in turn generate electricity. Computer networks are especially interesting because they can be exploited in two ways. Firstly, a denial of service can be created preventing communication; and secondly, they can be used as a tool to shut down or disrupt other types of infrastructure.

Because of the inherent interdependency existing today, infrastructure is becoming an attractive target for a number of different actors. We are also starting to see instances of weakness in the chain of interdependence through a number of well-publicized and unintentional incidents. Vulnerabilities exist

because of a number of factors, including the increasing commercialisation of many utilities. The result of this leads to the creation of potential single or critical points of failure [1]. If these can be successfully exploited, the vulnerability is increased further because of interdependence. In a worst-case scenario, if enough single or critical points of failure in a given nation's infrastructure could be targeted and exploited simultaneously, there could be a full breakdown of national infrastructure.

In the sections that follow, we begin by looking at what constitutes critical infrastructure. This definition will vary from one nation to another, so we begin by trying to define infrastructure items that should be included as a minimum. We also attempt to show how other infrastructure items can be assessed to determine if they should be added to the list. Section 8.3 examines some of the common threats to infrastructure, while section 8.4 takes a look at some of the different types of vulnerabilities that can be exploited in an attack. In general, the attacks fall into three categories: software attacks, physical attacks and attacking other infrastructure that the targeted service depends on. In this section a lot of attention is given to computer networks and electrical energy distribution systems. Special emphasis has been given here because so many other forms of infrastructure depend heavily on both of these items. Energy distribution systems and computer networks are also highly dependent on each other. In section 8.5 we briefly look at what preventative measures can be taken to protect infrastructure before presenting a conclusion in section 8.6.

8.2 Critical Infrastructure

In May 1998 the US government released a white paper entitled "The Clinton Administration's Policy on Critical Infrastructure Protection: Presidential Decision Directive 63" or PDD 63 [2]. This white paper recognized the increase of potential vulnerabilities in critical National Infrastructure (NI) and aimed to encourage a public-private partnership to help reduce these vulnerabilities. The paper included examining three steps, which we have listed below, as well as going on to propose making plans to detect and react to major attacks. The three steps are as follows:

1. Draw up an asset inventory of infrastructure with the potential to be included on the critical NI list. We will discuss guidelines for selection shortly in section 8.2.1.

2. Conduct a vulnerability assessment for the additional infrastructure identified in the step above. This subject has been covered to some extent in Chapter 7.

3. Perform a risk management analysis so that vulnerabilities identified in the step above can be minimized in a cost effective manner.

These three steps represent a strategy that can be followed so that weak points in critical infrastructure can be identified, ranked by priority and addressed. We begin looking at this strategy in section 8.2.1 below by considering what to list in a critical NI inventory. Note, that this is only one aspect of the security cycle, which is to "Protect, Detect and React". In this chapter our discussion is limited to protection; however, any good security system as a rule must also incorporate operational parts that detect and react.

8.2.1 Critical NI Inventory

There is no right or wrong way to determine what constitutes an addition to the NI inventory. However, some existing publicly available documents can be used to provide guidelines. In January 2000, the US Critical Infrastructure Assurance Office (CIAO) released a document that discussed practices for securing critical information assets [3]. It also specifically considered a "CIAO Infrastructure Asset Evaluation Survey", which can be used to help identify critical assets in the context of PDD 63. The survey included a checklist from several different domains, including:

- Evaluating the asset in terms of essential national security missions. This includes considering how such missions would be dependent on the asset and what the implications would be if the asset were to be lost.

- The value of the asset in helping to maintain order. PDD 63 requires identification of assets that help state and local governments maintain public order.

- Ensuring orderly functions of the national economy. For example, does the asset protect sensitive economic data, and could it be misused to undermine the economy?

- Maintaining general public health and safety. This domain considers assets that manage regulatory controls over dangerous substances and diseases with the objective of protecting the general public's well being.

- Delivery of minimum public services. This considers the role that assets play in delivering minimum public services that are mandated by law and that are needed to sustain general public welfare.

- Dependency of other government programs on the department/agency's asset. This considers governmental assets and the impact that other agencies would experience if the asset were to be lost.

- Ensuring the delivery of essential private sector services. This domain examines the extent that the asset supports essential private sector services and whether or not it contains sensitive information, data and technology.

Once a checklist of this kind has been applied to the assets of a national infrastructure it becomes easier to rank them in order of importance. This ranking then allows a cut-off line to be drawn at an agreed level. The assets that are above the cut-off point in the list can then be put forward for a vulnerability assessment.

Generally, national infrastructure systems can be grouped into three main categories:

- **Basic Utilities**: These include all the items that we generally need in everyday life - i.e. national electricity supplies, oil / gas pipelines and storage, petrol refineries, water supplies, telecommunication networks, transport systems and guidance systems.

- **State Activities**: This would include government functions, government agencies (including the military), and national emergency services.

- **Commercial Activities**: This would include the networks from financial, business and news organizations.

As a minimum, the following should be identified in a NI list:

1. The national electrical distribution grid system and key power stations.
2. Major national telecommunication providers.
3. Gas and oil pipelines and distribution systems.
4. National water supplies.
5. Petrol and oil refineries.
6. Air traffic control and the Instrument Landing System (ILS).
7. Guidance Systems, including GPS.
8. Emergency Services (fire, police, ambulance etc).
9. Hospitals.
10. Financial networks – i.e. stock exchanges and EFT systems.
11. Media and entertainment networks (TV, satellite, radio and the Internet).

12. Wireless telecommunications, including GSM and satellite telephones.
13. Government networks for communication with key civil agencies.
14. Military command, control and communication networks.
15. Educational and research agencies.
16. Key private business networks.

It is worth noting that computer networks now play a key role in almost all areas of infrastructure listed above and it is for this reason that National Information Infrastructure (NII) is such an important subset of NI. Another interesting point is that, although the Internet has become commercialised, it is used by organizations in all three categories listed above. Therefore the Internet can be considered part of a Global Information Infrastructure (GII).

8.2.2 Vulnerability Assessment

This step involves finding and documenting vulnerabilities in critical assets. At this point it should be recognized that this can be a time consuming task and the assistance of a task force of industry experts will be required if this step is to be carried out properly. In October 1998, the Critical Infrastructure Assurance office commissioned the development of a Vulnerability Assessment Framework [4]. This document provides guidelines for gathering data, assembling appropriate expert task forces and assessing and prioritizing vulnerabilities.

8.2.3 Risk Management Analysis

After vulnerabilities have been identified, nations can perform a risk management analysis for each vulnerability associated with their critical assets and infrastructure dependencies. During this analysis, attention is usually paid to both threats (8.3) and vulnerabilities (8.4). It is likely that the vulnerability list will be long and that there are insufficient financial and manpower resources to address all of the issues at once. Improvements can be prioritized from the results of the risk management analysis. This will probably show that there are some "quick wins" that can be addressed immediately, quickly and with little cost while offering significant improvements. The remaining vulnerabilities can then be addressed in a manner appropriate to the available resources.

8.3 Threats

Threats posed against critical information and infrastructure vulnerabilities can be generally classified into two categories: those that occur naturally as result of an accident (Y2K, for example) or national disaster, and those that occur intentionally from a variety of actors. Common threats include the following:

1. **Terrorism**: Terrorists could exploit vulnerabilities in critical infrastructure by purchasing low cost commercial off-the-shelf computer equipment. For example, they might roam freely carrying a laptop and mobile telephone and use the equipment to attack computer networks responsible for switching a national electricity grid. This type of attack will become increasingly attractive to terrorists because of the potential to cause complete chaos and generate a lot of publicity without necessarily taking lives. Additional advantages come from the fact that the perpetrators expose themselves less and are less likely to be apprehended. Because of advances in telecommunication technologies these types of attacks can even be carried out from almost any location globally, making it extremely difficult to apprehend the perpetrators (if they can be identified) because of national jurisdictional boundaries. Note that the use of high explosives would also be an option for terrorists. While this puts the terrorist at greater risk of being apprehended, there is a more devastating and longer lasting effect after the attack. For example, if a building is destroyed and replacement equipment is available, it will still take some time for a suitable replacement building to be found.

2. **Information Operations (IO)**: As technology has advanced, there has been a revolution in military affairs. Information warfare techniques have been developed to attack both military and civilian systems. IO techniques can be used in an attempt to cripple a targeted nation's critical infrastructure, and their effect will depend on the extent the nation relies on its infrastructure and the measures it has taken to protect it. This means that developing nations will benefit from mastering IO techniques for two reasons: firstly, the cost of obtaining necessary equipment and tools is low; and secondly, they are not as dependent on their own infrastructure and counter attacks will not be very effective. For advanced nations, IO techniques are more likely to be used covertly in conjunction with special operations. Full IO attacks are less likely because retaliation is likely and the attack may be seen as a precursor to a physical invasion.

3. **Criminals**: These individuals would probably only attempt to disrupt critical infrastructure for objectives such as extortion. Exceptions might arise when the temporary disruption of critical infrastructure helps facilitate crime – for example, it would be far easier to rob a bank if there

was no way of alerting the police to the robbery. Consideration also needs to be given to hackers and "script kiddies" in this category. While these actors are not generally out to commit a specific crime, their actions are usually illegal. On the one hand, a hacker might probe (causing no damage) critical systems for purposes such as experimentation, while on the other hand script kiddies might disrupt or destroy critical systems for motives as simple as having fun.

4. **Major Bugs & Glitches**: The most notorious example of this type of threat was the Y2K bug. This threat has now been passed, but it does not mean similar types of threats will not occur again in the future. The Y2K threat was caused unintentionally through a programming oversight and it is likely that other unintentional threats will arise through oversights in the future.

5. **Disasters**: Examples include natural disaster such as fire, flooding, storm damage and earthquakes. These can threaten to physically destroy sites where critical infrastructure is housed. Locations where single or critical points of failure have been identified are particularly at high risk.

8.4 Vulnerabilities

All of the actors and entities described in the previous section pose threats to critical infrastructure. Yet, to pose a threat, information about the location and nature of vulnerabilities needs to be known and understood. In this section, we present an overview of some of the vulnerabilities that exist in key infrastructures such as computer networks, electrical distribution grids, gas and oil pipelines and telecommunication networks. Computer networks (and the Internet) are presented first and in considerable detail because of their fundamental importance. Remember that if vulnerabilities in computer networks can be exploited, so can all the other critical information systems that rely on them.

8.4.1 Computer Networks and the Internet

Although the Internet has only come into widespread use in the last 10 years, it has now become a standard medium for the transmission of personal data, multimedia and messaging. With the emergence of e-commerce technologies, some organizations are now completely dependent on the Internet's ability to deliver potential clients and new business. In general, organizations also tend to rely heavily on the Internet for internal and external communications.

The Internet has evolved considerably from the military-run ARPANET of the 1970's into the commercialized network that we have today. Originally, the ARPANET was designed with enough redundancy to continue operating in the event of potential nuclear attack. The same cannot be said of today's Internet, which is more anarchical and contains more critical points of failure. The US Department of Defense originally specified design constraints that would lead to a redundant and reliable network. Clearly, these constraints have not been observed and there may be several possible reasons for this. One such reason might be due to financial/budgetary constraints – i.e., the costs of the components required to implement new Internet connectivity were so high that redundancy was sacrificed to reduce the number of components required, hence reducing the total cost. Another reason could be that a more self-sufficient network could be incompatible with the business plans of the organizations implementing new, or upgrading existing, Internet connectivity.

Because of the widespread (including military) usage of the Internet, the impact of a partial or total Internet crash would be huge. To date there has not been a total Internet crash; however, there have been instances or both deliberate and unintentional partial failures resulting in short term outages. In the case of the former, some organizations recently experienced outages caused by the spread of Internet enabled viruses such as "Melissa" and "The Love Bug". These viruses were targeted at a particular e-mail/messaging software product. Within a few hours of this type of virus being released, affected organizations found their e-mail systems overwhelmed and inoperable. Large international organizations were particularly vulnerable to this sort of attack; not only did the virus spread extremely quickly, but it was also especially difficult to contain as geographically distributed branches kept re-infecting each other.

One example of an unintentional partial failure occurred on the 17th July 1997 when corrupt information was uploaded to the Internet's root domain servers [5]. The databases were corrupted to the extent that the top-level domains .com and .net were unable to be resolved. Until the problem was fixed it was not possible to send e-mail or browse these web sites unless the numerical IP address was known. Fortunately, in this case the problem was rectified in around four hours when valid information was re-uploaded to the servers.

In general, the attacks against the Internet depend on the ability to exploit single or critical points of failure. In principle, if enough critical points of failure could be identified and attacked simultaneously, it would be possible to crash most or large parts of the Internet [6]. In August 1997 an article was published on Hotwired.com that described 50 ways to crash the Internet [7]. Some of the main types of attack discussed included DNS attacks, router attacks, IP attacks, user level attacks and attacks based on weaknesses existent in other types of infrastructure. We will now look some of these in more detail.

8.4.1.1 DNS Attacks

The Domain Name System (DNS) allows humans to use user-friendly names such as http://www.company.com instead of having to remember a more complicated IP address such as 192.168.1.252. When a domain name is used, the user's computer (client) sends a message to a DNS server to resolve the IP address associated with the given name. Once an IP address has been returned, the client then uses the IP address to connect to the given site. The DNS resolution process is usually hidden from users, and accordingly most users would not even know the IP address of high profile web sites such as cnn.com. It is also worth remembering that some new users might not have even heard of an IP address.

DNS is a distributed database containing many servers in an inverted tree structure with a root [8] node at the top. This allows local DNS servers to control segments of the overall database, and at the same time, data from across the whole database is made available through client-server technology. Users are usually assigned two local DNS servers from their ISP, one acting as a primary server and the other as backup. When a DNS server cannot be reached, Internet applications can no longer resolve DNS names automatically. Effectively this means that the user either needs to know the IP address of another DNS server or the IP address of the site that they wish to visit. As we have already said, the likelihood of a user knowing the IP address of the site they wish to visit is low. Furthermore, if the user wants to determine a given IP address, they will need to find a working DNS server so that a DNS lookup [9] can be performed.

Failure of local DNS servers will cause a local outage either until the problem is rectified or until the users are given the IP address of a non-local DNS server. If the whole DNS system were to become unavailable, the whole Internet would become unavailable to most users. This is one of the Internet's main vulnerabilities and ways to exploit it include the following:

1. Attacking the base operating system of DNS servers and crashing the DNS server application software. This type of attack would tend to be localized, as it would be difficult to attack all DNS server hosts simultaneously.

2. Selective DNS entries could be forged to make certain strategically important sites disappear. This type of attack may take some time to discover, as the effect would not be immediately apparent. For example, users using a proxy server will still be able to view cached versions of affected sites web pages. It is also worth mentioning that this technique can be used to redirect users to "alternative" sites. For example, it would

be possible to make users think they are viewing a well-known news site when in actual fact they are viewing a pirate site with altered content.

3. Cause an Internet wide breakdown by corrupting the information database of the root level DNS servers. As we have already discussed, the effectiveness of this type of attack has already been proven when the .net and .com domains became temporarily unavailable after a corrupted database was unintentionally uploaded to a root domain server [10].

4. Conduct a distributed denial of service attack (DDoS) against the root and heavily used DNS servers. If this type of attack could be sustained for long enough, there would be a slow down in service and it would be difficult for law enforcement personnel to trace the perpetrators.

5. Root DNS servers are geographically distributed; however, some countries [11] do not have their own root DNS server [12]. There are 13 root servers and most of these are housed in the USA at military and educational sites. Notable exceptions include Britain, Sweden and Japan. There is a danger that some types of attack, such as DDoS and forging database entries, could be tailored to make entire countries "disappear" from the Internet.

The DNS system is a classic example of a standardized technology. Because the Internet is a global network, standardization is necessary, yet at the same time this leads to fundamental weaknesses that can be exploited. When DNS was designed, no one could have imagined it would be used on the scale that it is today. More importantly, it certainly was not designed to be a critical part of national infrastructure, or to withstand the types of attacks we have described above. Ultimately there may be a need for a new type of standardized DNS system that supports several different IP address resolution systems. The idea would be to produce enough redundancy that it would be very difficult to take all of the different systems down at once. It is also worth noting that organizations using the Internet for private communications, such as the military, could well have their own private equivalent of DNS that is far more robust and has additional security through obscurity.

8.4.1.2 Router Attacks

Routers perform packet switching and forwarding and represent one of the Internet's most fundamental building blocks. Originally, a distributed network of routers allowed traffic to be routed several different ways between any two given points. There was also a fair amount of variety in the brand of routers being used and no single manufacturer had attained market dominance. Today the commercialisation of the Internet has led to the development of high-speed

backbones that a large percentage of all Internet traffic passes through. Moreover, there are now fewer brands of routers in widespread usage and there are a small number of big players (such as Cisco Systems) supplying new hardware. These factors make routers attractive targets to groups and individuals wishing to disrupt Internet services. The following list summarizes some of the main ways that these devices can be exploited:

1. If key routers have not been adequately secured and administrative access can be gained, router configuration information could be changed to deny access to specific organizations; or it could be completely removed, rendering the router temporarily inoperable. When administration access is gained, the intruder can block access to the real administrator. In this case the real administrator would have to gain physical access to the router in order to perform a reset [13], which can take some time for routers that are normally administered remotely. Note, that while this type of attack may have been popular in the past, it is unlikely that it would be particularly effective today apart from the exception of isolated instances where the administration team was careless.

2. Routers use an operating system just like a computer. This is commonly referred to as an Internet Operating System or IOS. In other chapters we looked at the ways computer operating systems can be compromised through vulnerabilities and tools such as "root kits" and "backdoors". In theory, the same is true for routers as they are essentially a dedicated purpose computer running a specialized operating system. If a vulnerability can be found for a particular version of IOS, all routers running that software will be vulnerable. Serious vulnerabilities allowing administrative or equivalent access could be exploited in two ways. Firstly, affected routers could be locked and crashed causing a denial of service. Secondly, the routers could be changed subtly such that the change would go unnoticed for some time. This technique could be used for targeting individual organizations.

3. This exploit runs along the same lines as the previous exploit, but considers how to plant vulnerabilities in an IOS software version. One approach is to work for a company developing the software and add hidden functionality during the software code development. This clearly highlights the importance of carrying out background checks on employees working on such critical projects.

4. Some routers are more critical than others are. For example, loss of a router on a high-speed backbone will have a greater impact than that of losing a regional router that processes smaller amounts of data traffic. In

countries where the Internet has seen a lot of commercial development, it is likely there will be a small number of key routers providing the high speed and bandwidth that we have come to take for granted. If these key routers can be identified and taken out of service simultaneously, it follows that a widespread slowdown of Internet services would be experienced. As soon as the key routers became unavailable, traffic would be re-routed over slower segments that do not have the capacity to deal with the large amounts of traffic. For this reason one could consider these key routers as potential single or critical points of failure.

A Brief Hypothetical Example

If we suppose it was possible to incorporate a backdoor or "root kit" into a version of a common brand of router IOS, some interesting questions need to be answered. Firstly, what functionality should the backdoor have? Secondly, which routers should be identified for exploitation if the intention is to cause a widespread Internet crash?

One approach to designing a backdoor would be to build in functionality that can be used later and on demand. If the intention is to cause widespread denial of service, the extra functionality could be used to forward the data packets as normal, but with a blank or null payload. This method would still allow the perpetrators to use the existing high-speed network and remain in control of it.

To determine which routers are critical to any given nation requires extensive research or inside knowledge. In cases where the perpetrator has no inside knowledge, Internet based tools can be used to carry out some of the necessary research. One such tool is called "Bing", which is a variant of the well-known "Ping" tool. Bing [14] is a point-to-point bandwidth measurement tool that can be used to determine the raw throughput between any two given points in a connection. This type of tool could be used from a large number of geographically isolated locations to build statistics of traffic routing patterns and high-speed network segments.

8.4.1.3 IP Attacks

The Internet is controlled using an IP level message protocol called ICMP (Internet Control Message Protocol). This is used for purposes such as redirecting data packets and telling hosts the rate at which data packets should be transmitted. ICMP is easily abused. However, these types of attacks are now well understood and are less likely to have as significant impact as the type of attacks we have already discussed. These attacks are also likely to have an

impact that is localized rather than widespread, and for this reason the overview that we present here is brief.

ICMP messages can be faked for a number of purposes. Firstly, ICMP quench messages can be faked so that the target host sends out data packets more slowly. The effect is a slowdown in service. Secondly, "Host unreachable" messages can be faked making hosts wrongly believe the target with which they wish to communicate is unreachable. An obvious target for this type of attack would be a DNS server. Thirdly, ICMP redirect messages can be faked, forcing data traffic to take a route that would be longer than normal. The idea here is to create unnecessary extra traffic congestion.

8.4.1.4 User Level Attacks

Until recently, user level attacks did not pose a big threat. This has changed since the release of Internet enable viruses and the widespread propagation of public domain utility programs and scripts. Examples of user level attacks include the following:

1. Releasing public domain (freeware) software with hidden functionality on the Internet. It is fair to assume that if the user used the Internet to download the program, they must have Internet access. When pre-compiled software is downloaded and executed from an unknown source there is no way of knowing if the program will do what it claims to do until it is run. This makes it very easy to add hidden functionality that the user cannot see when the program is executed. Users are generally willing to take this risk because the software is free. One example of hidden functionality would be to periodically ping a prominent target web site. If enough users downloaded and ran the program at the same time, the target web site would experience bandwidth flooding from an excessive amount of ping requests, thus making the target unavailable to legitimate users,

2. The power of e-mail based viruses such as "Melissa" and "The Love bug" have already been demonstrated. This type of virus could be modified so that it propagates in the same way, yet causes every affected host to ping prominent DNS servers. Again, as in the previous example, the extra traffic would result in an effective denial of service and hinder Internet access.

8.4.1.5 Other Types of Attack

There are other forms of attack that are not as technical as those that have been described above. These generally involve turning attention to the services that the Internet is dependent on for its successful operation. The most obvious of these are the provision of electrical power, physical security and the telecommunication companies that operate key Internet backbone links. Methods to exploit these services include the following:

1. Identifying the physical location where key infrastructure representing single or critical points of failure is maintained and causing its destruction. For example, a terrorist group might use high explosives to destroy part of a high speed Internet backbone link. Other techniques could involve arson and vandalism. High-speed routers are expensive, and organizations maintaining backbone links are likely to carry a minimum inventory of spare devices due to the high cost and the speed at which the hardware technology is evolving. This means that in the case of physical destruction, new hardware will probably need to be ordered, resulting in a long service outage.

2. Internet infrastructure needs electricity to operate. Generally, we would expect key infrastructure to be protected by uninterruptible power supplies (UPS) and for network operators to have emergency power generating equipment. If the electrical supply can be disrupted, there will clearly be a serious outage until the supply can be restored. To achieve this, perpetrators would need to examine each instance individually and attempt to identify weaknesses and single or critical points of failure in the local electrical supply system. The concept of this type of attack has already been demonstrated in July 1997, when a major Internet router went offline due to a power failure [15]. This outage was apparently caused because the operator failed to provide an adequate amount of power protection. Once the device went down, Internet traffic was re-routed via other routers, which quickly became overloaded, causing parts of the network to become so slow they were effectively unusable.

3. Many of the new high speed Internet links use underground or submarine fiber optic cables, and are owned and operated by large telecommunication companies. These links can also represent single/critical points of failure and can be targeted accordingly. In some instances, telecommunication companies build all of their fiber optic cables into one large bandwidth trunk cable, with the intention of saving money. The downside of this practice is that, if the whole cable becomes severed or damaged, the provider will not have enough bandwidth available using alternative cabling. Cutting or severing a cable in a remote location could

be a particularly effective form of attack. Firstly, little physical security would be protecting the cable and secondly, it would take longer for the cable to be repaired. Accidental damage to fiber optic cables has already demonstrated this concept. In one instance, a work crew cut through a fiber optic trunk cable in the Californian desert running between Los Angeles and Las Vegas [16]. This incident led to the loss of approximately 500 high bandwidth (45Mbps) data traffic lines and slowed Internet traffic considerably.

8.4.2 Basic Utilities

8.4.2.1 Electrical Distribution Grids

In many countries the electricity supply industry has been privatised to encourage market competition and to ultimately pass on the resulting cost reductions to consumers and shareholders. In general, a national electricity grid purchases energy from power generating companies to meet the load demand of its users, which are usually regional electricity service providers or organizations such as state agencies and the military. The load on a national grid continuously fluctuates, depending on things such as the time of day, the time of year, weather conditions and end user habits. The latter item could be as trivial as 100,000 television viewers switching their electric kettles on simultaneously during a commercial break. To meet the energy demands placed on the national grid, the organization controlling it switches various different energy providers in and out of the grid in real time. This can even include purchasing power from energy sources in other countries. The full electricity supply chain is usually made up of the following infrastructure components:

1. **Power Generators:** These are the companies who are responsible for running and managing the power stations that generate the electrical energy. In many countries a variety of generating options exist, including coal, oil, gas and nuclear powered power stations, hydroelectric schemes, solar power and windmill farms. Power generators also compete with each other by offering competitive energy unit pricing to the organization running and controlling the national grid.

2. **National Grid Infrastructure**: These are the physical components that make up the national grid, including overhead pylons, high voltage cables, substations, switchgear and the computer network used to control the grid.

3. **National Grid Control Centres:** These are the organizations that control, manage and switch different power generators on and off the grid. They monitor the power generator prices and the usage demand placed on the grid and use a combination of the cheapest generators to meet the required energy demand. Because demand and pricing is constantly changing, a computer network is used to operate the grid switchgear in real time. This network is also used for monitoring and reporting power levels, infrastructure status and for performing fault diagnosis. In general, national grids usually have some degree of redundancy built in so that supplies can be re-routed in the event of a localized failure; however, some single or critical points of failure may still exist.

4. **Regional Service Providers:** These are usually small regional consumer energy providers. Typically, they will draw enough energy from the national grid to sustain the needs of their clients.

Electrical energy is perhaps one of the most important building blocks in today's society. We depend upon it to support almost all aspects of our every day lives, and most of us have already experienced the amount of inconvenience that even a minor power cut can cause. Even more importantly, other types of basic infrastructure are highly dependent on electrical energy and there would be a serious knock-on effect in the event of a sustained power outage. Examples would include:

- Not being able to pump fuel. Most petrol stations use electric pumps that would become inoperable. Cars and many forms of public transport would run out of fuel, quickly causing a breakdown in the national transport system. Cleary, this would have repercussions for food distribution networks.

- Many computer networks would be unavailable after a few hours. Backup power supplies protect most network infrastructure; however, these are only designed to serve short outages of a few hours. More critical networks may be served by backup generators designed to operate through longer outages. In these instances, networks can continue to operate as long as there is fuel to power the generators, which brings us back to the point above about pumping fuel.

- Telecommunication networks will have backup power supplies, but as described in the point above, they will only be able to function as long as there are enough fuel supplies.

- Electricity and computer networks are used in the control and operation of gas and oil pipelines. In some instances a large proportion of a country's

electrical energy may be generated by gas or oil fired power stations. In this situation the interdependency goes in a full circle.

Electrical distribution systems make attractive targets to a range of actors with different motives. For example, a terrorist organization might wish to disrupt a national grid to cause economic damage, whereas a military organization might want to disrupt the system as a precursor to an invasion. In order to disrupt the system, knowledge of how to exploit weaknesses or vulnerabilities is required.

Vulnerabilities can come from two sources. Firstly, inherent vulnerabilities will exist through weaknesses built into the system, such as single or critical points of failure. Secondly, external vulnerabilities will exist because of the dependency on other types of infrastructure where inherent weaknesses can also be exploited. Vulnerabilities can also be exploited anywhere along the energy supply chain, making the network very difficult to defend. We will now look some of these more closely:

Power Generators

Disrupting power generation means that no power is supplied to the grid. It is generally difficult to simultaneously disrupt all power generation because power stations are geographically distributed and use a variety of fuel sources. That said, individual power generators can be attacked and techniques include physical destruction and locking fuel supplies. Note that attacks against generators are less likely to occur than some of the easier attacks described below.

National Grid Infrastructure

Because this infrastructure is so geographically distributed, it is the most difficult to defend. Grid supplies can be disrupted by simply physically attacking the pylons carrying the high voltage cables. It is likely that there would be enough redundancy in the network to cope with one off random attacks, as routine failures have to be dealt with on a daily basis. The same cannot be said for coordinated attacks where the perpetrators have inside knowledge of the grid topology. Topology information can be used to identify single or critical points of failure, and if enough of these can be found and exploited simultaneously, it would be possible to disrupt a large proportion of the entire grid.

National Grid Control Centres

Unless a backup centre exists, these centres can be considered to be single or critical points of failure because they have the capability of controlling part or all of the national grid system. Attacks can include physical destruction of the centre, occupying the centre and taking over its operation, and more advanced techniques that exploit the computerized control system. The latter example would involve breaking into and gaining control of the computer system. Exploits could then involve switching the grid to deliberately overload switch-gear at identified single or critical points of failure. This probably constitutes one of the most serious threats to the grid, as it would cause widespread outages that would take a significant amount of time to repair. The computer systems could also be crashed and the hard disks could be wiped - this would effectively leave the grid in the last state that it was left in before the attack and mean that it could not adapt to meet changing supply and demand requirements.

Regional Service Providers

These organizations and their infrastructure can be attacked using most of the techniques that we have already described above. The main difference is that the effect of the attack will be localized. This doesn't mean that the potential for an attack is less likely, rather that some regional service providers will be at greater risk than others. For example, a terrorist organization might decide to simultaneously attack all of the sub-stations providing electrical supplies to the financial district of a major city. If high explosives were used for this purpose, the resulting outage would be long in duration and cause a large amount of financial damage.

8.4.2.2 Gas Pipelines

Pipelines play an important role in delivering natural gas to large cities. They generally cover large distances and in some cases may be exposed to physical attacks, particularly when they pass through large uninhabited landscapes. The cost of constructing pipelines is also high, so, as a result, some cities might receive their gas from a single line. This means that potential single or critical points of failure exist and can be exploited. Physical attack is one option; however, in order to be effective some knowledge of the pipeline network topology is needed so that a suitable location can be chosen.

Control and monitoring of pipelines in remote locations is made possible through the use of computers and telecommunication networks, which in turn are both dependent on electrical power. Two more potential opportunities for

attack arise here. Firstly, the computer network could be compromised and taken over. This would then give the perpetrator the opportunity to shut down key critical pipelines. Alternatively, an attempt could be made to destroy the pipeline by overriding safety systems and causing a pressure build-up by opening a certain combination of valves. Again, detailed inside knowledge of the topology would be required to do this. Secondly, the pipeline's computer network or electrical supply could be targeted. Failure in either of these systems would probably cause a safety system to kick-in and shut down the pipeline.

In the case of an accidental shutdown being caused by vulnerability in the system, the impact on other types of infrastructure would be relatively low and a sustained outage could be dealt with in most cases. One exception would be in the case of a failure in a gas pipeline supplying a gas-fired power station. This would have an impact on the electrical grid. At the opposite end of the scale, if a long-term outage occurred in the electrical supply system or tele-communication network that the gas pipelines depend on, the impact would be high, resulting in an immediate shutdown in service.

Note that oil pipelines are subject to the same vulnerabilities as gas pipelines.

8.4.2.3 Telecommunication Networks

Most countries now have two types of phone network: the traditional terrestrial hardwired phone system and a mobile phone network. In the case of the latter, coverage is getting close to 100% in some countries. Both types of network are dependent on electrical power and computer based switching systems. In general, more redundancy is built into telecommunication networks than other forms of infrastructure; however, this does not mean that single/critical points of failure cannot be identified. Different types of telecommunication infrastructure are used under different circumstances depending on geographical position and the anticipated bandwidth utilisation. In general, most large cities have a local fibre optic ring to cater for commercial activities and are also connected to a fibre optic backbone that links them with other major cities. Other types of network infrastructure include the use of microwave radio links, coaxial cables, digital radio concentrators and submarine cables. The latter is now generally a fibre cable and is used for intercontinental links. Mobile phone cells also form the basic infrastructure that makes the use of mobile phones possible. Mobile phone networks depend heavily on the terrestrial phone network, and in many cases are even run by large cable telecommunication providers.

Redundancy in terrestrial telecommunication networks allows calls to be routed via alternative routes dependent upon utilisation levels. Some routes are shorter than others and the most direct route is usually chosen first. Different links will

often have different bandwidth capabilities and this means that a large proportion of traffic will have to be carried by high bandwidth links. This information can then be used to identify potential choke points representing potential single or critical points of failure. For example, two large cities might be connected by two separate fibre cables, yet both cables might pass through one single telephone exchange. If the exchange were to be destroyed, calls might still be able to be rerouted through other exchanges, but the other lower bandwidth links might not be able to cope with the resulting amount of demand.

Dependency on electricity supplies applies to telecommunication networks in the same way that it does to computer networks – we discussed this in section 8.4.1. As with any other type of infrastructure that uses a computerised control system, there will also be potential for network-based attacks. While these may be possible, additional inside information will be required (such as the locations of systems controlling potential single or critical points of failure) in order to make the attack have a serious impact.

Some exchanges will have strategic importance. These are usually exchanges that provide international gateways or locations through which more than one key high bandwidth link passes. In the case of the former, all such exchanges can be identified and attacked simultaneously to cut international communications. For the latter, if these key exchanges were to be taken out simultaneously, there could be the potential for a national telecommunication service failure.

Satellite communications also form an important part of many countries' telecommunication infrastructure. Some telecommunication companies have their own domestic satellite system, and many rent capacity from worldwide satellite systems such as the INTELSAT consortium. In the case of domestic satellite fleets, the satellites could be used to carry other services such as secure defence signals, ground to air communications and air traffic control systems, remote oil and gas pipeline monitoring, the Internet and radio and TV services. These satellites are controlled from ground stations, which may be susceptible to physical attack. The satellite control channels are encrypted on new satellites; however, some older satellites do not use encryption and could potentially be put out of action with the proper equipment. Even in cases where encryption is used, foreign military organisations could have the capability to disrupt normal operations.

8.4.2.4 Transport Systems

All transportation systems are dependent on fuel sources. In some cases, such as rail and air transport, there is an additional dependency on control systems that are required to maintain safe operation. Like other types of basic infrastructure,

transport systems also have vulnerabilities that represent single/critical points of failure. These can be inherent within a given transport system or externally based through dependencies on other types of infrastructure.

Inherent attacks can be either physical or remote aimed at key control systems. Physical attacks might include blowing up key rail and road bridges, destroying air traffic control and key railway signalling centres, attacking fuel distribution centres such as oil refineries and destroying key parts of control networks. Remote attacks could include electronically breaking into and disrupting key control systems. These would include air traffic control, railway signalling, oil pipeline control systems and traffic control centres in large cities.

While some transport systems use alternative forms of power, the majority use oil-based fuels such as petrol and diesel. This makes oil refineries and fuel distribution networks some of the most important parts of transport infrastructure. Fuel consumers are dependent on petrol stations, which in turn depend on receiving supplies from oil refineries on a frequent basis. Between August and September 2000, demonstrators protesting about increasing fuel prices exploited key weaknesses in the fuel distribution system. The protest started in France, but by September had moved to Britain. In the case of the latter, a small number of organised protesters blockaded oil refineries around the country. Tanker drivers did not make their usual frequent deliveries, and within a few days the nation's petrol pumps were dry. This also meant that national food distribution systems were starting to break down and the normal operation of emergency services was being threatened. In this case, the protesters were successful for three reasons: firstly, they had the element of surprise; secondly, the protestors had insight into the location of critical points within the fuel distribution network; and thirdly, the protesters had an efficient and organised communication structure. In the case of the latter, protests were coordinated through the Internet and mobile phones were used for communications.

The impact of the European fuel crisis was high and exposed serious vulnerabilities in the distribution network. In this case, the crisis was caused by peaceful demonstrations. Laws can be used to protect the network against this sort incident in the future; however, this does not mean that the vulnerabilities will have been removed – i.e. the network will still be susceptible to attacks from terrorists and other entities. This crisis took a few days to develop and the full impact could have been avoided with state intervention.

One other weakness in the fuel supply chain is that petrol stations require electricity to pump the gas from underground tanks into the consumer's vehicles. A sustained electrical outage would render most pumps inoperable. The pumps could be converted to manual operation, but this would take time and would probably result in fuel supplies being rationed in the short term.

8.4.2.5 Water System

Clean water supplies are essential for maintaining hygiene and sustaining life. A sustained failure in the delivery system would cause widespread drought, crop failures, encourage the spread of disease and cause a breakdown in food production supply chains. As with other types of infrastructure, the greatest impact will be caused when vulnerabilities representing single/critical points of failure are either deliberately or accidentally exploited. Water network infrastructure usually includes pipelines, pumping stations, purification and sanitation plants and holding reservoirs. Pressure is required to move water supplies through the pipelines; pumping stations fulfil this role. Dependencies on other infrastructure include electrical and fuel supplies for pumping, telecommunication links for monitoring infrastructure in remote locations and computer networks for control.

Water supply systems could be attacked in a number of different ways, including:

- Paralysing control networks and attempting to disrupt the supply by causing flooding. This would assume that there would not be enough time to allow personnel to physically reach the affected infrastructure and take manual control.

- Poisoning a water supply. Clearly, this would kill large numbers of people; however, once word was out no one else would use the water supply again until it was known to be safe. Poison could be added anywhere in the supply chain and in the event of an incident it would be difficult to know exactly what happened, or if the incident was deliberate or accidental. In short, it would take a long time to restore the supply.

- Physically attacking water supply infrastructure. Examples here would include blowing up dams or pumping stations. If enough potential single/critical points of failure could be identified and exploited simultaneously, it might be possible to cause a full breakdown in water supply.

Other forms of infrastructure are not immediately dependent on water supplies; however, water should be taken extremely seriously as our very existence depends on it heavily.

8.4.3 Commercial Activities

Commercial activities are very much dependent on the availability of basic utilities. In particular, a high degree of dependency is placed on computer networks, telecommunication systems and the supply of electricity. As soon as these services become unavailable there will usually be an almost instantaneous impact. In the two sub-sections that follow we look at two different categories of commercial activity: business networks and financial networks.

8.4.3.1 Business Networks

Most organisations are highly dependent on national infrastructure and can only survive outages for short periods before serious financial damage or competitive loss occurs. In general, electrical supplies are the most critical. Some (not all) buildings have backup power supplies or generators designed to cut in when power cuts occur; however, in general they can only be used for a few hours. In some cases they may only support devices that are considered to be critical building infrastructure, and where backup generators do exist, there may only be enough fuel supplies to last a few hours. In the event of a sustained power outage most organisations would need to suspend operations. The loss of power would mean key IT systems could not be used, and today this would mean that most employees would be unable to do their work. Key building functions such as elevators and air-conditioning would also be unavailable, making buildings unfit for occupancy. In some cases where water is pumped electrically, the water supply could even be unavailable.

8.4.3.2 Financial Networks

Financial network can really be considered as a sub-category of business networks. The main difference is that the operators are likely to go to greater lengths to ensure that their networks are less likely to be affected by a partial breakdown in infrastructure. When assessing vulnerabilities two different categories of financial networks need to be considered:

1. **National Stock Exchanges and International Inter-bank EFT Systems.** These networks represent critical infrastructure upon which a nation's economy may depend heavily. Most organisations running and using these networks carried out vulnerability assessments and developed disaster recovery plans and tests in the run up to the year 2000. This meant that many organisations specially developed service level agreements with their infrastructure service providers, explicitly specifying how long they could expect normal services (such as

communication and data networks) to be available in the event of problems such as power cuts occurring. In principal, this meant business continuity plans could be put into effect quickly in the event of issues arising due to the Y2K bug. While the recovery plans may still be valid, it is important to remember that they were only designed to withstand outages that were only expected to last for a few days at the most. Many of the plans also did not consider the impact of a full simultaneous breakdown in infrastructure.

2. **Consumer Based Financial Networks.** This category encompasses financial systems used by the public such as ATM networks and credit card verification systems. These systems are far more vulnerable to electrical and telecommunication network infrastructure failures. Most financial terminal equipment such as ATMs and credit card readers are located in locations that are easily accessible and convenient for the public. The down side to this is that is too expensive and difficult to build power and network redundancy into all of these locations. In the event of a major power and telecommunication failure, these networks would become unavailable almost immediately, and in a sustained outage there may well be wider implications as bank branches would quickly run out of available cash reserves.

8.4.4 State Activities

State activities can be separated into two categories: those that depend on basic infrastructure, such as civil services, agencies and departments; and, those that can take advantage of their own private (classified) infrastructure, such as military organisations. The former can be considered in the same way as financial networks (see above) – i.e., because of their importance, more precautions are taken compared with a normal business organisation, so that services can be maintained during short-term outages. In the case of sustained outages, these organisations will be impacted heavily and will quickly become ineffective.

Military organisations have to be considered separately because they may have access to private classified forms of infrastructure. Examples include satellite and high frequency communication systems, alternative electrical supply systems (such as nuclear power), private fuel depots and transportation systems. Obviously, by design, different types of classified infrastructure will be in use in different countries. Not all countries will be able to afford a full private infrastructure. Even those that do have a substantial network may still choose to use public infrastructure to some extent to save costs and resources during times of peace.

Any form of private/classified infrastructure will still be susceptible to the some of the vulnerabilities that we have discussed throughout this chapter. Potential single or critical points of failure will still exist; however these are most likely to arise in the event of coordinated physical attacks. Classified private infrastructure has one major advantage over public infrastructure: because details concerning its operation are classified, it is afforded a certain degree of security through obscurity.

8.5 Prevention

From what we have already presented, it should be clear that there is a need to protect critical infrastructure from attack. One of the first steps in prevention is for a nation to define and identify what constitutes its critical national infrastructure. Some of the items for this list will be obvious, whereas others may have to be identified by conducting an asset inventory followed by a vulnerability assessment – this has been already discussed in section 8.2. A vulnerability assessment will need to identify potential points of single failure and take into account the interdependency that exists between different types of infrastructure. Ideally, the task force assembled to conduct the vulnerability assessment should have representatives from a military information operation group, so that a balanced issues list can be compiled. Afterwards, a risk management analysis can be conducted to determine which issues should be addressed first.

Because of the size and complexity of most countries' infrastructure this is an enormous task to carry out, and accordingly will be time consuming, labour intensive and expensive. The vulnerability assessment process is extremely important and will ultimately allow controls to be put in place that will minimise the effects of attacks. Some governments have already mandated this process and more are likely to in the future. There will also be a need to establish government departments or agencies that are responsible for monitoring and providing warnings of threats and investigating incidents where foul play is suspected. In some countries these already exist, and one well-known example is the National Information Protection Centre (NIPC)[17] in the US.

8.6 Conclusion

National infrastructures have been modernised to take advantage of efficiency and cost savings made possible by standardised technology and the globalisation process. The result is that different types of infrastructure share common technology and have become highly interdependent on each other.

We have shown that two main types of vulnerabilities exist: those that are inherent to a given form of infrastructure; and, those that are external – i.e., vulnerabilities that can be exploited in services upon which the infrastructure is dependent. These vulnerabilities can be exploited accidentally or deliberately, and we have already considered the threats that are posed by different actors in section 8.3.

Potential single/critical points of failure can exist in all types of infrastructure. In general, most types of infrastructure have just enough redundancy to be able to cope with random failures and accidental incidents. However, this does not hold for deliberate coordinated attacks. In these instances there is a very real threat that a long-term outage could be experienced that would have immediate and serious consequences for other forms of dependent infrastructure. In a worst-case scenario, a full infrastructure breakdown could occur if enough critical points of single failure could be targeted simultaneously.

Of the different types of infrastructure we have considered, electrical energy supplies are arguably one of the most fundamentally important. This is because almost every other form of infrastructure is dependent upon it either directly or indirectly through computerised control networks.

Effective attacks will require detailed inside information regarding infrastructure network topology. In some countries this can even be found on the Internet [18], and, assuming it was accurate, would provide valuable intelligence information to the perpetrators. In the future, serious consideration should be given to classifying or restricting this information. Some military organisations have a clear advantage in this area because they have access to private infrastructure where the specification and topology is classified, making it far more difficult to attack.

Knowledge of vulnerabilities in infrastructure is being raised through isolated and well-publicised incidents. Foreign states and terrorist groups are most likely to have the motivation to instigate a full infrastructure attacks; however, the latter is more likely because foreign states would fear reprisals. In short, a whole nation's infrastructure could be attacked anonymously, with a combination of low cost commercial off-the-shelf equipment and high explosives. Therefore, it is vital that this threat is taken seriously and addressed.

References and Further Reading:

---

[1] Critical points in infrastructure exist, where a failure would lead to an actual disruption of service due to a lack of redundancy. A single point of failure exists where a disruption at a single critical point results in a service outage. The definition of critical point(s) of failure is less narrow and takes into account that a disruption would have to occur simultaneously at two or more critical points to cause a service outage.

[2] The Clinton Administration's Policy on Critical Infrastructure Protection: Presidential Decision Directive 63, May 1998. For more details see

http://www.ciao.gov/CIAO_Document_Library/paper598.html.

[3] Practices For Securing Critical Information Assets, Critical Infrastructure Assurance Office, January 2000. This paper can be found at

http://www.ciao.gov/CIAO_Document_Library/Practices_For_Securing_Critical_Information_Assets.pdf.

[4] Vulnerability Assessment Framework 1.1, Prepared Under Contract for the Critical Infrastructure Assurance Office by KPMG Peat Marwick LLP, October 1998. This Framework can be downloaded from:

http://www.ciao.gov/CIAO_Document_Library/vulfrmwkass1_1.pdf.

[5] Partial Failure of Internet Root Nameservers, Daniel Pouzzner, The Risks Digest Volume 19, Issue 25, Friday July 18th 1997. For more details see

http://catless.ncl.ac.uk/Risks/19.25.html#subj1.

[6] In this sense "crashing" the Internet would mean that working parts of the Network would become overwhelmed with re-routed data traffic and in effect become unusable.

[7] 50 Ways to Crash The Net, Simson L Garfinkel, Hotwired Archives, 18th August 1997. For more details see

http://hotwired.lycos.com/synapse/feature/97/33/garfinkel0a_text.html.

[8] Here root refers to the DNS server at the top of the tree, a similar concept to the root directory in the Unix file system.

[9] Although DNS lookups are performed by most Internet applications, they seldom record or reveal the actual IP addresses to the user. If there is a requirement to determine a specific IP address, a DNS lookup can be performed manually by querying a DNS server.

[10] Net Cannot Work by Man Alone, Rebecca Vesely & Tim Bark, Wired News (http://www.wired.com), 18th July 1997.

[11] Australia is one such example. For more details see the reference below.

[12] How Hackers Could Crash the Internet, Nathan Cochrane, Industry News: The f2 Network IT section, 22nd August 2000. This article was found at

http://it.fairfax.com.au.

[13] Many routers have sophisticated security functionality that allows network administration access to be configured very strictly. In cases where the router has been compromised and administrative access has been gained, the intruder can lock the administrator out such that control can only be regained by connecting a computer directly into the back of the router and performing a hard reset.

[14] More details about the "Bing" project can be found at

http://web.cnam.fr/reseau/bing.html.

[15] Broken Glass Sharp Tempers, by "Toxic", Wired News (http://www.wired.com), 22nd July 1997.

[16] This was a second example as detailed in the reference above.

[17] According to the NIPC's web page "The National Infrastructure Protection Center (NIPC) serves as a national critical infrastructure threat assessment, warning, vulnerability, and law enforcement investigation and response entity. The NIPC provides timely warnings of international threats, comprehensive analysis and law enforcement investigation and response." For more details the NIPC web page can be found at http://www.nipc.gov.

[18] See for example, Research Paper 18 1997-98, Thinking about the Unthinkable: Australian Vulnerabilities to High-Tech Risks, Dr Adam Cobb, Foreign Affairs, Defence and Trade Group, 29 June 1998. This document can be downloaded from the Parliament of Australia's Parliamentary Library website, using the following url: http://www.aph.gov.au/library/pubs/rp/1997-98/98rp18.htm.

Chapter 9 – Law Enforcement Issues

9.1 Introduction

Information technology has advanced to such an extent that it has been integrated in to almost all aspects of business and society and plays a key role in most people's everyday lives. As information technology usage has increased we have seen the emergence of computer related crime, which is now a growth industry. These crimes are a relatively new phenomenon, and because they are usually highly technical in nature, they are frequently misunderstood. As such, they pose new challenges to law enforcement personnel, who, in some countries, are struggling to come to terms with the impact of new technology.

Computer related crimes exist in many different forms and can have serious financial consequences for the victim. The extent of this depends on the type of crime, the methods used and the victim's dependence on IT. For example, the consequences of a financial institution's on-line banking service being hacked will be quite different from those of an on-line merchant suffering from a denial of service attack. In the first case, although this is still a very serious crime, it is unlikely there will be a large financial impact to the financial institution, as online services will only represent a very small proportion of its activities[1]. The most serious issue here is adverse publicity and damage to reputation. In the second case, the on-line merchant is totally dependent on Internet connectivity and is likely to suffer from serious losses for every minute that customers cannot reach its web site.

At a recent computer crime conference, James Robinson, the United States Assistant Attorney General for the criminal division at the Department of Justice, noted that computers are being used to commit crimes in three main ways[2]:

- "First, a computer system can be the target of an offence." One of the most common examples of this occurs when a hacker gains unauthorised logical access to a computer and threatens the system's integrity and confidentiality. Other crimes in this category include theft of services running on the target computer and attacks on its availability through denial of service techniques.

- "Second, a computer can be used as a tool to committing criminal behaviour." Because of the increasing use of IT, traditional types of crimes are adapting and moving to the Internet. Examples include the dissemination of illegal material, Internet based fraud, and the purchasing of stolen or prohibited goods.

- "Third, a computer can be incidental to an offence, but still significant for our purposes as law enforcement officials." For example, a terrorist group might store contact information on their computers. Similarly, we might expect to find illegal images stored on a paedophile's computer.

In section 9.2, different examples of computer crimes that span these three categories are presented. Here, the specific objective is to provide the reader with an understanding of the technical nature of these crimes.

At this point we need to add our disclaimer by stating that the writers of this text are not legal experts. Different types of law (i.e., civil, common and Islamic) will apply in different jurisdictions, as will different rules for the admissibility of evidence. In this chapter we primarily concentrate on providing legal and non-technical personnel with a practical understanding of the tools and mechanisms needed to understand computer crimes and to successfully gather evidence in a reliable manner.

Many countries are now spending significant amounts of resources[3] in combating computer related crime and redeveloping their legal and policy framework to take account of computer network technology. In March 2000, President Clinton's Working Group on Unlawful Conduct on the Internet produced a report[4], which recommended a three-phased approach for tackling computer crime. Specifically, the report stated that:

- *"First*, any regulation of unlawful conduct involving the use of the Internet should be analysed through a policy framework that ensures that online conduct is treated in a manner consistent with the way offline conduct is treated, in a technology-neutral manner, and in a manner that takes account of other important societal interests, such as privacy and protection of civil liberties;

- *"Second*, law enforcement needs and challenges posed by the Internet should be recognized as significant, particularly in the areas of resources, training, and the need for new investigative tools and capabilities, coordination with and among federal, state, and local law enforcement agencies, and coordination with and among our international counterparts; and

- *"Third*, there should be continued support for private sector leadership and the development of methods – such as "cyberethics" curricula, appropriate technological tools, and media and other outreach efforts – that educate and empower Internet users to prevent and minimize the risks of unlawful activity."

By April 2000, the member and signatory states of the Council of Europe agreed to a draft convention on cyber-crime[5], which begins to address some of the items listed in the first two points above through international cooperation.

In reality, these goals are going to be difficult to achieve because of the truly global nature of Internet based crimes and differences in local laws. Cyber-crimes can now be committed from virtually anywhere in the world where there is Internet access, and computer network technology can even allow individuals to route their data traffic through intermediary countries where computer crime laws do not exist and international cooperation is unlikely. In such cases territorial jurisdiction can pose serious problems to law enforcement personnel and sovereignty issues can hamper the investigation process. Even in countries where good cooperation exists, the speed at which computer crimes can be committed, coupled with different time zones, means that law enforcement personnel need to be on standby 24 hours a day in order to react to incidents in a timely fashion.

A wide variety of tools and techniques are available to law enforcement personnel for gathering evidence; these are the subject of section 9.3. Interestingly, some of these tools have been developed by the hacking community and are now being used against them extremely effectively. In this section consideration is also given to the investigative process and determining the scope of compulsory measures such as search and seizure.

Once technical and initial legal challenges have been overcome and the location of evidence has been determined, caution needs to be exercised to ensure that the evidence can be used in court. After evidence has been gathered it needs to be accounted for from the time it was created to the time that it is presented in court. This account is necessary to prove beyond any reasonable doubt that its integrity is still intact. In cases where computer generated log files are used as evidence, the behaviour and security of the logging mechanism may also be required to be available for examination by the court. Processing and storage of evidence is examined in section 9.4. Consideration is also given to methods for differentiating between reliable and false evidence.

In section 9.5 we bring some of these ideas together in a case study that examines a real case of computer crime. This example is the well-publicised case of the famous hacker, Kevin Mitnick, and, although this case is now more than five years old, it highlights some of the technical and legal challenges faced by the investigative team. In particular, this example demonstrates the value of a combined technical and legal team.

Finally, a conclusion is presented in section 9.6, where we attempt to summarise some of the most salient points presented in this chapter.

9.2 Types of Computer Crime

In the last section we saw that a computer can be:

1) The target of a crime;
2) The tool used to commit a crime; and
3) Incidental to an offence.

When considering the different type of computer crimes, one often finds that they fall in to more than one of these categories. For example, if a hacker gains unauthorised access to an online banking system, the bank's computer is the target of the crime; the hacker's computer is the tool used to commit the crime; and any pertinent evidence stored in between these two machines (say, at an ISP) is incidental to the offence. This has been summarised for each example of computer related crime presented in this section in the table below.

Type of Computer Crime:	Computer can be:		
	The Target	The Tool	Incidental
Breaches of Logical Security	X	X	X
Denial of Service Attacks	X	X	X
Computer Virus Releases	X	X	
Copyright Infringement			X
Computer Fraud		X	
Illegal Content			X
Obstructing the course of Justice	X		

9.2.1 Breaches of Logical Security

Logical security can be compromised by a number of different actors, including hackers, crackers, script kiddies, disgruntled employees and state-sponsored operatives. Note that the different roles these actors can play have already been discussed in chapter 7.

A wide range of public domain tools can be freely downloaded from the Internet and used in attempts to gain unauthorised logical access. These have also already been presented, and were the subject of chapter 5. Once an intruder has

gained unauthorised logical access a number of crimes can be committed, including:

- Replacing web pages with alternate versions, hence, causing adverse publicity and damage of reputation to the victim.

- Theft or alteration of data. For example, stealing credit card numbers from a database system or changing the prices on an on-line merchant's web site.

- Deliberate malicious damage such as the deletion of data and the destruction of operating systems rendering systems useless.

- Unauthorised electronic eavesdropping of system messages and transactions. This action could be perpetrated by a foreign state conducting information operations, or by an intruder seeking information for financial gain.

If the victim of a logical security breach is security conscious and has taken precautions to protect themselves from attack, there should always be some form of audit trail that can be followed to start tracing the intruders. It is acknowledged that in extreme cases, the intruders may be able to successfully disable all monitoring and logging mechanisms. Even when audit trails do exist, the source IP address from which the intruder appears to originate from is often false or belongs to another system that has also been hacked. In general, smart attackers pass through several systems before attacking their target to hinder law enforcement personnel's search for them.

According to the CERT Coordination Centre[6], the number of incidents reported in 1999 was almost 10,000. This showed a sharp increase from almost 4,000 in 1998 and approximately 250 in 1990. Clearly, with the increased proliferation of hacking tools on the Internet the number of logical security attacks is rising sharply. Note that these statistics only include reported incidents and consideration also needs to be given to the attacks that are not reported through fear of adverse publicity.

9.2.2 Denial of Service Attacks

The various types of different denial of service attacks and their method of operation have already been discussed in chapter 5. While theft, failure or destruction of system components can lead to a denial of service, these threats are generally well understood and the impact can be minimised easily. The same cannot be said for network-based denial of service attacks. Bandwidth or application denial of service techniques can be particularly devastating to on-line organisations since they are completely dependent on Internet connectivity for their activities.

While denial of service techniques can be hard to prevent, they can be detected by examining router and firewall traffic patterns. In some cases the impact of attacks can be minimised by blocking the source IP address at the incoming router. Router and firewall monitoring software can show the originating source IP address and port number of the incoming data packets. While this information can be used to trace the perpetrators, consideration needs to be given to the fact that many denial of service programs allow the attacker to fake the source IP address and port number. Note that this is perfectly feasible because the data traffic is, in effect, one-way.

Recently, distributed denial of service attacks have been reported. This is especially worrying as it is difficult to prevent this type of attack. Because the attack is distributed, any information obtained from routers and firewalls will be hard to interpret, thus making the perpetrators more difficult to trace. For more details see chapter 5.

9.2.3 Computer Virus Releases

Computer viruses have been around for some time, but it only recently that they have attracted lots of media attention. New types of viruses have been written to take advantage of global Internet connectivity. Recent examples include the "Melissa" and "Love Bug" viruses, which were spread worldwide by e-mail and caused a denial of service for many international companies. For more details on the history of viruses and operating principles, see chapter 5.

While anti-virus software can protect computer systems from known viruses, they cannot provide protection against new strains. Most anti-virus software uses an anti-virus definition engine or database, where profiles of known viruses can be stored. Once a new virus is released it must be contained and examined before updates can be made to anti-virus software. In general, anti-virus software updates are often released within days of a virus outbreak; however, affected organisations can experience serious financial consequences in the

meantime. Admittedly, it is hard to estimate the total cost of a virus outbreak, but if one considers that a virus can cause the whole of an affected organisation to be without its IT for a day, loss of revenue or productivity can be significant. A recent study even clamed that yearly global cost of virus outbreaks has been estimated to be in the order of 1.6 trillion US dollars[7].

As software products offer more functionality to users, there will be more opportunity for the creation of new strains of viruses. On the positive side, the creators of the "Melissa"[8] and "Love Bug"[9] viruses were caught; however, it is likely that we will continue to see just as virulent viruses in the near future.

9.2.4 Copyright Infringement

Computers can be involved with copyright infringement in a number of ways. Probably the most well known example is software piracy; however, recent developments in computer hardware and Internet speed mean that audio CD's and DVD movies may be copied and exchanged easily.

Software piracy is a big problem, and according to the Business Software Alliance (BSA)[10] more than 38% of software in use worldwide is illegally copied. The organisation's web site also claims that piracy costs the software industry $11 Billion US Dollars in 1998. At present most software is provided on CD, and it can easily copied using a CD burner or be installed on more than one computer. The biggest impact occurs when organisations purchase a single user licence for a piece of software, and then deploy it across the whole organisation. In these cases several hundred illegal copies may be in use. Honest organisations can also get caught out with multiple user software licences if they do not take measures to ensure that all installed software is paid for.

With increasing network speeds, pirate software can also be exchanged over the Internet. Pirate software sites can be found relatively easily using standard search engines and searching using the keyword "warez". Subcategories of "warez" include "appz" (applications) and "gamez" (games). Online piracy can also be quite sophisticated and, in some cases, dedicated "warez" client-server programs can found on the Internet. These work by communicating on a non-standard (hence, harder to detect) TCP port number often using a ratio system, where the user must upload a file before they can download something. In general, pirate sites are only on-line for a few hours because the pirate software is usually up-loaded to a compromised server belonging to a third party.

In 1999 a study on global software piracy was conducted by the International Planning and Research Corporation on behalf of the BSA and the Software and

Information Industry Association (SIIA)[11]. This study showed that the world piracy rate has decreased steadily from 49% in 1994 to 36% in 1999. The report also suggested that this might be due to a number of factors, including increased governmental cooperation, more awareness among businesses, and increases in user support making customers more willing to pay for software. Clearly, the current figures are far from ideal, and it is likely that software houses will look for a technological answer to this problem in the future.

The music and video industry is now also facing threats from computer related piracy. Recent developments in compression technology have made formats such as MP3 music and DVD video possible. The MP3 format can compress a normal audio CD (approximately 650Mb) by a factor of ten. As a rule of thumb, one hour of audio can be recorded as a 60Mb MP3 file. This means that an average pop music single can be compressed into a MP3 file approximately 5Mb in size. Files of this size can easily be exchanged over the Internet, and many websites are dedicated to this purpose. While some of these sites promote emerging bands and artists, others contain links to sites where pirate copies of music have been uploaded. The MP3 format has become so popular that an American company (Napster Inc) has produced dedicated client server software that allows users to exchange MP3's with each other on-line. At the time of writing Napster Inc is currently involved in a legal battle with the Recording Industry Association of America (RIAA) after being accused of facilitating wholesale music piracy[12]. This is an interesting case because, on the one hand, Napster claims that its users are merely sharing music for non-commercial purposes, while, on the other hand, the RIAA see Napster's service as copyright infringement and a substitute for the sale of music.

DVD movie discs are protected with an encryption system called CSS to deter piracy. Towards the end of 1999 a Norwegian teenager and two of his associates wrote a piece of software called DeCSS. The software allowed movies to be downloaded onto a computer hard disc to be viewed at a later time. It is now being claimed that consumers can use this software to make unauthorised copies of DVD movies and then exchange them through the Internet using compression technologies[13]. If this is the case, it will not be long before DVD piracy catches up with MP3 piracy.

9.2.5 Computer Fraud

Traditional forms of fraud have adapted to take advantage computer and Internet technology. Computers can be used as a tool to facilitate fraud in a number of ways[14], including:

- Setting up an electronic commerce web site and failing to deliver the promised goods, or delivering items of inferior quality. Remember that it is much harder for the average consumer to spot an on-line fraudster than one in the flesh.

- Hiding the origin of funds obtained through criminal activity by using money laundering techniques. For more details on this subject see chapter 6.

- Posting fake information on the Internet for purposes such as financial gain. This can range from begging under false pretences to trying to influence share prices.

In the last instance, one case has already been publicised in which an individual successfully influenced the share price of a telecommunication company. An e-mail was posted to a message board saying that the company was going to be taken over by an Israeli company, and a link was provided to what appeared to be a Bloomberg news service web site[15]. In reality the story and the website were fake; however, the telecommunication company's stock rose more than 30% because of the news, thus causing financial losses to investors who purchased the stock at inflated prices. In this instance the perpetrator was tracked by his IP address and sentenced.

Instances of computer related fraud are becoming more common and are likely to be a problem for a considerable time to come. Not surprisingly, governments are starting to address this issue by encouraging consumers and business to report instances of on-line fraud. For example, in the U.S. an Internet Fraud Complaint Centre (IFCC) has been set-up through a partnership between the FBI and the National White Collar Crime Center (NW3C)[16]. The IFCC even has a web site that allows victims to file their complaints on-line.

9.2.6 Illegal Content

Illegal content can be difficult to detect unless law enforcement officials spend exhaustive amounts of time trawling the Internet for it. In most cases law enforcement personnel are likely to be made aware of its existence after a complaint has been made. Examples of illegal material include images of child pornography, bomb and drug manufacturing recipes, web sites selling stolen goods or promoting the use of prohibited substances, and racist / hate e-mail. This list could be endless.

The good news is that once law enforcement officers are aware of a problem, the perpetrators can be traced relatively easily. In cases where illegal content has been posted on web servers, the offending server can be traced through its IP

address and DNS records. Where e-mail is used as a transport medium, it should be relatively straightforward to trace the sender of the message.

9.2.7 Obstructing the Course of Justice

With networked computer systems there are several ways law enforcement efforts can be hampered, resulting in an obstruction of the course of justice. This could range from the deliberate deletion of data (containing evidence of criminal activity) to turning a blind eye to criminal activity, instead of reporting it to law enforcement personnel. Consider the following two examples.

1. Internet service providers could delete log entries of Internet access transactions in order to protect criminal identities. They could also turn a blind eye to content of web pages that exist on their servers, and withhold evidence when compulsory measures are issued.

2. Companies could cause obstruction by instructing their employees to delete evidence, such as e-mail, in case it is seized for use in court. Since e-mail was used as evidence in the recent Microsoft antitrust trial in the US, there has been a lot of media speculation that companies will instruct their employees to delete all unnecessary e-mail. This raises other questions, such as how long e-mail should be legally kept for, and the implications that it will have in terms of data storage.

The skill needed to successfully obstruct the course of justice depends on the individual involved and the circumstances of each case. Instead, it is more useful to consider each instance of obstruction. Deletion of data requires specialist knowledge to completely remove all traces of the fact that it has been deleted. At the opposite end of the scale, turning a blind eye to criminal activity, or not fully cooperating with law enforcement personnel, requires no specialist knowledge.

9.3 Gathering Evidence and Technical Challenges

Once criminal activity is suspected or has been reported, the next step in the incident response procedure is to gather evidence. Exactly how this done will vary from one incident to another; however, in general a two-phased approach is required. In the first phase, information can be gathered from the crime scene and examined using public domain tools, search engines and registries that can be found on the Internet. The second phase usually takes account of information gathered in phase one and uses compulsory measures, such as search and seizure, to target sources of pertinent evidence. During active information gathering, a number of useful tools are available to law enforcement personnel. Interestingly, some of the most effective evidence gathering techniques involve the use of public domain hacking tools in a controlled and legal manner (for more details see section 9.3.2).

Caution needs to be exercised during evidence gathering, and it is advisable to seek legal advice from the outset, even if law enforcement personnel are not involved in the early stages of the investigation. In some jurisdictions there may be advantages for organisations that have suffered from criminal activity to hire external consultants to investigate the crime and gather evidence. This step can be advantageous because law enforcement personnel may need a warrant to investigate, whereas system administrators and hired consultants may not[17].

If it is necessary to use compulsory measures, a certain amount of preparation is required. Apart from applying for legal permission, careful consideration needs to be given to determining the scope of search warrants. For example, how does one search the computer system of an international organisation if we are not sure which country the data resides in? Another common question is how do we protect innocent third parties (such as ISP's) if the seizure of any computer systems containing incidental evidence will cause them to suffer financial hardship? Preparation is also an important issue. With the speed at which data can be transmitted, deleted or altered, it is important to arrive at a crime scene ready to gather evidence immediately.

9.3.1 Information Gathering

Considerable time can pass between an incident first being reported and the incident response team first arriving on the crime scene. Depending on the type of crime being committed, what happens to the target computer system during this time can determine the amount of evidence that can be successfully gathered. In cases where systems have suffered from a logical break-in, turning the computer off can result in evidence being lost from the computer's memory.

Similarly if a system is left turned on and plugged into the network, the intruder may be presented with an opportunity to delete evidence. In general, law enforcement personnel should consider the following steps when responding to computer crime incidents:

1. What type of crime has been committed? If a breach of logical security has been reported, it is important to ensure the computer remains switched on and no one gains access to it. Consideration also needs to be given to the seriousness of the crime. In cases where critical systems have been hacked the network cable should be unplugged from the back of the machine to prevent the intruder doing any more damage or covering their tracks.

2. Is there enough evidence available to be able to trace and apprehend the perpetrators immediately? This would be an ideal scenario, but it rarely occurs.

3. Is the crime likely to be a one-off offence, and are the perpetrators likely to try committing the crime again? In cases where the crime is likely to be repeated, tools can be deployed to collect additional evidence when the offence is repeated. See section 9.3.2 for more details of how such tools can be used.

When computer crimes are network based there are some obvious starting points. Firstly, every computer must have and address on the network it is connected to in order to be able to receive and transmit data packets. In the case of the Internet this is an IP address. Moreover, most organisations protect their systems from the Internet by employing sophisticated routers, firewalls and intrusion detection systems. These devices can produce log files of reported exceptions and unauthorised connection attempts that explicitly list source and destination IP addresses and port numbers. For more details see chapter 4.

Obtaining a source IP address of a perpetrator is a good starting point. Once this has been achieved there are many resources available[18] on the Internet that can be interrogated to find out who owns the IP address. This process should at least be able to determine the ISP or organisation that owns the IP address. Where web sites have been established with illegal content, the location of the server can be found by interrogating DNS records so that the real IP address can be determined. The location of the server can then be determined in the same manner as before.

In practice tracing smart perpetrators is not always easy. Often after a source IP address has been successfully identified, it is discovered that the system it belongs to has been the victim of a breach of logical security and used as a proxy server, thus making the audit trail harder to follow. Even when

individuals can be traced back to a specific ISP it can be hard to determine their true identity without engaging in pro-active monitoring. Where pre-paid Internet and mobile phone accounts are used, the user's anonymity is strengthened and makes them very hard to trace. In this case the location of the user's mobile phone would need to be determined as the crime was being committed.

9.3.2 Evidence Gathering Tools

A wide range of tools can be used to gather evidence. The key here is to know when to use which tool and to have a good understanding of how the tool works. As we said earlier, a number of hacking tools in the public domain can also be used for gathering evidence. Some of these will now be discussed; however, if any of these tools are used, care should be taken to ensure they are obtained from a trustworthy source and that functionality cannot be questioned in court. In general, the safest way to obtain such tools is to download the source code, examine its functionality, and then compile it locally.

Keystroke Recorder

This is exactly what its name suggests. If keystroke recording software is installed on a computer, it records the keystrokes that are typed on the keyboard by the user. Some keystroke recorders are also capable of transmitting the keystrokes recorded to another computer over a network or the Internet[19]. If permission can be obtained to covertly install a keystroke recorder on a suspect's computer, it could be used to monitor the suspect's activities and reveal any username and password combinations typed.

Another advantage of this tool is that it can also provide a means to defeat any cryptographic techniques used by the suspect. Typically, private encryption keys are protected by a user password, and as we have already seen a keystroke recorder can be used to capture passwords if the user happens to type them. All in all, keystroke recorders are exceptionally useful tools and can yield good evidence. Once installed, a keystroke recorder would have little impact on the target computer's performance and would be very difficult for the user to detect.

Packet Sniffers

Packet sniffing can best be described as the computer network equivalent of a telephone wiretap. In essence this tool[20] can be used to capture any data packets passing by the computer on which it is installed. One of the nice things about packet sniffers is that they can be programmed to monitor selectively. Specific source and destination IP addresses can be targeted to reduce the

amount of information gathered. In addition, monitoring can be refined even further by explicitly specifying certain port numbers.

The data payload of captured packets can yield a range of evidence. For example, if a packet sniffer is used to monitor port 25 (E-mail), the data payload of the captured packets could be reconstructed to show the e-mail message. In this instance, a packet sniffer could be used to trace the location of an individual exploiting vulnerable mail servers to send hate e-mail. The principles of packet sniffing are covered in more detail in chapter 5.

When this type of monitoring is used, consideration needs to be given to whether or not users can have a reasonable expectation of privacy, and if the organisation can justify monitoring data traffic such that it outweighs the user's right to privacy.

Recovery of Deleted Data

Safe deletion of data is difficult to achieve. In most instances when a file is deleted the index point of the file is removed from the file access table (FAT) and the data still remains intact. In reality, the file is only removed when the operating system overwrites the deleted file with new data. The main reason for using this technique is speed. If a file were actually overwritten with a series of digital ones or zeros to delete it, the disk would need some time to perform this operation, especially if a large amount of data was being deleted. By removing the index marker on the FAT, the file appears to be deleted to the operating system and better disk speed performance is obtained. This means that if data has been deleted, but not overwritten, there is a good chance that it can be recovered. Commercial and public domain tools[21] are now available for several operating systems, and these can offer a quick and convenient way for investigators to recover evidence that has been deliberately or unintentionally deleted.

Techniques exist that can be used to recover data that has already been deleted and overwritten. Magnetic force microscopy (MFM) is a recent technique for imaging magnetisation patterns with high resolution and minimal sample preparation. In effect this means that techniques like MFM can be used to recover several layers of overwritten data. In simple terms, if a one or a zero is written to a disk, the disk records a one or zero. In reality the effect is more like obtaining 0.95 when a zero is overwritten by a one, and 1.05 when a one is overwritten by a one. Thus, with some basic electronic equipment, the actual disk levels can be recovered and analysed to reproduce previously recorded signals. Further information on this subject (including more technical details) was presented in a paper by Peter Gutmann at the Sixth Usenix Security Symposium Proceedings[22].

Techniques like MFM also have implications for the recovery of data or illegal content protected by encryption schemes. A lot of research has been conducted into cryptographic security, but very little has been done to find methods to safely delete the original plain text forms of the encrypted data. Official guidelines for magnetic media sanitation are often outdated and have not taken account of new recording densities and other sophisticated techniques that were introduced in the early 1990's. Gutmann also notes in his publication that many guidelines on media sanitation are classified. In any case, while encryption technologies have gained a lot of public awareness, the opposite is true of secure deletion techniques.

Password Cracking Tools

When hard disks are being examined for evidence gathering purposes some files may be found that are protected by a password. Examples include word documents, excel spreadsheets and zip files. Most of these applications offer primitive security mechanisms in comparison with modern day encryption programs, and a number of password cracking tools[23] can be used to unlock the files with relative ease.

Other Tools

Any number of tools can be used in different situations, and we couldn't possibly hope to cover all of them here. In general, most hacking tools can be useful to law enforcement personnel, and these have been discussed at length in chapter 5. That said, there are some tools we want to briefly touch upon here. These are "rootkits" (backdoors) and operating systems that can run completely in memory.

A "rootkit", or backdoor, provides a means of gaining unauthorised logical access to a system through its network connection. Examples include Back Orifice for Windows 9.x, Netbus[24] for Windows NT and Linux Root Kit 4[25] (LRK4) for the Linux platform. Two potential uses for these tools exist. Firstly, a backdoor could be installed on a suspect's system and used to covertly monitor activity over a period of time. Secondly, if an organisation suffers from recurring logical attacks, a backdoor could be set up on a deliberately weakened system acting as a honey pot. In this instance, one would wait for the intruder to take over the backdoored system and then gather evidence by covertly monitoring their activity.

A compact operating system that can run in computer memory is useful because it can be stored on floppy disks and run on any compatible computer that can be booted from a floppy disk. One example is the Trinux[26] operating system. It comes complete with a set of network monitoring tools, including a packet sniffer, and can be used to turn a standard desktop computer into a network

monitoring station in a few minutes. This type of tool is portable, and means that packet sniffers can be deployed in several different network locations quickly and easily.

9.3.3 Compulsory Measures

Obtaining permission to exercise compulsory measures, such as search and seizure, can pose difficulty to law enforcement personnel. In most cases, evidence of criminal activity will be stored on a computer system's hard disk or any other magnetic storage media that might be in use. These are the items that search and seizure warrants should target; however, network connectivity can make their physical location hard to determine. Moreover, even when the location of a system has been determined, data may be geographically distributed over several networked systems.

Another problem arises when the decision has to be made whether to search or seize a system. Here, consideration needs to be given to the protection of innocent third parties. In cases where incidental evidence of computer crime is stored on the systems of innocent third parties, seizure of equipment can lead to loss of productivity and cause financial losses. This is especially true in the case of ISP's, where the business is totally dependent on IT.

Searching a magnetic media, i.e., a hard disk, can be a lengthy and time-consuming process. If these devices are searched at a crime scene there is a danger that vital evidence could be overlooked. This evidence could then be deleted before the law enforcement personnel have a chance to return. From an evidence-gathering point of view, seizure is preferable because the system remains in an unchanged state and can be searched at leisure.

Hard disk search or seizure doesn't necessarily mean that the target system cannot continue to be used. One approach is to take an image of the system's hard disk. For example, the hard disk could be removed from the target machine[27] and copied using a dedicated piece of hardware designed for copying hard disks quickly. The hard disk could then be returned to the target system so that normal operations can continue. This image taken would represent a snapshot in time of the target system, which could be searched later without the fear of vital evidence being deleted. If it were suspected that vital evidence has already been deleted, the seized hard disk would need to be retained for further forensic examination in the hopes that the deleted data could be recovered. In these instances, an image of the system can be made using a second hard disk. This can then be substituted for the hard disk in the target system.

Determining the scope of a search warrant can be a complex task. A balance has to be drawn between making sure all evidence is collected and protecting the privacy of the owner of the target system. Because of the nature of distributed networks, law enforcement personnel are facing some serious challenges. For example, if a warrant is obtained to search one machine, and it is logically connected to a second machine in a different location, does the law enforcement officer have the right to search for evidence on the second system? When hard disks are searched it is quite likely that evidence of other crimes will be uncovered. This cannot be helped as, in general, an exhaustive search has to be made to ensure that no evidence pertaining to the initial compulsory measure is missed. Here the question is: Can the evidence of different crimes be considered admissible in court, or do the law enforcement personnel need to obtain a second search warrant? In the case of the latter, the problem is that all evidence could be deleted before another warrant is issued.

When arriving on a crime scene to seize evidence, one of the first steps should be to remove operators from the target systems and disconnect any network cables from the back of the machines. This simple step ensures that the console operative or any other third party with network access to the machine cannot delete evidence. Consideration should also be given to physical security at the site and any schemes that could be used to delete evidence. An example of the latter could be a magnetic field around the door of the computer room that ensures any media removed from the room is deleted.

9.3.4 Other Practical Considerations

In cases where an intruder is making repeated attempts to gain unauthorised logical access to a system, the local law in some jurisdictions may restrict the actions that system administrators can take to protect their systems. In some circumstances monitoring can even be illegal and treated as a criminal offence. When monitoring does occur it is important to have warning banners that notify potential intruders unauthorised use is prohibited and all unauthorised actions will be monitored.

Another problem is that system logs can be treated as hearsay evidence unless they are produced and monitored on a regular basis. Where this does occur, the logs may be treated as business records; however, the organisation may be required to demonstrate proof that logging is part of the business process by providing documentation of their monitoring policy that clearly explains what is audited and why. This highlights the importance of the need for the documented standards and procedures, which we presented in chapter 7.

The use of most of the tools and techniques discussed in this chapter rely on the operative having well developed technical skills. In reality it is rare to find good technical skills and legal knowledge in the same resource. The current accepted best practice is to use a combined team of technical and law enforcement staff. If legal advice is sought from the outset, technical staff can perform most of the evidence gathering and identify perpetrators. This then allows law enforcement personnel to apply for search warrants so that perpetrators can be apprehended in a timely fashion.

Current levels of computer crime mean that law enforcement personnel have to be selective in which cases they pursue. In doing this consideration needs to be given to the type of offence, the difficulty of the suspect being traced and the likelihood of evidence being recovered. For example, for petty crimes it would not be worth going to the trouble to use techniques such as magnetic force microscopy. However, it would be worth using this technique if a terrorist was apprehended with a laptop computer.

9.4 Processing and Storing Evidence

Evidence pertaining to computer crimes will mostly exist in electronic form. It will need to be accounted for from the time of its creation until it is presented in court so that its integrity cannot be questioned. Similarly, after use in a trial it will need to be archived in a secure manner in case it is needed again for an appeal process. These issues are presently addressed by storing evidence on non-erasable media such as compact discs. While this method is better than nothing, there are some disadvantages: Firstly, the evidence will not be available to other law enforcement personnel using on-line search systems. This means that opportunities for cross-referencing the evidence and comparing it with other crimes may be lost. Secondly, the integrity of the evidence will depend on the physical security over the compact disc. Dishonest law enforcement personnel could take the CD, upload it onto a computer, delete or add false evidence, and then burn the information back on to another CD.

Once evidence has been gathered, there are good reasons to consider storing it on an online system, such as a data warehouse, that can be accessed by other law enforcement personnel:

- First, a large number of crimes are conducted across national boundaries and this may be one of the best ways for law enforcement agencies to share evidence and cooperate with each other.

- Second, the evidence will be able to be reviewed by other law enforcement personnel investigating different crimes. Evidence of other crimes is usually observed when perpetrators of computer crime are apprehended.

- Third, there will be an opportunity to distribute the evidence processing to regional centres of excellence. This can result in evidence being processed quickly and without bias.

The disadvantage of an online system is that evidence could be altered or deleted; however, PKI technology can be used to ensure the integrity of evidence. This is the subject of section 9.4.2. In addition, regular archives of the system can be made to non-erasable media.

9.4.1 Differentiating Between Reliable and False Evidence

Under most circumstances, evidence that has been gathered by law enforcement personnel in real time can be considered highly reliable. The same cannot always be said of evidence that has been gathered retrospectively, and this can

also apply to evidence that is presented by cooperative third parties. The main problem is that electronic evidence can be easy to fake. For example, it could be easy to create fake logs records detailing access to a system, or ISP, that never took place. E-mail is also notoriously easy to fake, and there are many insecure mail servers on the Internet that can be exploited for this purpose.

In most cases the authenticity of electronic evidence should be fairly easy to prove because several traces are made for each transaction. In cases where it is suspected that evidence has been spoiled or faked, inconsistencies will be revealed when different traces of the transaction are cross-referenced. For example, in the case of fake ISP access logs; cross-referencing the supposed access times against data call telecommunication records would reveal the inconsistencies. Similarly, fake e-mail can be detected by checking the e-mail header information (often not displayed to the user) to see if the machine that sent the e-mail has the same address as the supposed sender.

9.4.2 Potential Uses for PKI

Public Key Infrastructure (see chapter 3) offers an alternative to storing evidence on non-erasable media. It also offers the security mechanism upon which a secure on-line evidence sharing system can be built.

After all the pertinent evidence has been gathered from a crime scene the records can be signed with a digital signature. A copy of this signature can then be supplied to the defendant and relevant investigating authority. In effect, this signature provides a mechanism to detect if the evidence has subsequently been tampered with. If the evidence is stored in an on-line system, its integrity can be verified at any time by creating a new digital signature and comparing it with the original one. If these do not agree, the evidence has not been preserved in its original form.

There will be other issues with setting up an on-line system in which evidence can be shared, especially if information is shared across national boundaries. Mechanisms will be required to authenticate cooperative entities and to ensure that any information exchanged is kept confidential during transit. PKI offers all of these mechanisms and could even cater for using the Internet as a virtual private network. Clearly, this would have substantial benefits for developing countries.

9.5 An Example: The Kevin Mitnick Case

On Christmas day 1994, Andrew Gross (at this time a graduate student) discovered that the computer system assigned to computational physicist Tsutomo Shimomura had been compromised. This attack was reported to have used two techniques: SYN flooding and TCP hijacking. At the time this was interesting because TCP hijacking had been theorised, but had not been reported before. This attack was very technical in nature, and was reflected in the log files that were preserved in a secure location on Shimomura's machine. Extracts of these log files are provided and discussed in an intrusion detection book by Stephen Northcutt[28]. This should be referred to for more technical details of the attack. In another book on intrusion detection, Rebecca Bace[29] examines the same case from more of a legal perspective; this book has been used as our source for this information.

News of the Christmas day attack was well publicised. Not long after, a user of the San Francisco based ISP, the Well, found several large files in his account with Shimomura's name on them. After seeking legal advice, the Well hired Gross and Shimomura to investigate the incident. Evidence was found of log tampering and credit card numbers that were believed to have been taken from another ISP called Netcom.

It was assumed that the intruder would return, and a packet sniffer was used at the Well to capture and log data packets originating from Netcom. When Netcom was contacted with regard to the incident, full cooperation was received and Gross and Shimomura were formally requested to help.

Tap and trace orders were obtained to determine the number used to dial-up to Netcom; however, it was soon discovered that the telephone switch had been hacked. Further investigation revealed that the intruder was actually gaining access from a cellular phone switch and was using this to route to the hacked switch. Gross and Shimomura correctly assumed that the intruder was using a cloned mobile phone and waited for him to go online again. Kevin Mitnick's location was then determined by triangulation when Gross and Shimomura drove around the cell with radio direction finding equipment. Finally, Mitnick was arrested after the FBI executed a warrant.

In her book, Bace noted several interesting legal points about this investigation, including the following: First, the log files obtained from the compromised system were stored in a secure location where they couldn't be tampered with. Second, because Gross and Shimomura were hired as security consultants, they were not required to obtain a court order for their monitoring activities. Thirdly, traces were performed with the full permission and cooperation of the ISP's

involved. If permission had not been obtained, this could have been considered an illegal act.

9.6 Conclusion

Computer related crime is a growth industry, and an increasing number of incidents are being reported. These include technically complex hacking style offences as well as forms of traditional crime that have been adapted to take advantage of new technology. At the same time law enforcement agencies must adapt to come to terms with this shifting paradigm, and are facing a number of different legal, technical and operational challenges.

Legal challenges will vary from one jurisdiction to another, and this may hinder the investigation of crimes that are committed over national borders. This is especially true when the perpetrators of crime have enough legal knowledge to go "jurisdictional shopping" and route their data traffic through a country where there are lenient or no computer crime laws. When dealing with incident responses, it is advisable to seek legal advice from the outset and use a combined technical and legal team to investigate. Consideration also has to be given to way evidence is obtained if it is going to be used in court. Technical experts may have to provide evidence that any mechanisms used to generate technical log files are trustworthy. They could also be required to explain the content in court and provide the legal defence access to mechanisms used to generate logs. Clearly, this can have an impact on defence organisations[30] that may wish to prosecute perpetrators, but be using classified monitoring techniques.

Technical and operational challenges have been considered in detail in this chapter. Specifically, we have tried to show that:

- Hacking tools used in conjunction with compulsory measures make effective investigative tools.
- Some intrusion detection methods can be out of line with legal restrictions for the monitoring, tracing and intercepting of criminal attacks.
- In many circumstances deleted data can be recovered. Even where it has been deleted and overwritten, techniques like magnetic force microscopy can be used to recover evidence.
- It may be relatively easy to recover encrypted or password protected data. Password cracking tools can be used to decode password-protected files, and in some cases magnetic force microscopy can be used to undelete plain text versions of strongly encrypted data.

- Enough of an electronic audit trail usually exists so that it possible to differentiate between reliable and false evidence. This can also make detecting obstructions of the course justice easy to prove.

- There are safe and reliable means available to store electronic evidence and share it with other law enforcement agencies. In this instance we examined the use of PKI for this role.

Looking to the future, certain types of crimes promise to cause big problems: Distributed denial of service attacks and virus outbreaks. These have been singled out because they are capable of causing widespread destruction and massive financial losses. They are also two of the hardest types of crime to investigate.

Law enforcement agency's success in apprehending perpetrators will not only depend on how well they are addressing legal, technical and operational issues; it will also depend on international cooperation and having rapid response teams available 24 hours a day to ensure that incidents are dealt with quickly.

1 At present Internet banking is only used by a small proportion of customers and the value of transactions is often restricted to modest sums compared to those used in inter-bank wire transfers.

2 Remarks of James K. Robinson (Assistant Attorney General for the Criminal Division at the United States Department of Justice in Washington DC) at an international computer crime conference entitled "Internet as the Scene of Crime", Oslo, Norway, May 29-31, 2000.

3 See for example, White House Task Force to Study Net Crime, Jeri Clausing, New York Times on the Web (Technology, Cybertimes), August 10 1999.

4 The Electronic Frontier: The Challenge of Unlawful Conduct Involving The Use Of The Internet, A Report of the President's Working Group on Unlawful Conduct on the Internet, March 2000. This report can be downloaded from http://www.cybercrime.gov.

5 Draft Convention on Cyber-Crime (Draft No 19 - Declassified Public Version), European Committee on Crime Problems (CDPC) & Committee of Experts on Crime in Cyber-Space (PC-CY), Prepared by the Secretariat Directorate General I (Legal Affairs), Strasbourg, April 27 2000.

6 The CERT (Computer Emergency Response Team) is a federally funded research and development centre at Carnegie Mellon University. It provides incident response services to sites that have been victims of attack as well as publishing security alerts and information. Its web site can be found at http://www.cert.org.

7 Study Finds Computer Viruses and Hacking Take $1.6 Trillion Toll on Worldwide Economy, Excite.com (news), July 7, 2000.

8 Melissa Virus Suspect Caught, David Kocieniewski, The New York Times on the Web (Technology Section), April 3rd 1999.

9 Philippine Officials Charge Alleged "Love Bug" Creator, CNN.com (Insurgency on the Internet), July 29, 2000.

10 The Business Software Alliance is an international organisation that aims to stop and prevent software piracy. Its web site can be viewed at http://www.bsa.org.

11 1999 Global Software Piracy Report: A Study Conducted by International Research Planning Corporation (Management Consultants) For the Business Software Alliance and Software & Information Industry Association, May 2000. This report can be downloaded from the BSA's web site.

12 Recording Industry Faces Music with Napster Case, CNN on-line(technology computing section), July 31, 2000.

13 Norwegian Teenager Appears at Hacker Trial he Sparked, Carl S. Kaplan, New York Times on the Web (Technology, Cybertimes), July 21, 2000.

14 See for example, Chain of Foolery on the Internet, Tina Kelley, New York Times on the Web (Technology, Circuits), July 1, 1999.

15 This example can be found in the Report of the President's Working Group on Unlawful Conduct on the Internet, March 2000. For more details see reference number 3.

16 For more details, visit the IFCC's web site at http://www.ifccfbi.gov.

17 See for example, Intrusion Detection (Chapter 9, Legal Issues), Rebecca Gurley Bace, Macmillan Technical Publishing, USA, Published 2000.

18 See for example the SAM SPADE web site at http://www.samspade.org and doing an IP block search for the target IP address.

19 A more in depth treatment of keystroke logging programs can be found in "Hacking Exposed" by McClure, Scambray & Kurtz, Osborne McGraw Hill, 1999.

20 One such tool is "Sniffit" which runs on Unix and Linux platforms. It can be downloaded from a number of web sites including http://www.sniffit.com.

21 For example, see Directory Snoop from Briggs Softworks. A trial version of this program may be downloaded from http://www.briggsoft.com.

22 Peter Gutmann, "Secure Deletion of Data from Magnetic and Solid-State Memory", Department of Computer Science, University of Auckland. The paper was published for the first time in the Sixth Usenix Security Symposium Proceedings, San Jose, California, July 22-25, 1996.

23 Links to tools of this type can be found be visiting http://www.neworder.box.sk.

24 See http://www.netbus.org.

25 Linux Rootkit IV can be downloaded with instructions for use from http://www.rootshell.com.

26 Trinux is a Linux variant and can be obtained from http://www.trinux.org.

27 In these instances care would also need to be taken to ensure that any evidence is removed from the target system's memory before it is shut down.

28 Network Intrusion Detection: An Analysts Handbook, Stephen Northcutt, New Riders publishing 1999.

29 Intrusion Detection (Technology Series), Rebecca Gurley Bace, Macmillan Technical Publishing, USA, Published 2000.

30 Intruders targeted the US Air Force's command and control research facility, Rome Lab, in 1994. In this case, although the intruders were apprehended, the US Air Force did not allow free access to mechanisms or data and according to Bace (see reference #29) this did not help the prosecutions case.