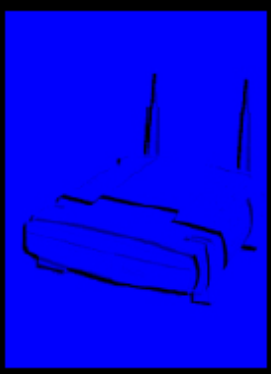
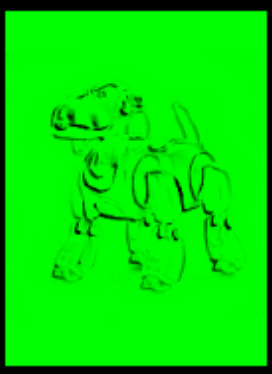
Zoom In

Advances in Networks, Computing and Communications 3

Proceedings of the MSc/MRes Programmes from the School of Computing, Communications and Electronics

2004 - 2005





Edited by

Paul S Dowland
Steven M Furnell

Advances in Networks, Computing and Communications 3

Proceedings of the MSc/MRes Programmes from the School of Computing, Communications and Electronics

2004 - 2005

Editors

Dr Paul S Dowland
Prof Steven M Furnell

School of Computing, Communications & Electronics
University of Plymouth

ISBN: 978-1-84102-179-9

Printed in the United Kingdom

Preface

This book is the third in a series presenting research papers arising from MSc/MRes research projects undertaken by students of the School of Computing, Communications and Electronics at the University of Plymouth. These one year masters courses include a significant period of full-time project activity, and students are assessed on the basis of an MSc or MRes thesis, plus an accompanying research paper.

The publications in this volume are based upon research projects that were undertaken during the 2004/05 academic year. A total of 29 papers are presented, covering many aspects of modern networking and communication technology, including security, mobility, coding schemes and quality measurement. The expanded topic coverage compared to earlier volumes in this series reflects the broadening of our range of MSc programmes. Specifically contributing programmes are: Network Systems Engineering, Communications Engineering and Signal Processing, Robotics, Information Systems Security, Web Technologies and Security, Computing, Computer Applications, E-commerce and Interactive Intelligent Systems

The authorship of the papers is credited to the MSc/MRes student in each case (appearing as the first named author), with other authors being the academic supervisors that had significant input into the projects. Indeed, the projects were conducted in collaboration with supervisors from the internationally recognised research groups within the School, and the underlying research projects are typically related to wider research initiatives with which these groups are involved. Readers interested in further details of the related research areas are therefore encouraged to make contact with the academic supervisors, using the contact details provided elsewhere in this publication.

Each of the papers presented here is also supported by a full MSc or MRes thesis, which contains more comprehensive details of the work undertaken and the results obtained. Copies of these documents are also in the public domain, and can generally be obtained upon request via inter-library loan.

We believe that these papers have value to the academic community, and we therefore hope that their publication in this volume will be of interest to you.

Prof Steven Furnell and Dr Paul Dowland

School of Computing, Communications and Electronics
University of Plymouth, May 2006

About the School of Computing, Communications and Electronics

This new School was formed from a merger between the School of Computing and the Department of Communication and Electronic Engineering in August 2003. It is a large multifaceted School with interests spanning across the interface between computing and art, through software, networks, and communications to electronic engineering. The School contains 61 academic staff and has 1100 students enrolled on its portfolio of taught courses, 104 of which are at MSc level. In addition there are 78 postgraduate research students enrolled on a variety of research programmes, most of which enjoy sponsorship from external sources.

The bulk of the staff in the School are housed in the Portland Square building, a purpose built state of the art building costing over £25million and situated near the centre of the historic city of Plymouth on the University campus. The laboratories are located in the newly refurbished Smeaton Building, and the Clean room for nanotechnology also recently refurbished courtesy of a Wolfson Foundation grant is situated in the nearby Brunel Building. All buildings are a short walk from each other, enabling a close collaboration within our research community.

This School sits alongside two other Schools in the Faculty of Technology, the School of Engineering (the merged School of Civil and Structural Engineering and Department of Mechanical and Marine Engineering), and the School of Mathematics and Statistics. There are research and teaching links across all three schools as well as with the rest of the University. The closest links are with the Faculty of Science, principally the Centre for Computational and Theoretical Neuroscience which started in Computing, and Psychology through Artificial Intelligence and Human Computer Interaction research.

Prof Phil Dyke
Head of School

Contributing Research Groups

Fixed and Mobile Communications

Head: Professor M Tomlinson BSc, PhD, CEng, MIEE
E-mail: mtomlinson@plymouth.ac.uk

Research interests:

1) Satellite communications
2) Wireless communications
3) Broadcasting
4) Watermarking
5) Source coding and data compression

http://www.tech.plymouth.ac.uk/see/research/satcen/sat.htm
http://www.tech.plymouth.ac.uk/see/research/cdma/

Network Research Group

Head: Professor S M Furnell
E-mail info@network-research-group.org

Research interests:

1) Information systems security
2) Internet and Web technologies and applications
3) Mobile applications and services
4) Network management

http://www.network-research-group.org

Signal Processing and Multimedia Communications

Head: Professor E Ifeachor
E-mail eifeachor@plymouth.ac.uk

Research interests:

1) Multimedia communications
2) Audio and bio-signal processing
3) Bioinformatics

http://www.tech.plymouth.ac.uk/spmc/

Centre for Interactive Intelligent Systems

Head: Professor E Miranda & Professor A Cangelosi
Email: eduardo.miranda@plymouth.ac.uk

Research interests:

1) Natural language interaction and adaptive systems
2) Natural object categorisation
3) Adaptive behaviour and cognition
4) Visualisation
5) Semantic web

http://www.tech.plymouth.ac.uk/Research/computer_science_and_informatics/

Interdisciplinary Centre for Computer Music Research

Head: Professor E Miranda
Email: eduardo.miranda@plymouth.ac.uk

Research interests:

1) Computer-aided music composition
2) New digital musical instruments
3) Sound synthesis and processing
4) Music perception and the brain

http://cmr.soc.plymouth.ac.uk

Centre for Robotics and Intelligent Systems

Head: Dr G Bugmann
Email: guido.bugmann@plymouth.ac.uk

Research interests:

1) Cognitive systems
2) Social interaction and concept formation through human-robot interaction
3) Artificial intelligence techniques and human-robot interfaces
4) Cooperative mobile robots
5) Visual perception of natural objects
6) Humanoid robots

http://www.tech.plymouth.ac.uk/socce/ris/

Contents

SECTION 1 Network Systems Engineering

SECTION 2 Communications Engineering and Signal Processing

SECTION 3 Information Systems Security & Web Technologies and Security

SECTION 4 Computing, Computer Applications, E-commerce & Interactive Intelligent Systems

Section 1

Network Systems Engineering

Security Usability: A Survey of End-Users

A.Jusoh, S.M.Furnell and D.Katsabas

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

This paper examines the reasons why security technologies are not used correctly by users. It is a major concern in a computer security world if users fail to use tools available to protect their own system and data. To investigate this issue, a survey was conducted to assess the extent to which security features in popular applications are understood by users. A total of 313 respondents were successfully gathered, and it was discovered that users may be simply ignorant of security, or they may not be getting enough information on using the related technologies. The main contributing factors have been discovered which includes users, interface and computer application itself.

Keywords

Usability, Security, Users, Interface, Human Computer Interaction

1. Introduction

Security is an important factor that should be considered when connected to the Internet. Even if users have nothing stored in the system that they consider important, their computer can be a weak link, allowing unauthorised access to the organisation's network and information. The design of interface plays a big role to connect users and the usable functions within the system. In a usability study of PGP 5.0, Whitten and Tygar (1999) stated that security software will not be satisfactory if users do not know how to use it. They had worked on a good user interface but it was still insufficient to give information in an effective manner to non-technical users. The security on wireless network attracted Balfanz et al. (2004) to look at different versions of Public Key Infrastructure (PKI) in terms of providing usable security for users. The traditional PKI deployment was complex, with installation requiring a total of 38 steps within 140 minutes. This was in complete contrast to a user-friendly version, which required just four steps and an average of 1 minute and 39 seconds. It is not surprising if users give up using technologies that require a lot of time and effort to make them functional.

The usability of the systems becomes an issue because many users find it difficult to follow the steps to enable security features in their system. Users should know what type of security features they want to apply and they must be able to find the desired functionality. They also have to be provided with maximum information on how to determine the level of protection and when they should apply the features. A survey has been conducted to gather information about the usability of security. The

purpose of the survey is to focus on users' understanding of the security features available in a range of popular applications: Internet Explorer (IE), Word and Outlook Express. This paper investigates and analyses the factors contributing to why security technologies are not used correctly.

2. Determination of Security Technologies Usability

The survey (entitled *Assessing the Usability of Computer Security Features*) was distributed in paper and web format. Total of 313 responses were gathered, with an almost equal split between male and female. Most of the respondents were aged between 17-29 years, with degree or post graduate education. Thus the majority of respondents were educated persons within an age group that grew up with information technology. From the total responses, it shows nearly all respondents were frequently using a computer either in office or home. Furthermore, most of the respondents classed themselves as either intermediate or advanced users, and generally had a good knowledge of threats and security technologies. In terms of the named threats, 302 respondents were aware of *Viruses*, *Spam* (288), *Spyware* (282), *Hacking* (280), *Worms* (275) and *Phishing* (69). With the percentages of more than 80%, most respondents claimed to know the role of Firewalls, Auto-Updates and Virus Protection. More than 70% of respondents are aware of the security features available in Internet Explorer, but the pattern changes for the other applications, with only 58% of respondents aware of security in Word, and less than 40% awareness for Outlook Express.

The sub-sections that follow examine the responses observed in relation to each of the applications under consideration.

2.1 Internet Explorer

Although users claim they are aware of security features in IE, many of them still do not understand the description of the default security level setting (see Figure 1). They are aware about it but did not spend their time to explore further and get to know the features available.

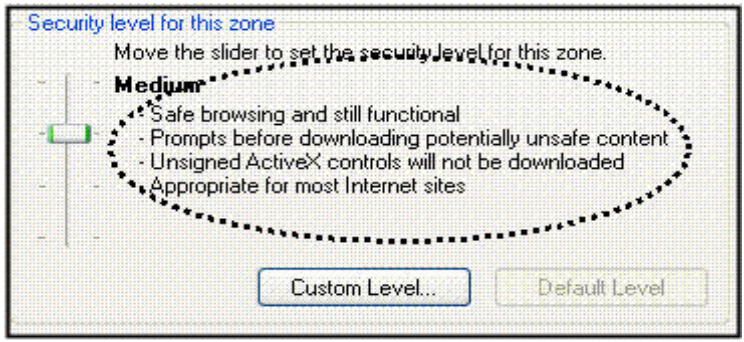


Figure 1: Security Level setting in Internet Explorer

Option	Respondents	
	Number	Percentage
Yes	197	63 %
No	105	34%
Nil	11	3%
Total	313	100%

Table 1: Understanding the Security Level setting

Furthermore, only 59% of respondents understand the difference between trusted and restricted sites. It is a quite large percentage of respondents who do not know or are not sure about the difference. Tyler (2001) explains the meaning of the both sites:

- Trusted Sites: Users should only use this Web sites pages if they believe it is safe and it will not upload harmful content to the computer. The default security level for this site is Low.

- Restricted Sites: Use this zone for Web sites and pages that users access but do not completely trust because it suspected the sites may send potential harmful content to computer. The default security level is High.

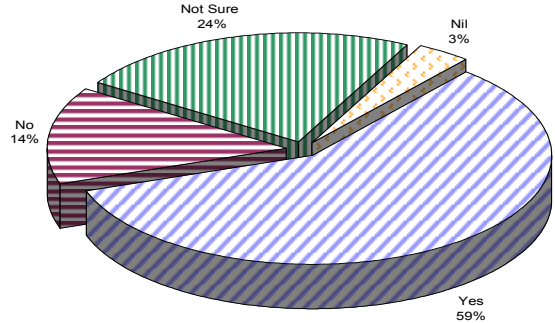


Figure 2: Respondents' understanding of the 'Content Zone'

The same scenario happened when they were asked about ActiveX (see Table 2). Although many of them have heard of it, almost half of these (47%) do not actually know what it means. More than 50% of the respondents said they do not know how to decide if they presented with the option of running active content. It shows that they are exposing themselves to a risk, as their decision could harm the computer.

Option	Respondents	
	Number	Percentage
Yes	128	41 %
No	179	57%
Nil	6	2%
Total	313	100%

Table 2: Knowing how to make decisions with ActiveX content

2.2 Microsoft Word

Word is a popular Microsoft application used by many of the respondents on a daily basis. However, from the result of the survey, it is quite surprising that many do not know about the security features available within the application (see Table 3). As they are not aware, it is reasonable that not many of the respondents use the security features available for their documents.

Security Features	No. of aware respondents
Password	194
Encryption	52
Digital Signature	29
Other	37

Table 3: Awareness of security features used in Word documents

An interesting finding in relation to Word security features is how users interpret the password functionality. Figure 3 shows the dialog box that is presented when a document is password protected to prevent unauthorised modification. Respondents were asked to indicate what they believed this dialog to mean, with the options and their responses listed in Table 4 (the correct answer is that '*The document cannot be changed without password*'). 61% answered correctly, 24% of respondents misunderstood the dialog, and another 13% were not sure. 37% of respondents cannot answer correctly although most of them claimed they were using password as the main protection.

Figure 3: A Password dialog in Microsoft Word

Option	Respondents	
	Number	Percentage
The document cannot be opened without a password	75	24 %
The document cannot be changed without a password	191	61%
Not Sure	41	13%
Nil	10	2%
Total	313	100%

Table 4: Understanding of the Password dialog from Figure 3

Apparently respondents might use the security technologies inappropriately since they could not understand them properly. Because the 'open read only' sentence is at the same line as 'Enter password to modify', users thought they also need password to enable them to open and read the document.

As Sutcliffe (1995) has indicated, the issue of HCI design analyses what people do with the computer systems and their interfaces to understand the user's task and the requirements. Designing of application should help user to fulfil and match the system characteristics. Sutcliffe identifies seven of basic principles of HCI which are consistency, compatibility, predictability, adaptability, economy and error prevention, user control and structure interface.

2.3 Microsoft Outlook Express

The worst results were in relation to Outlook Express, where most of the answers show respondents know nothing about security features available in it. More than 60% of them do not understand the Digital ID, Encryption, the Advanced Security setting and also they are not using the Sign and Encrypt option when sending messages. From the comments added by some of the respondents, they said they rarely use the application and some did not use it at all. Since they do not use the application, they do not notice the use of security features available to protect them. However, they should also acknowledge the use and function of each security features available in case they need to use them in the future.

Option	Digital ID	Encrypt	Advanced Security
Yes	82 (26%)	87 (28%)	85 (27%)
No	209 (67%)	196 (63%)	211 (68%)
Nil	22 (7%)	30 (9%)	17 (5%)
Total	313 (100%)	313 (100%)	313 (100%)

Table 5: Understanding the options available for Digital ID, Encrypt and Advanced Security settings

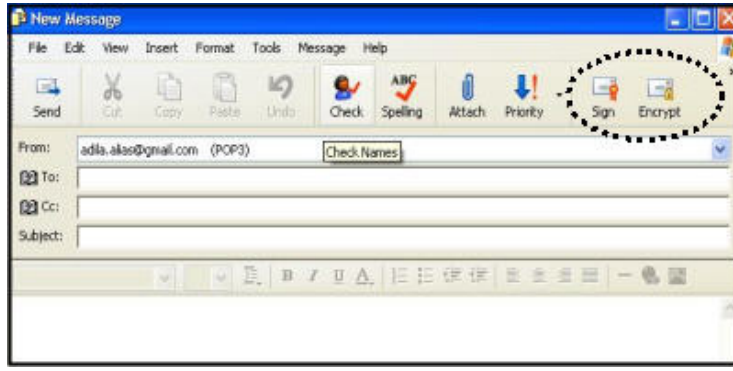


Figure 4: Sign and Encrypt options while Sending Message

3. Contributing Factors

Factors of users, interface and the application itself all play a part in determining the usability of security technologies. The tendency of users to ignore, neglect and take things for granted will lead to a very bad situation for the future of security technologies. Users with a bad behaviour will contribute nothing to the usability of security technology when they could not use the tools correctly. They were aware of the security features available in each of the applications, but awareness is just not enough to make computer safe without taking any further action. Those who have an IT department in their company may normally leave all of the computer responsibilities to the IT personnel. However, they should not forget the fact that all the basic settings and security features are often the user's own responsibility. Users should learn how to protect their own computers so that they are safe from any threats. They should realise they are the main contributing factors that could harm the computer and probably affect the system of an organisation.

An unfriendly interface presents a big problem to the usability of security technologies. If users interpret the interface wrongly, they will get a wrong idea about the security features as well. As a result, they may use the security technologies inappropriate way. Wording used in any application must be as simple as possible and easily to understand by users. Furthermore, the layout of the application must also be designed in such a way that it will not be so complicated to be understood.

The application itself should give more help functions to users to make it easy to understand how to use the security features. Many of the current help functions are not helping users to understand certain features much better. If users utilise the 'help' system in the hope that it will explain what something means, they will often be very disappointed as the help function will give only a very brief description. An example is the context-sensitive help in Microsoft Word when choosing the encryption type. It is not helping much in giving clues to users on how to determine the type of encryption. It explains only basic information on what the page is all about. The '?' function should play an important role in helping users understand the features available more detail and precise. If the information could be delivered in a very short, compress and efficient, it will help users a lot in understand the security technologies in a correct way.

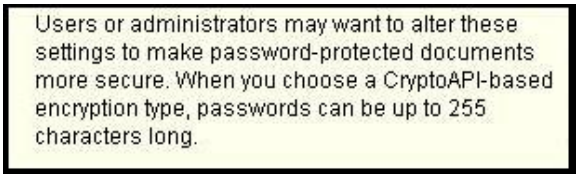


Figure 5: Context-sensitive help for the Encryption Type in Microsoft Word

4. Conclusion

This paper has discussed the usability of security technologies based on a survey distributed to 313 respondents. The features in Internet Explorer, Word and Outlook Express have been analysed to obtain assessments for each application. The results give insights into how users perceive the security technologies based on their awareness. It seems that some of the applications are consequently lacking in terms of usability since users do not understand them.

The contributing factors have been recognised in an attempt to understand and clarify why the security technologies are not used correctly, and consist of users, interface and the application. If users could use the security technologies in appropriate and right manner, the usability of security will not be an issue anymore in any computer application.

5. References

Balfanz, D., Durfee. G., Smetters, D.K. and Grinter, R.E. (2004), "In Search of Usable Security: Five Lessons from the Field", *Security and Privacy,* Vol. 2, No. 5, pp. 19-24.

Faulkner, C. (1998), *The Essence of Human – Computer Interaction*. Prentice Hall, Cornwall, UK.

Sutcliffe. A. G (1995), *Human – Computer Interface Design* (2nd Ed.) Macmillan Press Ltd. London, England.

Tyler, D. (2001), *Windows XP – Home and Professional Editions*. SYBEX Inc. USA.

Whitten, A and Tygar, J.D. (1999), "Why Johnny can not encrypt: A usability evaluation of PGP 5.0", in *Proceedings of the 8th USENIX Security Symposium*, Washington, D.C., USA, August 23–26, 1999.

Super Fast Retransmit: A Proposal to Improve the Performance of Short-Lived TCP Connections

A.J.Tarr and B.V.Ghita

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: research@alextarr.co.uk; info@network-research-group.org

Abstract

This paper presents a method of improving the performance of TCP (Transfer Command Protocol) based short-lived connections, such as web transfers. This paper presents the methods used to examine the characteristics experienced while communicating with a number of Internet sites, and finds that constancy in the RTT (Round Trip Time) and throughput can be observed for periods extending to a number of days. The research identified a temporal correlation between the three-way handshake and subsequent measurements of the RTT; by using this correlation, the algorithm is able to detect the signs of a packet loss and retransmit the affected segment before a standard implementation of TCP is able to recover the lost segment. Following the proposal of the algorithm, simulations were performed to study the performance of the proposed algorithm. In summary, it was found that use of the Super Fast Retransmit algorithm provided enhanced performance showing the average duration of a connection was at least 8% lower compared to TCP Reno.

Keywords

IP Network Performance, Short-Lived TCP Connections

1. Introduction

While network connections have experienced exponential increases in performance, Internet usage has also grown to become the major source of traffic over IP networks. Web transfers, *i.e.* HTTP connections, can be characterised by the small amount of data they transfer, with the majority of connections just a few kilobytes in size; these transfers therefore typically last for less than a second and are termed "short-lived connections". Although TCP and its component algorithms perform well, the increased demand to detect and utilise larger path capacities over ever shrinking time periods leaves considerable room for improvement within TCP.

The research presented in this paper aims to investigate the current state of the Internet by collecting a set of traces for web transfers and establishing the trends which can be observed. The findings from this investigation will then be used as the basis to propose an algorithm to enhance TCP which will improve the performance of connections, especially those of a short-lived nature. The proposal that results from this paper will therefore be an important part of improving the performance of the Internet, especially when used for web-browsing activities, with the analysis adding to the understanding of the dynamics of the Internet.

The paper is organised as follows: section 2 presents the background literature and current algorithms in TCP. Section 3 details the method used to collect traces of short-lived connections, with section 4 detailing the analysis of the characteristics that occurred during the captured traces. In section 5 the Super Fast Retransmit algorithm is proposed. Section 6 introduces the tools that will be used to validate the proposal, with section 7 presenting the findings of the simulations revealing the performance improvements offered by the proposed algorithm. Finally, section 8 presents the overall conclusions and ideas for future work.

2. Related Work

The notion of path constancy was introduced in (Zhang, 2001), where the research focused on how well past values can be used to predict future path characteristics. The research explored three definitions of constancy: mathematical, operational and predictive; whereby a path could display any combination of these, e.g. appearing statistically non-constant while being entirely predictable in its future performance. The current work uses the concept of mathematical and constancy when examining the collected traces. The work, based upon data collected between 1999 and 2001, identified constancy which existed for period in excess of several minutes; throughput displayed change free regions of up to 20 minutes, with delay constant for periods up to 30 minutes; the current investigation re-examines the levels of constancy seen in 2005.

Modern TCP implementations are based upon the algorithms presented in (Stevens, 1997). One of the algorithms, Slow Start, controls the initial transmission in the period before the first loss occurs; the algorithm exponentially increases the amount of in-flight data until the network capacity is detected. TCP detects losses using a timeout that occurs when a segment has not been acknowledged within a period of time determined by the observed RTT; to reduce the time taken to detect a pack loss, Fast Retransmit uses the receipt of three duplicate acknowledgements to signal that a retransmission is required. Importantly, Fast Retransmit therefore requires at least three packets to be transferred following the lost segment.

3. Trace Data Collection

To investigate the behaviour of TCP connections a collection of traces was obtained upon which further analysis could be performed. The analysis will answer two important questions: "What level of constancy is experienced in the Internet in 2005?" and "Are there any correlations between the characteristics of a connection?"; by answering these questions an improved understanding of the behaviour of short-lived TCP connections can be gained.

The collection technique aimed to recreate a realistic usage scenario; this was achieved by monitoring HTTP transfers with using automated requests which provided a predictable stream of connections, and the ability to examine the behaviour of a site over a prolonged period. The homepage of a selection of sites were requested using the *wget* (Wget, 2005) tool at one minute intervals, with the packet streams captured using *tcpdump* (Tcpdump, 2005), and the resulting files

were stored for post-processing. The selected sites represented a range of genres *e.g.* shopping, news *etc*. in the UK and US. The sites used were: BBC, BBC News, CNN, DigitalSpy, Exeter University, Google UK, Tesco, The Register and Yahoo UK.

The sites used were referenced by IP Address rather than DNS host name; this allowed the specification of a single end-point and avoid the effects of load-sharing, which might otherwise bias the results. To ensure the RTTs observed were accurate and had not been skewed by the presence of a caching proxy server, the RTTs were also sampled using *ping*. The times returned by *ping* matched closely with the RTT measured from the HTTP requests, with the majority of connections exhibiting a ratio between TCP and *ping* RTTs of between 1.0 and 1.1.

To ensure the results reflected a wide range of situations, over 250,000 traces were collected between April and June 2005. With the traces collected from two locations, both running Linux and both locations were on the campus network at the University of Plymouth, which is connected to the Internet via JANET. The first location was a workstation in a student laboratory, while the second was a monitoring station on the network backbone.

4. Analysis of Collected Traces

Based upon the collected data, the *tcptrace* (Osterman, 2005) tool was used to perform per-flow analysis, with the resulting characteristics of each connection extracted into a CSV file. The output from *tcptrace* was then used to establish the period for which constancy could be observed, and presence of any trends or correlations in the characteristics of the traces.

4.1 Path Constancy

To asses the constancy of a path, the behaviour of two key characteristics: RTT and throughput, were examined; these are critical in affecting the overall performance which an end-user will experience. The cumulative distribution of the RTT and throughput for four of the sites monitored is presented in Figure 1.

As can be observed in Figure 1 (left), the majority of connections experienced relatively constant RTTs, the level of constancy can be confirmed with standard deviations of 1.8ms, 5.5ms, 3.63ms and 3.6ms for BBC News, CNN, Google and Tesco respectively; these values illustrate the small variation in RTTs which the paths experienced. In all cases, the RTTs were almost identical for the duration of the trace collection with few outlying results observed.

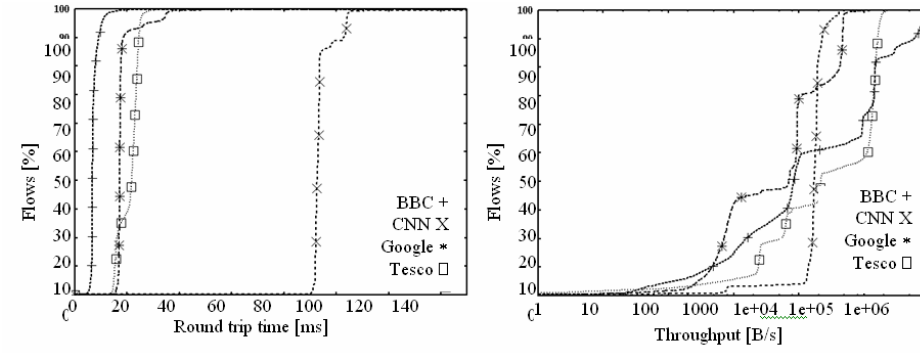


Figure 1: Cumulative distribution of four sites taken from data sets ω_2 to ω_4 showing: left - RTT; right - throughput.

The throughput, as shown in Figure 1 (right), experiences increased variability over time when compared to that experienced for the RTT. Most sites demonstrated clustering of throughput around a number of distinct levels over the observation period, with these levels occurring concurrently for the entire length of the trace, rather than at different throughputs for disjoint periods. The high levels of variation can be confirmed by observing the standard deviations of the sites, the values returned were: 241kBytes/s, 6kBytes/s, 19kBytes/s and 68kBytes/s respectively for BBC News, CNN, Google, and Tesco.

4.2 Throughput Analysis

As observed in the previous section, the throughput of the captured traces occurred at a number of levels for the duration of the data collection sessions, in this section the cause of these different levels are investigated. The investigation focussed on the Yahoo UK site which exhibited 11 main throughput levels causing a layered appearance over time, see Figure 2 (left).

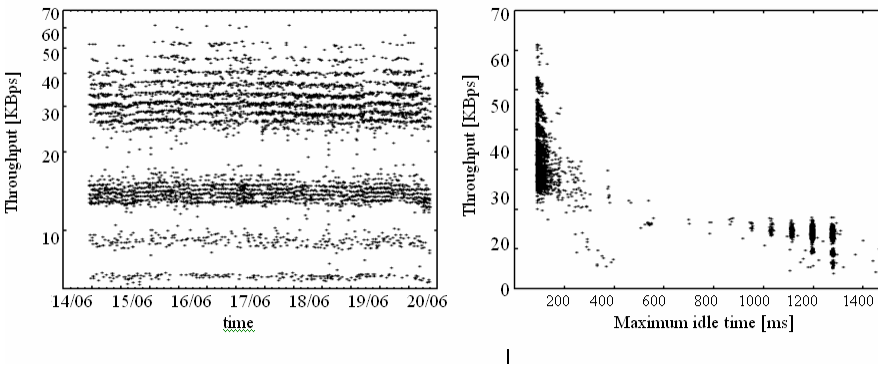


Figure 2: throughput for Yahoo UK shown: left - over time; right - against the maximum idle-time for the connection.

Examining the number of packets *tcptrace* reported the workstation as receiving identified variation between 22 and 32 packets. Part of this variation is due to a fluctuating page size, which accounts for between 22 and 25 packets being observed.

Examining the throughput, a connection receiving 25 or fewer packets achieved between 20 to 60kBps; while connections receiving more than 25 packets experienced throughputs of less than 20kBps. Therefore, if more than 25 packets were received, some of these had been retransmitted. The characteristics from *tcptrace* revealed that each connection sent, on average, nine duplicate acknowledgements to the server; this illustrates the high level of either re-ordering or loss of the studied path.

By examining the maximum idle-time (the longest period between two packets in the same direction), it can be observed that a connection generally achieves a lower throughput the longer the maximum idle-time, see Figure 2 (right). When a loss occurs, the maximum idle-time represents the time taken to detect, and resend the segment. It could therefore be observed that a large number of connections experienced an idle-time greater than 500ms, therefore indicating that a timeout had triggered the retransmission rather than Fast Retransmit. As a connection which experiences a timeout achieves significantly lower throughput, the detection of lost packets is therefore a key factor in determining the performance of a connection.
These investigations illustrated how TCP and the Fast Retransmit mechanism can fail when insufficient packets exist to maintain the return data flow. The Fast Retransmit algorithm requires three duplicate acknowledgements to be received to trigger a retransmission; therefore, if less than three packets are sent following a lost packet, Fast Retransmit is unable to recover the loss and a timeout must be used to correct the loss.

4.3 Three-Way Handshake Observations

Examining the RTT statistics reported by *tcptrace, a* link between the RTT for the three-way handshake, and the subsequent RTTs experienced over the connection was observed. The ratio between these values was computed and plotted in a cumulative distribution, as shown in Figure 3; analysing the ratios reveals that more than 80% of connections demonstrated how the three-way handshake value accurately predicted the average RTT within 10%. Further, 98% of the connections experienced a maximum RTT of three times or less than the original three-way handshake predictor. It was therefore concluded that the three-way handshake could be successfully used to predict the subsequent RTTs that a connection will experience.

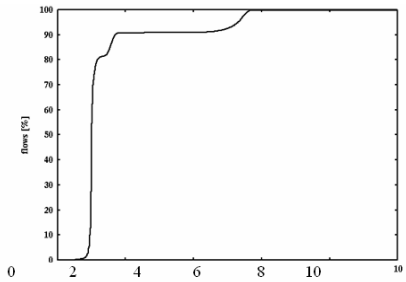


Figure 3: ratio between three-way handshake RTT and subsequent RTT for set ω_2.

Figure 3 shows a number of connections with a ratio between five and six; these values all originated from the DigitalSpy site which contains a forum, and were the result of a large (more than 15 times the other RTTs for that connection) delay between the HTTP request and response. As a result, this site was excluded from further investigations as it was not typical of the values obtained for the other sites.

5. Super Fast Retransmit Algorithm

The proposed algorithm improves the performance of short-lived connections by detecting a loss more efficiently than either Fast Retransmit or the standard TCP timeout mechanism. The proposed algorithm uses the relationship between the three-way handshake RTT and subsequent RTTs as a basis to estimate the period in which a packet should be acknowledged; if an acknowledgement has not been received, the first unacknowledged segment is retransmitted.

The Super Fast Retransmit algorithm functions as follows:

- Step 1: Establish the three-way handshake RTT to determine the timeout period.
- Step 2: When sending a segment a timer set to expire after the timeout period is started.
- Step 3: For each new acknowledgement, the timer is either cancelled if outstanding data is cleared, or restarted if unacknowledged segments still exist.
- Step 4: If the timer expires, the first unacknowledged segment is retransmitted. The timer is not restarted until a new segment is transmitted.

It is obviously unrealistic to expect all packets to be acknowledged within a single RTT, therefore the proposed algorithm includes a parameter to determine the algorithms timeout period, known as the "multiplication factor". The largest ratio between the three-way handshake RTT and the maximum RTT was found to be three, therefore this value is used. The three-way handshake RTT will be multiplied by the multiplication factor to determine the timeout used to generate a retransmission. For example, the Yahoo traces had a mean RTT of 84ms, thus the Super Fast Retransmit algorithm would set the timeout to three times 84ms, thus 252ms; this would therefore at least halve the maximum idle-time experienced by a connection.

As the algorithm can potentially generate unnecessary retransmissions if the RTT should alter over the life of a connection, another parameter determines a maximum volume of data that the proposed algorithm functions over, this is known as the "cut-out limit". Based upon the average page size of 72kBytes observed in (Ghita, 2003), the algorithm uses a cut-out limit of 80kBytes. Following the successfully acknowledgement of the 80kBytes of data, the algorithm is disabled and can generate no further retransmissions.

6. Validation Environment

To asses the improvement in performance offered by the proposed algorithm, a number of simulations were performed using the *ns-2* network simulation tool (ns-2, 2005). Within *ns-2,* the *PackMime* (PackMime, 2005) traffic generation tool was used to produce aggregate Internet packet traffic over a link. *PackMime* uses a stochastic model to control the connection initialisation rate and a heavily-tailed HTTP response size distribution. This provides a simulation environment similar to that experienced on the Internet.

The simulations were all based upon an underlying 5Mbps bottleneck link, chosen to allow realistic volumes of aggregate traffic yet providing suitable simulation durations; the delay was set to either 10ms or 60ms representing UK and US sites respectively, with the delays having been verified through *traceroute*. *PackMime* starts connections at a specified rate where the selection of value is used to control the utilisation seen over the bottleneck link.

The Super Fast Retransmit algorithm was implemented by modifying the implementation of TCP Reno (named Full-TCP), included with *ns-2*. This allowed observation of the difference in performance between the modified implementation and the standard TCP Reno provided with *ns*-2 to establish the effect to the proposed algorithm on performance.

7. Simulation Results

The first simulation performed aimed to observe the general effects of utilising the Super Fast Retransmit on an environment with a 10ms link delay, and a utilisation of 50% over the bottleneck link, chosen to represent a moderately congested link. The simulations were performed twice, once using the original TCP Reno from *ns-2*, then with the modified TCP Reno implementation; therefore allowing the performance to be compared.

The results of the first simulations proved extremely encouraging: the average connection duration under TCP Reno was 297ms; while when Super Fast Retransmit was used, the average connection duration dropped to 268ms. This equates to an 8% improvement in performance when using the proposed algorithm. Further simulations confirmed the behaviour of the algorithm under a variety of scenarios and to the validity of the parameters, *i.e.* cut-out limit and multiplication factor, which control the behaviour of the algorithm.

7.1 Network Conditions

In the simulated network conditions, the effect of different network characteristics upon the performance improvement offered by Super Fast Retransmit were investigated. The impact of delay was initially examined, by considering two scenarios to represent a connection to a UK and US sites, using one-way delays of 10ms and 60ms respectively. As can be seen in Figure 4 (left), the algorithm

improved performance under both scenarios, although a larger improvement was observed for the 10ms delay. Both simulations occurred with 25% utilisation.

Comparing the simulations using TCP Reno and Super Fast Retransmit, the proposed algorithm improves the average connection duration by 39% for a delay of 10ms, and 8% for a delay of 60ms. The significantly lower improvements for a network with a delay of 60ms can be explained as the timeout used by Super Fast Retransmit is set at 360ms, compared to 60ms on a network with a delay of 10ms. Therefore, on the 60ms network, the proposed algorithms timeout period is closer to the standard timeout mechanism used by TCP, this reduces the time benefit of using the proposed algorithm compared to TCP Reno.

The effect of background traffic levels on the performance of Super Fast Retransmit was also investigated by generating utilisation levels of 25%, 50% and 75% over the backbone link in the simulation. The results, as shown in Figure 4 (right), indicated that the algorithm improves the performance of a connection under all scenarios. The largest improvement in performance was observed for 50% utilisation; where the connections are experiencing numerous losses due to congestion, but critically the link is not completely saturated thus by the time Super Fast Retransmit retransmits the lost segment, sufficient capacity exists to allow the segment to transmit successfully. The resulting improvements in performance, based upon the average duration of a connection, showed Super Fast Retransmit was 23%, and 38% and 8% faster for utilisations of 25%, 50% and 75% respectively.

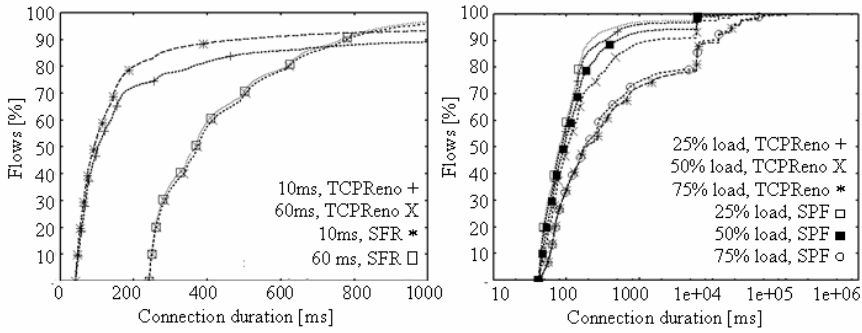


Figure 4: The effect upon the performance improvement of Super Fast Retransmit under varying: left - network delays; right - traffic levels.

7.2 Algorithm Parameters

The proposed algorithm contains two configuration parameters, the multiplication factor and the cut-out limit; the setting of these critically determines how well the algorithm functions. Simulations were therefore performed to identify the optimum configuration of the algorithm. The results identified that the highest performance increase occurred when a multiplication factor of two was used. This finding is the result of lower variations than normally occur due to the absence of application layer interactions within *ns-2* and *PackMime*. Surprisingly, the simulation reported the second best value as a multiplier of four with practically identical average connection durations but considerably fewer retransmissions caused by Super Fast Retransmit.

The cut-out limit responded as expected, with a larger limit allowing more connections to benefit for their entirety from the proposed algorithm. Importantly, as the cut-out limit increased, the relative performance improvement decreased, therefore it was concluded that a limit of 80kBytes represents the optimum value based upon the average web page size.

7.3 Summary

In summary, the proposed algorithm significantly improves the performance of short-lived connections, even under situations such as long delay links and high traffic loadings. The results presented here indicate the algorithm reduces the average duration of a connection by more than 8%; and, importantly, the performance is never lower than as for TCP Reno.

8. Conclusions and Future Work

This paper used the analysis of a large collection of trace data to establish the characteristics of short-lived TCP connections. This identified that current TCP implementations are unable to recover from a lost packet when less than three packets follow the packet which is lost, therefore insufficient information is returned which would allow timely recovery from the loss. The analysis of the collected traces revealed that constancy is evident over extended time periods, which extend to several days, suggesting that current Internet backbones have sufficient capacity to avoid congestion occurring; critically to this research, the relationship between the three-way handshake and subsequent RTT values has been identified.

Based upon these findings, an addition to the current TCP algorithms, named Super Fast Retransmit, which causes the first unacknowledged segment to be retransmitted following a period equal to three-times the original three-way handshake RTT. To validate the proposed algorithm, a series of simulations were performed under the *ns-2* network simulator to examine the algorithms performance under a variety of network conditions. The results of these experiments indicated a decrease in the average connection duration that was in excess of 8%, with a maximum observed improvement of 39% over a moderately loaded link.

Future work should aim to improve the proposed algorithm when multiple packet losses have occurred, a scenario which is currently handled poorly by the proposed algorithm; for example this problem is particularly evident in long lived connections where halve the congestion window is lost at the end of slow start.

Also, additional examination of the improvements in performance the Super Fast Retransmit algorithm provides must be performed over a live network; this will allow the results of the simulations, as presented in this paper, to be confirmed under a real scenario.

9. References

Ghita, B.V., Furnell, S.M., Lines B.M. and Ifeachor, E.C. (2003) , "Endpoint study of Internet paths and web pages transfers", *Campus-Wide Information Systems*, Vol. 20, Issue 3, pp90-97

ns-2, (2005) "The Network Simulator - ns-2", http://www.isi.edu/nsnam/ns/ (Accessed 17 July 2005)

Osterman, S., (2005) "Tcptrace Home Page", http://www.tcptrace.org/ (Accessed 21 January 2005)

PackMime, (2005) "Web Traffic Generation in NS-2 with PackMime-HTTP", http://www-dirt.cs.unc.edu/packmime/ (Accessed 27 July 2005)

Paxson, V., (1999) "End-to-End Internet Packet Dynamics" *IEEE Transactions of Networking*, Vol. 7, Issue 3, pp272-292

Stevens, W. (1997) "TCP Slow Start, Congestion Avoidance, Fast Retransmit and Fast Recovery Algorithm", RFC2001

Tcpdump, (2005) "Tcpdump Public Repository", http://www.tcpdump.org/ (Accessed 21 January 2005)

Wget, (2005) "GNU Home Page", http://www.gnu.org/software/wget/ (Accessed 21 January 2005)

Zhang, Y., Duffield, N., Paxson, V., and Shenker, S., (2001) "On the Constancy of Internet Path Properties", *Proceedings of ACM SIGCOMM Internet Measurement Workshop,* November 2001

Security policies for small and medium enterprises

A.Kanellos, V.Dimopoulos and N.Clarke

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Over the last few years, the advances in information technology have brought many changes to the business environment. More and more businesses now try to take advantage of the technologies and applications that promise to help them improve many of their tasks. Companies of whatever the size are now becoming increasingly reliant on the Internet to fulfill many of their basic functions and ultimately to remain competitive in the demanding marketplace. Many of them are now conducting transactions over the Internet using email and other applications to make some of their tasks more efficiently and to improve their relationship with their customers. But as they increase the level of their connectivity to the Internet and the applications surrounded by the web, they increase the potential of a malicious security breach that could harm the business continuity and that could result ultimately in significant losses. Hence some sort of protection measures are now considered necessary for businesses of any size with the security policies regarded as the first and most important of them. Controversially to that aspect small and medium enterprises are found through major surveys like the DTI and the CSI/FBI security surveys to lack sufficient protection since the majority of them reported not to adopt or having difficulties to adopt security policy practices. In general, the findings from various surveys are quite discouraging for the small and medium firms as far as security protection is concerned. Because of this situation some sort of solution was investigated and provided by this paper by taking into consideration the factors that minimise the adoption of security policies in the SMEs. In addition, the solution described takes into account the potential of using baseline guideline documents.

Keywords

SMEs, information security policies, information security

1. Introduction

In the recent past, a number of surveys were conducted worldwide on the security issues taking into investigation businesses of any size and subject. Such surveys that were conducted by well known authorities like the DTI Survey 2004 from the Department of Trade & Industry in the UK and the CSI/FBI Computer Crime and Security Survey 2004 Computer Security Institute in the United States indicated the behaviour of SMEs towards the use of security policy practices.

The surveys indicated that the information security is related to the size of the company and, being more specific, the level of approach to security practices (like the use of security policies) shown by a company decreases as the size decreases. Small and medium sized companies are found to be missing security practices and most importantly security policy documents. There exist, though, some factors that

hinder the small and medium enterprises from adopting adequate security practices. These factors as well as others were investigated in order to identify the specific needs of the small and medium sized companies towards security procedures and to provide consequently proper solutions for them. This paper presents an investigation into the lack of security practices in SMEs identifying the reasons for the minimised adoption of security policies and proposing some solutions that aim to improve the security practices within a SME organisation.

2. Security survey findings

As a start point in this paper the findings around the security practices in small and medium companies are going to be presented. The findings below that were collected from major security surveys illustrate the security behaviour in small and medium sized businesses and, as a result, in what extend they adopted security measures and especially security policies. The surveys were conducted on the businesses of any size but since the majority of them were small and medium businesses we can state that their results can give us a clear idea about them. The surveys that were used were the following:

- DTI Survey 2004 - Information security breaches survey 2004 technical report
- CSI/FBI - Computer Crime and Security Survey 2004
- Ernst and Young - Global Information Security Survey 2004
- Deloitte Touche Tohmatsu Org. - 2004 Global Security Survey
- 2005 Australian Computer Crime and Security Survey
- 2002 Information Security Magazine Survey – Does size matter?
- E21-MagicMedia - E-security survey report

A major finding from the above surveys was that the majority of small and medium enterprises lack documented security policy practices. Only the 39 percent of the small and medium enterprises adopt a security policy according to the E-security survey report. In addition to the effect on the use of security policy, the size of the organisations was found to be also important in the security awareness of the organisation. More specifically, we can say that the larger the organisation, the more mature its security awareness. (Melek, 2004)

Moreover, outsourcing of security functions seems to be a more common finding in larger organisations rather than in the smaller ones. Only a very small percentage of the small and medium sized enterprises considered outsourcing of security functions with the lack of budget reported as their great obstacle. Figure 1 below illustrates the percentage of respondents according to the level of outsourced functions.

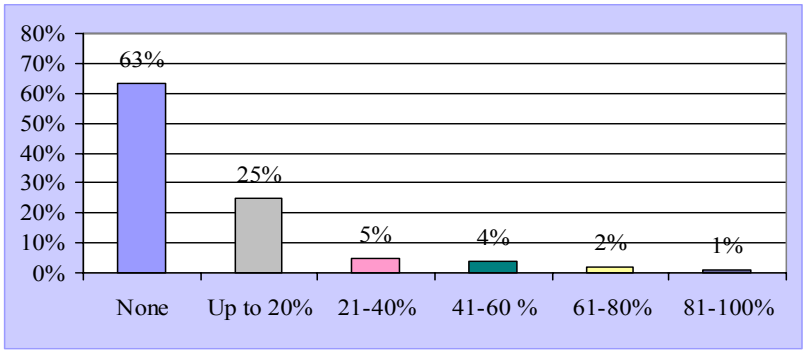


Figure 1: Security functions outsourced by businesses
Source: CSI/FBI - Computer Crime and Security Survey 2004

It can be easily derived that only a small percentage (25%) of respondents, and in fact SMEs, outsourced a respectable portion (up to 20%) of their security practices while the majority of them outsourced a very small percentage or no percentage at all of their security practices.

Another finding is that SMEs spend only a small portion of their IT budget on security. Most of the SMEs tend to spend a small percentage of their budget on security either because they have restricted budgets in general or because they do not feel that security is an important issue. So the amount spent on IT security is not adequate enough to support the security practices and provide an acceptable level of robust security (S. Timms *et al*. 2004). Figure 2 below gives a notion of the money spent on security.

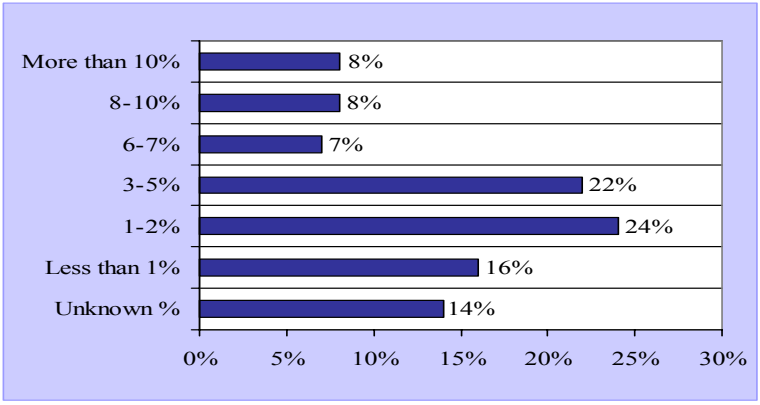


Figure 2: IT budget spent on security
Source: CSI/FBI - Computer Crime and Security Survey 2004

The more disappointing fact is that in addition to spending a small percentage of their annual budget on IT security, they also often feel that their organisations are adequately protected. And in addition to their minor investments on security and the poor security measures, they feel that they are adequate protected. Because of that

many of the enterprises failed to conduct a regular assessment and re-evaluation of their security policy even if an 82% of the organisations conduct some kind of security audits that could be very useful for something like this. (Gordon *et al*. 2004)

In addition to the above, the education and training process was found to be poor in many organisations despite that it is perceived valuable in many of the aspects of security like the network security, the security management, the economic aspects of computer security, the security systems architecture, the investigations and legal issues, the cryptography, etc. (Gordon *et al*. 2004) It is described that about 70 percent of the respondent companies failed to list employee training/ education as their important priority in the IT budget. (Ernst and Young)

The risk management process that is the main part in the process of creating a security policy and hence protecting the company's assets was also found to be missing in most of the enterprises and hence SMEs (Melek, 2004). There is even a percentage of them that failed to complete successfully the risk management process due to mistakes in the process. Because the small and medium enterprises in general lack personnel with sufficient expertise they are not able to recognise the assets and their importance as a first step and the security breaches that may consider a potential risk as a next step. On the contrary, 89 percent of the enterprises feel that having a risk management process within the organisation was either extremely or very important. (Melek, 2004)

3. Reasons for the minimised adoption

In the previous section we have identified through surveys that small and medium enterprises lack sufficient security protection due to the lack of security policies and we also described some reasons for this minimised adoption. In this section we are going to give a better description of these factors and provide some other too.

3.1 Lack of security policy guidelines for SMEs

The lack of security policies guidelines for SMEs can be attributed to two separate factors. First of all it is the fact that it is difficult for someone to find security policy guidelines that are written especially for SMEs and secondly it is the fact that even if they find some papers that claim to provide guidance for the SMEs these papers are not covering all the aspects of security. Most of the security white papers and any baseline guideline documents are mainly written with a general subject so that any size of business is supported. Moreover, by going through some of them it can be realised that they are more appropriate for use by security specialists with a good knowledge background on security. In addition to the difficulties in finding security policy documents for the SMEs, the few of them available contain a number of flaws. Having a look at papers that claim to provide guidelines for the security practices of the SMEs we can realise that these papers basically provide a detailed tutorial on implementing hardware and software security with other important aspects of security missing like the risk management and education process.

3.2 Budget limitation

One of the greatest hindering factors towards adopting security practices in the SMEs is the lack of sufficient budget allocated on security. It is indicated that smaller organisations are spending less on security than the larger ones while they spend a greater percentage of their IT budget on security as shown in Figure 3 below. This can become reasonable if we take into consideration the fact that the amount of security budget becomes smaller as the size of the company decreases. An average security budget for a small enterprise is $132,000 per year, $360,000 for a medium one, while for a large enterprise reaches the $1.3 million per year.

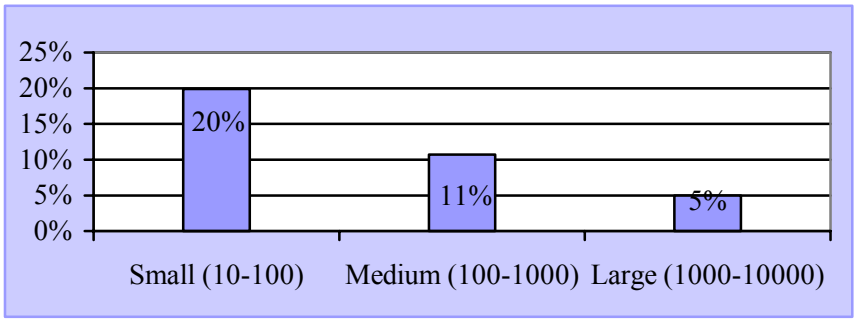


Figure 3: Percentage of IT spending on security according to business size
(source: Briney and Prince, 2002)

3.3 Lack of security specialists

Many businesses are too small to be able to justify in-house security specialists. According to the DTI survey 2004 three-quarters of UK businesses obtain external advice on security matters and only the 42 percent of them have somebody with IT qualifications. In some cases it is not only the fact that the enterprises cannot justify in-house security specialists, but moreover it is the security personnel that they feel that the environment in a small business is not what they are looking for. For whatever the reason, the fact is that small and medium enterprises are equipped with personnel that lack or do not have the adequate security qualifications.

3.4 Lack of incentive

It was found from observations and interviews that organisations may not adopt security procedures without the appropriate incentive. The senior management has to be convinced, in other words, that reducing risk by improving information security is worth the investment. And this is a difficult issue if we consider that the value of information security countermeasures is not visible since it is based on disastrous events. Making matters worse, there is no well-understood economic model for evaluating the benefits of using security procedures and policies as a mean of reducing threats. (Ernst and Young 2004, Timms *et al*. 2004)

3.5 Others

In addition to the above there are other less important reasons that may hinder the adoption of security policies in small and medium enterprises. Firstly, the pace of information technology changes acts as a hindering factor which makes matters worse for small and medium enterprises that already suffer from insufficient security budgets. The complexity of some provided solutions when considering the hardware and software can add difficulties when identifying the suitable measures for establishing the appropriate security policy. Finally, it is simply the case that the security policy procedures are being put aside for 'more important things' that acts as a preventive factor. (Ernst and Young 2004, Internet Industry Association 2003, Dimopoulos *et al*. 2004)

4. Proposed solutions

Now that we have identified some of the basic issues around the lack of security policies in SMEs it is important to introduce some sort of solutions so that the small and medium companies can adopt more widely security practices. In order to rectify this situation initially the limiting factors that are found in SMEs must be considered and then some sort of specific guidance in accordance with the use of baseline guideline documents must be provided. The paper we will simply try to present a simplified procedure for creating a security policy.

4.1 A simplified procedure for security policy

As a starting point the IT management can simply separate individuals into categories according to their needs and the level of access needed to the company's information system. For example, the following groups can be specified: users, administrators, guests, third parties that cooperate with the company, customers, etc. By doing this the obligations and tasks for each group can be dealt separately.

The next and most important step is the risk management process. According to this proposed solution the company's assets, that need to be protected and that may be targeted by a security threat, are initially identified and imported in a list. As asset we can define anything that needs protection including databases, information, personnel, facilities, applications, computer hardware and software, and communications systems. After that, it is time to identify the possible security threats that are associated with the assets and that may target them such as earthquakes, viruses, hackers, data destruction/ modification/ theft, theft of company property, fire, sabotage, fraud, or embezzlement. From information security surveys or other similar sources that can be found easily and free at the Internet (e.g. CSI/FBI – Computer Crime and Security Survey) the importance of the security threats can be also defined. A list that contains all the possible security threats sorted according to their importance must be produced at this point. Going forward it is now necessary to estimate the impact of the identified threats to the company's assets. Towards this some sort of statement phrases should be produced like in the list below. Together with each statement phrase the importance of the impact must also be added. The importance must be estimated according to the consequent losses for the enterprise

such as financial losses, business continuity disruption, incident response costs, data loss and disruption. (HKCERT Coordination Centre, Hamilton) The list should be sorted according to the importance field so that it looks like the following example:

Impact of security threats to business	Importance of impact
E-commerce server down due to DoS attack	9
Reveal of business data to competitors due to theft	7
Loss of business continuity because of attack on the company's network	5

Table 1: Example of list with statement phrases

Having done the above, we can proceed to produce any possible countermeasures. In order to define the possible countermeasures the security administration must first examine carefully the last two lists mentioned previously in order to realise which of the risks are worth considering and which not. An 'impact of security threat to business' that is highly rated, while the threat is also highly rated, introduces a risk of high importance that requires countermeasures to be introduced. As countermeasures we can define any of the security controls that can be employed to eliminate, reduce or mitigate the impact of a threat occurrence. The table above, for example, that shows that a DoS attack on the e-commerce server can cause significant losses and hence is of high importance indicates that intrusion detection software and firewall must be established. It is assumed that it has already been found that DoS attacks are frequent and hence regarded as important. (HKCERT Coordination Centre, Hamilton)

As a last step in the risk management process the senior management of the organisation must select which of the countermeasures will be implemented in the company's information system. The selection is based on cost issues and uses the ROI (Return Of Investment) analysis that is easily comprehensible by the organisation's senior management. According to this the benefits of the proposed security countermeasures specified by the security administration team are all introduced to the senior management in term of their investment and the return of investment costs. (Cisco Systems, Inc. 2004) In other words, for each of the countermeasures the cost for establishment (investment) and an estimation of the money saved on a long term basis (return of investment) are required. The senior management of the organisation can then decide according to the aforementioned costs which of the proposed security countermeasures are worth applying to the organisation and which not.

Once the countermeasures have been defined the next step is the security policy creation. The construction of the security policy document, which will specify what is acceptable and what not for everyone working inside the business environment, can be simplified by breaking down security into individual segments and dealing with each of them independently. Having the notion that security should be limited only to the absolutely necessary can also help in the simplification of the process. The responsible for the construction of the security policy should not worry about producing a well articulated security policy with their first attempt but must go

through the review and evaluation of the policy and possibly the construction of a new policy document.

5. Using ISO/IEC 17799

Towards increasing the security awareness and improving security practices in small and medium enterprises there is also the potential of using baseline guideline documents. The ISO/IEC 17799 document is one of them providing an extensive guidance on organisational aspects of information security management. The ISO 17799 code of practice can be a good starting point for the small and medium enterprises to implement information security. It can help the inexperienced personnel of the SMEs to identify the areas of the information security that need to be considered. Having identified the areas that need attention, the IT personnel will simply have to follow the recommendations on the corresponding sections. In addition, baseline guideline documents like the ISO code of practice or the NIST handbook are provided by organisations and institutions that may have a dominant position on national or international scale and their quality definitely reflect the expertise in security of the organisations they belong to.

In the small and medium enterprises, though, the ISO code of practice cannot be easily adopted and used widely by their personnel that may lack knowledge and experience on security issues. That is because the ISO as well as other baseline guideline documents just provide the framework for introducing security practices without going in depth to give details of specific countermeasures to follow. Hence we can say that the ISO is more appealing to security experts that can take the information provided by the document and implement security solutions according to the needs of the organisation.

As a way to help the small and medium enterprises in using more widely the ISO code of practice, the document must be initially minimised so that to include only the aspects that are necessary for the SMEs. Some of the sections provided in the ISO 17799 can be regarded as optional and can put aside for other more important. For example the sections “information classification” and “security in job definition and resourcing” are some of the sections that can be regarded as optional. Going further the ISO code of practice can be enriched with more details in each of its sections so that is more comprehensible by the personnel of the SMEs that may lack sufficient knowledge and experience on security. Detailed steps for the proposed security solutions together with explanations of information security terms (e.g. what we define as asset, security policy, threat, etc.) is the extra information needed so that the ISO code of practice can become more appealing for the personnel of the small firms.

The task of providing a simplified version of the ISO code of practice for the small and medium firms is now a necessity and must be taken seriously into consideration by national and international organisations on security. But until this task is accomplished, there are some other ways that SMEs can improve their practices by making use of the ISO. For example, SMEs can hire experts on the ISO 17799 and use them according to their needs. The experts on the ISO code of practice can

greatly enhance the security practices with their experience and deep knowledge on the subject. While the use of experts on ISO 17799 can help to improve the security practices within a SME, a wide range of software tools referred as ISO17799 toolkits can offer additional help towards using the document. Such tools are entirely based on the ISO17799 document and make easier the process of establishing security procedures within an organisation by using an easily approachable graphic user interface (GUI).

6. Conclusions

This paper presented the issues around security policies in small and medium enterprises. The results from various surveys conducted worldwide were found to be discouraging indicating that the majority of the small and medium companies do not follow or have difficulties in following the proper practices to secure their information system. Only a few of them reported that had a security policy in place while the money spent on security in general was minimal. For this disappointing situation there were though some reasonable factors. The lack of personnel with the proper skills and competencies is found to be the most important reason that hinders SMEs from adopting adequate security measures. It is reasonable to expect the qualified and experienced on security to seek for a career in a larger enterprise with more appealing opportunities and salaries than a small enterprise. In addition, the lack of sufficient budgets that can be justified by their small size makes less possible the chances of having security practices and particularly security policies.

Coming to a conclusion, some sort of solutions that are specific for the SME needs should be provided as these proposed in this report. The potential of using the ISO/ IEC 17799 code of practice must also be taken into consideration as a valuable and worldwide accepted solution for enterprises of any size. While the help from experts on the ISO 17799 or even the use of ISO toolkits can enhance this potentiality, a simplified version of the ISO, so that is more appealing for the SMEs, is definitely the best solution.

7. References

Briney, A. and Prince, F. (2002), “Information Security Magazine Survey – Does size matter?”, http://infosecuritymag.techtarget.com/2002/sep/2002survey.pdf (Accessed 12-Jul-2005)

Cisco Systems, Inc. (2004), “Cisco IP Communications Security Policy Development and Planning Guide”, http://www.cisco.com/warp/public/cc/pd/nemnsw/callmn/prodlit/ipsug_wp.pdf (Accessed 06-Jul-2005)

DTI (2004), “Information security breaches survey 2004 technical report”, http://www.dti.gov.uk/industries/ information_security/downloads.html (Accessed 01-Mar-2005)

Ernst and Young (2004), “Global Information Security Survey 2004”, http://www.ey.com/global/download.nsf/Austria/2004_global_info_sec_survey/$file/2004_Global_Information_Security_Survey_2004.pdf (Accessed 17-Mar-2005)

Gordon, L.A., Loeb, M.P., Lucyshyn, W. and Richardson, R. (2004), "CSI/FBI Computer Crime and Security Survey 2004", http://www.itsecurity.com/papers/insight2.htm (Accessed 12-Mar-2005)

Hamilton, C.R. (2002), "Risk Management & Security", http://www.riskwatch.com/Whitepapers/Risk_ Management_and_Security_11-07-02.pdf (Accessed 19-Mar-2005)

Hong Kong Emergency Response Team (HKCERT) Coordination Centre (2005), "Information security guide for small businesses", http://www.hkcert.org/secguide/eng/sme_guideline_e.pdf (Accessed 01-Jul-2005)

Kai, S.C., Chanson, S. and Wong, J. (2002), "E21-MagicMedia - E-security survey report", http:// www.e21magicmedia.com.hk/esecurity2002/pdf/e-Security%20Survey%20Report.pdf (Accessed 24-Jun-2004)

Melek, A. (2004), "Deloitte Touche Tohmatsu Org. - 2004 Global Security Survey", http://www.deloitte.com/ dtt/cda/doc/content/dtt_financialservices_SecuritySurvey 2004_051704.pdf (Accessed 15-Jun-2005)

Zuccato, K. (2005), "2005 Australian Computer Crime and Security Survey", http://www.auscert.org.au/ images/ACCSS2005.pdf (Accessed 16-Jun-2005)

Authentication based upon secret knowledge and its resilience to impostors

L.Zekri and S.M.Furnell

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

This paper presents an assessment made on an alternative to the present password and PIN-based methods of user authentication. In the recent years, many alternative authentication methods emerged, but none of them seems to have been a major breakthrough. Nevertheless, two techniques emerged as potentially efficient: image-based authentication and cognitive question and answer techniques. Even if the viability of these techniques has been proved, little research has assessed the resilience of the methods to impostors. Therefore, an environment has been created to test the robustness of the alternative techniques. The evaluation comprises both a theoretical and a pragmatic analysis to rate the robustness of the methods. The results show that the methods are vulnerable in different ways, with PassImages susceptible to phishing and shoulder surfing, whereas cognitive questions can be targeted via social engineering.

Keywords

Security, Authentication, Graphical Passwords, Cognitive Questions, Robustness.

1. Introduction

Passwords are the most commonly used method for identifying users in computer and communication systems. They were introduced first in the early 1960s as an authentication solution with the emergence of the first multi-users operating systems. Since that date, users relied more and more on passwords as computer and networks spread, and especially the Internet. According to a study (Danchev, 2005), almost 99% of the home users rely heavily on passwords as a basic form of authentication to sensitive and personal resources.

On average, we need to remember about 10 passwords, and this average increase up to 16 passwords for IT workers (Brostoff, 2004). Moreover, we can add to our list of passwords a list of PIN (Personal Identification Number) codes. This huge amount of passwords and PINs that we have to remember threatens our security. For security reasons, the best password would be a random one. However, it is widely recognised that, in order to remember all these passwords, we tend to use dictionary words or other words that have special meaning for us. We also tend to use the same password everywhere, allowing a hacker that discovered a password in one account to gain access to others.

Because humans live in an environment where their sense of sight is extremely active, we have acquired an amazing ability to treat and store large amounts of graphical information easily. These recent years, many studies tried to exploit this ability in the context of user authentication. Three of them, performed by Irakleous et al. (2002), Papadopoulos (2001), and Charruau (2004) were the building blocks of this research. The subject of these studies was to assess the viability and the efficiency of image-based authentication, and also some other secret-knowledge techniques. None of them assessed the robustness of the methods to impostors. In order to be adopted in a large scale, the potential alternative techniques should be first assessed from a security point of view. It is from this perspective that this study has been developed.

The paper begins by presenting the weaknesses of the existing password-based approaches, as well as an outline of the previous attempts to utilise image-based methods as an alternative. It then discusses the adopted methodology, the implementation of the tests conducted on the alternative techniques, and the results observed from the experimentation. The implications of these results are then discussed, leading to the suggestion of future research directions in the concluding section of the paper.

2. Background

Passwords were a viable way to implement an authentication process that could work with the rudimentary and simple command line interface. By using passwords, it was simple for system designers to provide an efficient authentication, and it was easy for users to adopt it.

However, it is the users themselves that often compromised the password-based protection. Simple passwords are easy to remember, but vulnerable to attacks, whereas complex passwords are more secure, but hard to remember. Many studies have been carried out during the past decades investigating the roots and the impact of the weaknesses of password-based methods on authentication. A study, conducted in the early 70's (with a population of 3829 users) presented quite interesting results: about 15% of the passwords had length no more than three characters and 85% of the passwords were dictionary words (Morris and Thompson, 1970). Another study (1990) conducted on a population of 15,000 persons showed that 21% of the passwords have been cracked in less than a week (Klein, 1990).

The problem with passwords is due to the fact that they rely on a precise recall of the secret information, which is not a strong point of human cognition. Thus, other secret-knowledge based authentication techniques have been developed recently, and they rely on recognition rather than a precise recall of memory. Image-based authentication consists on the recognition of previously seen images, a skill at which humans are amazingly talented. Many studies have shown that humans can remember and recognise thousands of pictures very rapidly. In an experiment, a sequence of about 2,600 pictures was presented to an audience which realised, in a second step, an interesting 90% recognition rate (Standing et al, 1970).

In addition, it seems that image-based authentication is viable and efficient. In fact, two earlier studies established the high authentication rate of these alternatives. The first one involved 27 persons and presented a 63% authentication rate (Irakleous et al. 2002). In the second one, from a total 911 trials, the users were able to authenticate 867 times. This gives an authentication rate of 95% (Charuau et al. 2004).

In addition to the graphical authentication techniques, researchers have developed other alternative methods. For instance, cognitive passwords are based on specific questions where the answers depend on user's opinions, interests and life history.

The purpose of this study is to assess the robustness and the resilience of some alternative techniques to impostors. The two chosen methods to be assessed are PassImages and Cognitive questions. In fact, three previous studies assessing the effectiveness of alternative methods have previously been performed by Irakleous et al. (2002), Papadopoulos (2002) and Charuau (2003). The first study established that a technique based upon associative questions (i.e. using word association as the basis for challenge-response pairs) suffer from a low authentication rate (4%) and thus it will not be considered for further research. In the opposite side, it established that successful authentication methods could potentially be cognitive questions techniques (59% of successful authentication) and PassImages (63%). Thus, and according to these results, this study will mainly focus on the robustness of these two techniques.

3. Methodology

The goal of this project is to assess the robustness of both PassImages and Cognitive Questions (see Figure 1 & 2). Thus, what should be considered is the potential of these methods to be compromised by impostors. To do so, tests have been conducted to investigate their resilience to would-be masqueraders, to determine whether the replacement methods might be more easily compromised then traditional passwords. To be as realistic as possible, the first challenge was to perform in a realistic way the impostors' behaviour. Thus, two persons were asked to behave like impostors by contacting the legitimate users in different ways and trying to obtain information that could help impersonation.

To assess the robustness of PassImages and, in a second step, cognitive questions, two approaches have been adopted: a theoretical analysis and practical tests. The aim of this dual analysis is to provide both mathematical analysis as well as pragmatic conclusions, and then try to compare them.

The aim of the theoretical analysis is to measure the ability of an impostor to guess the secret knowledge. Davis et al. (2004) conducted a study on graphical password schemes in which they used a measure that indicates this ability: termed the "guessing entropy". In order to compute it, the computer randomly selects a set of "test PassImages". Then the computer generates a random PassImages and checks whether it belongs to the test set. Then it selects another random one and so on. Once

all the PassImages of the test set are guessed, the computer returns the total number of attempts he performed. It then performs the same operation with another non-overlapping set of PassImages, until all the passwords distribution is covered. Finally, it calculates the average number of attempts, which is the average entropy.

Figure 1: PassImages authentication

Please answer carefully the following questions:

1. What is your mother's maiden name?	Kefi
2. Where were you born?	Tunis
3. What is your favourite colour?	Blue
4. What was the name of your best friend at school?	Mourad
5. What is your favourite music?	Jazz
6. What is your favourite food?	Coucous
7. What was the name of your first pet?	
8. Which primary school did you go to?	La Goulette
9. What is your favourite sport?	Volleyball
10. Where was your first house?	La Goulette
11. What make was your family's first car?	Peugeot
12. How old were you when you had your first kiss?	17
13. What is your favourite film?	
14. Where was the first place you remember going on holiday?	Rafraf
15. What was your favourite subject at school?	Maths
16. What is the most important part of your body?	
17. What is your favourite type of animal?	

Figure 2: Cognitive questions authentication

In addition to the theoretical study, it was decided that a practical test on authentication methods robustness would be performed in order to measure, in a pragmatic manner, the ability of an impostor to guess the secret knowledge. The first test consisted of physical and virtual meetings between would-be impostors and their targets. The challenge in physical meetings was to remember the conversation and especially details that can be useful for guessing the secret knowledge. Virtual discussions were performed with the help of Internet messengers, and are easier to remember since all the discussions are stored in a historic of conversation.

The second test was phishing. When setting up their passwords, an email containing all the secret knowledge was sent to the users and they were supposed to save it. The phishing strategy was to send them another email asking them to send back the subscription email. The trap was that the destination address was a forged one.

The third test was shoulder surfing. As its name implies, this refers to watching over people's shoulders as they perform authentication. The assessed users were asked to authenticate themselves with a person sitting behind them. The test was performed once with each of the users who agreed to the request.

Results of these tests would be compared with the theoretical ones in order to determine whether the theoretical robustness of some PassImages corresponds to real difficulties when trying to guess robust PassImages.

4. Results

The results of the statistical analysis were obtained with the help of a simulation made with a computer (Table 1). In order to get any significance from the numbers presented in this table they should be compared to the number of possible passwords. In fact, in order to select a PassImage a user should select the first image of his PassImage from one hundred ones, then the second image of his PassImage from the following ones and so on. According to a calculation made in this manner with the help of a computer, there are 1,192,052,400 possible PassImages for a maximum guessing entropy of 596 $*10^6$. The results show that the worst 10% of passwords can be guessed only after 980,297 attempts, and the worst 25% ones after more than four millions attempts. This result states that with an adequate brute force attack tool, the attacker would spend hours and even days to abduct the authentication process.

G (Average)	286990167,2
G median	143490375
G 10%	980297
G 25%	4901487

Table 1: Guessing Entropy for PassImages

It is also interesting to note that the average Entropy G^{Avg} is higher than the median G^{Med}. This means that there are many good passwords in the dataset that, by increasing the average number of guesses, they make it harder for an attacker to guess the PassImages.

Concerning the practical assessment, forty-three users received two emails asking them to send back or to communicate personal data. The results show that twenty-five persons succumbed to phishing (58%), from which twenty persons (46%) sent back the subscription email they received few weeks ago and sixteen (37%) persons dropped in the trap of the second email (asking them to re-authenticate to avoid account expiration). Finally, eleven persons succumbed to both traps (25%).

For a social engineering attempts, in order to be considered as successful, the impostor should be able to correctly answer at least 75% of the target's cognitive questions. In fact, cognitive questions techniques are considered too time consuming (Papadopoulos, 2002) and usually, the administrators implements a system that selects randomly from three to five cognitive questions, in spite of listing the twenty questions. By acquiring the answers for at least 15 questions, the impostor could typically expect to come across a set of questions he can answer. Forty-three users participated in this test, and were divided into categories as follow: 18% were close relatives, 28% were friends, 30% were colleagues or classmates and 23% had a non-significant relationship with the impostors. The most interesting thing here is the fact that the stronger the relationship with the user is, the more successful is social engineering. The success rate grows from 33% for strangers to 38% for colleagues and flatmates, then 66% for friends, and finally 87% for relatives. Globally, social

engineering is an efficient method to compromise the robustness of cognitive questions techniques, with a success rate of 53%.

The "shoulder surfing" test consisted of evaluating the ability of an impostor, sitting behind the legitimate user, to capture the password when the users are authenticating themselves, either with PassImages or with cognitive questions. Despite the small number of users assessed (20), some interesting conclusions emerge from the results. Eleven (55%) PassImages were held by the impostors, when only three successful attempts (15%) to retain answers of cognitive questions have been realised. The inefficiency of shoulder surfing with cognitive questions is obvious.

5. Discussion

According to the findings, PassImages and cognitive questions are unequally sensitive to the different attacks performed to assess their robustness. When social engineering seems to be efficient to bypass cognitive questions method (with a success rate of 53 %), its effectiveness to compromise the robustness of PassImages is less evident. In fact, and according to the comments of the impostors, trying to link the user's choice of images to their hobbies, activities or their past is simply inefficient. The impostors observed: "for instance, a user declaring he plays volleyball avoided the volley ball and selected the tennis one. He was not a smoker but selected a lighter and a cup of coffee. His favourite meal was a French meal made with peppers and minced meat, but he selected a picture of a lemon. Examples of such a contradictions between the users' profiles and their selections are various." In a rare case, an impostor successfully guessed the PassImages of his sister ("I just thought that for PassImages she would has chosen an easy context. I thought about selecting the images referring to his breakfast, and I succeed!").

In the other side, PassImages seems to be far more vulnerable to shoulder surfing (55% of success rate) than cognitive questions (15%). This can be explained by the fact that, in order to remember answers to cognitive questions, the impostors need to retain a large amount of information in a very short time. In addition, as the only five questions were displayed randomly each time, the information caught by the impostor is simply incomplete. It also shows that it is more easily to remember graphical information than the written one, which works in favour of shoulder surfing and against the technique. However, what must be remembered with this study is the fact that shoulder surfing was performed once with each user. In the real life, a diligent impostor may perform shoulder surfing more frequently in order to get all the information he needs.

Finally, we can conclude that the two techniques are complementary. In fact, in actual means of authentication, and more especially on Internet, many service providers portals and banks (e.g. yahoo, Gmail, HSBC) have adopted cognitive questions (for password recovery) coupled with traditional passwords in order to authenticate their customers. What could be suggested is to use a more efficient PassImages plus cognitive questions strategy.

Another important objective was to accept or disprove the hypothesis stating that the theoretical analysis of the dataset could be enough and avoid the need to perform practical assessments on authentication methods.

The success rate for guessing PassImages with practical tests was 58% (phishing), 55% (shoulder surfing) and 2.5% (social engineering). When balanced with the total number of participants for each test, the overall success rate of the practical tests is 30.25%. According to the theoretical study, the guessing entropy $G^{25\%}$ corresponding to the effort that an impostor should provide in order to guess the 25% weakest PassImages is 4,901,487 attempts. It is clear that the two persons acting as impostors have made far less efforts in order to achieve their results. In fact, what could be reproached to the theoretical study is the fact that, by its nature, it is a quantitative test and not a qualitative one: it is based on the analysis of the information the dataset contains, and not the way of extracting it. In addition, the statistical distribution of the images has shown that the most recurrent image represents only 2% of the overall selections which implies that no weakness have been revealed by the users' choice of a PassImages. In practice, the user's behaviour was insecure, careless and naïve in a general manner, and the theoretical analysis could not uncover it.

Finally, PassImages can be considered more robust than passwords in many aspects. First, passwords are more vulnerable to social engineering than PassImages and cognitive questions. Secondly, passwords are not more robust to phishing than the two alternative techniques: phishing relies more on the carelessness of the users than the authentication technique itself and we can state that all the listed authentication methods are equally vulnerable to phishing. Thirdly, unlike passwords, PassImages are more robust against sniffing: there are no alphanumeric data entered, and even if the hacker guesses the mouse movement, he could not repeat identically the authentication process since the images are refreshed randomly every time. By contrast, cognitive questions rely on alphanumeric data and thus, they are as vulnerable to sniffing as passwords unless encrypted. Nevertheless, PassImages are more vulnerable to shoulder surfing than passwords. In fact, by its nature, graphical information is more easily memorable than alphanumeric one, and this was established when comparing the vulnerability of PassImages to shoulder surfing confronted to the cognitive question one.

6. Conclusion

The experimentation produced many interesting results on the robustness of PassImage and cognitive question techniques. On one side, the theoretical assessment states that the number of attempts that an impostor should perform in order to get access to the system is astronomical and then the system can be considered as secure enough against brute force attacks. On the other side, the findings that emerged from the practical tests have shown that the theoretical study by its own is not sufficient to assess the robustness of authentication methods. In fact, the average success rate to bypass the process was a significant 30.25%. This

reflects the fact that, by its nature, the theoretical study is based on the analysis of the information that the dataset contains, and not the way of extracting it.

Despite the fact that the methods remove many weaknesses compared to passwords, they still suffer from some inherited ones. PassImages techniques are easily compromised with shoulder surfing when cognitive questions seriously suffers from the effects of social engineering. Moreover, both methods are prey to phishing. Nevertheless, the complementary nature of these techniques has been identified, and further studies should consider this aspect.

Another aspect of assessment could also be performed. In our study, only two impostors were volunteers to assess the robustness of the methods. Providing an environment on which more participants will be willing to act as impostors is an issue to consider for future researches.

Finally, this study suffers from a missing direct comparison of the assessed authentication methods compared to passwords. Future works should consider this issue in their conception.

7. References

Brostoff, S. (2004), *Improving Password System Effectiveness*, PhD Thesis, Department of Computer Science, University College London.

Charruau, D. (2004), *Assessing the Viability of alternative authentication methods*, MSc Thesis, University of Plymouth, UK.

Charruau, D., Furnell, S.M. and Dowland, P.S. (2004), "PassImages: an alternative method of user authentication", in *Advances in Network and Communications Engineering 2*, ISBN: 1-84102-140-7.

Danchev, D. (2005), "Passwords - Common Attacks and Possible Solutions", www.windowsecurity.com (accessed 03/09/2005).

Davis, D., Monrose, F. and Reiter, M.K. (2004), "On user choice in graphical password schemes", *Proceedings of the 13th USENIX Security Symposium*, San Diego, August 2004.

Irakleous, I., Furnell, S.M. and Dowland, P.S (2002), "An experimental comparison of secret-based user authentication technologies", *Information Management & Computer Security*, vol. 10, no. 3, p103.

Klein, D. (1990), "Foiling the Cracker: A Survey of, and Improvements to, Password Security", *Proceedings of the USENIX Second Security Workshop*, Portland, Oregon, 1990, p.3.

Morris, R. and Thompson, K. (1979), "Password Security: A Case History", *Communications of the ACM*, 22(11), Nov 1979.

Papadopoulos, I. (2002), *User Acceptance of Alternative Authentication Technologies*, MSc Thesis, University of Plymouth, UK.

Real User Corp., (2005), “Technology and products”, www.realuser.com, access [10/12/2005].

Standing, L., Conezio, J. and Haber, R. (1970), “Perception and memory for pictures: Single-trial learning of 2500 visual stimuli”, Psychonomic Science, Vol 2, p73-74.

Network Security Audit

D.Liu and B.V.Ghita

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

With the increase of broadband connections users, the number of home computers has increased importantly. As a consequence security issues have gained in importance in this domain. Most of these new computer users are novice and do not have the knowledge to understand exactly the repercussion of their actions in term of security on their machines. Software companies have developed several products to protect these stand alone computers. Some of them are designed to produce security audits which evaluate the security risk of the Personal Computer (PC).

Unfortunately, even with these audit programs, users do not become aware of the danger they can face on Internet. This project has developed a security audit tool which is intended for novice computer users. This tool's objective is to evaluate the materiel security level and the behaviour security risk of the user. Moreover, to be sure of the users' understanding, this tool also contains some explanation and demonstration elements, which show them how a malicious person can exploit their lack of prudence.

Keywords

Security, Audit, Ports, Novice users, Key logger, Password, Antivirus.

1. Introduction

Security has become a critical issue for modern companies, they are actually spending important amount of money to prevent any malicious persons to access their data. With this increasing level of protection, hackers are turning away from banking and high technology companies, and are more targeted into small organisations or ordinary single PC/home networks which are less secure.

Aware of this, software producers have developed very efficient products adapted to ordinary Internet home users like personal firewalls and personal antivirus. Unfortunately on the other hand their use generates a wrong feeling of security. Most of time, these tools are not configured properly by users, who do not have the right knowledge. In this case the computer is still unsecured.

In this project, the author is going to try to make them understand the risks they can face on Internet network by using simple demonstrations and explanations,

First of all, a software application should be built to execute technical tests of the computer and produce a physical security audit. To be able to provide external and

internal tests, the application will have a client server configuration. The server side of the application will be put online, and will execute external tests and attacks. The client part of the application will be run on the local host and will offer local tests possibilities.
This program must be as easy as possible to use in order to make it accessible for every ordinary computer user. Unfortunately a technical audit is not enough to evaluate the security level of a computer, human behaviour is also a key element. To assess the user, a MCQ will be integrated to the software application. In this questionnaire some questions will focus on the existing security element of the PC, and some others will deal with the user's behaviour in front of different events: for instance a website which asks the user to download and execute a JavaScript program.

Combined with the technical audit results, the user will be able to have a very accurate and customised security audit based on both the machine security configuration and the user behaviour. Moreover, to provide a better understanding of the MCQ questions, they will be illustrated by screenshots and schemas.

Once the audit results have been given, the application will be able to help the user to understand them. It is important to keep in mind that ordinary PC users have really few knowledge to exploit theses results. For the third part of the project, some informative pages have be included and linked to the audit results to provide explanation and give advices to the users to avoid problems which have been detected.

2. Existing audit software

Before starting the conception of this project's audit software, it is important to have an overview of existing audit tools used by network administrators. Here can be found the analysis of four of them.

2.1 Nessus

It is maybe the most famous one. It is a free vulnerability scanner based on client-server architecture. Normally it runs on UNIX like systems, but recently a windows version has been adapted: Tenable NeWT Security Scanner (Nessus, 2005). It is this version tested. The functionality of Nessus is very similar to the audit tool this project should produce. The client part of Nessus allows the interaction between the user and the machine, by sending to the server the user's instructions. The server receives the user's information, then runs the appropriate command (attack test) and finally sends the result to the user. This program has three main positive points: first of all, it is free; secondly, it is coded as a plug-in, making him easy to update; and at last, it contains a wide range of options to parameter the vulnerability audit.

2.2 ATK: attack tool kit

It is also a free audit software, it is based on a mix between a vulnerability scanner and a exploiting frameworks (ATK, 2005). This application has two main benefits:

firstly it is very easy to exploit and uses schemas representations to explain to users its functionality; secondly, it provides advices to avoid the security hole if vulnerability has been discovered.

2.3 GFI LAN scanner

It is a complete security software. Once launched, it will perform an external vulnerability scan and also an internal security audit (GFI, 2005). This internal security scan will check in the current computer or on every host of the LAN (Local Area Network) the installed softwares, their patches, the passwords used, the USB connections, the register entries, the shared folders, the wireless access points, etc…

2.4 MBSA (Microsoft Baseline Security Analyser)

It is a security tool created by Microsoft for Windows based computers. "It scans for common misconfigurations in the operating system, IIS, SQL, and desktop applications, and can check for missing security updates for Windows, Internet Explorer, Windows Media Player[…]" (Microsoft, 2005). MBSA is directly connected to the Microsoft Vulnerability database, which gives him the advantage to be constantly up to date. Moreover, it provides to users clear explanations and solutions to discovered problems.

3. Important Issues

All security risks and protection elements are well known from professional security administrators, but on the other side, ordinary people can have some problems to understand them. Even if they use a familial version of security audit software, they may not be able to interpret its results.

Software designers try to make their product as simple as possible to affect as many users as possible; but unfortunately, because of the complexity of the network security area, it is very difficult to create something really comprehensible for any ordinary user. For instance, with Nessus even an intermediate knowledge level user will need about ten minutes to realise all the possibilities of this program and how to exploit it efficiently.

Moreover today's security audit programs can be very complete from a technical point of view. Unfortunately they do not take into account the human behaviour factor which is the weakest link in the network security chain.

The last issue of this project which makes it different from other commercial audit softwares concerns the usage because all existing softwares need to be installed on the computer they are auditing. This obligation can be a serious problem for very novice computer user who do not know how to install, or for users who do not have an administrator account on the current computer or who do not have the right to install any programs on it.

This project tries to offer appropriate solutions to these issues. Firstly, it will not require any installation obligation. Thus it will provide to any users no matter the computing knowledge level the possibility to run easily a security audit on the current computer. Secondly on top of all technical audit elements, this project will contain some tools which will evaluate the security level of users' human behaviour. Finally, the users' understanding is a key element in this project, they will be provided clear explanations about how to use the different audit elements, what are they auditing, how to avoid the problems….

4. The audit programs of the project

After the study of exiting audit programs, some audit tools have been chosen to be integrated in this project. They can be grouped into two categories, the ones which analyse the material security risk and the ones dealing with the human behaviour risk.

4.1 Material audit elements

4.1.1 The port scanner

It is based on a client/server architecture. The client is launched from the audit web page on the user's computer and then sends a request to an external server which will scan the current PC's ports. The user will be given the choice between a "quick" scan and a "long" scan It has been decided to add this option because a complete port (long scan) scan takes about 20 minutes, and it is not possible to impose a time consuming test. The "quick" scan takes about 30 seconds, and it checks 31 "well-known" ports. Once the user has clicked on the start button of the user interface, the client program will be run; it will send a scanning request to the server. Then, the server will analyse the request, find out the IP address of the user from the connection settings and finally run the appropriate IP scanning program. There are actually two scanning programs, one for each type of scan (quick and long), but their functioning principle is the same. They try to create a socket connection with the given port in a specific timeout. If this try fails that would mean the port is closed, inversely if it success it would mean that the port is opened. For the "well-known" ports scanning, the timeout is set to one second, and for the complete one, it is set to 200 milliseconds. The longer is the timeout, the more it can be certain that the port is closed. To scan 65535 ports, it is not possible to set important timeout; otherwise the scanning process will take hours.

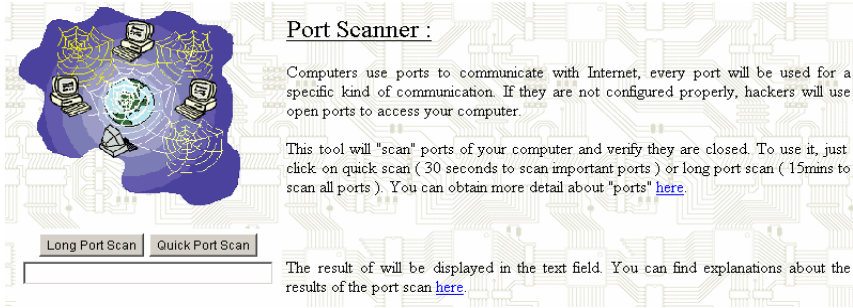


Figure 1: the port scan interface, with a scan result displayed.

The scanning result is recorded in a table and transferred to the server who will convert it into a single data flux and send it back to the client program. Once the client has received the result it will display it in the user interface of the audit page.

4.1.2 The Antivirus tester

It will analyse the reactivity of the antivirus installed on the machine. Recently some online antivirus have been developed, they can perform a complete hard disk scan of people who connect to their WebPages. But they do not perform real tests which can audit the reactivity of the installed antivirus. To test the reactivity of user's antivirus; the project uses elements of EUCAR virus test file. "This test file has been provided to EUCAR for distribution as the EUCAR Standard Anti-Virus Test File", […]. It is safe to pass around, because it is not a virus, and does not include any fragments of viral code. Most products react to it as if it were a virus (though they typically report it with an obvious name, such as "EICAR-AV-Test")" (EUCAR, 2004). The tester creates an "eicar.exe" file on the disk on the user's machine then copies the following 68 bytes ASCII characters in it: X5O!P%@AP[4\PZX54(P^)7CC)7}$EICAR-STANDARD-ANTIVIRUS-TEST-FILE!$H+H*.

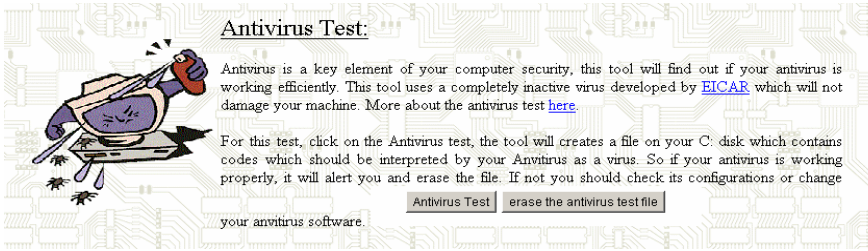


Figure 2: the user interface of the antivirus test tool.

Normally, the antivirus should alert the user that an infected file has been detected. If nothing happens, that would mean that the antivirus is not properly configured or not efficient. Finally the tool will give to the user the possibility to delete the file.

4.2 Human behaviour assessment tools

4.2.1 The password analyser

Depending on the password entered by the user, the analyser will estimate the amount of time needed to crack it. There are several free password crackers which can be used to show to users how easy it is to crack a weak password. But their main defect is the delay required to crack it. If the password is not a dictionary word, it may take from a few minutes to hours to break it, and users can not wait so long. So instead of integrating one of them, the author has decided to develop a time estimator which gives a quicker response.

The user is asked to enter a password, and then the tool import a list of English dictionary words from a specific webpage (http://dcool75.free.fr/mot.txt). Once the list is imported, it will compare the user's password with its content. If there is a match, it will inform the user that only a few seconds are required to crack this password.

If there is no match, the audit tool will process to the structure analysis of the password; depending on characters it is built with; the program will calculate an estimation of time needed to crack it with any ordinary password crackers.

This estimation is based on the password's length, and the type of characters it contains. Basically, if the user's password contains capital letters, numbers, and special characters, it would take approximately 4320 minutes to find out one character. If it contains only lower case characters, the cracker will need about 4 minutes per character. If it has some number on it, 30 minutes are required for each character. And finally if it contains numbers and upper case characters, it will take about 150mins to find out a character. These numbers come from statistics the author has done with L0phtcrack 5.04 a very popular password cracker.

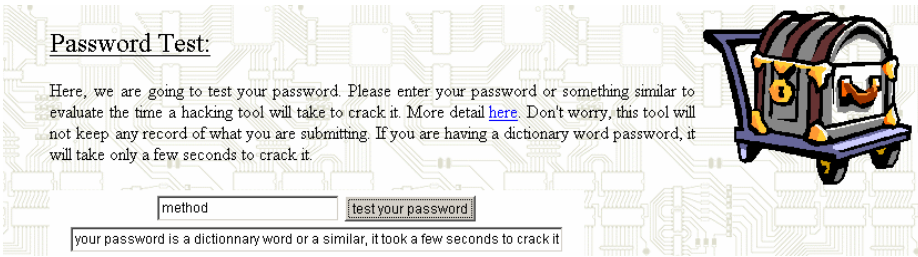


Figure 3: the result of the password tester when the password is a dictionary word.

4.2.2 The MCQ

Here the users have to answer some questions about their behaviour when they meet some specific situations. Most of the existing security audit softwares do not take into account the effects of human aspect on the security. Here depending on the

users' answers, a specific amount of points will be added to the risk counter. Once all questions have answered, the final risk counter will determine in which category the user belongs to: low security risk, medium risk, high and very high risk.

4.2.3 The key logger

It is more a demonstration tool than an auditing tool. It is important to show to users how a real hacking tool works. So they will realise by themselves how dramatic it could be if a hacking tool of this kind was installed on their computer. In order to avoid any installation requirement to use it, it has been preferred a simple design with an easy usage instead of complex codes similar to commercial key loggers which would require manipulation from the users. So the program included into the audit page has nothing to do with a real key logger except the concept. It does not run in the background of the computer to capture every keystroke but clearly asks the user to write a few words in a designed square of the audit page. When the user has written 10 characters the JAVA program will create a socket connection with the key logger server, and everything that is in this square will be sent then, the server will process the data. Here to give a quick demonstration to the user, the server will send the data back to the audit page which will display it in a pop up page.

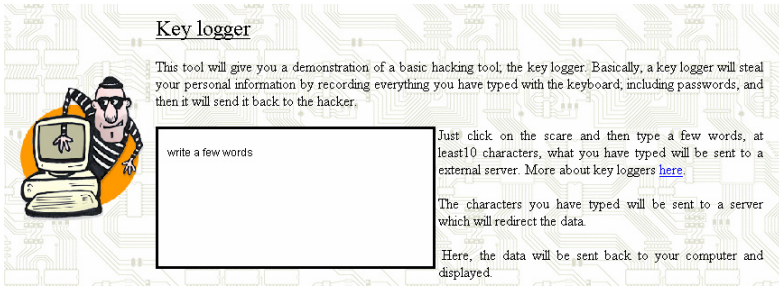


Figure 4: the interface of the key logger where the user is invited to write a few words in the indicated area.

5. Achievement

The final security audit program of this project has been integrated in an online web page (http://dcool75.free.fr), when existing audit softwares require an installation on the computer it is testing. So the accessibility is far simpler, the users have just to go to this page in order to test the security of their computers.

Another good point of this project is that it has been designed to be used by very novice users. When existing audit softwares perform only their tests and provide a final result, this project's program gives clear explanations about the audits elements, why they have been chosen, how they are functioning, how to exploit the results and what are the security risks.

The last important difference of this security audit project with other programs is that it is commercially neutral. An important part of existing audit programs are designed by security software companies, and sometimes they try to make the audit result

worse to encourage people to buy their products. It is frequent to see in an audit result: “Your computer has a very high security risk, if you want to fix the problem buy the following products of our company”.

Here, the audit project has no commercial purpose and always advice the users to download freeware security programs.

6. The survey

A survey has been added in the main audit page, in order to collect users’ opinions about this project and its programs. It does not contain a huge number of questions, but it will give the possibility to know if this project has reached its main objectives. In addition, it will provide a very interesting comparison from the user point of view of this audit program with commercial security audit softwares.

A part of the novices users questioned had not fully exploited the audit programs which compose this project. After investigations, it turns out that some of them had a firewall which blocks the client/server communication of the port scanner and key logger; and others did not have properly modify the JAVA security policy on their computer. That means that the project has only reached partially one of its main aims which is to create an audit project accessible to everybody.

The general opinion of users is that they really appreciated some elements they consider to be real innovations. The antivirus tester and the password analyser for instance were real success, similar tool do exist but they all require installations in order to be run.

A new problem has also been raised by this survey. Some users explained that the manipulations required to change manually the JAVA security policy may discourage novice users. To modify a security file on the hard disk of the computer because a web page says so is not something every user will do. Most of users will certainly be afraid of this step because their do not know the origin of the web page, and it they are novice users they will not know exactly what they are doing. That is why it is a key element in future improvements to bypass this step by using another coding language which can be less strict in terms of security or by developing some script which will make the modifications of the policy file automatically each time the audit page is loaded by the web browser and will erase the modifications once the user has closed the page.

Additionally the analysis of expert and intermediary users’ answers has proved that they were able to use the audit programs without any technical problem. Most of them have found this project simple to understand and easy to use comparing to existing security audit softwares. Thus, it is reasonable to say that this project is more interesting than existing audit softwares in terms of usage simplicity, but some improvements are still required to improve the understanding of novice users.

7. Conclusion

By studying some existing security audit programs, this research has identified some key issues which could be improved for very novice computer users in terms of understanding and manipulations.

This project has developed an online webpage which contains a set of security audit tools and clear explanations texts. This project brings to novice users a simpler overview of their computers' security via technical audits and human behaviour assessment. Moreover this research gives to the end users the possibility to have a better understanding in the domain of computer security thanks to its explanations texts which describe in detail each security element broached in audit the webpage.

In order to make this project technically more complete and competitive with existing commercial audit programs in the future, some new audit programs can be added. For instance, inspiring from Microsoft Baseline Security Analyser, an upgrade analyser can be added to the project. It would analyse the upgrade level of the Microsoft products: Microsoft Windows, Microsoft office, Microsoft Internet explorer; with the latest version published by the Microsoft web site.

8. References

ATK: Attack Tool Kit, (2005), http://www.computec.ch/projekte/atk/, accessed in September 2005

Eucar Online, (2003), *The Anti-Virus Test File,* www.eicar.org/anti_virus_test_file.htm#dl, (accessed in August 2005)

GFI, (2005), http://www.gfi.com/lannetscan/ accessed in September 2005

Microsoft, (2005), http://www.microsoft.com/technet/security/tools/mbsahome.mspx, (accessed in March 2005)

Nessus, (2005), www.nessus.com/about/, accessed in September 2005.

The use of the Internet by the elderly in France

M.P.C.Blin and A.Phippen

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Information and Communication Technology (ICT) is more and more present in the daily life because a lot of services and products are being offered electronically. Some people, mainly the disabled and the elderly, are often excluded from these new technologies and a digital divide is established between the generations. This digital divide can be greater if people live in rural areas. Of course, the government takes some initiatives to try to decrease this gap. They develop Internet Public Access Points (IPAPs) which welcome everybody who wants to attend initiations to learn to use a computer. The work presented here aims to define if the elderly using the Internet are numerous and to identify what are the main factors that enable some elderly people to use the Internet. This paper includes two main parts. The first is about the use of computers and the Internet by people who are over 50 years old and living a rural area. The second is on the actions taken by the governments to decrease the digital divide.

Keywords

The Internet, computer, elderly people, digital divide, Internet Public Access Point (IPAP)

1. Introduction

Technological innovations take an important scope near the general public, so it would be interesting to know if the widespread of ICT has not excluded the elderly. A first research showed that three main factors have an impact on the use of computers: the age, the fact to live in a rural or an urban area and the social category. Ageing is considered as the main handicap in the acquisition of ICT. Thus, a study was leading on the use of a computer and the Internet by the elderly in a rural area to know if the digital divide is real. The government has developed a plan to try to decrease it. Several IPAPs have been setting up to help people to access ICT.

2. Position of France in computer's equipment

A study carried out in 2004 by Gfk France (European marketing study institute) reveals that 43% of French households were equipped with a computer at the end of 2003. In comparison with others European countries France is a backward country. Indeed, the Netherlands are the most equipped in computers with 67% and England is better equipped than France with 45% (Beky, 2004).

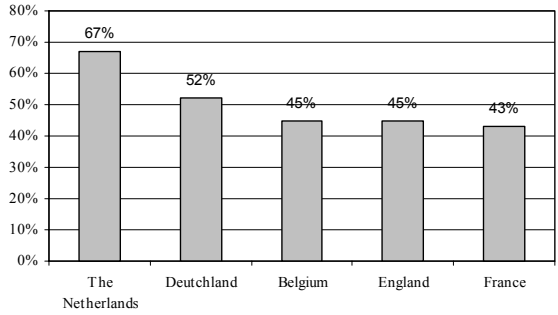


Figure 1: Position of France about computers' equipment at the end of 2003 (Beky, 2004)

The access to ICT is not the same for each group of the population. Indeed, there are some groups which are more or less excluded. Three factors, which have an influence on the use of computers and the Internet, were distinguished (Carboni, 2003; Bouchayer, Gorgeon et al, 2002):

- The age: most of Net surfers are people who are less than 25 years old (59%). The more population is aged, the less they used new technologies, only 19% of people who are 65 years old and over use the Internet.
- The location of people in a rural or an urban area. More than half of people living in Paris think that the Internet is important in their life against 35% of people who live in provincial France.
- The social category, which is a few linked to the fact that people live in a rural or urban area because the job with responsibility are more often located in cities. 57% of the executives use the Internet against 30% of the workers.

3. Research method

A first literature review was carried out to design a survey on the use of a computer and the Internet by people aged over 50 and living a rural area. The survey contained questions for the users and the non-user. There were questions on the use of a computer and the Internet, the duration and the frequency of use, the main activities on the Internet, the encountered difficulties, and the factors that prevents some people to use the Internet… People have been asked in the Mayenne department, located in the north west of France. Mayenne is rather a rural area with less than 300,000 inhabitants. People were asked in the streets during several markets which took place in the town centre of Mayenne. About 170 answers have been collected. In this study as many people aged of 50-59, 60-69, 70-79 and over 80 years old were asked. The level of diploma, the incomes and the social category, which could be factors on the access to new technologies, were not taken into account. This could be another interesting study to lead.

3.1 Have used a computer

Figure 2 shows that only 38% of people have already used a computer. If we consider the Europe, according to the Older Population Survey, 40% of the European aged over 50 have already used a computer in 1999 (Kubitschke, 2001). Since this study, we can suppose that the percentage of users have increased. If the age is taken into account, it can be seen in Figure3 that the more people are aged, the less they have used a computer. Besides, anybody over 75 years old has used a computer. About 30 people living in retirement homes answered the survey. All answered that they never use a computer. These people have over 80 years old and this can explain why they never use a computer. Indeed, the Internet is a recent technology; it is available for the large public for a few more than 10 years and at its beginning few people used it. The elderly who use the Internet are more often some who used it at work. The employment could be a key factor linked to the use of the Internet. Indeed, people who did not use ICT when they were in the professional environment do not use them when they are no longer in activity (Östlund, 1998).

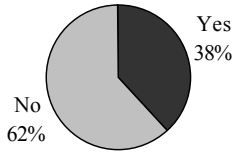


Figure 2: Use a computer

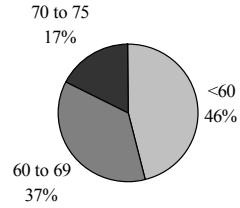


Figure 3: People using a computer according to the age

3.2 Own a computer/ an Internet connection

Among asked people, a quarter have a computer and 80% of people owning a computer have an Internet access. Thus, it can be assumed that people have a computer mainly to surf on the Web. In the autumn term 2004, the French average of people owning a computer was of 45.1%, the Mayenne department is behind in the use of a computer (Mediametrie, 2004).

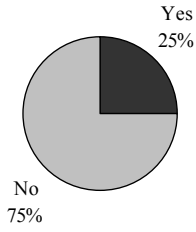


Figure 4: People owning a computer

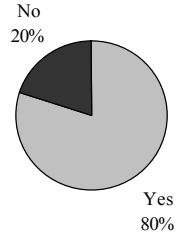


Figure 5: People having an Internet access

According to the Figure7, 33% of people aged 60-69 years old have a computer. According to the INSEE (National Institute for Statistics and Economic Studies is a "General Directorate" of the French Ministry of the Economy, Finance, and

Industry), at the beginning of the year 2004, 28% had a computer. Thus, people of Mayenne are above the national average. Figure 6 shows that 46% of the 50-59 years old own a computer compared to 56% of the French average for the same age group. People of Mayenne are below the average. About 7% of people of 70 years old and over have a computer compare to 9% of the national average (Frydel, 2005). This age group is also below the average. Overall, people living in Mayenne are less equipped of computers if we compare the averages of this study and of the INSEE study.

More or less the same results are obtained if we consider the ownership of an Internet access at home. However, the study reveals that people living in Mayenne are more equipped of Internet access than the global French population. Indeed, the study reveals that 63% of the Net surfers have a high speed Internet access against 50.1% of the national average (Mediametrie, 2004). This can be explained by the fact that the Mayenne department is rather well equipped in high speed connections. At the end of the year 2005, 97% of the department would be covered. Thus, Mayenne would be among the first department which provides an almost total cover (Conseil Général de la Mayenne Direction de la Communication, 2005).

Figure 6: Own a computer 50-59 years old

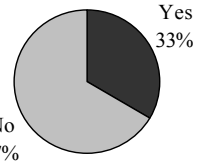


Figure 7 : Own a computer 60-69 years old

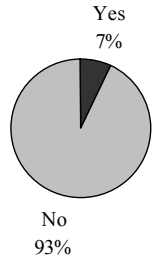


Figure 8: Own a computer - > 70 years old

According to the following pie charts, about 20% of people live in villages between 500 to 1,000 inhabitants and only 9% of people using a computer live in such a village. About 30% of people live in communes between 1,000 to 2,000 inhabitants and 40% of people using a computer live in such a commune. About 45% of people live in towns of more than 2,000 inhabitants and 47% of people using a computer live in such a town. It can be deduced that the majority of people, who answered the survey and who use a computer, live in communes between 1,000 and 2,000 inhabitants. In France, a town is defined by 2,000 inhabitants. So, the study reveals that in the Mayenne department, there are more people who use a computer living in villages than in towns.

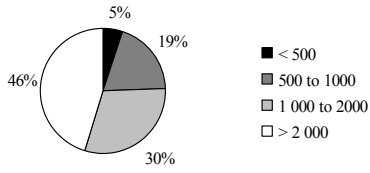


Figure 9: Percentage of asked people according to the size of the town

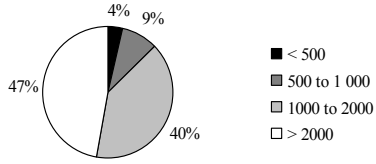


Figure 10: People using a computer according to the size of the commune

3.3 Type of difficulties

People encountered difficulties at the beginning of the use of the Internet. The difficulty the most spread is not to find relevant information. Then, lot of people have problem of connection, essentially with 56ko modem because they often are disconnected and sometimes cannot access the Internet. This problem is rather technical than a real user problem. Advertisement and spam overrun people's email box and people can be lost by advertisement which display when they surf on the Web.

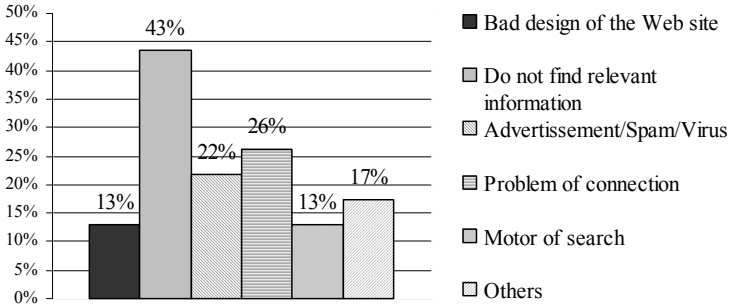


Figure 11: Type of difficulties

3.4 People wanting to use a computer

Very few people want to learn to use the Internet (16%). Generally, the more people are aged, the less they want to begin to use the Internet (84%). There is no distinction between men and women and if people live in a rural or an urban area, but only a distinction according to the age.

3.5 Factors that prevents people to use a computer

A first research reveals that old people can encounter the following problems (Viriot-Durandal; Coutty, 2004): fear of the innovation and the complexity, lack of skills, lack of confidence, fear of the failure in the learning, manipulation of the keyboard and the mouse, difficulties with vision, difficulties with mobility (rural isolation), training facilities can be inappropriate or inaccessible, … In the study, we find more or less the same factors.

The first factor that prevents people to use a computer is the fact that people are not interested in using this technology because they do not see the usage. About 54% of people answer this. This result agrees with the poll realised by the CREDOC (A research centre for the study and observation of life conditions) which reveals that 50% of the 60 years old and over think that computers are useless in the daily life. Indeed, according to the CREDOC study, 29% of the 60-69 years old see an interest to the use of a computer and only 26% of the retired people (Viriot-Durandal). Perhaps if the elderly knew what the Internet could bring them, for example find out more about their personal interests and hobbies, they would change one's mind. More people will think that there is an interest to the use of a computer. Indeed, if people do not see an interest to use something they will not use it. Of course, there are IPAPs with coordinators to help people to use computers. But if people are not interested, they will not come. We need to find something that could interest them and introduce the computer like a tool to achieve their needs.

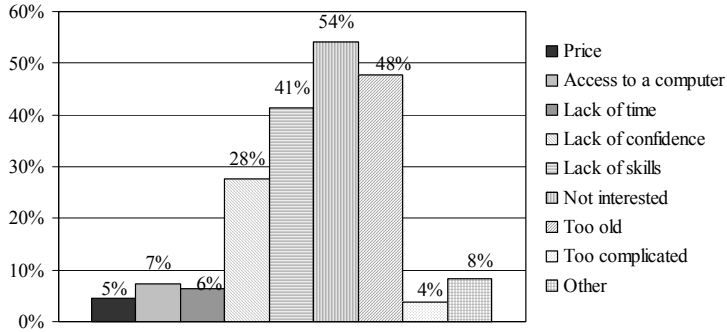


Figure 12: Factors

The second most spread answer is "I am too old to use a computer" with about 48%. People think that the use of a computer is not of their generation and that it is too complicated to begin at over 60 years old. However, an experience realised by Marquié, Jourdan-Boddaert and Huet in 2002 shows that the elderly underestimate their own capabilities to use new technologies. Indeed, we know that people hesitate to do something when they think that they are incompetent. If we want to encourage the elderly to use new technologies, we need to help them to find confidence in their capabilities to do it (100 fenêtres sur Internet).

About 41% of people assert that they do not use a computer because they do not have the necessary skills. Indeed, if we compare a telephone or a TV (which were also new technologies) to a computer, it seems that a computer and its software are technologies really more difficult to use. In the case of a phone, the same procedure is repeated at each call (hang up, dial, talk and hang back up), so the technique is transparent and easy to learn. With a computer and the Internet, the technique is more complex and the learning can be longer. Certainly, the use of a keyboard and a mouse can be difficult at the beginning. They have in particular problems with some keys of the keyboard such as make stress mark, write in capital letters or erase mistakes. It is also difficult to use the mouse because the cursor move too fast at the beginning and when they want to click in a particular place the cursor go away. This

learning can be long. Moreover, software is more difficult to use because people need to learn, to understand and to remember the different functions and this in a continuous way because new versions are provided. However, there are always difficulties when someone begins to use a new technology but the skills can be gained after some practises.

28% of people say that they lack of confidence and think that they do not succeed in using a computer correctly. Thus, they prefer not to use it. The "other" category regroups people who have problems of vision, problems of memory, and people who think that there are too freedom in the use of the Internet. From people who answered that they are interested in using a computer, the main factors that prevent them to use it are: the lack of skills, the access to a computer and the price.

3.6 Promote the use of ICT

For the moment, France is behind in the use of the Internet by elderly people. Thus, the authorities need to carry on promoting the Internet otherwise some people could be passed through a great number of services. Many initiatives have been taken into place on local and national level to favour the Internet access to everybody. Since 1998, majority of regions and communities have set up IPAPs. Thus, in 2004, more than 3,000 IPAPs mesh the country. It is poor compared to the UK with more than 6,000 online centres (DUI, 2004). These premises are needed to communicate the benefits of ICT. Indeed, the computer is currently the most important way to go to the online world and to offer services. These will continue to develop. It is mainly the elderly who need these services more than younger people. So, if they do not take part to technologies, they will be excluded of the society (Bradbrook and Fisher, 2004).

4. Discussion

The survey reveals that 40% of people who have 50 years old and over have used a computer. It can also be noticed that the more people are aged, the less they have used a computer. Indeed, the greatest number of users has less than 60 years old. People over 75 years old have never used a computer (Figure 3). This can be explained by the fact that in the past few people used a computer in their work because it was a new technology and in rural areas new technologies arrive later than in urban areas. Indeed, some surveys have stated that the professional environment is often a key factor linked to the use of a computer. Thus, if the elderly did not use a computer when they worked, they will not begin to use it when there are in retirement. It can be assumed that in the future years more elderly people will use a computer and the Internet because people will use it during their active live.

The survey also shows that few people, who do not use the Internet, want or wish to use it. Only 16% would like to begin to use the Internet (Fig 12). This trend does not depend on the gender or if people live in a rural or an urban area, but depend rather on the age. The more people are aged, the less they want to begin to learn to use the

Internet. The three main factors that enable them to begin to use the Internet are: Not interested/Not see the usage (54%), too old (48%) and lack of skills (41%)

However, it seems that the elderly who want to learn to use a computer and the Internet wish to attend training. 64% of people mainly people who think that they lack of skills would like to go in an IPAP to attend initiations. Thus, IPAPs seem important in the acquisition of the knowledge and the practises for people who want to begin to use a computer.

The interview, given by M. Roussel, shows that even if people have a good education (lecturer at the university) they do not necessarily use a computer. It would seem that **the use of a computer is neither a matter of age nor of diploma but rather of individual wish**. It is right that the family or friend circle is perhaps a factor which favours the use of ICT.

To extend this research the following could be undertaken:

- Lead the same study in Paris or a big city to compare the results,
- Take into account in the survey the level of diploma and the social category to know if there is a link with the use of a computer,
- Find a “hook” to interest the elderly to ICT. If this interest is found, it will be easier for them to begin to use a computer because they will have the wish. Indeed, if someone does not want to do something, he (she) will never do it.
- Find new interfaces or improve the existing interfaces (the keys of the keyboard could be larger and the mouse must be easier to handle).
- Design easier software with only useful functions (browser with only the file, edition and view menu)

5. Conclusions

As show some studies, few elderly people use a computer or the Internet and the more they are aged, the less they use it. As revealed by the Ipsos study, particularly older citizens are at risk to be left behind on the "Information Highway". In the Mayenne department, only 40% of people aged over 50 have already used a computer. The rate of computer equipment is weak, 25% against 45.1% of the global French average. The main factor that enables people to use a computer is that the elderly are not interested in this technology. Thus, this group of population is excluded despite the fact that the government promotes the use of ICT by setting up IPAPs through the country in order to help people to reach new technologies. The premises are well present but people, who are not interested in, do not come to these places. The benefits of ICT, which should focus on the interests of the elderly people, have to be proved. If an initial “hook” is providing, then it will be easier to introduce the computer and the Internet like tools to achieve their needs.

6. References

100 fenêtres sur Internet, “Chapitre 1 - Vers quelle société de l'information ?”, http://www.mshs.univ-poitiers.fr/laco/Pages_perso/Rouet/Textes/rapport-100fenetres/chap1.pdf

Bradbrook, G. and Fisher J. (2004), *Digital Equality: Reviewing digital inclusion activity and mapping the way forwards*

Beky A. (2004), *NetEconomie* “Equipement PC : Les Français rattrapent leur retard européen”, http://www.neteconomie.com/perl/navig.pl/neteconomie/infos/article/20040121180111

Bouchayer F., Gorgeon C., Rozenkier A. (2002) “Les techniques de la vie quotidienne - âges et usages”, *Mission Recherche-DREES*, Ministère de l'Emploi et de la Solidarité, p93, http://www.sante.gouv.fr/drees/ouvrage-mire/ouvr10.pdf

Carboni F., (2003), “Les Français et Internet” written for the *altema newspaper*, http://www.altema.com/Dossiers/Univers/internet.html

Conseil Général de la Mayenne Direction de la Communication, (2005), “Internet & nouvelles technologies : haut débit en Mayenne : un des meilleurs taux de couverture de France”, http://www.lamayenne.fr/front.aspx?SectionId=9&PubliId=4324

Coutty M., (2004), “La fracture numérique entre les générations se réduit” written for the *Monde newspaper*, http://www.globalaging.org/elderrights/world/2004/fracture.htm

DUI - Délégation aux Usages de l'Internet, (2004) “Programmes territoriaux”, http://www.internet.education.fr/acces/regional.htm

Frydel Y., (2005), *étude INSEE*, No1011, “Un ménage sur deux possède un micro-ordinateur, un sur trois a accès à internet”, http://www.insee.fr/fr/ffc/docs_ffc/IP1011.pdf

Kubitschke, L., (2001) “Older People and Technology - Preliminary Results form the SeniorWatch Surveys”, http://www.stakes.fi/cost219/prockubitske.doc

Mediametrie, (2004) “L'audience de l'Internet en France en Octobre 2004”, http://www.mediametrie.com/resultats.php?resultat_id=71&rubrique=net

Östlund B., *Linkôping University*, Sweden, (1998) “Profil des utilisateurs des technologies de l’information et de la communication chez les personnes âgées”, http://www.cnav.fr/4presse/themes/pdf/theme6/technologie/profil.pdf

Viriot-Durandal, J.P., lecture teacher at the University of Franche-Comté, “Seniors et nouvelles technologies”

Biometrics for Mobile Devices: A Comparison of Performance and Pattern Classification Approaches

M.Krishnasamy and N.L.Clarke

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Mobile devices have become indispensable tools nowadays. With growing technologies, applications and services are being added to the mobile devices all the time. Its usage in business and enterprises need it to be secure from unauthorised access. The extent of protection currently available is not adequate for the services that are employed in mobile devices. Biometrics have the privilege of providing secure authentication through utilising the unique characters of a person. Reports on the theft and loss of information and the wide acceptance on biometric authentication paved the way in its research on mobile devices. Several performance issues are to be considered when implementing biometrics in mobile devices

This paper focuses on the comparison of different pattern classification approaches employed in Face, Fingerprint, Keystroke and Signature biometric techniques and their effect on the performance on these devices. A detailed study on different algorithms employed in each technique has been performed. Most of the algorithms that are used for authentication follows similar approach regardless of the techniques and are broadly categorised between statistical and neural network approaches. Processing time in each approach is spent for feature extraction and classification and the storage for holding these features. Neural network techniques performs authentication with higher accuracy but require huge memory capacity and longer training time which makes it infeasible to be employed in mobile devices. Statistical approaches although consumes less processing time than neural networks, still requires considerable processing time to perform authentication in real time.

Biometrics is a future technology which can provide secure authentication. Biometrics in mobile devices will become practical if the developments in technology in mobile architectures and software are implemented fully on these devices.

Keywords

Mobile Devices, Neural Network, Biometrics, Statistical, Pattern Classification

1. Introduction

Mobile device technology is one of the rapidly growing technologies in the past few years with some 1.5 billion devices currently in use all over the world which is more than three times the number of Personal Computers (PC) (Prensky, 2004).This exponential growth is due to its widespread emergence and rapid adoption of handheld computing devices. Mobile devices have become an inherent part of the business environment with its usage on online banking, share dealing, micropayment

and m-commerce. M-commerce which has predicted revenue of $554.37 by 2008 (Telecom Trend International, 2004) is a major application that has been facilitated by the features on mobile devices. It also has the promising future in business to consumer market. As more and more services are added to the mobile devices with the substantial increase in the number of devices, security risks on those device has become more apparent. So the success of m-commerce and services depends on how security is implemented on these devices.

Security in mobile devices is provided using Personal Identification Numbers (PIN), a widely used method for authenticating users. Some of the drawbacks on using this method are that they can be forgotten by the user, can be easily guessed, stolen or cracked. PINs are also weaker in providing strong authentication that is needed for the services and applications available in current 3G phones. There are several incidents published on the loss of information or theft from PDAs, phones and converged devices (CSO Online, 2004). It was reported that in 2002, a mobile phone was stolen every three minutes on average in the UK which made the UK government to ask the mobile manufacturers to improve the security of the mobile devices that will make them difficult to use after stolen (Techplus, 2002).

In addition to that, a survey on mobile phone subscribers showed that about 45% of the mobile phone users responded that PIN is an inconvenient method of authentication and 81% of users were interested in improving security in the mobile devices and also want an authentication method which is different from the traditional way of using PINs (Clarke *et al*, 2003A). Smart cards are another alternative to PINs which is based on what the user has, but it can also be lost or stolen. From the aforementioned survey, with 81% of respondents believing on higher security, authentication using biometrics is one of the solutions to it.

2. Biometrics and Mobile Devices

Biometrics is the process of identifying an individual using his/her physiological or behavioural characteristics which are unique to that individual. It is an excellent candidate for identity verification which uniquely differentiates a user by authenticating him with the characteristic possessed by him. Physiological techniques measure the physiological characteristics of an individual, such as fingerprint which are unalterable and remain relatively stable over time. Behavioural techniques such as signature and keystroke measure the behavioural characteristics of a person. Behavioural characteristics however tend to vary to a greater degree over time due to a variety of reasons.

Performance of biometric devices is governed by the False Rejection Rate (FRR), the rate at which an authorised user (genuine user) is rejected from the system and the False Acceptance Rate (FAR), the rate at which an unauthorised user (impostors) are authenticated to the system. Equal Error Rate (EER) is used to compare between the FAR and FRR which determines the accuracy of the system with lower the EER the more accurate the system.

2.1 Biometric System

A biometric system is a device which uses a single or multiple biometric techniques to identify an individual and provide access to the system. Centralised and distributed are two types of architectural design used in biometric systems. In centralised system the feature values extracted from the raw data are sent to the common system which will perform the comparison with the stored features and produce the output whereas in distributed systems extracted features are compared locally on the device and the result is produced. Even though centralised systems has the advantage of being supervised and maintained easily they consume higher communication bandwidth and also has a greater risk of a system-wide failure when compared with the distributed systems (Kung *et al*, 2005).

2.2 Biometrics in Mobile Devices

Mobile devices can be used both in centralised and distributed systems. Studies on mobile devices in centralised systems have been carried out by (Massachusetts Institute of Technology, 2004; Tsai *et al*, 2003; Weinstein *et al*, 2002; Hazen *et al*, 2003; Clarke *et al*, 2003B; Clarke *et al*, 2004), where the data was collected and sent though the network for verification, authentication is performed if it matches with the stored features on the database. Implementing authentication as a distributed system in a mobile device is not popular and requires considerable development in technology on the mobile devices to be adapted widely. This paper focuses on evaluating the processing and storage requirements needed to implement a distributed mobile device biometric authentication. Personalisation is one of the features that can facilitate distributed biometric authentication in mobile devices because it allows the user templates to be stored on to the local device.

The performance of biometric authentication on a mobile device will depend on the pattern recognition and classification algorithms used for enrolment and verification of the extracted data. Algorithms which require more complex functions for execution will consume more processing time, more power and storage which are the critical components in a mobile device, so the success of biometric authentication will depend on the algorithms employed. Statistical and neural network approaches are widely used pattern recognition methods in biometrics.

3. Mobile Specific Biometric Approaches

Several biometric techniques are available which can be used to recognise a user. The following techniques are chosen for its employability in mobile devices. Different algorithms used in each technique are discussed in relation to their performances.

3.1 Fingerprint Recognition

In fingerprint recognition, authentication can be performed using a sensor device which senses the fingerprint of a user to provide authentication. Some of the main

characteristics of a fingerprint image are area, resolution, number of pixels, geometric accuracy, contrast and geometric distortion. Minutiae based method, k-Nearest Neighbour (k-NN), SVM and Backpropagation neural network algorithms are compared and the results were concluded based on the studies of (Yao *et al*, 2001) in Table 1.

Algorithm	Minutiae	k-NN	SVM	Neural Network
Accuracy (%)	95	87.9	88	86
Advantages	Reduce minutiae's will reduce processing time	Simple Algorithm Less processing	Requires less training Accuracy is high	Efficient output once trained
Disadvantages	Accuracy reduced with few minutiae's	Execution time increases with k	Requires considerable development	Requires more time for training

Table 1: Performance Comparison of Fingerprint Recognition Algorithms

3.2 Face Recognition

Face recognition is the process of identifying an individual from the images of their faces using the extracted features which are stored in the database. In mobile devices, face authentication can be performed by capturing the image using the built in camera. The presence of input devices like this facilitate easier authentication. Eigenface method, Elastic Graph Matching (EGM), Support Vector Machines (SVM) and Backpropagation Neural Network are the extraction and classification algorithms that are compared on face recognition and the results were concluded based on the studies of (Ho-Man, 2003; Jun Zhang *et al*, 1997) in the Table 2 (calculated using a 1400 MHz desktop PC)

Algorithm	Eigenface	EGM	SVM	Neural Network
Storage (ORL) (MB)	5	1.5	38	32 KB
Accuracy (%)	80.3	81.5	95.5	91.5
Avg. Running Time (Seconds)	2.1	16.3	6	1.4
Advantages	Provides lossless data Less execution time	Higher accuracy in less lighting conditions, face positions and expressions Uses only key point of the image	Higher accuracy in less lighting conditions	Faster authentication time More accurate
Disadvantages	Lower accuracy in varying light intensities, scale and orientations	Longer computational time	Longer time to train	Longer training time More data for trained Affected by lighting small variations

Table 2 Performance Comparison of Face Recognition Algorithms

3.3 Keystroke Recognition

Keystroke dynamics is the behavioural way of authenticating a user by analysing the way a user types on the keyboard input and identifying a rhythm pattern which can vary with time. It is considered to be a most attractive biometric authentication scheme for its transparency to the user. There is no requirement of additional tool or hardware need to implement authentication where the keypad itself acts as a tool to authenticate the user. k-NN and Neural networks are algorithms that are compared and the results were concluded based on the studies of (Wagacha, 2003; Cho *et al*, 2000; Clarke *et al*, 2003B) in the Table 3.

Algorithm	k-NN	Neural Network
Accuracy (%) – EER	14.2	11.3
Advantages	Easy to program without the need for optimisation or training Accuracy can increased by increasing k	Highly accurate (low EER)
Disadvantages	Execution time is more when k is high and more data is applied	More sample data to get trained Retraining when new user added

Table 3: Performance Comparison of Keystroke Recognition Algorithms

3.4 Signature Recognition

In Signature authentication the dynamic characteristics like speed, acceleration, direction, pressure, etc are compared. In a mobile device, on-line signature recognition can be implemented by writing using a pressure sensitive pen on the touch screen of the mobile device. Similar to keystroke, signature authentication does not require any additional hardware; instead the stylus of the mobile device can be used. Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) algorithms that are compared (Griess, 2000).

Algorithm	HMM	Neural Network	DTW
Accuracy (%) - EER	1 - 4	4	2 - 3
Advantages	Can model wide range of variation Increase in state increases accuracy	Higher Accuracy	Finds the exact points during matching
Disadvantages	Longer training time when the states increased	Requires more time to be trained	More computation Suffer from warping forgeries

Table 4: Performance Comparison of Signature Recognition Algorithms

4. Discussion

In biometrics, the authentication time is the time required for the system to process the request made by the user to authenticate into the system which depends on the algorithm used. For a more accurate output the algorithm employed will perform more calculations which in turn consumes more time and for a less accurate output the algorithm consumes lesser time for authentication so there needs to be a trade-off between the accuracy and execution time.

Biometrics application on mobile devices is currently on development with number of studies being performed on it. This paper analysed the four different biometric techniques and their algorithms used to perform authentication. From the study it is revealed that each different technique has some features that facilitate to be implemented in mobile devices.

Face recognition in mobile devices are performed by (Massachusetts Institute of Technology, 2004; Tsai *et al*, 2003; Weinstein *et al*, 2002; Hazen *et al*, 2003) used dedicated servers which consumed more time in transmitting the data over the network and this can be improved by performing the processing in mobile device itself. SVM and EBGM approaches have the advantage of performing face detection even in less lighting conditions which will make them suitable to be used in mobile devices, where the authentication needs to be performed with variable background environment which depends on the user location.

Experiments performed by (Clarke *et al*, 2003B; Clarke *et al*, 2004) used mobile keypads interfaced with a desktop PC to provide the necessary processing and suggested that the neural network patterns performed well in producing keystroke recognition. Although neural networks performed authentication in higher accuracy, it requires more training time, so k-NN can only be possible in mobile devices.

Fingerprint biometrics for mobile devices needs a dedicated hardware such as sensors to be fabricated to the mobile device hardware. Dedicated processors or chips are being developed to perform fingerprint recognition in mobile devices. Minutiae method can be a possible solution in mobile devices where a reduced minutiae has the capability to perform accurate and faster authentication

Signature recognition has the capability to be used in mobile devices such as smart phones and PDAs which has a touch screen and digital pen that allows the user to sign on the device to perform authentication. HMM method has the advantage of modelling wide range of variation with increased accuracy and less execution time.

Algorithms are available in each technique with low computation time and higher accuracy that can be used for authentication. But practical implementation requires issues such as mobile device architecture needs to be considered for authentication in real time. As computation of the algorithm becomes more complex there will be more processing done by the registers and other processing components in the mobile device and this will also have a significant influence on the battery life time if those algorithms had to run many times.

Neural networks had more success in producing accurate output with less FAR and FFR in most biometric techniques and it also produced a quicker authentication when trained. But training requires a large number of features with huge storage. Accuracy can be increased with multiple layers which increases processing time for training as the complexity of the algorithm is increased. Since current mobile devices come with a moderate storage capacity, storing will not become a big issue but the processing time in neural networks make it impossible to be used in mobile devices. Mobile devices are personal devices so it is wasteful to store the features of other users or impostors to make the comparison, so neural network techniques are not recommended for biometric authentication in mobile devices.

Most of the pattern recognition algorithms require floating point operations for their complex mathematical functions but the current mobile devices include processors which are unable to perform floating point calculations. They perform floating point operations by converting them to fixed point numbers which results in longer processing and execution time. One of the method to reduce processing and execution time is by choosing an algorithm which requires less arithmetic and floating point operations like k-NN algorithms where there is no arithmetic operations involved instead the result can be obtained using the shift and logical operations. Other method can be employing code optimisation, in this approach codes are written by making changes in the algorithms implementation that can utilise the processor efficiently in order to save energy and reduce execution time when applying the complex algorithms. Dynamic voltage scaling is another approach that can change the processing frequency and voltage at run-time to reduce the energy consumption.

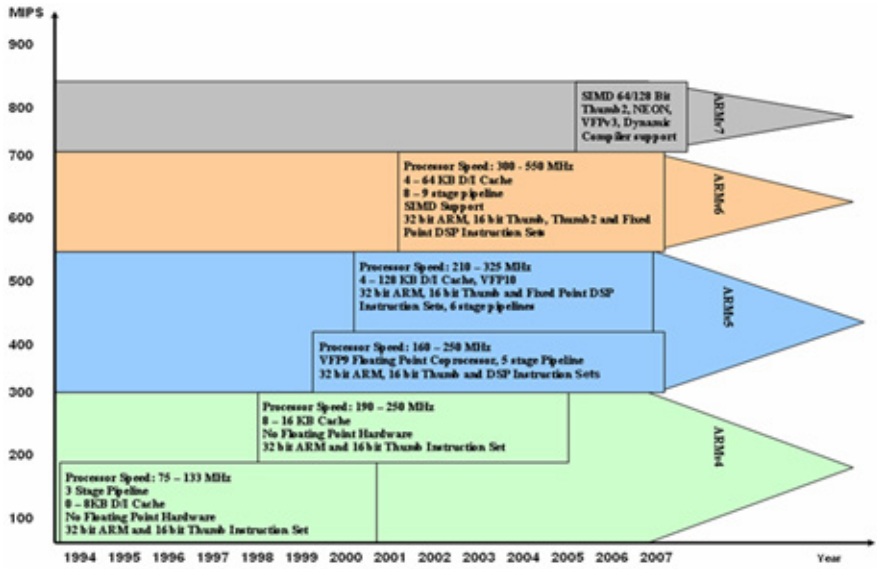


Figure 1: Trends in Mobile Technology

The technology trend in mobile architecture of ARM (ARM, 2005) is given in Figure 1. The figure show that although there is a gradual increase in processor speed over the years, a considerable change in instruction set has been taken place such as an addition of DSP and Thumb instructions which improved the programme flow and its efficient code size improved the power and performance of the mobile devices. NEON technologies on ARMv7 will be implemented in future mobile devices which can provide an extensive set of new instructions to provide an accelerated output when compared to all the other previous technologies (3x performance when compared with ARMv5 and 2x performance of ARMv6 on DSP applications). Floating point coprocessors are being added along with increasing MIPS. The addition of SIMD instructions can also increase the performance of software applications and increasing pipeline stages are being taking place which can facilitate parallel processing. These new technologies will allow biometrics authentication to succeed in mobile devices.

Multi-model biometrics can also be another solution to reduce the processing time where multiple less accurate, low processing algorithms from different techniques can be implemented, providing an accurate and reliable authentication with less EER so that the combined computational time is less than the computational time for executing a more accurate algorithm. This can be facilitated by inbuilt facilities like camera, sound input, touch screen and keypad in the mobile devices.

5. Conclusion and Future Work

Secure mobile communication enables users to use application such as mobile payment and finance, mobile ticketing, mobile voting and location based services with increased convenience and confidence. In providing security the current form of authentication is a cheap solution but it suffers from a number of security weaknesses and biometrics is a strongest approach when compared to all other forms of authentication.

Even though biometrics seems to be a perfect solution, issues of performance and its ability to provide authentication in real time has to be considered before implementing them in the mobile devices. Applications that perform biometric authentication will become feasible if those devices are equipped with large memory storage and high speed processor that has low power consumption. The memory storage is being increasing to facilitate more services but the issue of processing and power requirement affected by the complexity of the algorithms still remains an issue on mobile devices.

Neural network approach requires larger processing time and memory during its training phase which makes them infeasible to be employed in mobile devices. While in statistical approaches, algorithms with less mathematical operations or using architectural and algorithmic optimisation on the codes has been suggested to improve the processing time.

As this study analysed various algorithms and their performance in terms of the processing, it provided an efficient starting point for further research on its deployment in mobile devices. In future, the study on the instruction sets and architecture of the processor employed in the devices will allow writing software codes with code optimisation that can evaluate the processing time and storage requirements for each algorithm practically and determine its efficiency.

Mobile device are personalised devices and authentication can be preformed using one to one verification, but the success of this approach will depend on the accuracy of the algorithms which is used to perform authentication. Technological trends discussed showed that there are improvements being taking place that can facilitate biometric authentication using statistical approaches and also the usage of neural networks can become viable on the development of cell processor technology that facilitates parallel processing. So biometrics in mobile devices will become a practical reality within a few years.

6. References

ARM (2005), "Processor Core Families", http://www.arm.com/products/CPUs/families.html, (Accessed 1 September 2005)

Cho, S., Han, C., Han, D.H. and Kim, H. (2000), "Web-Based Keystroke Dynamics Identity Verification Using Neural Network", *Journal of Organisational Computing and Electronic Commerce*, 10(4), pp295-307.

Clarke, N.L., Furnell, S.M., Lines, B.M. and Reynolds, P.L. (2003A), Keystroke dynamics on a mobile handset: A feasibility study, *Information Management & Computer Security*, Vol. 11, No. 4, pp161-166.

Clarke, N.L., Furnell, S.M., Lines, B.M. and Reynolds, P.L. (2003B), Using Keystroke analysis as a mechanism for Subscriber Authentication on Mobile Handsets, *Proceedings of the IFIP SEC 2003 Conference*, May, Athens, Greece, pp 97-108.

Clarke, N.L., Furnell, S.M., Lines, B.M. and Reynolds, P.L. (2004), Application of Keystroke Analysis to Mobile Text Messaging, *Proceedings of the 3rd Security Conference*, 14-15 April, Las Vegas, USA.

CSO Online (2004), "Managing and Securing Mobile Devices", http://www.csoonline.com/analyst/report2794.html, (Accessed 1 September 2005)

Griess, F.D. (2000), "On-line Signature Verification", http://www.cse.msu.edu/cgi-user/web/tech/document?ID=449, (Accessed 12 September 2005)

Hazen, T.J., Weinstein, E., Kabir, R. and Park A. (2003), "Multi-Modal Face and Speaker Identification on a Handheld Device", *Proceedings of the Workshop on Multimodal User Authentication*, December, Santa Barbara, California.

Ho-Man, T. (2003), *Face Recognition Committee Machine: Methodology, Experiments and A System Application*, MPhil Thesis, The Chinese University of Hong Kong.

Jun Zhang, Yong Yan and Lades, M. (1997), "Face Recognition: Eigenface, Elastic Matching, and Neural Nets", *Proceedings of the IEE*, 85(9), pp 1423-1425.

Kung, S.Y., Mak, M.W. and Lin, S.H. (2005), *Biometric Authentication: A Machine Learning Approach*, Prentice Hall PTR, ISBN: 0-13-147824-9

Massachusetts Institute of Technology (2004), "MIT Project Oxygen", http://oxygen.lcs.mit.edu/H21.html, (Accessed 1 September 2005)

Prensky, M. (2004), "What can you learn from a cell phone? – almost anything!", http://www.marcprensky.com/writing/Prensky-What_Can_You_Learn_From_a_Cell_Phone-FINAL.pdf, (Accessed 1 September 2005)

Techplus (2002), "Why the interest in mobile phone security?", http://www.tekplus.com/TP0039A02V01.html, (Accessed 1 September 2005).

Telecom Trend International (2004), "Mobile Commerce Takes-off", http://telecomtrends.net/pr_MIIS-1.htm (Accessed 1 September 2005).

Tsai, Y., Fu, R., Huang, L., Huang, C. and Liu, C. (2003), "Handheld Person Verification System Using Face Information", In *7th International Conference on Digital Image Computing: Techniques and Applications*, 10-12 December, Sydney, Australia.

Wagacha, P.W. (2003), "Instance-Based Learning: *k*-Nearest Neighbour", http://www.uonbi.ac.ke/acad_depts/ics/course_material/machine_learning/kNN.pdf, (Accessed 1 September 2005)

Weinstein, E., Ho, P., Heisele, B., Poggio, T., Steele, K. and Agarwal, A., (2002), "Handheld Face Identification Technology in a Pervasive Computing Environment", http://cbcl.mit.edu/projects/cbcl/publications/ps/pervasive-2002.pdf, (Accessed 1 September 2005)

Yao, Y., Frasconi, P. and Pontil, M. (2001), "Fingerprint Classification with Combinations of Support Vector Machines", *Proceeding of the 3rd International Conference on Audio and Video based Biometric Person Authentication*, 6-8 June, Sweden, pp 253-258.

Security issues in Globus Toolkit 4

P.Coste and P.J.Brooke

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: coste@et.esiea.fr, info@network-research-group.org

Abstract

The aim of this paper is to present a review of the Globus toolkit security environment. The Globus toolkit is a development tool for grid computing applications. The goal of such applications is to share resources of several geographically disparate computers within a single entity: a "grid". These computers communicate using networks like the Internet. As a consequence the grid applications have to be secured. The first part of this paper explains the fundamental concepts behind grid computing. Then, the actual security concepts on which Globus is based are appraised and their issues spotted. Finally, the results of the implementations are reviewed. This paper provides the reader with an overview of the common problems encountered when implementing a Globus application and the potential threats. It reviews the actual security infrastructure available in the Globus toolkit environment.

Keywords

Grid computing, security, Globus, issues, mechanisms, Grid Security Infrastructure

1. Introduction

Grid computing is an emerging field. This form of computing connects several computers together into a single entity: "the grid". Today, grid computing is used successfully in applications like SETI@home or folding@home but also by scientists at the CERN with the Large Hadron Collider Computing Grid, which will be able to handle the 15 petabytes of data produced by the collider.

All connected computers use a network to communicate. This network may be shared with other users who should not have access to the grid. This paper studies the mechanisms implemented within the Globus toolkit to secure its grid applications.

The Globus toolkit is a major application in the grid computing area and its mechanisms are among the mechanisms commonly used by the grid computing community to secure applications. Globus is now at version 4 and this paper is an assessment of both the security concepts of the Globus toolkit and the actual implementation of security mechanisms in applications.

2. Background

Nowadays, some applications need huge calculation power. An alternative to mainframes and supercomputers has been developed: this is known as grid

computing. The goal is to gather together the resources of multiple computers in order to form a single entity that aggregates all those resources. Generally, the resources shared are processing power and/or storage capacity; grid computing uses parallel processing to deal with the tasks to be done. In parallel processing a task is typically split into several sub tasks which are distributed to the computers of the grid. Each computer will perform the calculation for its subtasks and return the results. The final answer is reconstituted from each result obtained when all the sub tasks have finished (Joseph and Fellenstein 2003). Similar aggregation can be done with storage and other resources. The aggregation of many computer resources can produce a huge drive which allows the storage of voluminous data other the network.

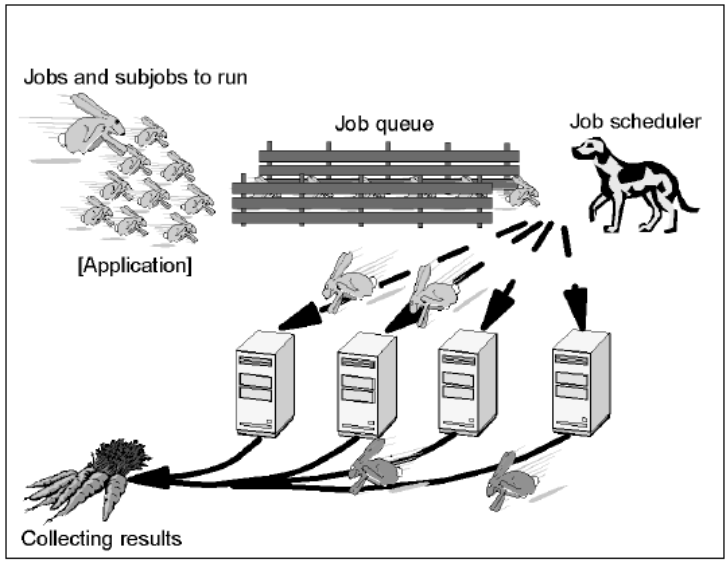


Figure 1: An application is one or more jobs that are scheduled to work on the grid (Berstis, 2002)

Grid computing is frequently used at the Internet scale and computers can be connected in a single grid even if they are separated by thousands of kilometres. As a consequence a network is used in order to connect all the computers together. This network can be a private network, but in the case of the Internet, it is shared with many other users. Some of these users may be part of the grid but others may not. Unauthorised users should not have access to information exchanged within the grid, especially sensitive information is exchanged within the grid. Grid applications can implement security features to prevent unauthorised users from accessing the grid and reading protected data.

The Globus toolkit 4 conforms to the Open Grid Service Architecture (OGSA) which relies on stateful Web Services (WS) (Sotomayor, 2005). A WS is an application with a standard interface that interacts with other applications using the network. An example of a WS is a weather report application that will reply with the weather report of a specific region when queried with specific parameters defined in the interface. In this case, the parameters could be a post code and a date. A typical WS

does not keep track of previous invocations. A stateful WS on the other hand uses a resource to store information that can be recalled or modified in future invocations. WS uses Simple Object Access Protocol (SOAP) message to communicate (Sotomayor, 2005).

The Globus toolkit is composed of two main categories of elements: the WS elements and the pre WS elements. In the present paper, the WS elements will be studied since they are today most used for grid applications because they are more flexible for programming web applications and they use standard interfaces and SOAP messages to communicate. As a consequence, the discovery and conversation are easier with a web service.

The Grid Security Infrastructure (GSI) implements the security mechanisms in the Globus toolkit to protect the WS developed (Globus, 2005). The GSI is an attempt to solve the three major concepts of security: Authentication, Authorisation and Accounting or the "AAA" (Stell, 2004):

1. Authentication means establishing the user's identity unambiguously.
2. Authorisation means giving the users access to the components/resources they are allowed.
3. Accounting means logging what have been done in the environment for future usage.

Furthermore GSI concentrates on three requirements (Globus, 2005):

- To establish secure communications between elements of the grid. A Secure communication is at least an authenticated conversation; it may also be encrypted and/or signed.
- Since several physical organisations may be crossed by the grid, security must be supported within all the organisations and may not be centrally-managed.
- Users must sign on only once for convenience even when using several components. "Single sign on" implies the delegation of the security parameters for some jobs that may require multiple resources and/or locations.

3. Security concepts within GSI and their issues

The following section will describe the security mechanisms in use in GSI and the associated issues.

3.1 Public key cryptography

GSI relies mainly on the RSA algorithm (named after its creators: R. Rivest, A. Shamir and L. Aldeman). This algorithm is an asymmetric algorithm. That is to say that the key used to encrypt and the key used to decrypt the message are different. The key used to encrypt can be publicly available. Nevertheless, it is impossible to recover the message without the private key. Furthermore, the private key cannot be

recovered from the public key. The major advantage of public key cryptography is that the decryption key does not travel on the network and is thus better protected. Default Globus keys are 1024 bits long and use an exponent (e) of 65537. These parameters are relatively secure nowadays. Nevertheless appropriate rights on the files which store the keys have to be selected to restrict access. Menezes et al. (1996) explain in detail the RSA mechanisms. The attacks against RSA can be found in the paper by Boneh (1999).

3.2 Digital signatures

GSI also uses RSA to digitally sign the messages. This procedure is considered as a proof that the right person created the message and that it had not been tampered with. A mathematical hash of the message is performed with the MD5 algorithm. It is then encrypted with RSA. The receiver of the message can decrypt the hash of the original message and compute the hash of the message received. If the two hashes are identical, the receiver can be sure that the message received has not been altered. Further details about MD5 can be found in the RFC 1321 (Rivest, 1992).

Wang et al. (2004) have identified a way to provoke collisions with the MD5 algorithm: two different messages will have the same message digest after the MD5 hash has been computed. Randomly chosen numbers added to the certificate may prevent this attack but unfortunately the X509 certificates used in the Globus toolkit do not have this functionality.

3.3 X.509 certificates

In order to solve the problem of authentication, GSI uses X509 certificates (Globus, 2005). The certificates combine a cryptography system with information about the users. Certificate systems are based on a Certificate Authority (CA) and trust. The certificates are issued by a third party: the CA, which certifies that the user is the person he claims to be. The CA collects information on the user and then issues a certificate, signing it with the CA's private key. When verifying the certificate of a user, the CA public key is used to check that the certificate has not been tampered. If the CA is trusted, the identity of the user with a valid certificate is accepted.

Nevertheless Lenstra et al. (2005) have constructed a valid pair of X.509 certificates that have identical MD5 signatures. As a consequence, a false version of a signed certificate which will appears to be valid can be issued by an attacker using a certain type of public key as described in their paper. The other fields can be chosen arbitrarily which is even more dangerous. The solution could be to migrate from a MD5 signing hash function to a SHA-1 hash function but this implies that X.509 certificates used today must be revised to include SHA-1 digital signatures. Other certificates like Verisign certificates already use SHA-1 signatures. Unfortunately, most CA are still using MD5 hash functions to sign their certificates.

3.4 Proxy

Globus uses a proxy in order to enable delegation and single sign on (Globus, 2005). When users want to log in the grid, they are asked for their passphrase to check if

they are the persons they claim to be. Once their identity has been confirmed, a proxy is created.

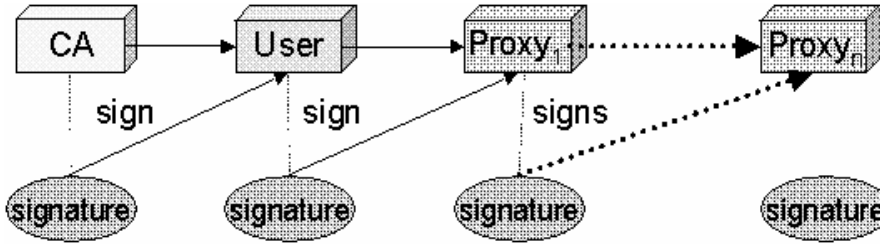


Figure 2: relation between CA, User and proxy (Globus, 2005)

This proxy will act on the behalf of users when they need to access a service to perform an operation. It is also used when delegating a task to another service.

Lock and Sommerville (2002) underline that when a proxy certificate is created by Globus, it does not contain the passphrase of the user and the key is unencrypted. Furthermore, when a proxy is created, it is valid for a defined period of time. If the user leaves the host, the proxy is still valid until the end of the period. An attacker may be able to use this old proxy to log in on the grid with a valid identity.

3.5 Conversation modes

There are two modes of secure conversation in GSI between the WS and the clients:

- Transport level security
  This implementation is based on HTTP over SSL (HTTPS) but it has been modified to enable proxy certificates. The Globus toolkit 4 (GT4) implementation does not support delegation of proxy certificates. This implementation provides privacy (i.e. encryption) and integrity to the data. The SOAP messages do not have any security information. The encryption and digital signature are done at an upper level by SSL.

- Message level security
  This is configured within the SOAP messages and not at the transport layer. There are two sub mechanisms available:
  - GSI Secure Conversation
    The security context is established before any data is transferred by sending dedicated SOAP messages. Once established, this context will be used for the entire conversation. A shared secret key is exchanged. Then all future operations use the shared secret key to sign and/or encrypt the data.
  - GSI Secure Message
    Here the security context is specified in each SOAP message exchanged for a request and a response, each time with the proper parameters. No additional messages are generated and external keys can be used.

3.6 Authorisation modes

Authorisation is mainly managed in GT4 with "gridmap files". A gridmap file is a file where the users authorised to access the grid / the service / the resource are listed. Each time a user wants to access a component with a gridmap file, this component will check if the user is in the gridmap file. This structure allows for an accurate granularity of control but unfortunately it is not very convenient since the administrator must modify the files one by one for any change he wishes to perform. This becomes an even more difficult task if the administrator has to manage several hundreds of users and hosts.

Alternate authorisation systems using Security Assertion Markup Languages (SAML) can be used such as the Community Authorisation Service (CAS) and the Privilege and Role Management Infrastructure Standards (PERMIS). In this approach, the authorisations details are stored in a database. A CAS proxy is generated for each user: this proxy will have credentials for the user and also the list of resources available to this user. This proxy is signed by the CAS service. When users connect to a component of the grid, the component will check the authorisations available in the CAS proxies. If the users are authorised, the service will grant them access. In Globus, a postgres database is generally used. In this database, users are assigned to groups. Each user is given rights (e.g. read, write …) on the desired services. CAS works using a specific service: the CAS service. This service is used to administrate the CAS policies for each user. A user logging in the grid must authenticate himself to the CAS server. This will create a CAS proxy. This proxy will be used by the user to communicate with the others services of the grid. The proxy contains the rights of the users on the services embedded in a modified GSI proxy. The proxy is signed by the CAS server which has to be trusted by the different services to grant or refuse the access to the users.

4. Globus Toolkit 4 WS implementation issues

During the project, the different mechanisms proposed by GSI have been implemented in a GT4 WS. The results are described in this section.

The security has been configured separately in the client and the server and different mechanisms may be chosen. Nevertheless, the client needs to satisfy the minimal level of protection asked by the server and conversely the server needs to satisfy the requirements of the clients in order to establish a communication.

At the server side, security parameters can be configured at several levels:

- At the container level
  The container is the entity containing all the WS available in the server. The parameters set at this level will be applied for all the web services.
- At the service level
  For each WS, different parameters can be set. Furthermore specific parameters can be chosen for each method remotely accessible of the WS.
- At the resource level

Each WS may have several resources that can be separately configured from a security perspective.

The structure described above provides fine tuning of the grid and the grid application security, and allows for better granularity of the overall security in the server side.

The first issue concerning security is common for the three methods of conversation between a client and a service. They all induce an overhead of packets and/or data that obviously slows down the transfers. In addition, when encryption is used, there is also an additional delay due to the encryption mechanism. The study by Shirasuna et al. (2004) confirms that security mechanisms slow down the connections. Among the security mechanisms, GSI Transport is the fastest and this fact has been verified during this research when testing the different communication mechanisms. Among the message security mechanisms, GSI Secure Conversation is faster if the communication has more than one invocation, because the security context is negotiated once. On the contrary in GSI Secure Message, the context is set up each time introducing more overhead. The only problem with GSI Secure Conversation comes out if the number of clients increases. GSI Secure Conversation does not scale well if numerous clients are connected, because for each client the security context has to be negotiated creating a lot of overhead.

The authentication methods are much more problematic. Actually, CAS is not implemented for WS since no interfaces are available to include CAS authorisation method in WS code contrarily to others authentication methods. From all the other methods available, the gridmap files method is certainly the most flexible; the implementation is easy and works well. Unfortunately it might have scalability issues: if a container has many services and users, administrating all the gridmap files will be a hard task. At the moment there is not a large choice of authentication methods.

The delegation of credentials is also an issue because it is only available with GSI Secure Conversation. On the other hand, the delegations parameters can be set entirely and easily using the security descriptors files which is very convenient.

5. Conclusions

The study of the security concepts supporting the Globus toolkit has shown that the choices made for the technologies used to secure the Globus toolkit are on overall good except for the MD5 hash problem. Nevertheless the implementation of the security mechanisms is much more problematic. The first and major problem comes from the lack of documentation and examples concerning the way to implement the security options. The documentation available is either too unclear or incomplete, and essential steps that may be obvious for an experienced developer are often missing. In addition, some mechanisms are partially implemented (e.g. delegation only available with SecureConversation, gripmap files only available in the server side). Globus is still a tool in development and is not yet mature enough for intensive

use in production with strict security requirements. Nevertheless, alternative systems using Globus as a root may be an efficient way to solve the issues of Globus.

6. References

Berstis V. (2002), *Fundamentals of Grid Computing*, IBM Redbooks paper, www.redbooks.ibm.com/abstracts/redp3613.html, (Accessed 10 September 2005)

Boneh D. (1999), "Twenty years of attacks on the RSA cryptosystem", *Notices of the American Mathematical Society* (AMS), Vol. 46, No. 2, pp. 203—213

Globus (2005), "GT4 Security: Key concepts", www.Globus.org/toolkit/docs/4.0/security/key-index.html (Accessed 1 September 2005).

Joseph J. and Fellenstein C. (2003), *Grid Computing*, Prentice Hall, Upper Saddle River, ISBN: 0-13-145660-1

Lenstra A., Wang X. and de Weger B. (2005), *Colliding X.509 certificates, version 1.0* www.win.tue.nl/~bdeweger/CollidingCertificates/CollidingCertificates.pdf (Accessed 20 August 2005)

Lock R. and Sommerville I. (2002), *Grid Security and its use of X.509 Certificates*, Lancaster DIRC, www.comp.lancs.ac.uk/computing/research/cseg/projects/dirc/papers/gridpaper.pdf (Accessed 20 August 2005)

Menezes A., Oorschot P., Vanstone S. (1996), *Handbook of Applied Cryptography*, CRC Press, Boca Raton, ISBN: 0849385237

Rivest R. (1992), "The MD5 Message-Digest Algorithm", *RFC 1321* www.ietf.org/rfc/rfc1321.txt (Accessed 1 September 2005)

Shirasuna S. , Slominski A. , Fang L. and Gannon D. (2004), *Performance Comparison of Security Mechanisms for Grid Services*, www.extreme.indiana.edu/xgws/papers/sec-perf.pdf, (Accessed 1 September 2005)

Sotomayor, B. (2005), *The Globus Toolkit 4 Programmer's Tutorial, version 0.1.1*, gdp.Globus.org/gt4-tutorial/, (Accessed 1 September 2005)

Stell A. (2004), *Grid Security: An evaluation of authorisation infrastructures for Grid Computing*, labserv.nesc.gla.ac.uk/projects/etf/MScProj.pdf, (Accessed 1 September 2005)

Wang X., Feng D., Lai X. and Yu H. (2004), *Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD*,eprint.iacr.org/2004/199.pdf (Accessed 20 August 2005)

Implementing a network operations centre management console: Netmates

R.Bali and P.S.Dowland

Network Research Group, University of Plymouth, Plymouth, United Kingdom
email: info@network-research-group.org

Abstract

Network Management & Intrusion Detection Systems (NMIDS) are an important part of any network security architecture. They provide a layer of defense which monitors network traffic for predefined suspicious activity or patterns, and alert system administrators when potential hostile traffic is detected. Commercial NMIDS have many differences, but information systems departments must face the commonalities that they share such as significant system footprint, complex deployment and high monetary cost. Netmates - Network Monitoring & Attack Evaluation System, which is based on Snort was designed to address these issues. It features a near real-time snort alert monitor, providing many ways to indicate that the network may be experiencing an intrusion attempt including audio / visual warnings, email warnings, etc.

Keywords

Network Management Monitoring Intrusion Detection Console Real-time NMS NMIDS IDS

1. Introduction

Netmates is an implementation of a range of selected software to monitor (in real-time) the network and security breaches in the form of alerts generated by Snort. Other combinations of such tools like Snort (Snort, 2005) & ACID (Acidlabs, 2005) And Snort & BASE (BASE, 2005) lack something or other. Snort is great for identifying suspicious traffic and ACID is great for digging in to the details there was a need for something that was a little higher overview and able to sound alarms if certain conditions were met. For instance, if the network is attacked 50 times in a 2 minute period. Netmates does not replace Snort or ACID but rather it compliments them. This paper discusses Netmates, based upon Snort which is here being used as rules-based traffic collection engine, in turn as a NMIDS

Netmates fills an evident gap in the domain of network security: It is a robust implementation of cross-platform, lightweight network management and intrusion detection tools that can be deployed to monitor TCP/IP networks and detect a wide variety of suspicious network traffic as well as outright attacks. It can provide users with enough data to make informed decisions and take proper precautions and actions to face the suspicious activity. The beauty of the system is that it can be rapidly deployed to cover the potential hole in the security as and when it is detected,

while the commercial equivalents depend on OS or firmware update. This is possible due to its open source nature and wide user base of its core detection engine - Snort.

It is possible to easily deploy Netmates on any machine running windows as host operating system. Because Netmates is a pre compiled, pre configured and portable linux system which is an ideal NMIDS solution that is available as a virtual machine based on Microsoft Virtual Machine Technology (Microsoft Corporation, 2004). This makes deployment of such a powerful tool quick and easy where other propriety solutions strive to score anywhere closer. It can comfortably fit on a CD, or can be downloaded over the network. Its ready and working is just five minutes, only needs Microsoft Virtual Machine installed and configured to use Netmates as a virtual hard disk. Also the Netmates console to view the alerts with audio visual aids.

2. Architecture

Netmates fulfills its objectives i.e. light weight Network Monitoring, Intrusion detection and wireless sniffing by deploying the following key techniques in the form of add-ons to snort which come as a separate package or as a plug-in. These can be either installed together to form a single software or multiple instances on one single machine or in a distributed environment logging to same database.

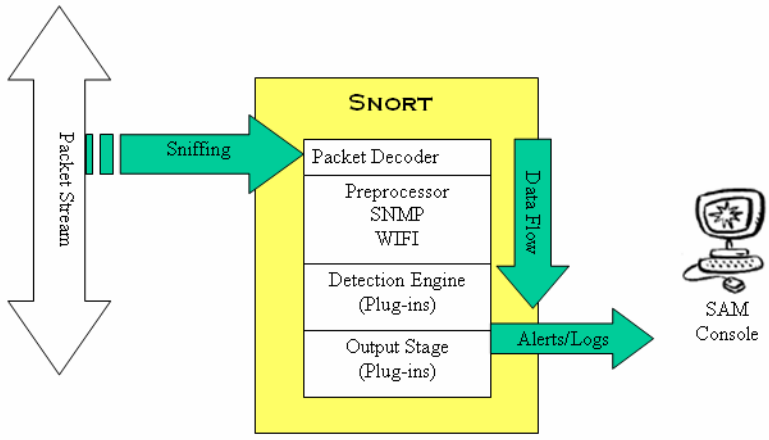


Figure 1: Netmates Data Flow

2.1 Snort

At its most basic level, Snort is a simple sniffer which means that it captures the network traffic for further analysis. It does so by listening to the network traffic in 'promiscuous mode' which allows it to capture all the data that passes over the network regardless of its destination address. Inherently, if a packet of information is addressed to another computer all other network cards will reject this data, which saves system resources.

Snort emerges more powerful when it is properly configured to detect intrusion using a combination of rules and pre-filters. It can also recreate data streams and analyse

them for any number of signatures that give some indication of a possible network attack

Most of the intruder activity has some sort of bit pattern also called signature. Information about these known patterns can be used to create Snort rules. There are many known vulnerabilities that attackers want to exploit. But at the same time these can be used as weapons to find out if someone is trying to exploit them. These bit patterns can be found in the header parts of a packet or in the payload. Snort's detection system is based on rules which in turn are based on intruder bit pattern. Snort rules can be used to check various parts of a data packet. Rules are applied in an orderly fashion to all packets depending on their types. The detection engine is programmed using a simple language that describes per packet tests and actions. Snort rules are written in an easy to understand syntax. Most of the rules are written in a single line but can be extended to multiple lines using backslash character at end of the line.

The rules are then used to generate an alert message, which is logged to the Snort MySQL (MySQL AB, 2005) database. The database can be simultaneously accessed by Snort Alert Monitor SAM (LookAndFeel, 2002) to produce a real-time output. The database grows quite rapidly if the rules are incorrectly configured or specifically in attempt to gather more packets for later analysis.

2.2 SNMP: Network Management Integration

The SnortSNMPplugin (Cyber Solutions 2005) which has been used here enables snort to send Simple Network Management Protocol (SNMP) alerts to Snort alert database. The alerts can be traps – information broadcasts which do not get any acknowledgements or informs where the alert will be acknowledged by the receiver. This adds significant power to the NMS by allowing it to monitor the security of the network. This makes it possible for the snort sensor to exploit the features that are built into existing network management systems.

An SNMP notification carries information in the form of a set of name-value pairs. The names are, Object instance Identifiers (OID). Managed Objects (MO) are observables that are used by NMSs. To report any network link or status update (e.g. sensor location, alert message, attack source etc). MOs are uniquely identified by their OIDs. The MOs and their OIDs are defined in Management Information Base (MIB) modules.

The OIDs are organised in the form of a tree - the "global naming tree". Each node in this tree has an identifier and a label. The identifier is unique among the siblings of a node. The concatenation of the identifiers of the nodes on the arch starting at the root and ending at a node is the OID of that node. The organisation that has been assigned a node in the "global naming tree" is in charge of the sub tree rooted at that node. Organisations in turn may delegate the administration of sub tree(s) of the tree in their charge.

Snort.org has been assigned the unique node numbered *10234* under the enterprises node of the global naming tree. The OID of this node is 1.3.6.4.1.10234, the

concatenation of the identifiers of the nodes on the arch starting from the root node to the node assigned to snort. (Going by the labels of the nodes the OID will look like iso.org.dod.internet.private.enterprises.snort). All Snort related MIBs can be defined under this node.

2.2.1 The MIB implementation

In the simple case we just want to send SNMP alerts to a NMS or a Network Security Manager. This is simple because snort does the detection and calls the SNMPplugin module to generate the corresponding SNMP alert packet with the appropriate OID-value pairs. The SNMP alert packet is then logged to the database.

2.2.2 The Actual communication

The actual communication between the snort and the NMS will use the Simple Network Management Protocol. Both the versions of SNMP are supported i.e. SNMPv2C and SNMPv3. In order to set up Snort for generating SNMP alerts it is required to set up the snort.conf with the appropriate parameters for SNMP alert generation. The SNMPTrapGenerator output plug-in requires several parameters. The parameters depend on the SNMP version that is used. More information on setting up Snort can be obtained from Cyber Solutions SNMP Snort Guide. (Keeni, 2005)

2.3 Wireless Sniffing

Another add-on is Snort-Wireless (Snort Wireless, 2005) which is an attempt to make a scalable 802.11 intrusion detection system that is easily integratable into an IDS infrastructure. It is completely backwards compatible with Snort 2.0.x and adds several additional features. Currently it allows for 802.11 specific detection rules through the new "wifi" rule protocol, as well as rogue AP, AdHoc network, and Netstumbler (NetStumbler.com, 2005) detection.

2.3.1 WiFi Protocol Rules

Snort at present does not contain direct support for rule based detection of anything below the IP layer. It is possible in Snort 2.0.x to match byte patterns in a packet, but it is not very straightforward and is very time-consuming to write detection rules this way. Rules for detecting particular 802.11 frames are specified using the following syntax:

<action> wifi <src mac> -> <dst mac> (<rule options>)

2.3.2 RogueAP Preprocessor

The RogueAP preprocessor detects both rogue APs and AdHoc networks. To configure it, for the APs BSSIDs and channels it has to be specified that they operate on in the snort.conf file using the ACCESS_POINTS and CHANNELS variables.

2.3.3 The AntiStumblerPreprocessor

NetStumbler is a tool for Windows that detects Wireless Local Area Networks (WLANs) using 802.11b, 802.11a and 802.11g. It is potentially used for war-diving. The AntiStumbler preprocessor attempts to detect Netstumbler like traffic. It does this by keeping track of probe request frames sent with NULL SSID fields.

2.4 Console

Data is useless without some manageable method for review. If we simply expect to be able to sit down and read each entry in a log file, we will be quickly overcome with pages and pages of alerts, warnings, and even regular user activity. With Netmates console, one can quickly and easily target the important information.

Snort Alert Monitor: Snort Alert Monitor (SAM) which has been modified and bug fixed for this project is a Java-based console that can be used to get a quick look at the Snort alerts in MySQL database. It runs as a Java-console, so it's platform independent. The frequency of the updates from snort's MySQL database can be tweaked to get a near real-time view of incoming Snort alerts. SAM is freely available under LAF General Public License. (lookandfeel, 2002)

The console has many ways of grabbing attention. The first is the rather large stop light in the top left corner of the screen. The second is by playing a specific sound when a particular threshold is reached. The third way we can be notified is that an email can be sent to a specific person or group of people.

Figure 2: Network Monitoring & Attack Evaluation System Console

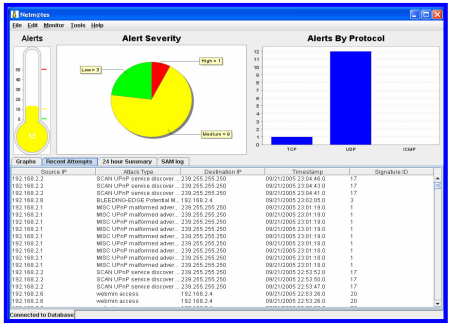


Figure 3: Snort Alert Monitor Recent Attack Attempts

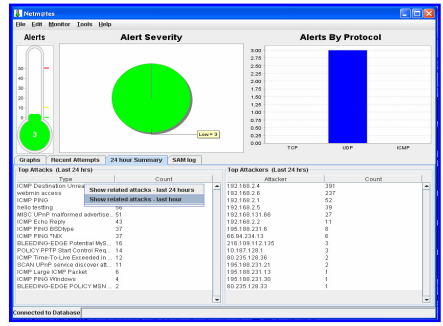


Figure 4: Snort Alert Monitor Summary

2.4.1 Alert Severity

Sometimes it's difficult to predict how serious the alert is. To help us determine the severity of the outage, the reasons administrator may be alerted have been divided into three basic categories: Green, Yellow, and Red alerts:

Green Alerts are **OK**. They are a result of a less than ten alerts in last five minutes. This threshold value can be altered by changing the value of alertlevel.medium=x to desired number of alerts per five minutes. Any value below this number will be considered to be of low sensitivity.

Yellow Alerts are **cause for concern**. Yellow alerts are characterised by ten or more but less than 50 alerts (these default thresh hold values can be changed). The idea is that when the number of alerts (in the past five minutes) is equal to or above this value, the alert level is then set to medium, and the traffic light will flash yellow.

Red Alerts are **serious outages**. This is a numerical field (alertlevel.high=x) that represents the threshold for the medium alert level. The idea is that when SAM receives more than fifty alerts in five minutes, the alert level is then set to high, and the traffic light will flash red.

SAM will warn the administrator if the connectivity to the snort database server is lost, but it can not tell why. Some research might be needed to determine the failure of the link or server.

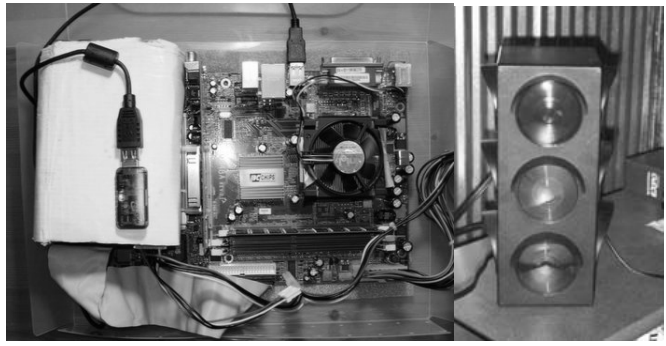


Figure 5: External Stoplights interfaced with Netmates

2.4.2 Stoplights

It is possible to connect the some kind of external visual alarm system in the form of spot lights. As seen in the Figure 5, a multi colored alarm can indicate the current status of the attacks, the color or level of the attack can be determined on the same principles discussed in section 2.4.1. This has to be done by using some kind of low level program. This feature has not been implemented yet, but there are many possibilities of further developing the alert and alarm system.

3. Conclusion

Netmates would be the most admirable tool for security professionals, powered by Snort which is an icon of intrusion detection software, proves how effectively it can implemented using its modular approach in which Snort applies rules and preprocessors, this program can be enhanced by anyone with even a basic understanding of security.

The uniqueness of Netmates is that it puts together a working model that features the functionality of Network Management and Intrusion Detection in a single box, complemented with a java based Console which is so robust and just the right desirable for network security personnel.

4. References

Acid Lab (2005), “Analysis Console for Intrusion Databases (ACID)”, http://acidlab.sourceforge.net, (Accessed 01-Aug-05)

BASE (2005), “Basic Analysis and Security Engine”, http://sourceforge.net/projects/secureideas, (Accessed 25-Aug-05)

Cyber Solutions (2005), “SnortSNMP”, http://www.cysols.com/contrib/snortSNMP/index.shtml, (Accessed 01-Aug-05)

Keeni G.M. (2005), Cyber Solutions, “Snort-SNMP Guide”, http://www.cysol.co.jp/contrib/snortsnmp/snortSnmpGuide.html, (Accessed 22-Aug-2005)

LookAndFeel (2005), “Snort Alert Monitor”, http://sourceforge.net/projects/snortalertmon, (Accessed 01-Aug-05)

Microsoft Corporation (2004), “Microsoft Virtual PC 2004”, http://www.microsoft.com/windows/virtualpc/default.mspx, (Accessed 23-Aug-2005)

MySQL AB (2005), “The World's Most Popular Open Source Database”, http://www.mysql.org, (Accessed 25-Aug-05)

NetStumbler.com (2005), “NetStumbler”, http://netstumbler.com, (Accessed 25-Aug-2005)

Snort (2005), “Snort - the de facto standard for intrusion detection/prevention”, http://www.snort.org, (Accessed 01-Aug-05)

Snort Wireless (2005), “Snort – Wireless”, http://snort-wireless.org, (Accessed 01-Aug-05)

Security and Risk Analysis of VoIP Networks

S.Feroz and P.S.Dowland

Network Research Group, University of Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

This paper address all major issues related to VoIP security, and provides detailed technical information about VoIP. The focus of this paper is to highlight, discuss and introduce security issues relating to VoIP networks given the expansion in the usage of VoIP within large corportations. This paper discusses current threats and future security measures related VoIP.

Keywords

Security measures, vulnerabilities, risks, solutions and best practices.

1. Introduction

Voice over Internet Protocol (VoIP) is developing telephony solution that brings voice and data traffic together on the same IP-based network. Telecommunication networks are now getting replaced with data communication networks and voice signals are now getting transferred over data networks by converting them into data packets. In VoIP calls are transmitted over an IP network instead of using PSTN. Because Internet network is getting widely available at various high bandwidth place of world VoIP that's why VoIP is becoming the best option.

The focus of this paper is to highlight, discuss and introduce security issue regarding VoIP networks, since VoIP is spreading rapidly and getting adopted by every other multinational and end user. There are rapidly increasing security threats taking place. This report discusses current threats and future security measures related VoIP.

2. VoIP Over view

Voice over Internet Protocol (VoIP) provides a communication between people and continuous access to networked services in such flexibility. A VoIP technology deals with the routing of voice and data between wired and wireless network. Many problems arise, such as poor service quality where data and voice packet shared the same bandwidth. The impact of security in the current environment of VoIP and the concerns related to its security QoS issues, protocol level security of several Threats as well as their impact on VoIP.

VoIP is a technology that is used to make telephone calls over the Internet using broadband connection using computer network instead of a regular phone

connection. VoIP converts the analog signals from the phone to digital signals so that the signals can travel over the Internet (Polaris, 2002).

Anyone can place a VoIP call by just picking up the phone and dialing the relevant number. The call from your local telephone provider is routed to your VoIP provider and through the internet the call goes to the other party's local telephone provider. In this way the VoIP connection is established from the calling person to the caller. Depending on your VoIP provider the call charges may be either flat or he may charge for local calls. Generally a flat per minute rate is charged from VoIP service providers. That means you can make local call, long distance calls or even international calls.

2.1 Advantages of VoIP

- If you are having a broadband internet connection then you do not need to maintain another line just for making phone calls. Since the same VoIP line can be used to dial any phone number. So, you can save a lot on your telephone bills.
- You can talk to as many people as you want at the same time without paying any extra charges and this facility is known as conferencing.

2.2 Disadvantages of VoIP

- If you think that you can replace your normal telephone with a VoIP connection then you need to be sure as many VoIP providers do not have back up power incase of power outages.
- Mostly through VoIP call do not connect to the emergency numbers.
- VoIP providers generally do not have directory assistance

3. Features of VoIP

3.1 VoIP – Cost effective

If we compare the regular calls charges of the PSTN i.e. Public Switched Telephone Networks especially for long distance calls then the VoIP calls are far cheap (Network world, 2005).

3.2 Quality of Voice in VoIP

VoIP is a very good alternative to PSTN in terms of bandwidth and better quality. But in practical scenarios it does not perform up to the level it guarantees. Since there is a single network maintained, organisations face a lot of data congestion issues. The voice signals need to be transmitted in real time but actually it does not happen and there are significant amount of delay in the packet delivery at the other end that results in voice breakage. Since, VoIP is an emerging technology; research is still going on to deliver better services to the consumers.

3.3 VoIP's legacy and privacy issues

Government rules of monitoring the lines incase of PSTN is absolutely different from the VoIP lines. Security of Call Detail Records (CDR) is one of the privacy issues which fall under the privacy Act of 1974 (Microsoft Tech Net, 2005). Many private VOIP service providers maintain CDR so that they can keep track of the billing, fraud, theft of the resources etc. So, the VoIP service providers keep these records for future purpose but the maintenance of CDR comes under the privacy and security issues of an individual.

3.4 VoIP's Vulnerability

Considering the fact that VoIP systems have some security concerns still the organisations deploy VoIP systems in order to get a better quality service at lower cost. The quality of service provided by VoIP systems is the most important factor in switching to a VoIP System. However, the organisations should realise that voice signals in the form of packets, traveling over the internet are highly susceptible to the amount of attacks as the core data networks. All the packets can be easily intercepted by any hacker and can be manipulated and re-routed. Denial of service and hijacking are major issues in VoIP networks. Even the operating systems are vulnerable as VoIP systems are installed on existing Operating systems and application having no or rather very less security protection.

VoIP requires some basic components and a signaling and transmission protocol for its deployment. The components include the customer premise equipment, Call processing and Management Application and Voice Handling Server.

3.5 Hardware and Software Requirement for VoIP System

In order to create a VoIP System you need to have a Computer with full duplex capable Sound card and a broadband internet connection. You also need the appropriate dialer software and headset with mike if you are dialing through keyboard. We need a duplex sound card else one cannot hear anything while speaking. This is the minimal requirement for a VoIP system but you need special cards with hardware accelerating capabilities like Quicknet and Voice Tronix (VoIP NEWS," Articles of VoIP). Operating systems like Windows or Linux are good enough for VoIP to take place.

Microsoft Windows NetMeeting provides some VoIP services and in Apple Macintosh they have something similar known as iChat. Even Linux has a lot of VoIP applications.

3.6 VoIP Communication

With VoIP communication coming into existence the internet technology has really changed. Now the voice packets are inserted into data packets using some real time protocol. Next thing is to use some signaling protocol to call the users. When the data packets have reached the destination then those packets have to be decompressed and the data needs to be extracted from the packets.

3.7 VoIP Components:

- Customer Premise Equipment
- Call Processing and Management Application
- Voice Handling Server

4. Security Measures to Threats

- Denial of Service (DOS)
- Toll Fraud
- Call Recording
- Eavesdropping
- Call Hijacking
- Message Integrity

This attack generally relates to IP issues that include VoIP, email, e-commerce and Domain Name service**.**

4.1 VoIP Security Issues

The popularity of VoIP increasing day by day the VoIP security issues are also increasing. Before VoIP came into existence, people were only considering the data security but now voice security is also important. Anyone can intercept the call and can easily gain access to that information if the voice packets are not encrypted.

4.1.1 Why security has been overlooked?

Currently there are not much cases heard about the breach in security of VoIP communication. Once people start thinking in terms of security automatically they would start investing in security infrastructure in order to protect their VoIP systems and VoIP network communication resources.

4.1.2 Security Challenges

Once, VoIP reaches to the masses, security will gain importance amongst VoIP service providers, with the happening of few incidents concerning security breach (Tyson and Valdes, 2004). If we talk about the organisation's usage of VoIP network, they have started feeling the lack in security infrastructure as the packets have to travel through an un-trusted medium known as internet. So, gradually there is a growing demand for security systems to protect VoIP network components from nasty attacks from any intruder in and outside their domain area.

4.1.3 Security and VoIP

The VoIP application running on the organisation data network. If an organisation is considering security planning for VoIP systems then they might consider the following:

Since we all know the potential of VoIP like lower costs and greater flexibility, we should be careful before deploying VoIP components into existing IP networks. In case the existing network is already congested and overburdened then the integration of a VoIP system would cause serious issues.

Generally people think that since the voice packets are also digitised they can be easily used over the existing data network architecture with similar security measures. But actually there is a lot more to VoIP security than the data security. NIST has also laid down some of the security guidelines for VoIP systems. (Rosen. B, 2005).

4.1.4 Is VoIP Scary?

VoIP without security is just like a person without mind. Mind controls the body and in VoIP security controls anyone entering your system. So, security is one of the preconditions for the deployment of VoIP system. Majority of the VoIP attacks are application based. Some of the indications of security issues in VoIP systems are dropped calls and hearing issues. Once the companies start broadcasting their SIP addresses in VoIP communications then VoIP security would be a major concern for most IT experts. According to Internet Security Systems (ISS), Cisco's VoIP is not designed with security in mind and have so many security flaws. An implementation flaw in Cisco's Call Manager that handles call routing and signaling, could allow an overflow in buffer that would grant an intruder to access the VoIP system and listen all calls routed through it (Internet Security Systems Inc. 2004"VoIP). ISS warns the companies using the new VoIP technology to take VoIP security seriously else they might loose enough money if some intruder steals some important information.

5. Quality of Service (QoS)

Quality of Service (QoS) refers to the quality of the usual or traditional telephone network compared to the quality of the voice in VoIP network. Although calls in VoIP systems are far cheaper than that of the usual PSTN telephone calls but still of VoIP cannot guarantee the equivalent quality of service which traditional PSTN offer then it is of no use. Any VoIP network should address these QoS issues before the VoIP system is actually deployed.

5.1 Different Protocols used in VoIP

- H.323 Standard
- H.323 Multipoint Control Units, Gateways, and Gatekeepers
- SIP

5.2 Benefits

H.323 products and services offer the following benefits to users:

- Since various companies have adopted H.323 as a standard for audio transmission over the internet. All products services developed by different manufacturers using the H.323 standard protocol can interoperate. H.323 conferencing clients, bridges, servers, and gateways support this interoperability.
- Different bit rates are used for formatting of the data with audio codecs that are provided by H.323. It is up to the users to choose the codec that is best supported by their computer and network selections.
- Audio-visual teleconferencing can be done with the support of T.120 with H.323

5.3 Components of SIP

- SIP Servers
- SIP User Agents

5.4 Best Practices for moving to secure VoIP

- **Network Architecture**: We need to have strong network architecture. By strong we mean that the architecture or the network design should be such in which we have separate networks for voice and data. Ensure that all the VoIP related communication is through some standard firewall.
- **Legal advisors**: You should regularly visit your legal advisor to verify about any possible new law or concern that the company needs to give some extra attention. You should be aware of any new law if passed that may affect your company at a later stage
- **Soft phones**: Always try to avoid using soft phones with headphones and special software's as the computers use data networks and that may interfere with the voice network.
- **Risk Analysis**: A proper risk analysis should be done before implementing VoIP in your company as to know the cons and the danger involved is equally important. Also, to know the cost of the implementation is essential.
- **Security Features**: There should be proper security environment to implement VoIP systems. Adequate security layers should be there in order to have a proper security enabled environment
- **Backup Power Supply**: There should be a very good backup power supply system for the office where VoIP systems is implemented. Even the backup power system should be provided for the individual instruments.
- **Physical Controls**: VoIP Networks should be encrypted and this is one of the most important characteristics of voice networks. The landlines can be easily tapped so there is no question of the VoIP networks being interpreted by intruders.
- **Emergency Services**: Dealing with emergency service is one of the major challenges in the implementation (E-911) of VoIP systems as all VoIP system will not be able to identify where the physical location of the office is and route the 911 calls to the right center
- **WiFi Security**: Now since the technology is changing we should consider the need of integration of VoIP systems with Mobiles, since to break the

security of WiFi systems is tougher than that of conventional landline phones. Security features of Wired Equivalent Privacy (WEP) offer very little or rather no protection as WEP security can be easily broken with some publicly available software's.

6. Conclusions

There are some specific additional security measures, mostly dealing with securing the signaling to set up VoIP sessions, but, in general, networks with good practices for IP security will have good VoIP security. Although solutions to some problems have o been proposed, designed and accepted; the research does not stop. This is due to the fact that technologies are an on-going subject evolves everyday. As this unique environment of VoIP develops and increase at rapid pace, new challenges and problems occurred. However, we must not ignore the security impact that relies on how we tackle the situation in handling the security issues. It is important to know the various threats in VoIP technology. Something that is will end up suffering due to poor security implementation. Usually some times is required for a new technology to gain adequate level of security. The awareness of risk factors described in this paper will help to prepare for VoIP and should help mitigate potential security breaches and raise internal security awareness within organisations to significantly reduce risks from unwarranted attacks in order to provide reliable operations and services in the VoIP. In order to meet user requirements and to satisfy user needs for reliable operations over VoIP, some sort of guidelines are needed. This paper has highlighted the challenges faced by user of VoIP environment and the various kinds of approaches that can be used to tackle those problems.

7. References

Internet Security System Inc. (2004), "VoIP: *The Evolving Solution and the Evolving Threat*", http://www.documents.iss.net/whitepapers/ISS_VoIP_White_paper.pdf (21/07/05)

Jupiter Web Network (2002), "Whitepaper: *Advantages of SIP for VoIP*", http://www.webpedia.com (02/09/05)

Network World (2005), "*VoIP security can not be ignored*" http://www.findarticles.com/p/articles/mi_qa3649/is_200508/ai_nl4879749 (25/07/2005)

Planchard, C. (2005), "*The Future of VoIP: Secure, Integrated Collaboration?*", http://www.tmcnet.com/usubmit/2005/Aug/1171625.htm (28/07/05)

Polaris (2002), "*A reference guide to all things VoIP*", http://www.voip-info.org, (12/06/05)

Rosen. B. (2005), "*VoIP and Frauds*", http://www.voipsa.org/pipermail/voipsec_voipsa.org/2005-February/000072.html (08/7/05)

Tyson and Valdes (2004), "*How VoIP Works*" http://www.computer.howstuffworks.com/ip-telephony.htm (10/07/05)

Attack Pattern Analysis: Trends in Malware Variant Development

U.A.Abu Bakar[1], S.M.Furnell[1], M.Papadaki[2] and G.Pinkney[2]

[1] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2] Symantec, Berkshire, United Kingdom
e-mail: info@network-research-group.org

Abstract

This paper presents an investigation into recent trends and patterns in malware variant development targeting the Microsoft Windows environment. This research focuses on three significant malware threats: Beagle, Netsky and Mytob; which were all successful mass mailing worms and unique in terms of their propagation techniques and functionality. The results from this investigation showed that mass mailing worms still prove to be the preferable propagation method, but other techniques are also required to ensure it becomes successfully widespread. Mass mailing worms also continues to prove successful in terms of their propagation speed and widespread distribution.

Keywords

Mass-Mailing Worm, Malware Analysis, Beagle, Netsky, Mytob

1. Introduction

Modern malware poses a major security threat to computer systems due to the speed it can spread over the Internet, exploiting the flaws and vulnerabilities in many systems. The threat is growing because malware is continually propagating in smaller time periods and malware routines are getting more complex. Having early detection and taking early prevention steps is better than having to clean up systems after they have been infected. One approach that may aid the process of detection and prevention is to better understand the malware attack patterns and trends. This may help malware researchers to expand the existing information on malware, and thus improved preventive actions could be taken.

The objective of the research presented in this paper is to determine if there are any trends and patterns in malware attacks from the perspective of variant development, focusing on the Microsoft Windows platform. Three malware examples are chosen for this purpose: Beagle, Netsky and Mytob; which have all been responsible for significant and widely spread malware incidents. To aid discovery of any trends or patterns that may exist in these malware attacks, this research proposes to answer the following investigation questions: 1) Is there a correlation between nature of malware development and time? 2) Is the creation of new malware based upon re-use of existing code and techniques, or is it a creation of entirely new techniques?

This paper is organised as follows: Section 2 presents the definitions of malware and overview of the current malware scene. Section 3 outlines data techniques and results from the analysis on malware variant development. In Section 4, the analysis results are used to derive the comparison and summary on discovery of trends in malware variants development. Finally, Section 5 presents the overall conclusions of this research.

2. Malware Overview

This section outlines some definitions in malware taxonomy related to this research and an overview of recent malware attack trends.

2.1 Definitions

Malware is short form for "malicious software", and refers to software that contains code which is typically designed to perform malicious activity or damage to single computer, networked computers or servers (Microsoft, 2004). Examples of malware are worm, virus, and Trojan horses. Malware is grouped according to its unique characteristics. However these days, it is rather difficult to provide perfect definition and categorisation of malware because some malware may show behaviour that fits into one or more category. The following description of malware is not exhaustive but they provide the basic definition (Szor, 2005; Microsoft, 2004):

Virus	A program that is written to enable replication of itself and it need host to infect as it does not function on its own. Each time the host program is executed; the virus is executed as well and reproduces itself by attaching to other programs.
Worm	A stand alone program that can propagate, replicate and distribute itself on networks with or without user intervention.
Trojan horses	A Trojan horse disguises itself as a program with some useful functions but also contains hidden code that, when executed, perform some malicious function.
Bots	An agent designed to seek and infect vulnerable machines, run silently in the background to open network ports and allow outbound connection, typically to IRC channel which is remotely controlled by attacker (SANS, 2003).

2.2 Current Malware Scene

The scope for malware analysis presented in this paper is on malware that was recent at the time of the investigation. Thus, this section presents the discussion on some notable trends in current malware attacks scene from the period of Q1 2003 until Q1 2005.

i. *Mass mailing worms continue to be dominant and prevailing threat.* E-mail continues to be the most effective vector for malware propagation as many successful worm outbreak uses e-mail as their propagation vector. Since the first major outbreak of a mass mailing worm, Melissa in 1999, this technique has been employed widely by other malware. 2003 saw mass mailing worms like Sobig,

Klez and Bugbear dominating the top of the malware charts and their propagation techniques were basically derived from previous mass mailing worms. The trend continued in 2004, where mass mailing worms like Netsky, Beagle and MyDoom were discovered and they all dominate AV (Anti-Virus) vendor's malware charts until the end of the study period. These worms are also derivative upon previous mass mailing worms (Symantec, 2004b; Trend Micro, 2005).

ii. *Increase threats to confidential information.* Recent trends showed that the threat to confidential data has increased significantly each year (Symantec, 2003; Symantec, 2004a; Symantec, 2004b; Symantec, 2005a). Malware are created with malicious intentions of stealing confidential information, such as financial information, passwords and login cache from the compromised machines. This trend also shows companionship between malware to serve this purpose. For example, Mytob utilises Mydoom as its main distribution technique, and uses SDbot to serve the purpose of stealing information.

iii. *Bots and its numerous variants.* The current breed of bot worms showed an enormous amount of variants and the number is still continuing to rapidly increase. During 2004 alone, Trend Micro documented 2,830 bot programs (Trend Micro, 2004).

iv. *Increment and commonness of blended threats*. Blended threats use multiple combinations of malicious codes and exploit multiple vulnerabilities (Szor, 2002). The numbers of malware that employ blended threats continues to increase and they also demonstrate more complexity in routines and are also targeting more new vulnerabilities. Recent malware outbreaks employing blended attacks have caused major damage and are always rated with high severity, with examples including Netsky and Mydoom (Symantec, 2004a; Symantec, 2004b; Symantec, 2005a).

2.3 Recent Successful Malware

The trends in recent malware scene have drawn the attention to three particular cases: Beagle, Netsky and Mytob. The following will discuss the background of these examples.

i. *Beagle* – Beagle or Bagle was first reported by Symantec on 18 January 2004. As a mass mailing worm, it propagates using its own SMTP engine, scans for e-mail addresses in local drives with various file extensions and uses spoofed sender addresses. Furthermore, Beagle employs social engineering techniques by using a pre-configured list of e-mail subject and message body, and spread itself in the e-mail attachment with various extensions such as .exe, .scr or .zip. It also propagates via peer-to-peer (P2P) and shared folders. Later variants also utilise its own DNS server if MX record is unavailable on local drives. E-mails sent by Beagle may clog communications and degrade network performance. Beagle also installs a backdoor and opens network ports to allow remote command execution, and subsequently attempts to connect to lists of websites via HTTP GET request and relay back information about the infected machine to its attacker server. Beagle also attempts to terminate AV and security related

services. Beagle also created mutexes which may be used by Netsky in attempt to prevent Netsky from executing in Beagle's infected machines. Later variants saw Beagle start to send copies of Trojan Tooso via its mass mailing capability. This Trojan attempts to disable security-related services, as well as degrade computer performance when it performs remote file download from pre-configured list of websites (Symantec, 2004b; Symantec, 2005a; Symantec, 2005b;Trend Micro, 2004; Trend Micro, 2005).

ii. *Netsky* – First reported on 16 February 2004 by Symantec, Netsky is a mass mailing worm that propagates via e-mail, LAN and P2P networks. Like Beagle, it also adds a value to a Registry key so that the worm runs automatically during Windows startup. Netsky spreads inside e-mail attachments and employs social engineering techniques to trick users into opening the e-mail and attachment. E-mails sent by Netsky contain a spoofed sender address with variable names for e-mail subject and message body. Attachments that contain the body worm are usually in .zip, .pif, .scr or .exe file. The file that contained the worm is usually compressed with various types of packers such as UPX, Petite, FSG and so on. Netsky also propagates via a known vulnerability exploit. After July 2005, no new version of Netsky has been discovered, but the infection from old variants from 2004 was still widespread (Symantec, 2004b; Symantec, 2005a; Symantec, 2005b).

iii. *Mytob* – Discovered by Symantec on 28 February 2005, Mytob demonstrate a new wave in worm creation where it combines the code and functionality of an early version of the mass mailing worm Mydoom and the IRC bot SDbot. As part of Mydoom, Mytob arrived as an e-mail that has a pre-configured list of subjects, message body and attachment with file extensions of .pif, .zip, .exe, .scr, .cmd or .bat. Mytob e-mailed itself by gathering addresses from local drives with various file extensions including .htm, .wab, .asp, and others. It also propagates via P2P networks and via two known Microsoft vulnerability exploits. As part of SDbot, it has backdoor capability to open ports and connect to pre-configured list of IRC channel in order to listen for remote command from the attacker which then performs action like update itself or stealing information from the compromised machines. Mytob became widespread within a short period of time, and with its ability to compromise machines and steal confidential information, Mytob was rated as high severity (Symantec, 2005a; Symantec, 2005b; Trend Micro, 2004; Trend Micro, 2005).

3. An Analysis of Malware Variants

Having identified the three worms as notable examples of recent developments, this section discusses the techniques used for data collection and analysis within the study, and will then examines the three worms in more detail and considers the characteristics that helped to make them successful.

3.1 Data Collection and Analysis Techniques

The analysis performed in this paper is done upon set of data that was collected from Chronological Virus List on www.secunia.com (Secunia, 2005). This source was selected because it provided comprehensive reports on new malware discovery everyday and organised it chronologically. The malware reports were basically derived from seven well-recognised AV vendors. All data in Secunia was grouped and indexed, and it contains information on the date a threat was reported and updated, its aliases, file sizes, severity ratings, description and links to its reporting vendor. The period for data collection was from 1 March 2004 until 1 August 2005. To verify the data collected, a number of random entries from the collected data are picked and information on the malware is checked with the original AV vendor website.

The data collection was performed by utilising a Bash script which automates the collection process. The script functioned by downloading the relevant page from Secunia and then stripping out the HTML and finally converting the output into a CSV file for easy manipulation. This data was then organised and split into multiple separate CSV files according to type of malware and its reporting vendor to improve the process of identifying trends. From the sorted data, type of information that could be extracted was the discovery date of new malware and its variants. From here, the selected data on Beagle, Netsky and Mytob was plotted into cumulative distribution plots and the distributional information is analysed. The analysis was performed using qualitative analysis where information and description of the worms' variants were gathered and compared to find trends.

3.2 Summary of Analysis Results

The results of analysis for each worm variants development are summarised in the following sections.

3.2.1 Beagle

The nature of Beagle development is shown in Figure 1, where the first few releases were in parallel with Netsky. The competitive growth was resulted from the effect of Netsky attempt to delete Registry keys used by Beagle and Beagle keep releasing more variants to extend distribution.

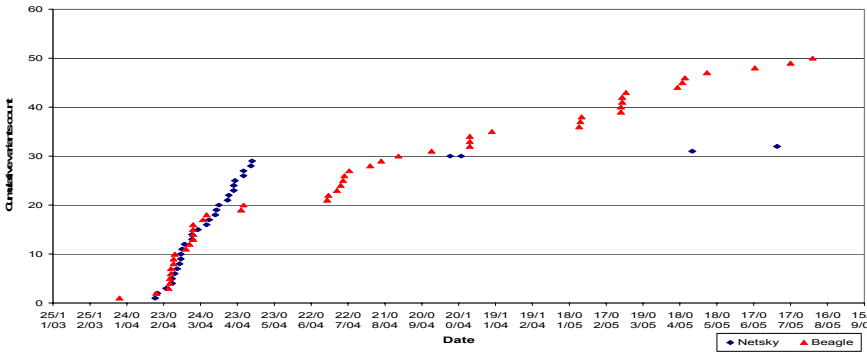


Figure 1: Development of Beagle variants versus Netsky variants

From the initial release, the trend sees clusters of variants released with rapid succession and the overall trend appears roughly linear throughout. Analysis on each variant clustering has shown correlation with its functionality, and each clustering shows improvement over time. The first cluster sees various techniques being tested, where it starts to include its own DNS server, terminate security-related services, create random ID for the infected machines, delete Netsky registry key and drop Trojan Mitglieder. The second clustering sees a similar technique, but with improvement in terms of added pre-configured list of e-mail subject and message body, and the latest clustering see that Beagle showing "new faces" in its pre-configured list of e-mail subject and bodies and it also sends out copy of Trojan Tooso via its mass mailing technique.

3.2.2 Netsky

Figure 1 shows the speedy release of Netsky in four months since its first outbreak. Note that the development was in parallel with Beagle, and the Figure is indicative of the competition between them, which explains why the curve of the both Beagle and Netsky early variants release was steep. Netsky's author was arrested in May 2004 and four variants discovered after that only reflect modification and re-use of the worm's code by somebody else. As a mass mailing worm that did not produce any more new iterations, Netsky proved to be very successful based upon its prevalence in malware charts. Figure 2 shows the number of infected machines for each Netsky variant since their first release and illustrate Netsky.P as the most successful variant of the family. The reason for this is because it employs powerful social engineering, built in SMTP engine, redundancy technique to retrieve SMTP server, exploit IE Incorrect MIME Header vulnerability and spread to P2P, LAN, FTP and HTTP server's folders, all in one variant.

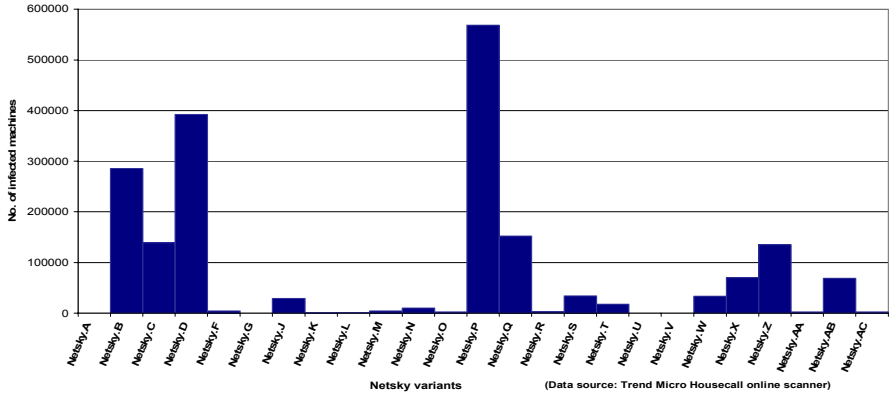


Figure 2: Number of infected machines by Netsky family

3.2.3 Mytob

Figure 3 shows the rapid development of Mytob in less than 6 months from its first outbreak.

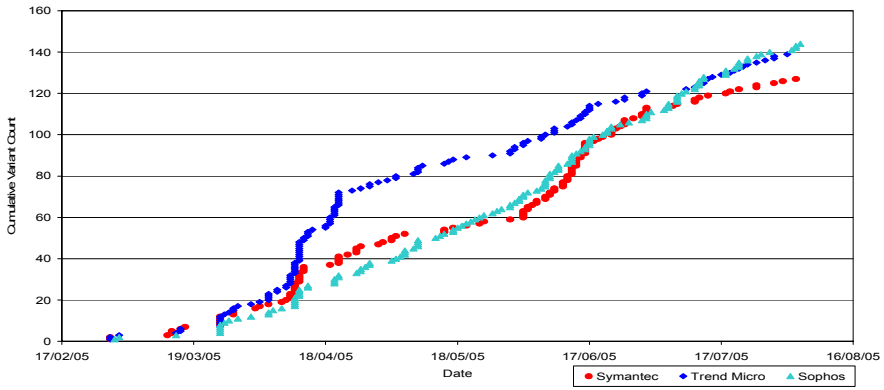


Figure 3: Development of Mytob variants According to Symantec, Trend Micro and Sophos

Since the first outbreak until 3 August 2005, Symantec has discovered 127 variants. The similar trend was also discovered by Trend Micro and Sophos. The rapid succession of variant release reflects that each new version does not show much adjustment from the previous one, but just enough to need new signatures to detect them. This also reflects the bot part in Mytob, which is to steal and relay information from the compromised machine back to its remote attacker and therefore it need to race against AV new signature deployment to ensure it compromised as many machines to make sure the number of compromised machines in the bot network is maintained. The success of this hybrid worm is mainly because of the technique it used, which is mass mailing and vulnerability exploit from Mydoom, and bot functionality from SDbot.

4. Discussion

Having analysed the three worms, it can be seen that the speedy development and the success of these worms proved that mass mailing is still the preferable technique for propagation. However, it is worth noting that using only the mass mailing technique is not adequate, but must be employed with other techniques as well to ensure the widespread of these worms. Success of these worms do not depend on the complexity of codes or technology, but mainly upon the right combination of techniques used to make the worms easily propagate, evade AV software detection and trick users into trusting the incoming e-mail, and open the infectious attachment. Table 1 summarises the characteristics that contribute to these worms' success.

		Beagle	Netsky	Mytob
Propagation technique	Mass mailing	✓	✓	✓
	P2P and shared networks	✓	✓	✓
	Vulnerability exploit		✓	✓
E-mail technique	Built-in SMTP engine	✓	✓	✓
	Social engineering	✓	✓	✓
	Password protected attachment	✓	✓	
Payload	Open backdoor	✓	✓	✓
	Attempt to terminate security-related services	✓	✓	✓
	Perform DoS/ DDoS	✓	✓	✓
	Steal information	✓		✓
	Dropped other malware	✓		✓
Vulnerabilities exploited	DCOM RPC Buffer Overflow			✓
	LSASS Buffer Overflow			✓
	IE Incorrect MIME Header		✓	
	IE XML Page Object Type Validation		✓	

Table 1: Main characteristics of Netsky, Beagle and Mytob

5. Conclusion

Although the result from this research does not provide exhaustive description about Beagle, Netsky and Mytob, it does serve to outline the general trend discovered based upon the analysis of worm variant development. This research also shows how

the functionality of the malware improved and evolved over time. The result derived from this research can be used as stepping stone for continuing this literature and also on other malware analysis. The findings will be useful to predict the evolution of new malware development behaviour. Therefore, if the new malware can be predicted, preventive action could be taken earlier and it the prediction were match, the damage which may caused by this new malware will be minimal as early preventive has been taken. In addition, the findings can be used to aid the quantitative part of malware analysis.

6. References

Microsoft (2004), "The Antivirus Defense-in-Depth Guide", http://www.microsoft.com/technet/security/topics/serversecurity/avdind_2.mspx, (Accessed: 1 September 2005)

SANS (2003), "Bots & Botnet: An Overview", http://www.sans.org/rr/whitepapers/malicious/1299.php, (Accessed: 1 September 2005)

Secunia (2005), "Chronological Virus List", http://secunia.com/chronological_virus_list/, (Accessed: 5 September 2005)

Symantec (2003), "Symantec Internet Threat Report Volume IV", http://enterprisesecurity.symantec.com/content.cfm?articleid=1539, (Accessed: 1 September 2005)

Symantec (2004a), "Symantec Internet Threat Report Volume V", http://enterprisesecurity.symantec.com/content.cfm?articleid=1539, (Accessed: 1 September 2005)

Symantec (2004b), "Symantec Internet Security Threat Report Volume VI", http://enterprisesecurity.symantec.com/content.cfm?articleid=1539, (Accessed: 1 September 2005)

Symantec (2005a), "Symantec Internet Security Threat Report Volume VII", http://enterprisesecurity.symantec.com/content.cfm?articleid=1539, (Accessed: 1 September 2005)

Symantec (2005b), "Search and Latest Virus Threats", http://securityresponse.symantec.com/avcenter/vinfodb.html, (Accessed: 4 September 2005)

Szor, P. (2002), "Blended Attack Exploits, Vulnerabilities and Buffer-Overflow Technique in Computer Viruses", http://peterszor.com/blended.pdf, (Accessed: 3 September 2005)

Szor, P. (2005), *The Art of Computer Virus Research and Defense*, Addison-Wesley, United States, ISBN: 0-321-39454-3.

Trend Micro (2004), "The Trend of Malware Today: Annual Virus Round-up and 2005 Forecast", http://www.trendmicro.com/en/security/white-papers/overview.htm#annualroundup2004, (Accessed: 1 September 2005)

Trend Micro (2005), “Outbreak Incidence and Prevailing Malware Trends: Q1 2005 Virus Roundup”, http://www.trendmicro.com/en/security/white-papers/overview.htm#q12005virusroundup, (Accessed: 1 September 2005)

Mobile Devices - Future Security Threats & Vulnerabilities

V.Sklikas and N.L.Clarke

Network Research Group, University of Plymouth, Plymouth, United Kingdom.
e-mail: info@network-research-group.org

Abstract

The success of the Internet technologies made telephony companies realise the advantages of the adoption of IP technologies over circuit switched networks. Now both the telecommunication and Internet technologies converge and integrate for the creation of an 'all-IP' wireless infrastructure, able to support mobile data and multimedia applications, resulting in making available all known Internet services to the forthcoming wireless networks. However, beyond the numerous benefits that arise, factors such as mobility, the compact size of the various mobile devices, the ease of their connectivity and their open nature increase the threats and the risks being posed, rendering the future wireless security an increasing problem. This paper reviews the threats introduced by traditional networking technologies and examines the way in which they could be adopted by the wireless technology, investigating possible threat scenarios and taking into consideration future technology capabilities.

Mobile technologies are a target for many threats that exist in traditional wired networks, in addition to many wireless specific threats. The underlying communications medium is open to intruders and is easier to eavesdrop as no physical access is required. Unauthorised access to a network through wireless connections, by bypassing any firewall protection, interception of unencrypted information, and the tracking of the mobile users are some of the most critical threats concerning the owners and the users of mobile networks and devices.

This paper introduces a number of network topologies and discusses the relatively advantages and disadvantages of implementing each. Generally security of mobile devices can be viewed from a central server or network centric perspective, managed by trusted third party authorities, offering security at the network and covering all security tasks needed on the devices with no end-user interaction.

Keywords

Mobility, security, wireless

1. Introduction

The two major fields that target in their convergence that will lead to a unified wireless IP infrastructure, are the cellular telephony and the Internet. Wired networks evolve into wireless and wireless networks evolve into wireless Internet Protocol (IP) networks, as the latter is a more suitable approach for supporting the forthcoming mobile data and multimedia applications. The workplace is being decentralised and the level of electronic mobility able to overcome the limits of traditional wired

devices, is given by new technologies for communicating, entertaining, and accessing information that introduce with incredible pace. Portability, flexibility and productivity increase, while the numerous mobile devices introducing, allow data synchronisation with network systems and application sharing between the devices. Remote users are allowed to synchronise personal databases and are provided access to network services such as wireless e-mail, Web browsing and Internet access, though, the receiving, modification and transferring of information over networks, while roaming, becomes feasible. Public access points are growing daily in number to serve mobile users and provide them with connectivity. The prerequisite to this scheme, so that Internet services become available to wireless networks, is the integration of the IP technologies into mobile devices.

On the other hand, until now we used to be concerned only about security associated to wired local networks especially due to the highly vulnerable IP protocol. Unfortunately the aforementioned convergence brings IP technologies and mobile devices together, setting new security problems. In more detail, wireless IP networks operate over the IP, resulting in them inheriting all known vulnerabilities of the Internet Protocol. Lack of strong network security provision and a number of flawed services are some of the IP's weaknesses that make it vulnerable enough to exploitation and specifically to misuse and spoofing. Additionally, the numerous mobile devices required and used to bring the wireless services to the end-user, have many critical weaknesses, making them vulnerable to security threats. Mobile communication and wireless technologies became a target for already existing weaknesses within fixed networks. However, the mobility offered through these technologies, the variety, the minimal physical size and the ease of mobile devices' connectivity make security risk levels to rise. Moreover, the open nature of the wireless technologies due to the airborne waves they use for data transport and some different protocol deficiencies than those introduced in wired networks, make some security aspects to occur in a different form than they occur in fixed environments. Current threats are able to compromise wireless vulnerabilities, making the future unsure for everything and everyone. But in which manner could wireless technologies be compromised? Which and how already existing vulnerabilities could be exploited in order for someone to hijack telecommunications? Which are the techniques already followed in wired networks and now adopted, modified and applied to mobile devices, for someone to serve his purpose? If the future is going to be so insecure, is there any security mechanism that could counteract the forthcoming threats and risks?

2. Possible Future Threats and Vulnerabilities

The numerous mobile devices, the enhanced wireless capabilities, the increased computational power, the integrated IP technologies and a majority of security unaware consumers that will be using these devices make a very risky combination for the advanced IP services that will be offered widely through the Internet. But how this combination renders the future mobile domain insecure?

2.1 How could Mobile Devices be compromised

The increased use of wireless devices, Personal Devices, PDAs and smart phones running an operating system in combination with the 3G technology that brings Internet services on these devices, increase the potential for attacks. Their open operating systems and the lack of antivirus and detection tools, make the devices capable of being infected by any kind of malicious code. Their data storage, their transfer capabilities and their support for executable files, add to the problem and will make them even more susceptible to worms, viruses and spammers, while they introduce new exposures for hackers and crackers to target. The numerous wireless access technologies they support open a new avenue, due to their open nature, and increase the potential for a number of attacks. Moreover, their compact size encourages them being stolen, while the lack of authentication mechanisms facilitates the unauthorised access to the devices.

Furthermore, mobile devices integrate office functionality making it more possible for Malware to widespread by exploiting their wireless ports. In addition, they are very often strictly associated with critical data and applications, but as mentioned before, not accompanied by any security facilities or data integrity mechanisms. The ultimate goal, though, will involve the collection, alteration or loss of financial and other confidential data. The devices' direct ties to systems that deal with purchases and other transactions (LLet and Hines, 2004), and the fact that the number of users that engage wirelessly in online banking increases, will make mobile devices a tempting financial offer to exploit.

Moreover, 3G technologies and the increased number of the multimedia applications, games and screensavers available for downloading, combined with the fact that providers usually allow all kinds of content to be sent to the handsets, will pose hidden risks. Malware will be able to spread through all these facilities infecting both network services, such as Short Messaging Service (SMS) or Multimedia Messaging Service (MMS), and mobile devices. Infected devices will make their numerous wireless access technologies available for Malware to distribute. In addition, the increased access technologies that will be supported within Wi-Fi networks, including wireless networking cards, wireless access points, and Bluetooth will increase the potential for attacks. The combination of the access technologies used by the devices and the enhanced Wi-Fi networks open handhelds up to a variety of attack scenarios.

Another issue concerns the variety of the mobile operating systems that ensures that the widespread Malware is minimised. The lack of a dominant operating system requires Malware to be specifically written for each individually. The most popular operating system to date, is Symbian, hence most of the attacks have been focused upon Symbian PDAs. But how long will this be for, until mobile devices adopt one common operating system?

Finally, technology improves and the devices are equipped with enhanced and accurate built-in capabilities, like cameras and microphones. This makes anything supposed to be safe and private within the physical perimeters of a user's location, office or room susceptible to eavesdropping. In addition, mobile IP and the numerous

location-aware services used by mobile devices, like GPS that has already started to be used, could provide unauthorised parties with information, which reveal end-users' location, resulting in the violation of end-user privacy.

2.2 Mobile Devices – An Attacking Tool

Mobile devices will not always be the victim; it will equally be the attacker. Intranets are expanding beyond the traditional enterprises' limits out into the Internet, through mobile and wireless access technologies increasing not only external but also internal threats (Greenfield, 2002). Threats within wireless networks could be born and occur in any form, with passive or active attacks, malicious codes or software tools that assist an intruder using a mobile device to compromise any security weaknesses. Mobile Malware will be able to cause widespread damage. Such devices will be possibly used within enterprises without being noticed, enabling someone to gain unauthenticated access in the network. Attached to network connections, they could constitute a backdoor threat from the outside world. Operating as hidden 'Zombies' (Gilbert, 2005), they could launch attacks and distribute Malware to other PCs or networks, aiming to the alteration or disclosure of any sensitive data or the affection of the network's reliability and availability through Denial of Service (DoS) attacks.

Furthermore, future handhelds with enhanced communication capabilities, processing power and executable files support, will open a new avenue for easily compromising various Internet services that are considered to be secure enough. Therefore, Voice over IP (VoIP) sessions will be spoofed, eavesdropped, or even disrupted. Voice packets will be possible to be recorded, or IP phones could become unstable and finally unavailable. Having physical access to the main servers will give the opportunity to disrupt any services integrated with VoIP applications, like unified messaging. In the same way, using a mobile device within the network to compromise and launch DoS attacks to the VoIP gatekeeper, could lead to limited bandwidth or no service availability. Every component of the VoIP network infrastructure becomes susceptible to distribution of viruses, DoS attacks and eavesdropping, since handhelds will be almost everywhere and without control within a company or a core network. Unfortunately, VoIP will become feasible to be compromised outside an intranet too. Hot Spots are increasing in number and such public access points are preferred by a number of uneducated end-users, resulting in increasing any potential of attacks, making such places risky too. Public access points will be using the Mobile IP, proved to be vulnerable, increasing, though, the possibility for VoIP ongoing sessions to be monitored and disrupted, while providing the adroit user with free voice services and charging the unaware consumer.

The conclusion is that mobile devices are going to become an increasing problem for everyday life, affecting it in all sections; financial attacks, thefts, and users' privacy. Thus, there is no need for more evidence that enhanced security mechanisms have to be proposed, designed and deployed as a counteract to the forthcoming problems.

3. Future Security Solutions

The need for mobile security mechanisms and improved device management is undeniable. A complete security solution should look for weaknesses, focusing independently on protocols, mobile technologies, software and the end user security knowledge, while it needs to commensurate with the threats posed against it and the cost of securing it. The ultimate security solution should rely on a centric framework that improves the network and the terminal security management.

There is a range of mechanisms that could enhance security on mobile devices. An appropriate BIOS configuration could constitute the primary level of security, involving authentication prior to system boot and locking mechanisms for preventing any configuration alteration. Enhanced solutions include third party audit tools installed on the devices that provide with log mechanisms able to monitor actions taken on them, while antivirus, firewall and detection tools increase security. In addition, technologies are improving in terms of reliability and accuracy and biometric devices (Silicon Trust website) will eventually become the preferred choice for user authentication, preventing identity fraud. Moreover, the numerous services that will be offered should be able to detect any kind of misuse by illegitimate users, making end-users feel confident about the services they use. Though, third party services that digitally certify the authenticity and integrity of an application (Meserve, 2005), could be considered an appropriate mechanism. End-to-end encryption methods imposed to all the network's nodes that the devices might connect to, advances the protection level. IPv6 and IPsec (Ford, 2005) that is contained in the new version of IP, is more viable than Secure Sockets Layer (SSL) when it comes to traffic from a large number of applications. Additionally, IPv6 would enhance security in the public access networks. Finally, the implementation of encryption mechanisms makes VoIP and Mobile IP related services more secure both within enterprises and public access points, eliminating the possibility of eavesdropping. Finally, the large number of mobile devices on the market, and the uneducated consumers who have little to no knowledge of the security threats that are posed, makes it an imperative need that enterprises and the market precipitate into technical education solutions that incorporate a level of education and security knowledge of the end users. Specific policies should be imposed, within enterprises and public access points, on the way mobile devices are used.

Unfortunately, common security solutions used to date are possibly insufficient for future security demands; on the contrary the above review and propositions, show that some advanced security is required; how this could be achieved or possibly approached, is from a centric perspective. The existences of centric frameworks introduce an advanced level of protection by using network resources.

One approach could be a network centric model that uses network resources in protecting the devices. A network centric model, suggests the provision of security solutions at the network and not at the client device side; the future Internet will be secure as long as it will rely on third parties that provide with the necessary services to filter and monitor the content that flows inside the core network, preventing from Malware's spread at the client devices. Third parties could use the appropriate

antivirus, firewall and detection tools to provide a gateway level protection, in order to successfully block any malicious or unsolicited code in the network. Centric solutions do not require any end-user interaction, facilitating amateur consumers.

The central server model implies a centralised management of security on the device. It could refer to a model that involves a third party agency or authority, equipped with a range of central servers, providing the corresponding security services, like password security, biometric information security, and the provision of sensitive data storage behind a firewall. A central management console interacting with the devices, could provide terminal monitoring and its security configuration, while patches, updates and any kind of fixes concerning the various operating systems that mobile devices use, are downloaded automatically. All security tasks taking place on the devices are time consuming and complex, discouraging amateur users who do not have the appropriate knowledge to engage in such processes; thus, the ultimate aim of a central server framework is to impose security services at a terminal level, with no end user interaction.

However, a number of disadvantages arise within each of these centric models. The nature of a network centric model renders it weak when it comes to any kind of failure (Karygiannis, 1998). The system becomes unable to monitor any component of the network and all the nodes being protected before, become available and easy to penetrate. On the other hand, although a central server framework is fail-safe, it is vulnerable to security weaknesses since an intruder can disable local security mechanisms on the terminal that report back to the central system; thus a device cannot always be trusted to diagnose itself with a host-based monitor alone (Karygiannis, 1998). Both the network centric model's and the central server model's weaknesses imply that there is a need for a hybrid model that inherits and combines most of their advantages eliminating any deficiencies. A hybrid model is illustrated in Figure 1.

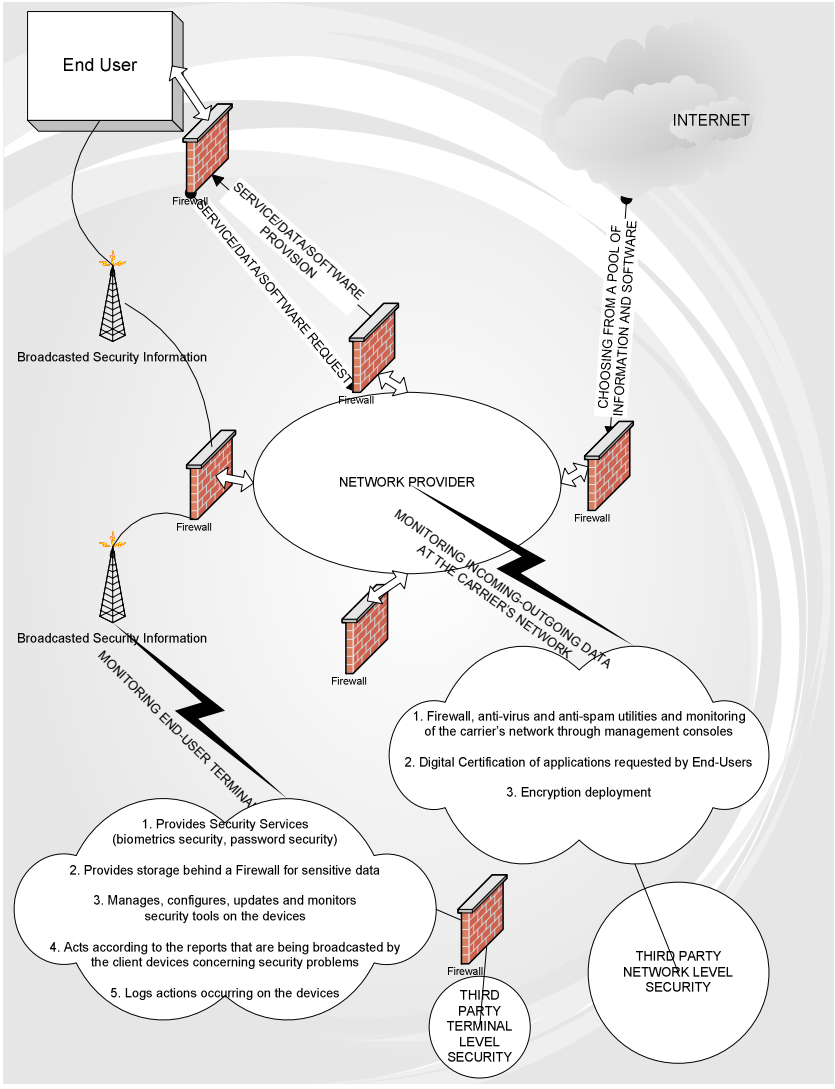


Figure 1. Hybrid Model

4. Conclusion and Future Work

The success of the Internet technologies made telephony companies realise the advantages of the adoption of IP technologies over circuit switched networks. Now both the telecommunication and Internet technologies converge and integrate for the creation of an 'all-IP' wireless infrastructure. Mobile devices introduce everyday, equipped with utilities and facilities to exploit the forthcoming networks. All known Internet services and multimedia applications will be delivered to mobile devices since future wireless networks are designed to support IP technologies. Unfortunately, an IP infrastructure inherits all known vulnerabilities occurring within

the traditional networks that operate over IP, in addition to factors like mobility, the variety of the devices that are used and their compact physical size. All these factors make it easy for current threats in traditional networks to be adopted and be applied to wireless technologies, in different forms due to their open nature. Mobile Malware will become an increasing problem for mobile devices, compromising network services to propagate. Attacks will be launched with the goal of financial gain and sensitive data disclosure.

The need for mobile security mechanisms and improved device management is undeniable though. A possible future security infrastructure should call for the following:

1. An end-to-end solution for all mobile devices within a carrier's network

    - Network centric security services maintained by trusted third parties;
      A gateway level security solution in the network is required to be able to
      flexibly filter the traffic. Intelligence lying on the network will always be
      better than any solution at device level.
    - Central server security approaches also maintained by trusted third parties

2. Mass coverage over all devices and geographic regions
3. Transparent security that requires no end-user involvement
4. Cooperation between network carriers, device manufacturers and technology providers, such as software companies that write applications for smart phones.

4.1 Future Work

Some of the issues that a future research could focus on, could be associated with the enhancement of the security mechanisms required to counteract the forthcoming threats, proposing the technologies and specifications that should be used.

Networks that consider the different access technologies including their advantages and disadvantages could be researched, combined with a study of the types of network infrastructures, able to protect mobile devices that might be applicable. A detailed report on suggested systems' configuration followed by network diagrams and simulations would be very useful for network administrators who focus on blocking the threats at the network rather than at the devices.

The detailed actions of trusted third parties, bearing in mind the nature of mobile devices and their technical requirements such as processing and memory needs should accompany a research on possible required equipment. Finally, the major topic, though, that should accompany the aforementioned possible future research topics, could negotiate with an investigation that defines the best choice and combination of future security frameworks and the proposed security mechanisms,

the equipment and technical specifications; it should best reflect in the most profitable way both for the authorities and the end users, financial issues and the quality of services that will be provided.

5. References

Ford, M. (2005), "Security and IPv6", http://www.ipv6.bt.com/tutorials/security.html, (Accessed July 2005)

Gilbert, A. (2005), "Botnets and Spyware still on the rise", http://news.zdnet.co.uk/internet/security/0,39020375,39208661,00.htm, (Accessed July 2005)

Greenfield, D. (2002), "New Public Network: Crystal Ball Gazers", http://www.networkmagazine.com/shared/printableArticle.jhtml?articleID=8703365, (Accessed 2 September 2005)

Karygiannis, T. (1998), "Network Security Testing Using Mobile Agents", http://csrc.nist.gov/mobilesecurity/Publications/Agents_PAAM98.pdf

LLet, D. and Hines, M. (2004), "Skulls program carries Cabir worm into phones", http://news.com.com/2100-7349_3-5469691.html, (Accessed July 2005)

Meserve, J. (2005), "Is your cell phone at risk?", http://www.networkworld.com/research/2005/041805-mobile-virus.html, (Accessed July 2005)

"HP Wireless Security", http://h200007.www2.hp.com/bc/docs/support/SupportManual/C00290881/C00290881.pdf, (Accessed August 2005)

"PKI: Future Trends", Silicon Trust website, http://www.silicon-trust.com/background/sp_pki_future_trends.asp, (Accessed August 2005)

Neural-based TCP performance modelling

X.D.Xue and B.V.Ghita

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Web users are expecting shorter response time when they are using Internet. However, Internet traffic is continued growth by adding more services and functions, the traffic will cause congestion problems and delays on Internet.

In this paper, we introduce the TCP protocol theory, the different between TCP long-lived and short-lived connection. And the neural network structure, algorithm and the parameters. Also described the current state of TCP short-lived connection traffic analysis and performance modelling. We explored and compared existing models for TCP long-lived and short-lived connection data transfer latency, the advantages and disadvantages are discussed.

This paper proposes to use mathematical model and neural network to predict TCP transfer latency for short-lived TCP connection for non-packet loss and packet loss situations, the results are compared by using the relative error and the overall comparison.

Keywords

Networks, TCP, Mathematical model, Neural network, Project, Paper

1. Introduction

Web users are expecting shorter response time when they are using Internet. However, Internet traffic is continued growth by adding more services and functions, the traffic will cause congestion problems and delays on Internet. Internet routers sometimes drop up to 5% of the incoming packets because of local buffer overflows (tcpipguide, 2004). For the TCP connections, it can be classified into two types: steady state and short-lived connections. This research will focus on the short-lived TCP connections by using steady-state flows extended mathematical model. On the other hand, in this project, the neural network simulation will be trained to estimate the short-lived TCP connections. The results will be compared with the mathematical model estimation results, and the conclusion will be made at the end of this research paper.

2. TCP, mathematical model and neural network

The Internet has become a more and more complex collection of network through its exponential increase in terms of users. With this increasing, there are also many

Internet applications, because such as the World Wide Web (WWW), usnet news, file transfer and remote login, have opted Transmission Control Protocol (TCP) as the transport mechanism. TCP is a very complex protocol, and the fast-changing network conditions make the development of an accurate TCP stochastic model to be a very challenging task. The TCP operates at the transport layer of the Open System Interconnection (OSI) network reference model. TCP connections drive the performance of Internet, because 90% of the Internet connections use TCP, 95% of the Internet traffic is carried by TCP (Garetto and Cigno, 1999).

2.1 TCP protocol

TCP is a full duplex protocol that each TCP connection supports a pair of byte streams, one flowing in each direction. TCP includes a flow-control mechanism for each of these byte streams that allow the receiver to limit how much data the sender can transmit. TCP also implements a congestion-control mechanism (Sinha & Ogielski, 1998). TCP also provides reliable data delivery. A key to provide reliability is that all transmissions in TCP are acknowledged. The recipient must tell the sender "yes, I got that" for each piece of data transferred. This is in stark contrast to typical messaging protocols where the sender never knows what happened to its transmission (Tcpipguide, 2004). This technique requires the TCP sender to assign unique sequence numbers to packets, also each packet sent is recorded until received the receiver sending back acknowledgements (ACKs) (Figure 1). On the sender side, the retransmission timer is also started whenever it sends a packet, if the packet is not arrival before the timer is expired, the sender assumes that the packet has been lost, and the TCP will arrange retransmission. The time between a sender sending data and receiving the acknowledgement is called the round-trip time (RTT) (Figure 1).

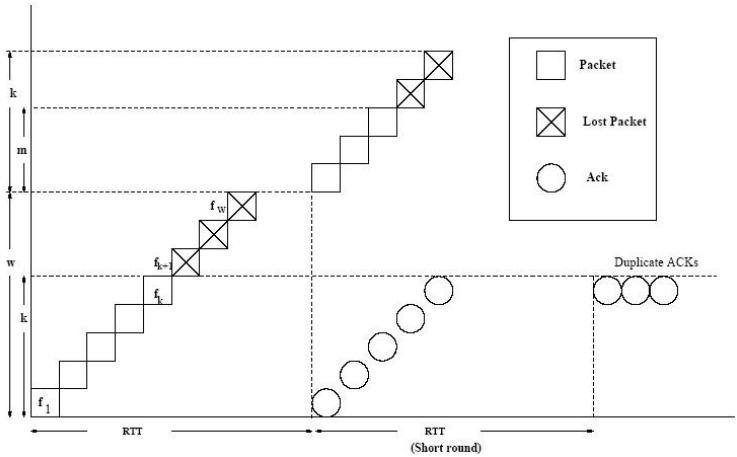


Figure 1: TCP connections (Padhye et al. 2000)

To utilise the network bandwidth effectively, TCP uses a sliding window flow control (Stevens, 1993) to send multiple segments at a time before it stops to wait for an ACK. TCP using flow control to regulate the sending and receiving rate, the sending rate depends on the receiver's processing capacity and buffer size.

Therefore, flow control is achieved by the receiver advertising a maximum window size (W_{max}) along with each ACK, thus limiting the transmission window size of the sender (Li, 2004).

This research aims to estimate TCP short-lived connections e.g. for Web transferring (smaller size date). However the mathematical modeling of long lived TCP connection is less suitable for the short lived TCP connections, because the whole date transferring process could be finished in the slow start period (Figure 1). In another word, viewing a short-lived TCP flows as an initial connection establishment handshake.

2.2 Cardwell-00 model

Neural network works on Cardwell presents the Cardwell-00 model in 2000, which is extended the steady-state results by accounting for the connection establishment phase and an approximate analysis of the initial slow start. This model added the observation that most of current TCP data transfers are short-lived and carry a small amount of data. There is a high probability for such flows, when they follow a path with low loss rate. To have a zero packet loss and, implicitly, to remain in slow-start for their duration. (Ghita, 2004). For the short-lived connection, in the case where the data transferring during the TCP short-lived connections with no packet losses (when the P value equal to 0), all data segment will be sent in the slow start phase. As the result, the model for these connections used only three inputs (Table1): the amount of data to transfer (Data), the estimated round trip delay (RTT) and the initial value of the congestion window (W1). The equation will be reduced to 1, which is the model for the time to send data segment in slow start. The data transfer latency E [T] is equal to below rather than the equation 2:

$$E[T] = E[Tss] + E[Tdelack] \quad (1)$$

On the other hand, in the case where the data transferring during the TCP short-lived connections with losses (when the P is greater than 0) had several loss characteristics that may have been added to the inputs. As the result, the model for these connections used five inputs (Table 1): the amount of data to transfer (Data), the estimated round trip delay (RTT), the data loss rate (the fast retransmit loss rate), the timeout loss rate, the initial value of the congestion window (W1) and the first occurrence of loss also need to be considered. The equation of this case is the equation 2

$$E[T] = E[Tss] + E[Tloss] + E[Tca] + E[Tdelack] \quad (2)$$

E [T]: the data transfer latency; E [T_{SS}]: the expected latency for the initial slow start phase; E [T_{loss}]: the expected cost for any RTOs or fast recovery that happens at the end of the initial slow start phase; E [T_{ca}]: the expected time to send the remaining data (the time spent in congestion avoidance) and E [T_{delack}]: the expected delay between the reception of a single segment and the delayed ACK for that segment.

2.3 Neural network

Neural network works on artificial neural networks, commonly referred to as 'Neural network', has been motivated right from its inception by the recognition that the human brain computes in an entirely different way form the conventional digital computer (Simon, 1998).

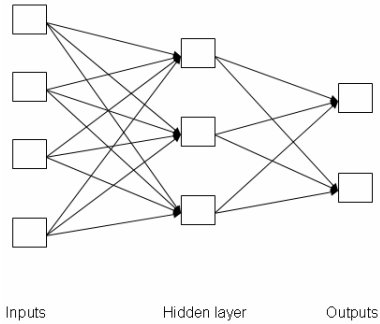


Figure 2: An example of a simple feed-forward network

Once a neural network is built to be of any use, the commonest type of artificial neural network consists of three layers (Figure 2): a layer of "**input**" units is connected to a layer of "**hidden**" units, which is connected to a layer of **"output"** units. Inputs and outputs correspond to sensory and motor nerves such as those coming from the eyes and leading to the hands. The hidden layer plays an internal role in the network. The input, hidden and output neurons need to be connected together.

This simple type of network is interesting, because the hidden units are free to construct their own representations of the input. The weights between the input and hidden units determine when each hidden unit is active, and so by modifying these weights, a hidden unit can choose what it represents. A neural network is the ability of the network to learn from its environment, and to improve its through learning. The network becomes more knowledgeable about its environment after each iteration of the learning process (Simon, 1999). This research will involve the neural network and mathematical model (Cardwell-00 Model), neural network (Matlab) can do different learning to estimate TCP connections and simulate the real network conditions also can be compared with the result of mathematical model, and the results will be more accurate.

2.4 Data pre-processing

The inputs in the equation 1, the first set of input for the mathematical model is including 1266 connection samples with non packet lost rate (P =0) and the inputs of mathematical model are Data, RTT, and Congestion window. The inputs in the equation 2, the first set of input for the mathematical model is including 5934 connection samples with packet lost rate (P > 0) and the inputs of mathematical model are Data, RTT, P, T0, and Congestion window.

3. TCP Short-lived connection estimation

The Cardwell-00 Model and the neural network as mentioned above will be tested in two stages below:

Stage 1 (Figure 3 and 4): Extract network parameters from TCP connections using TCP analysis software (e.g. tcptrace), those parameters are The amount of data to transfer (Data), the estimated round trip delay (RTT) and the initial value of the congestion window (W1) will be set as the first mathematical model inputs

The input parameters of the second mathematical model are the amount of data to transfer (Data), the estimated round trip delay (RTT), the data loss rate (the fast retransmit loss rate) (P), the timeout value (T0), the initial value of the congestion window (W1) and the first occurrence of loss also need to be considered.

Stage 2 (Figure 5 and 6): Input the resulting parameters, together with a performance estimate resulted from a mathematical model, into a neural network

Finally, the resulting performance estimate (i.e. throughput) for each connection, based on the trained neural network, will be compared with the current existing mathematical models and the actual performance value obtained for that particular connection in order to asses whether they are superior in terms of accuracy and robustness. Therefore, the comparison will be made in terms of relative error.

Factor	For equation 2	For equation 1
The inputs in the equation 1 and 2		
(RTT): Round-trip time	Between 1.13 and 5888.91 milliseconds	Between 1.13 and 23578.19milliseconds
(P): Data segment loss rate	The data segment loss rate is between 0 to 7%	N/A
(Data): Data segment size	Between 118 and 1460 bytes	Between 120and 1460 bytes
(T0): The average duration of the first timeout in a sequence of one or more successive timeouts	0 to 3588.177msec	N/A
(W): Congestion windows size	Between 240 and 64240 bytes	Between 240 and 59860 bytes
The mathematical result	Which is the input only for neural network estimation	
Target in neural network		
(Time): The data transfer time	Between 0.032478 and 103.4157 seconds	Between 0.016377 and 103.4141 seconds

Table 1: The inputs in the mathematical model and neural network

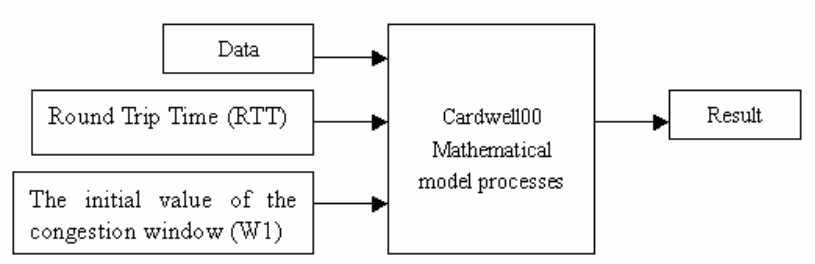


Figure 3: The mathematical model for TCP connections without packet loss

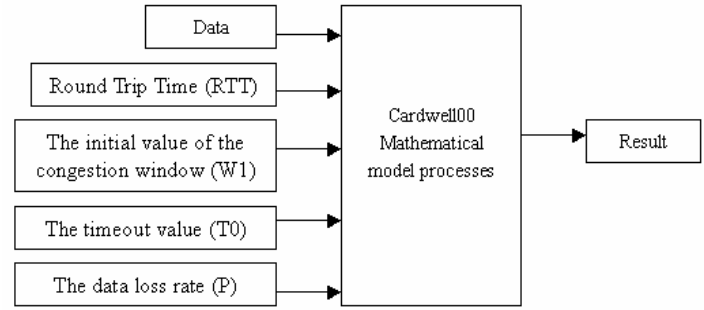


Figure 4: The mathematical model for connections with packet Loss

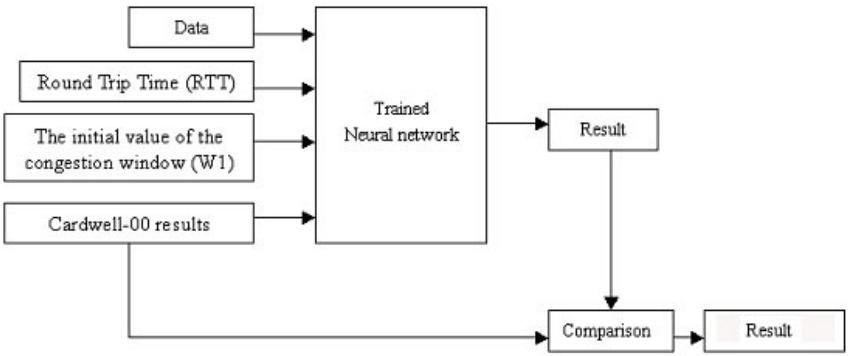


Figure 5: The neural network for TCP connections without packet loss

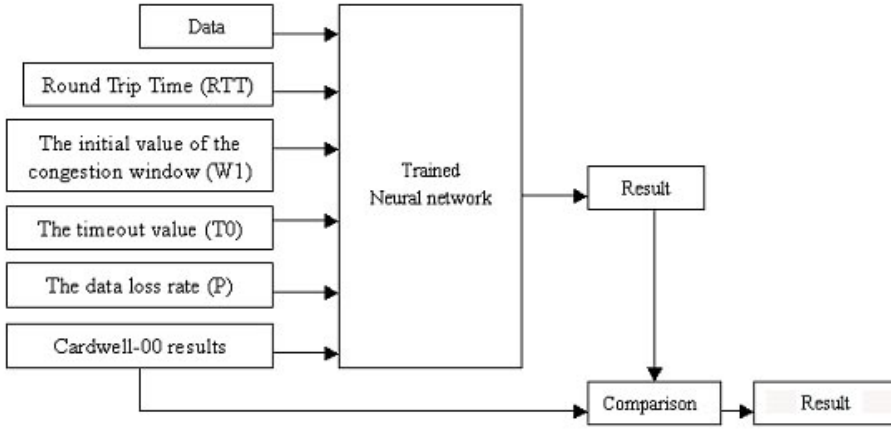


Figure 6: The neural network for TCP connections with packet loss

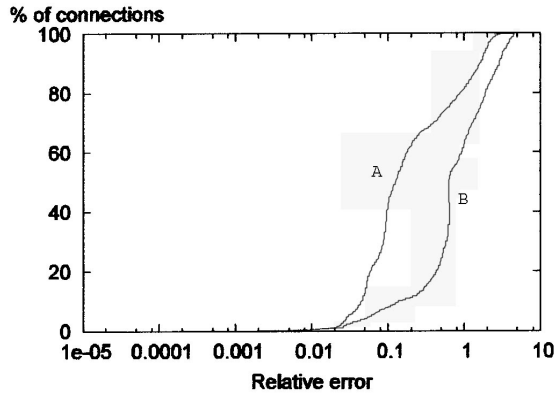


Figure 7: Cumulative distributions of the relative error resulting from using the mathematical model (B) and neural network (A) without packet loss

Figure 7 and Figure 8 compared the relative error produced by mathematical models and neural networks for TCP shot-lived with and without packet loss connections. As shown in the Figure 7 and 8 below that the neural network model outperformed the mathematical model. It shows that the relative error of connections estimated by neural network is around 5% - 20% less than the estimation of Cardwell-00 model. Although, the figure 8 shows that the accuracy of the neural network decreased, the neural network still performed better than the mathematical model. The average figures for mathematical model and neural network are presented in Table 2 and 3.
The next step is to compare the table shows that the average relative error, standard deviation of relative error and the correlation between the predicted values and the real values through the statistical results of the comparison, which are presented in Table 2 and 3 in order to estimate that how accurate is the mathematical model, the neural network model.

In Table 2 and 3 show that the neural network models perform on average better than the math model in both loss-free and loss estimation. It shows a 25% higher figure for the average relative error when compared with the results from 4-8-4-1 neural network for the connection without loss.

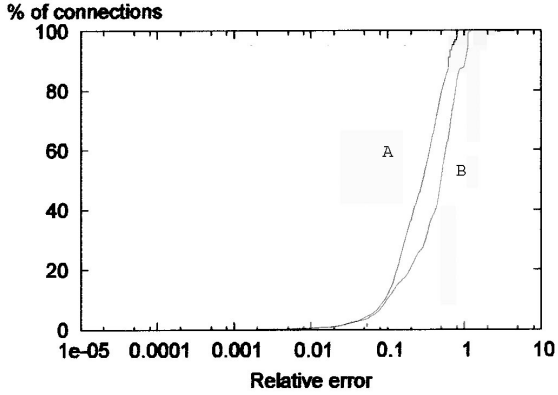


Figure 8: Cumulative distributions of the relative error resulting from using the mathematical model (B) and neural network (A) with packet loss

Model	Average relative error	Standard deviation of relative error	Correlation
Mathematical	0.596473	0.422372	0.732826
Neural network	0.352271	0.384017	0.783274

Table 2: Comparison of the resulting average figures for the short-lived TCP non-packet loss connections, using mathematical and neural network model

Model	Average relative error	Standard deviation of relative error	Correlation
Mathematical	0.557832	0.469931	0.448371
Neural network	0.530782	0.483257	0.498734

Table 3: Comparison of the resulting average figures for the short-lived TCP packet loss connections, using mathematical and neural network model

In Table 2, the neural network provided better accuracy than the mathematical model for the TCP short-lived connection without packet loss modelling results from mathematical model and neural network, and in the Table 3, the results from mathematical model and neural network model are very close; the different between the mathematical model and neural network model is smaller than the results in Table 2.

4. Conclusion and future work

We have presented our proposed neural network modelling TCP short-lived connections for different type of data sets (with and without packet loss). All the data

sets are collected from real world TCP transfers, and compared with the measured results from mathematical model and neural network.

For the TCP short-lived connections, we only separately estimated the data set as non-packet loss and packet loss conditions, to estimate the packet loss rate bigger than 10% and more should be the next step for this research.

For the neural network modelling, it only may improve the modelling accuracy of the method in some of the aspects. In another words, the proposed neural based model provides a better alternative to mathematical models in terms of accuracy. Therefore, it is hard to identify the error in neural network inputs. To estimate the TCP short-lived connections in different type of neural network with different MSE parameters should be the next step for this research. And this could improve the accuracy of the TCP short-lived connections modelling.

5. References

Garetto, M. and Cigno, R. Lo (1999) "*Modeling Short Modeling Short-Lived TCP*", http://portal.acm.org/citation.cfm?id=987127, (Accessed 10-04-2005)

Ghita, B.V., (2004) "Performance characterisation of IP networks", *PhD Thesis*, University of Plymouth, UK

Li, Y.j. (2000) "Modeling Web/TCP Transfer Latency", *Master Thesis*, the University of Calgary, Canada

Padhye, J., Firoiu, V., Towsley, and D., Kurose, J. (2000) "*Modeling TCP Throughput*", http://www.cs.ucsd.edu/~savage/papers/Infocom2000tcp.pdf, (Accessed 02-02-2005)

Stevens, W. (1993) "*The Protocols (TCP/IP Illustrated, Volume 1)*", Addison-Wesley Professional; 1st edition

Simon, H., (1999) "*Neural networks : a comprehensive foundation*", Englewood Cliffs, N.J., Prentice Hall

Sinha and Ogielski., (1998) "*A TCP Tutorial*", www.ssfnet.org/Exchange/tcp, (Accessed 17-06-2005)

Tcpipguide., (2003) "*the TCP/IP Guide*", www.tcpipguide.com, (Accessed 07-06 2005)

Section 2

Communications Engineering & Signal Processing

How will BT meet the challenge to provide multimedia services over its local loop?

A.S.Aloufi and C.D.Reeve

School of Computing, Communication and Electronics, University of Plymouth, UK
e-mail: kit.reeve@plymouth.ac.uk

Abstract

The aim of this paper is to give an overview of the deployment of fiber to the home (FTTH) technology and its economic issue. Currently ADSL is the leading broadband technology and it will be the leader for at least another decade. CATV technology could be in a strong competition with ADSL. Using FTTH, both technologies could be received simultaneously in a single cable. FTTH will offer the triple play technology, which will include data, voice and video with perfect quality. The main barrier of the deployment of FTTH is the high cost, which is the main reason causing delay in its distribution. In the UK, the deployment of FTTH would not be a reality before at least ten years although BT has started to experiment the technology to examine its performance and economic advantages. The best plan for BT is to start the deployment of fibre to the curbs, which will make it easier for a future deployment of FTTH. CATV companies could also improve their local loop by increasing the number of nodes and make the user per node lower then the current design. CATV companies have the upper edge in the case of availability of the use of fiber therefore it is easy for them to run it for the last drop to the homes.

Keywords

Communication engineering, FTTH, ADSL, CATV

1. Introduction

Since late 1970s the optical fiber prove itself as the saver for the communication industries to enhance the bandwidth and the quality of different networks. Fiber will shape the future of the different fields in telecommunication. It is unusual to hear about any communication project not including fiber as a backbone. Many experts agree that fiber would feature in the future of broadband. The main advantage of optical fiber is the availability of enormous bandwidth (Nicholas *et al.* 2004; Mark, 2002; Paul and Green, 2004). There are some other advantages such as its small size compared to copper, high immunity for interference, electrical isolation, security, reliability and flexibility. Many advantages of fiber are listed and elaborated in (John, 2003; Jeff, 2002; Govind, 2002). It is anticipated that FTTH will soon become a necessity rather than a luxury because of the increased requirement for high bandwidth served with high speed.

The main purpose of this study is to research and investigate the possibility of deploying fibre to the home in the local loop used by British Telecom in the UK. There is a huge demand for large bandwidth to support the future technologies

requirement and the potential to offer it with the current technologies is difficult. Optical fibre is one of the future promising technologies that can provide a real opportunity for Internet providers to meet the future challenges. Fibre to the home needs a revolutionary decision because of its high cost of deployment. Currently the telecommunication operators have the old infrastructure of copper and the cable operators have coax in their local access loop and the question is what is the chance of fibre?

It has always been a risk to develop equipment for access networks because of the equipments infrastructure. For BT it is extremely cost sensitive to run FTTH all around the UK which is the main barrier delaying the start of this technology. The question for BT is whether there is a perceived market to take the decision for such a large project? The answer depends on the customers whether they are interested in this service or not. On the horizon it seems that there is a very huge demand for large bandwidth to run the latest new technology especially for HDTV, video conferencing and distance learning. Being one of the largest telecom companies in Britain, BT has to meet the challenges for providing the latest technology to its consumers. The current bandwidth will not satisfy customers forever and in near future subscribers will need a higher bandwidth than currently offered with ADSL. The development of fiber for the access network seems to be very important for the future development and it is already available in many countries.

The aim now is to find a solution as to how BT could meet the challenge for providing the fiber for its local access network and what will be the cost to deploy fiber instead of the current copper.

2. BT ADSL and CATV broadband

Internet service providers (ISPs) are in a direct competition with the cable companies as the latter are offering the broadband service for customers using cable modems. BT might be in a competition with the cable companies such as Telewest and NTL . BT is leading the broadband service at this time but in the future there will be some change if the cable companies upgrade their access loop for better performance and they might become the leader in broadband services if they merge together. Currently the cable companies are laying in a mine of gold because of their deployment of fiber to the nodes just short distances from the subscriber. The other good thing about the cable companies is that most of their cables are laid in ducts and they have a ready to use infrastructure for fiber deployment. BT is not planning to distribute fiber to the homes in the short or medium term but it will be necessary in the future. According to a BT spokesman, the national operators have not “discounted the deployments of FTTH in the near future but it will be deployed in the UK starting in the green areas and new buildings” (Wilson, 2005). BT is running a trial of FTTH technology as a part of the 21CN program to test the technology cost and performance. The FTTH trial will offer the service for 1500 homes and businesses. This trial started in October 2004 and it will be running until September 2005 (btlc website, 2004; BT wholesale website, 2004). The future broadband leader will be the user of the latest technologies and the deployment of FTTH will be one of

the important issues in the near future. The use of wireless in the access network could play a very important role in providing Internet service especially in places where the deployment of FTTH is facing difficulties. It is very well known that wireless will not offer the amount of bandwidth offered by fibre but it could be used anywhere.

Figure 1 shows the basic operation of ADSL and figure 2 shows CATV architecture. Full details of the two technologies can be found Aloufi (2005).

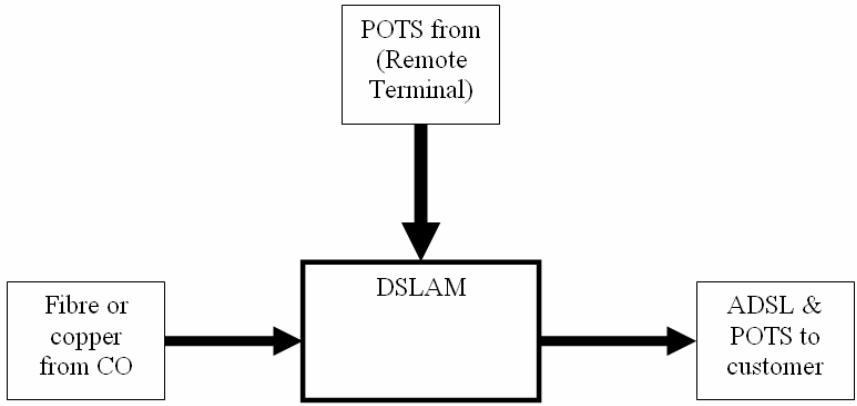


Figure 1: Basic of ADSL operation

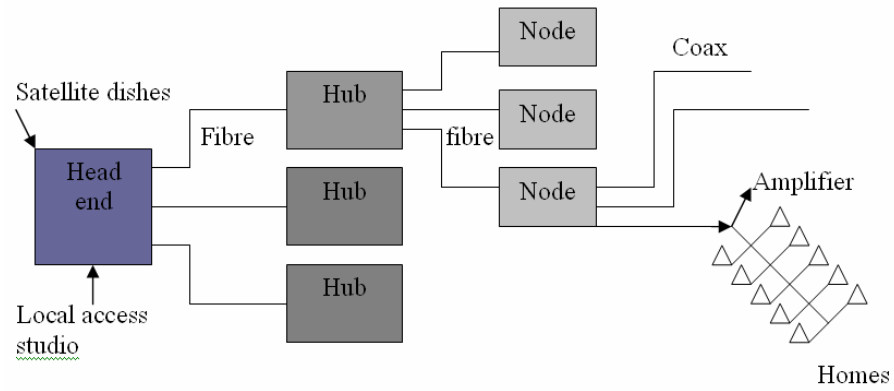


Figure 2: CATV architecture, adapted (Jeff, 2002)

3. BT networks

BT telecommunications network consists of two main segments. The first one is the core network, which consists of nearly 6000 exchanges most of which are connected to each other with a very high-speed optical fibre. The second segment is the access network consisting of twisted copper pair (Mayhew *et al* 2002), which connects the customers to those exchanges. The BT copper access network was originally designed to support the telephony service more than 50 years ago and they have been

upgraded to support the ADSL technology. The BT access network as most of the others consist of two types of flexibility points – the primary cross connect point (PCP) and the distribution point (DP). From the DP the twisted copper pair is distributed to the customer's homes and businesses. The distance between the DP's and the homes is in the range of 20-100 meters and each DP supplies 5-20 homes. The main problem with the last drop from the DP's to the customer is those wires buried in the ground with no ducts. In order to change those cables BT needs to build new ducts, which is very costly. The BT access network contains 121.7 million kilometres of copper wire (btlc website 2005) and 25000 kilometres of underground ducts (Mitchell, 2004).

4. Access network from copper to fibre

The main advantage of fiber over copper is its enormous bandwidth. It is a very big migration to meet the challenge of providing fiber as an alternative for copper in the access network. Fiber to the home has been a consideration from the earliest birth of the fiber technology; however, the implementation problem is more of an economical nature rather than technical. The deployment of FTTH is cost sensitive and requires a very high investment. Then what is the procedure and how can BT meet the challenge of providing fiber for its local access loop? It is clear that there is a very huge potential demand for high bandwidth for the fixed network in the near future.

BT could achieve this challenge in three steps:

- First: to deploy the fiber from the local exchange to the PCP's.

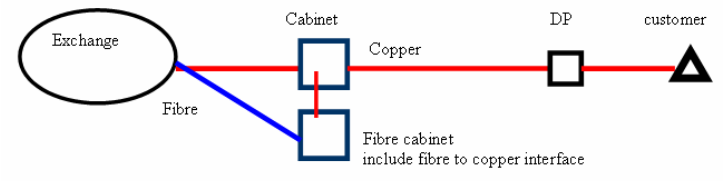


Figure 3: First step to FTTH

- Second: run the fiber from the PCP's to the DP's.

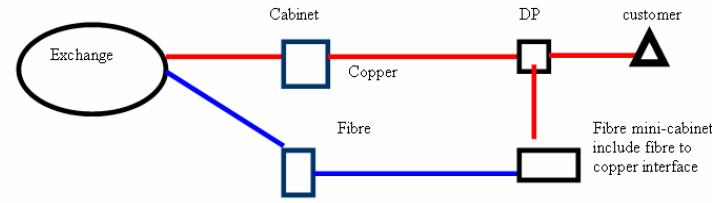


Figure 4: The second step to FTTH

- Third: the third step is to distribute the fiber to the homes as a last drop.

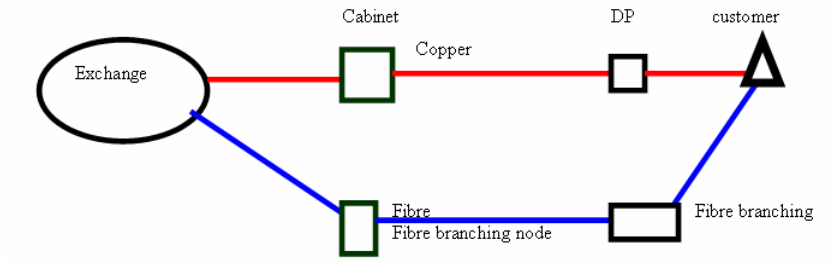


Figure 5: The last step to FTTH

The first two steps will be easier then the third one because of the availability of ducts to run the fiber. The third step will be difficult because of the high cost of building new ducts and building new infrastructures for some buildings. The above figures show one of the options to deploy the fiber gradually all the way to homes and buildings. The second option is to move directly from copper to fibre (Mayhew *et al* 2002). This will require a great investment and a huge amount of money.

5. The cost of FTTH in the UK

It is mentioned earlier that the deployment of fibre to the homes will be a cost sensitive project. BT is looking for the lowest cost investment and also for the best revenue to recover the large amount of money has been spent. The customer on the other hand will go for the lowest price and reasonable quality of service. The most important thing for the customer is to get a better service plus a reduced monthly bill. A study in the early 1990s shows that to run fibre to the homes in the UK will cost about £15 billion (Payne and Davey, 2002). There are no problems with the fibre and related passive equipment as the cost of them has dropped rapidly in the last few years. However, the infrastructure and labour works are expected to take the lion's share of the money.

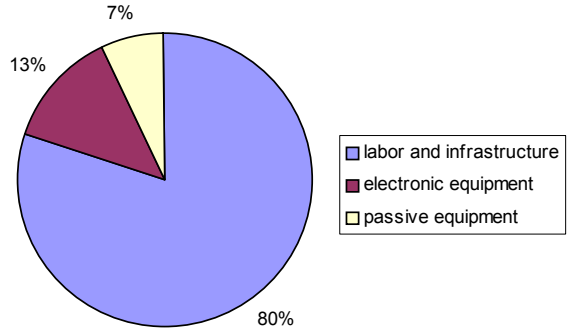


Figure 6: FTTH expected cost shares

It is anticipated that the total cost will exceed the £15 billion mark because of the increase in the labour cost. BT could make a strategy to reduce this cost and a list of solutions is provided below.

- BT needs to optimise the design of the network and choose the best relevant design.
- Making the best use of the existing infrastructure. Of course BT is going to use its old underground ducts but for places where there are no ducts, they need to find the possibility of using other services ducts (e.g. electricity, water, gas, cable ducts), as this will offer a readymade solution and thus reduce the cost.
- Better use by the subscriber by encouraging them to use the available services. BT will get better revenue by offering new services.
- It is a good idea not to use only the very skilled people everywhere. There is possibility to use non-skilled people in some locations. BT can run some short courses for the technicians to qualify them for the installation.

6. Future service revenues

Economically the disparity is great between one type of bit and the other (Nicholas *et al* 2004). The three services that will be provided are video, telephony and data. Some of the services will not be very acceptable immediately by customers and this might be frustrating for the provider. The ability to generate good revenue from the three services will depend on the customer acceptance. There are billions of people using the telephones and millions of them using the television cables. Most of the people will not think how much they pay for the bit even they understand it, they usually think how much they pay for the monthly bills. In the case of FTTH the subscriber does not have to pay for several services separately. Instead, only one bill will be issued for the triple play service and which will include data, voice and video services. Assuming all BT customers are going to have the technology with an average monthly charge of £80 per month, which will be £960 a year. Based on this data, the annual revenue for BT‘s 20,000,000 subscribers will be £$1.92x10^{10}$. This revenue will be divided between the cost of loan, customer care and marketing, maintenance, operations, billing and so forth. It will be several years until BT would be able to recover the cost. Of course more revenue per customer will provide a better income but this again depends on the customer use of the services provided by the new technology.

7. Discussion

There is no technical problem behind the deployment of fibre to the homes, but the cost is the main barrier. The limiting factors of ADSL (limited distance and bandwidth, old infrastructure of copper and the lack of advantages of fibre like smaller size, lighter, far more reliable, not affected by electrical interference, and not causing interference with other equipment) is pushing towards establishing a new technology for the future broadband and the most promising technology is optical fibre. CATV and ADSL are not expected to support the future broadband technologies in their current system. The broadband providers have to find a solution to meet the future challenge on providing the triple play service in their networks.

Referring to Aloufi (2005) the expected cost for BT to run the FTTH all around the UK is about £20 billion. Finding the best planning, encouraging the subscriber to use the available services could reduce this cost and give better revenue. BT realise that the FTTH will be a necessity in the future and that's why they are running it on an experimental basis in 1500 homes to test the performance and the cost of the technology. According to many BT managers there is no current plan in the short term to distribute the fibre to the homes. The easiest way is to start this deployment in the green areas, rural areas and new buildings. For the large city it will consume a lot of work and wireless stations using fibre backbone could be a good alternative in the future for the overcrowded cities in the UK.

8. Conclusion

This paper presents an overview about the future broadband technology and a discussion on how can BT meet the challenge to provide fibre in the local access loop. The increased demand for high bandwidth will put the death nail in the head of the ADSL in the near future. Economically delivered future increased bandwidth services need a reduction in the cost of unit of the bandwidth. The question is what is the possible way to reduce the unit cost? The best way to reduce the unit cost is to simplify the network by using the passive equipments technology. One of the facts achieved from this study is that the deployment of FTTH is not very largely constrained by technology rather it is mainly a commercial issue. In the case of deploying the FTTH the thing that is required by BT and telecom companies is both a reduction in the cost of the deployment and of course an increased gain of perceived revenue.

BT is in a competition with the CATV provider companies as the latter is in a position to provide the broadband service to the customers and their fiber already reaches near to the customers' promises**.** BT is still leading the broadband industry in the UK and it is anticipated that it will stay in its position especially with the high quality 21CN program, where BT is planning to migrate from the old PSTN system to the new VoIP. As a part of the 21CN program BT is running the technology of FTTH to 1500 homes to test the technology performance and to determine the commercial issues.

Cost is the main barrier delaying the use of the fiber. There will be a reduction in the future cost for fiber equipments but unfortunately the infrastructure building and labour will remain the main important considerations. Finally we can say that either high cost will make it hard to initiate or the future technologies will make it hard to stop.

9. References

BT wholesale website (2004), "*Questions and answer about 21st CN project*", http://www.btwholesale.com/application?origin=siblings.jsp&event=bea.portal.framework.internal.refresh&pageid=21cPressRelease&nodeId=navigation/node/data/our_business/media_information/21cn/21cn_news_releases (accessed 1/8/05).

btlc website (2004), "*BT to switch voice calls to IP as 21st century network takes shape",*http://www.btplc.com/ News/Articles/Showarticle.cfm?ArticleID=b80dbd58-a821-46d8-9852-81e1cab365dd (accessed 1/8/05).

btlc website (2005) "*The network story*", http://www.btplc.com/Thegroup/ Networkstory/HTML/slide.aspx_slide =1.html accessed 27/7/05.

Govind, P. (2002) "*Fibre optic communication systems*", 3rd edition, Wiley- Interscience, ISBN 0-471-21571-6.

Jeff Hecht. (2002) "*Understanding fibre optic*", 4th edition, Prentice Hall, New Jersey, ISBN 0-13-027828-9.

John, M. (2003) "*Optical fibre communications; principles and practice*", 2nd edition, prentice hall of India, New Delhi, ISBN 81-203-0882-4.

Mayhew, A.J., Walker, A.M. and Fisher, S.I., (2002) "*Fibre To The Home – infrastructure deployment issues*", BT technology journal. Vol20 No4, pages (91-103).

Mark P. (2002) "*Home networking*", IEEE television journals, pages (454-459).

Mitchell.J., (2004) "*getting fiber to the home*", University college of London, http://www.istmuse.org/Documents/NOC2005/Summer_School/John_Mitchell_FTTH.pdf (accessed 9/8/05).

Nicholas, J.F,. Iannon, P and Kenneth C. (2004), "*A view of Fibre To The Home Economics*", IEEE Optical Communications, pages (16-23).

Payne.D, Davey.R., (2002) "*the future of fibre access system*", BT technology journal. Vol20 No4, pages (104-114).

Paul, E., Green. Jr. (2004) "*Fibre to the home: the next broadband thing*", IEEE Communication Magazine, pages (100-106).

Aloufi. A., (2005), Msc thesis "how BT will meet the challenge to provide multimedia services over its local loop".

Wilson.R., (2005) "*No need for UK investment in FTTH*", www.electronicsweekly.com/article39277.htm (accessed 17/6/05).

Will all communication be wireless thus investment in fibre is a waste of time and money?

B.A.M.Khawaja and C.D.Reeve

School of Computing, Communications and Electronics,
University of Plymouth, Plymouth, United Kingdom
e-mail: C.Reeve@plymouth.ac.uk

Abstract

There's no doubt the world is going wireless – faster and more broadly than anyone might have expected in last 20 years, this paper will cover the research of current and future wireless technologies as compared to fibre technologies for both fixed and mobile communications in terms of cost, flexibility and mobility. The research has been found of great interest since wireless networks are becoming extremely popular especially wireless-LANs (Wi-Fi), they have specific usage segments in Metropolitan area networks (MANs), Wide area networks (WANs) and Personal area networks (PANs). Fibre optic communication systems are also becoming the basic necessity for reliable communication in LAN environment like Gigabit Ethernet and now 10Ge (10 Gigabit Ethernet). Fibre based systems are also very popular and considered reliable in telecommunications, CATV, as well as in SAN (Storage Area Networks) and MAN. In this research the costs comparison and technology trends for both fibre as well as wireless communication systems will be provided and in the end will draw conclusions on the knowledge gained as to likely dominant technologies either fibre or wireless? and if both then in what scenarios will they going to stay in a time frame of approximately next 20 years.

Keywords

Wireless-LANs, Wi-Fi, Metropolitan area networks (MANs), Wide area networks (WANs), Personal Area Networks (PANs), Gigabit Ethernet, 10Ge (10 Gigabit Ethernet), CATV, SAN (Storage Area Networks)

1. Introduction

Computer networks can be roughly divided into LANs, MANs, WANs and Inter-networks with their own characteristics, technologies, speeds and niches. The varying demands on communication networks are leading to the development of next generation networks. As the Internet become increasingly popular, geographic boundaries become meaningless. With a mouse click an Internet user can trigger an information flow between nodes that are thousands of kilometres apart. With this situation, data must travel ultra long distances and provide instantaneous connections; reliability and security are the main concerns in these types of communications. Fibre optic based wired networks as well as wireless networks are considered as Next generation networks because almost all different types of communications either in telecom or in Internet communications are now depending heavily on these next generation networks which are expected to be better equipped

than traditional networks to handle the traffic demands that are anticipated from today's as well as tomorrow's Internet-driven economy. Wireless communications involves the biggest paradigm shift in communications. The optical-fibre revolution was about bandwidth and information convenience. The market for wireless communications infrastructure continues to grow at a rapid rate. On the other hand security, frequency band allocation and multipath fading are the common problems with wireless communication systems.

2. Wireless Communications

Our age has given rise information junkies especially after the advent of Internet and modern wireless communications the people who need to be on-line all the time. For those mobile users, twisted pair, coax and fibre optics is of no use. They need to get their hits of data for their laptops, notebooks, and palmtop or wristwatch computers without being tethered to the terrestrial communications infrastructures. For all these users wireless communications is an answer. There are different types of wireless communications, the most common are Microwave communications, Infrared/millimetre wave's transmission, bluetooth transmission, Lightwave transmission, Satellite communications for satellite Internet, IEEE 802.11 Wireless LAN Technology / Wi-Fi, Wireless broadband / Wireless Local Loops (WLL) and Wi-Max. This research paper will cover some important and most commonly used technologies these days.

2.1 Satellite Communications for Satellite Internet

Satellite Internet access is ideal for rural Internet users who want broadband access. No telephone lines or cable systems are required for satellite based Internet access, but instead it uses a satellite dish for a permanent 2-way (upload and download) connection of Internet. Satellite Internet connections are permanent; user will never get a 'busy signal' when trying to connect. Upload speed in a satellite Internet link is about one-tenth of the 500 kbps download speed. Cable and DSL have higher download speeds, but satellite systems are about 10 times faster than a normal modem. The key installation-planning requirement is a clear view to the south, since the orbiting satellites are over the equator area (Tanenbaum, 2002).

2.2 IEEE 802.11 Wireless LAN Technology / Wi-Fi (Wireless-Fidelity)

IEEE 802.11 based wireless LANs are getting increasingly popular, more and more office buildings, airports and other public places are being outfitted with them. IEEE 802.11 defines an over-the-air interface between a wireless client and a base station (or access point), or between two or more wireless clients. Wi-Fi is the wireless way to handle networking 802.11 networking. The big advantage of Wi-Fi is its simplicity. One can connect computers anywhere in the home or office without the need for wires. The computers connect to the network using radio signals, and computers can be up to 100 feet or so apart. Currently three standards of IEEE 802.11 are there in the market, which are 802.11 a/b/g. Table .1 given below shows the Benefits/Limitations of Wi-Fi.

Standard	Benefits/Limitations
802.11b (Wi-Fi)	• Predominant standard for wireless networking for business and home LANs, as well as public hotspots. • Runs on three channels in 2.4 GHz spectrum • Transfers data at speeds up to 11 Mbps at distances up to 300 feet • Occasional interference with microwaves and cordless phones
802.11a	• Runs on 12 channels in 5 GHz spectrum • Transfers data at speeds up to 54 Mbps, but at distances up to only 50 feet • Not backward compatible with 802.11b, thus requiring all new wireless equipment if you're switching over • Few interference issues
802.11g (ratified Wi-Fi Standard)	• Runs on three channels in 2.4GHz spectrum (same as 802.11b) • Has the speed of 802.11a, but is backward compatible with 802.11b • More secure

A wireless hotspot is a connection point for a Wi-Fi network. It is a small box that is hardwired into the Internet. The box contains an 802.11 radio that can simultaneously talk to up to 100 or so 802.11 cards. There are many Wi-Fi hotspots now available in public places like restaurants, hotels, libraries and airports (Intel, 2005a; Intel, 2005b). Security is still an issue in Wi-Fi.

2.3 Wi-Max (IEEE 802.16)

WiMax can be called as an extension of Wi-Fi especially in terms of range; WiMAX is short for Worldwide Interoperability for Microwave Access (IEEE 802.16). WiMAX could potentially erase the suburban and rural blackout areas that currently have no broadband Internet access because phone and cable companies have not yet run the necessary wires to those remote locations, its an emerging technology that will deliver last mile broadband connectivity in a larger geographic area than Wi-Fi, enabling T1 type service to business customers and cable/DSL-equivalent access to residential users. WiMAX will enable greater mobility for high-speed data applications. With such range and high throughput (Intel, 2005a; Intel, 2005b). WiMAX is still not functional but it will hope to be functional by the end of 2005. WiMAX has the ability to replace cable and DSL services, because it provides universal Internet access just about anywhere. It will have the ability to connect automatically to the closest WiMAX antenna after turning the computer on.

3. Optical Fibre Communications

Fibre optics is used in many industries now days, most notably in telecommunications and computer network (LANs & WANs). The bandwidth and reliability of fibre optic systems are the main reason of their popularity with lower

attenuation and few repeaters for long distances. The different implementations of fibre in telecom networks include Last mile implementation (Fibre to the Home/Curb), between local exchanges, between secondary exchanges (between cities), Submarine cables (between continents) and Synchronous (SDH/SONET). On the other hand the implementation of fibre optics in LAN/WAN (generally called IT Arena or Data Transfer Networks) includes fibre over Ethernet (10 Gigabit Ethernet is used now a days in the market), fibre is also used in ATM, VoIP, SAN, FDDI based networks, Frame Relay and Fibre Channel in SAN. This research will cover the one, which are supposed to be the Future of fibre.

3.1 Fibre-to-the-Home (FTTH) / Fibre-to-the-Curb (FTTC)

In FTTC, the telephone company runs optical fibre, from the end office into each residential neighbourhood, terminating in a device called an ONU (Optical Network Unit), transceiver in a FTTC system remains outside the subscriber's home. On the other hand, FTTH is an access technology, which uses optical fibre infrastructure to deliver telephony, television and data services straight to the customer's premises. The technology paves the way for the launch of a range of next generation high-speed services, such as high-definition TV, faster music and video streaming. In FTTH, the transceiver is located inside the subscriber's home. FTTH is very expansive technology. The deployment of FTTH is possible through Passive Optical Networks (PON) because cost will be less due to the elimination of electrical devices from the system. FTTH deployment is already started in UK on trial basis as part of BT 21CN Project and in that trial, FTTH will be first deployed to 1,500 customers (PONForum, 2005; Global Telecoms Business, 2005).

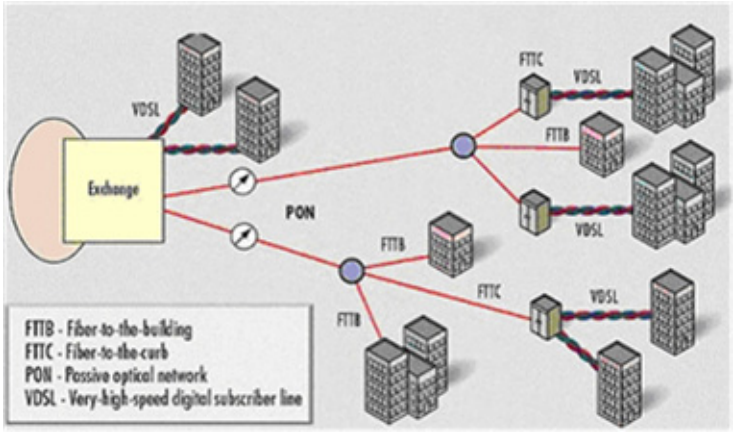


Figure 1: Showing Exchange to FTTH/FTTB Connectivity through PONs
Source: PONForum (2005)

3.2 10-Gigabit Ethernet in Local Area Networks (LAN), Metropolitan (MAN) and Storage Area Networks (SAN)

Ethernet has evolved to meet the increasing demands of packet-based networks. Gigabit Ethernet (802.3z) is also a version of Ethernet, which supports data transfer

rates of 1Gbps (1000Mbps). IEEE 802.3ae (10 Gigabit Ethernet - 10Ge) is different that it will only function over optical fibre (Tanenbaum, 2002; Intel, 2005c).

The 10Ge supports both single-mode and multimode fibre media, like Gigabit Ethernet, with links up to 40 km, 10Ge allows companies that manage their own LAN environments the ability to strategically choose the location of their data centre and server farms. Gigabit Ethernet is already being deployed as a backbone technology for dark fibre metropolitan networks. 10Ge now enables cost-effective, high-speed infrastructure for storage area networks (SAN). 10Ge can now offer equivalent or superior data carrying capacity at latencies similar to many other storage networking technologies, including Fibre Channel (Tanenbaum, 2002; Intel, 2005c).

4. Wireless as Compared to Fibre Optic Communication System

4.1 Microwave Systems versus Fibre Optic Communications Systems

In intercity transmission links, where it is difficult to dig and lay down fibre in the ground, microwave is the best solution. Before fibre optics, for decades these microwaves links formed the hearts of long-distance telephone transmission systems. Even now they are used in many types of communication links but there negative are they need line of sight communications and can be absorbed due to rain or interfere because of fog. In Intercity transmission, fibre has completely supplanted microwave systems as the new medium for intercity transmission. Fibre costs are typically more than microwave transmission setup but it has much higher transmission capacity and most importantly, distance between repeaters is much greater (Tanenbaum, 2002).

4.2 WiFi / WiMax Wireless Systems versus Fibre Optic Communication Systems

WiFi / WiMax as well as fibre optic systems, both are different in infrastructure, implementation, operations, flexibility and mobility. The comparison of both the technologies depends on their implementation in different scenarios like within a city, within an office building or between buildings or with in a university / hospital / airport.

4.2.1 Within a City Scenario / Within an Office Building / Within a Campus Scenario

WiMax is a step-up version of WiFi, it operates on the same general principles as WiFi, but it can covers wider, metropolitan or rural areas. It can provide data rates up to 75 Mbps/base station with typical cell sizes of 2-10 km. This is enough bandwidth to simultaneously support (through a single base station) more than 60 businesses with T1/E1-type connectivity and hundreds of homes with DSL-type connectivity (Intel, 2005b). On the other hand, fibre optic communication system in the same scenario will not be as mobile and flexible as WiMax solution but definitely the bandwidth and reliability of the system will be much higher than WiMax. The problem with fibre is its cost (mainly installations cost) for short distances and rigid

nature, which is not acceptable by mobile users but on the other hand for reliable communications, like for SANs, or backbone networks connectivity or connectivity security / surveillance, or in hospitals and in university campuses for data transmission and storage, the preferred media is fibre optic communication systems.

5. Cost Comparisons of Fibre Optic Communication Systems with Wireless Systems

Cost is a very important factor for any communication system for its installation, use, future survival and popularity. The only problem of fibre is its huge installation cost. On the other hand, for wireless systems, initial implementation cost is very low as compared to fibre optic communication system because of low equipment cost and there is no need of wires or infrastructure so, it save lots of money. Lower costs are also one of the main reasons of wireless systems popularity.

5.1 Fibre Optic Communication System Costing

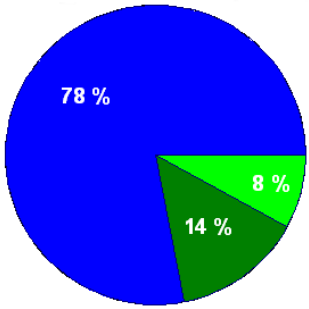


Figure 2: Costing Ratio Pie Chart of a Fibre Optic Communication System

Fibre Optic Installation and Labour Cost	75-80 %	78 % (Avg.)
Passive Instruments and Devices	7-8 %	8 %(Avg.)
Electrical Instruments	12-14 %	14 % (Avg.)

Table 2: Costing Ratio Table of a Fibre Optic Communication System

Costing is the main fear that catches the companies/institutions that wants to install fibre optic cables in their premises. The cost of fibre optic cable is as low as £ 940.62 per 1000 meters for Single Mode Fibre (SMF), the same 1000 meters Multi Mode Fibre (MMF) will cost around £ 850.75 (FTTH, 2002). Pie chart given in Figure 2 and Table 2 shows the cost comparison for every component of the system.

5.2 Case Study – Cost Comparison of Microwave and Fibre Optic

This is the cost comparison case study of Sauk County (2005) for the interconnection of their tower sites for high-speed data communications. There were two choices; microwave link and fibre optic, primary advantage of microwave system is low cost and flexibility. Of primary concern to the County are the licensing issues related to the use of Microwave and its cost. Fibre Optic Communication link is the other technology that would meet Sauk County's needs for tower interconnection. The advantages of fibre optic cable are the virtually unlimited capacity, low maintenance and long life expectancy. The case study shows the cost comparison of microwave system and fibre optic systems that although microwave system's installation and engineering cost was less (£ 728,637.04 - £ 815,638.48) than fibre optic

communication system (£1,653,043.25 - £1,910,131.46), but the annual maintenance cost of fibre communication system (£ 13,593.97 - £ 16,312.76) is less than microwave $ 31,500 (£ 17,128.40). On the other hand, fibre optic communication system life is three times greater than microwave system, which shows that, although the initial installation costs are high, but fibre optic communication systems are the right choice for long-haul communication (Sauk, 2005).

5.3 Costs for Wireless Systems

Wireless Systems are the one which do not need any specific infrastructure like wired system (Fibre optics, Coax systems), so because of no wiring, they save the installation and labour cost in the first place. Wi-Fi enabled Laptop / PDA is required by the user and for a PC user requires a Wi-Fi card for internet access, a decent WiFi card is easily available in £ 20-30 from the market. A Wireless router cost around £ 80-90 for the home network or a small business. A home user can easily setup a home wireless network for 2-3 computers with in £400. So, use of wireless will save lots of cabling hassle.

5.4 Wi-Fi Internet Access

WiFi is the highly adaptable technology in the last few years for wireless communications both in public places as well as in university campuses, corporate offices and home networks. There are different service providers throughout UK, who provide WiFi services like BT Openzone, T-Mobile and Megebeam etc. In Table 3 shows the cost and bandwidth comparison of these WiFi hotspots (Donoghue and Wearden, 2005).

Locations	Service Provider	Speed	Cost
Thistle Hotel, Tower Bridge	BT Openzone	2Mbps	£6/hour or £15 for 24 hours
Costa Coffee, Cowcross Street, Farringdon	BT Openzone	492Kbps	£6/hour or £15 for 24 hours
Benugo Sandwich Bar, Berwick Street	Broadscape	512Kbps	Half an hour free access when spend £2 (£4 at lunch time)
Starbucks, Wardour Street	T-Mobile	1Mbps	£5.50/hour or £16.50/day or £47/month
Internet Exchange, Covent Garden	Surf and Sip	1Mbps	£5/day or £20/month with 12 month contract or £30/month

Table 3: Costs and Bandwidth Comparison of WiFi Internet Access in London
Source: Donoghue and Wearden (2005)

6. Future Direction of NGNs (Next Generation Networks) & Conclusion

So, over the next 20-years, in both IT and Telecommunication sector, the future will be optical communications systems with the combination of wireless systems which will play a major role in communicating people, because the end users need is speed and reliability of service with security as well, but they also want flexibility, mobility with the same speed on their laptop as on the PC with wired connections. In future, it is quite possible that after the successful installation of FTTH/FTTB solutions, the home and office networks will be connected by wireless links, for the ease of use and mobility. The main concern is cost in fibre optic communication systems but if we look around us, although it has the high cost, but they are the only reliable, high speed communication systems available which are deployed everywhere in telecom sector and core IT networks, on the other hand, although wireless systems are considered unsecure, but still the world is moving towards them, the reason is flexibility, its human nature that we want ease in everything, and wireless systems gives natural attractions to us and that is why they are getting so popular these days.

The next 20 years scenario of communications systems will be according to this research is the complete takeover by fibre optic in telecommunications systems because cables and instruments costs are lowering down, and because fibre is reliable and they have longer life. All the telecommunication exchanges will be IP-based, and the whole setup will be replaced by routers. In IT-networks, all the core networks will be connected by self-healing ring based fibre networks, and use of fibre to the desktops will be common in office and universities. FTTH/FTTB will be deployed and from a single connection end user can get the services like, telephone, high speed internet access (Gbps data rates), <200 TV channels, and many more services. All the public places will have wireless Internet access for the end users; W-LANs will be commonly available in offices, stations, pubs, bars, university campuses are airports. Most of the computers, Laptops and PDA will be wireless enables, and the equipment for wireless network setup will be available cheap. Home networks will be wireless in next 20 years, and end user can use internet any where in the garage, living room or even in the toilet. Bluetooth devices will be cheaply available and will be in used. Places where it is difficult to deploy Internet setup will be connected with wireless links or satellite communications for Internet. The prices for satellite Internet will come down and it will be deployed in most of the developing countries with WiMAX like in Africa, Middle East, and Asian countries. The benefit of all this will directly go to the end user because, the high speed connectivity will be there either for end-users or for business users, the cost of the systems will be less, they will be secured and reliable because the link will be deployed using fibre optic from all over the area (from exchange to customers premises) for both business and home users.

Fibre optic systems in future will provide high speed, reliable links for local exchanges, office networks, Campus wide networks, LANs, SANs, MANs, surveillance systems. So, from the research it is clear that fibre and wireless systems both will going to stay as Next Generation Networks and we have to rely wireless as well as fibre for the current systems and for the future systems, we have to move on

to fibre based systems like FTTH/FTTB, with the combination of satellite communications systems and wireless systems.

7. References

Donoghue, A. and Wearden, G., (2005), "*Hot (spot) in the city - a Wi-Fi tour of London, Cost & Technology Comparison*", *ZDNet UK*, (Accessed on: 26-08-05) http://insight.zdnet.co.uk/ specials/wireless/0,39021194, 39115937,00.htm

FTTH Website, (2002), "*The time for fibre to the home is now*", (Accessed on: 08-08-05), http://www.ftthcouncil.org/documents/time%20for%20FTTH%20is%20NOW.pdf

Global Telecoms Business, (2005), "*How BT will implement 21CN, Global Telecoms Business*", (Accessed on: 12-08-05) http://www.globaltelecomsbusiness.com/default.asp?Page=7&PUB=45&ISS=14808&SID=500303

Intel Website, (2005a), "*Understanding Wi-Fi and Wi-Max as Metro-Access Solution*", Intel White Paper, (Accessed on 02-08-2005) http://www.intel.com/netcomms/technologies/wimax/304471.pdf

Intel Website, (2005b), "*Intel and WiMAX, Accelerating Wireless Broadband*", Intel Case Study, (Accessed on 10-08-2005) http://www.intel.com/standards/case/case_wimax.htm

Intel Website, (2005c), "*10-Giga Bit Ethernet Technology Overview*", Intel Technology White Paper, http://www.intel.com/network/connectivity/resources/doc_library/white_papers/pro10gbe_lr_sa_wp.pdf (Accessed on: 23-04-05)

PONForum, (2005), "*What is a PON, The Passive Optical Network Forum*", PONforum, http://www.ponforum.org/technology (Accessed on: 10-08-05)

Sauk County Website, (2005), "*Cost Analysis of Microwave link vs. Fibre*", http://www.co.sauk.wi.us/ data/tower/cost/micro_vs_fiber.htm

Tanenbaum A.S. (2002), *Computer Networks*, Fourth Edition, Prentice Hall PTR, ISBN: 0-1306-6102-3

Watermarking for Copyright Protection

J.P.Ashton and M.A.Ambroze

School of Computing, Communications and Electronics,
University of Plymouth, Plymouth, United Kingdom
e-mail: jamespashton@blueyonder.co.uk

Abstract

In recent years digital representations of images, audio and video have become increasingly popular and offer a number of advantages. The major disadvantage of digital media is that an unlimited number of perfect copies can be made and freely distributed. This poses a threat to content owners' rights and introduces the problem of copyright protection for such media. Digital watermarking has been proposed as a solution to the problem of copyright protection. The idea is that the copyright information is embedded into the media by means of a hidden watermark that is impossible to remove.

This paper proposes a blind spatial domain digital image watermarking scheme for copyright protection. The scheme is loosely based on the paradigm of spread spectrum communications whereby the narrowband watermark is spread to a wideband signal using an m-sequence and the spread spectrum watermark is directly embedded into the image data. The retriever uses cross correlation techniques to retrieve the embedded watermark. Although the proposed scheme is functional, producing an imperceptible watermark, it is not robust to common image processing techniques without embedding a watermark with such high amplitude that the image quality is degraded.

Keywords

Digital Watermarking, Spread Spectrum Communications, M-sequences

1. Introduction

In recent years digital representations of copyrighted material such as images, movies and songs have become widely available and offer many advantages. There is however one major disadvantage which is that an unlimited number of perfect copies of these types of material can be easily produced leading to large scale unauthorised copying. Furthermore, with the ever increasing popularity of the internet and the introduction of broadband internet access into many homes, distribution of these illegal copies of copyrighted material is very quick and easy. This poses a serious threat to the rights of content owners and undermines the music, film, book and software industries. In the past, using analogue techniques such as photocopying or recording an audio CD to a magnetic tape, there was a significant loss in quality of the media, making it obvious when a copy was being used. Nowadays, with the ability to make perfect copies of digital media and quickly distribute them via the internet, computer networks and mobile hard disk drives, how does a content owner know when a copy is in use or where it has come from?

Digital watermarking provides a solution to the problem of the copyright protection of digital images as it allows content owners to permanently embed their own details into images for proof of copyright ownership and also consumers' details when they purchase copies of the images, so as to trace the source of any illegal copies. Therefore, digital watermarking can help protect digital image owners' rights by enabling them to prove that they own the copyright to the image and also allowing them to trace any illegal image copies back to the consumer that originally distributed them, which then provides the option of prosecuting them.

The aim of this research was to investigate ways to embed copyright material into digital images in order to create a highly robust watermarking scheme. The watermarked image was subjected to some common image processing techniques such as scaling and JPEG compression so that the robustness of the watermark could be assessed and the watermark embedding process improved.

2. An Overview of Digital Image Watermarking

The basic idea of watermarking can be expressed as that of adding a watermark signal (copyright information) to some form of digital media (image) in such a manner that the watermark itself is secure and unobtrusive. At the same time the watermark signal must be partially or fully recoverable from the watermarked digital media when the correct key, required for recovery, is used.

In the case of robust invisible image watermarking the watermark must be imperceptible to the human observer. As a consequence of this requirement, the amount of allowable modification to the image data (i.e. the pixels or transform coefficients) must be very small when compared to their average amplitude. The size of this modification is referred to as the amplitude of the watermark. Therefore, in order to ensure imperceptibility of the watermark, the amplitude of the watermark must be small in comparison to the average amplitude of the image data.

The restriction discussed above results in a reduction in the robustness of the watermark. For a highly robust watermark the amplitude of the watermark is required to be large. To overcome this issue the watermark information is redundantly distributed over many samples of the image data allowing for a reduction in watermark amplitude whilst providing a substantial change to the image data that can be detected by the watermark retriever, in order to successfully retrieve the watermark. Watermarking in this way 'means that the watermark can usually be recovered from a small fraction of the watermark data, but the recovery is more robust if more of the watermark data are available for recovery.' (Hartung et al 1999) This method offers better resistance to attacks on the watermark and makes it very hard to remove without destroying the image since such a large amount of information in the image data is watermarked.

2.1 The Generic Digital Image Watermarking Process

The design of a digital image watermarking scheme consists of three main design steps:

1. Design of the watermark data to be embedded into the media data. The watermark data typically consists of the watermark information combined, in some way, with a key, so that it can be easily combined with the media data.
2. Design of the watermark embedder, which combines the watermark data with the media data to produce the watermarked media data.
3. Design of the watermark retriever, which detects or extracts the watermark information from the watermarked media data using the key used in the embedder and possibly the original media data.

The first two steps of the process are commonly regarded as one step as they can both be incorporated in the design of the watermark embedder.

The first step in the generic digital image watermark embedding process is to encode the combined watermark information and key into a form that can be easily combined with the image data. At this point, the use of a secret or public key can be decided, in order to enforce security if desired. Often, when watermarking images, watermarks are encoded as two-dimensional, spatial patterns. The watermark embedder then combines the seemingly random encoded representation of the watermark with the image data, which may be uncompressed or compressed. If the watermark embedding process is designed correctly, the resulting watermarked image appears identical to the original when perceived by a human, but will yield the encoded watermark information when processed by the watermark retriever.

In the generic digital image watermark retrieval process the retriever is passed the watermarked image and the key and depending on the method used, the original image and the original watermark. The output from the retriever is simply the retrieved watermark.

Most watermark retrieval processes require certain information to retrieve watermarks. This information can be referred to as a key, similar to keys used in cryptography. Keys can either be secret, in which case they are only made available to authorised people, or public, in which case they are made available to anyone. The level of availability of the key, in turn, determines who is able to read the watermark.

Generally, image watermarking systems use one or more secret keys to guarantee security of the embedded watermark against erasure and manipulation.

2.2 The Requirements of a Digital Image Watermarking Scheme

There are a number of characteristics that are important for a digital image watermarking scheme to possess. The importance of these characteristics may however be relative to a particular image watermarking application or technique.

The following list summarises the requirements for an invisible robust digital image watermarking scheme:

- **Data Payload:** The data payload of a watermark refers to the amount of information the watermark can carry. The amount of information required in an image watermark used for copyright protection could be high. Therefore, the data payload of the watermarking scheme must be high.
- **Security:** Once embedded, the watermark should be secret and unauthorised removal of the watermark should be impossible. This is usually achieved through the use of a secret key.
- **Robustness:** The watermark, once embedded, should remain in the watermarked image no matter what image processing or intentional attacks on the watermark should occur. The watermark should remain robust to the point where any attempts to remove it would destroy the image before the watermark is lost.
- **Imperceptibility:** The watermark embedded into the digital image must be imperceptible to the human observer and the watermarked image must appear identical to the original image.
- **Complexity:** The watermarking scheme should ideally have a low complexity. The complexity will however depend on the watermarking approach that is adopted which, in turn will affect the robustness of the watermark. For example, a transform domain watermarking scheme is generally more robust than a spatial domain scheme, but at the same time requires a higher level of complexity.
- **Tamper Resistance:** It should not be possible to modify the watermark in such a way as to create a different but valid watermark.
- **False Positive Rate:** The false positive rate of a watermark retrieval scheme is the probability that it will identify an unwatermarked image as containing a watermark. The false positive rate of the scheme should be as low as possible.
- **Watermark Retrieval:** Retrieval of the watermark should not require the use of the original image or the original watermark as this would not be practical for a copyright protection system.

Whilst all of the above requirements are desirable for a digital image watermarking scheme to possess, this is very hard to achieve in practice. Many schemes have been proposed over recent years that possess different mixtures of the above requirements but none that satisfy them all. Examples of such schemes can be found in Caronni (1995), Cox et al. (1996 and 1997) and Ruanaidh et al. (1996).

When designing a digital image watermarking scheme certain requirements will be more desirable than others and compromises will have to be made. Two of the key requirements in an image watermarking scheme are robustness and imperceptibility. As discussed previously, an increase in robustness results in a reduction in imperceptibility and an increase in imperceptibility results in a reduction in robustness. Therefore, this is a key trade off in the design of a digital image watermarking scheme and must be carefully considered. The optimum trade off will depend on the particular application.

3. Methodology

A digital image watermarking scheme can be viewed as a hidden communications channel in which, the signal being transmitted is the watermark and the image is noise. In this situation the noise power is significantly higher than the signal power and therefore, the signal to noise ratio (SNR) of the communications channel is very low, making it very difficult to successfully receive the transmitted signal. Spread spectrum communications techniques offer a means of successfully transmitting data through a noisy communications channel in which the noise power is higher than the signal power. For this reason a spread spectrum watermarking approach was selected for use in the proposed watermarking scheme. A spread spectrum watermarking approach also offers the following advantages:

1. Theoretically it is possible to embed the watermark with low enough amplitude for the watermark to remain imperceptible.
2. The embedded watermark is a series of pseudorandom noise (PN) sequences with statistical properties similar to those of white Gaussian noise and therefore, the watermark, in theory, can not be estimated from the watermarked image without knowledge of the PN sequence.
3. The watermark can be embedded into the perceptually significant components of the image (due to advantage 1), which theoretically, increases the robustness of the watermark to image processing techniques. This is due to the fact that many image processing techniques modify or discard perceptually insignificant components of the image and leave the perceptually significant components unchanged. Therefore, much of the watermark would remain in tact. Furthermore, an attacker would have to modify perceptually significant components of the image in order to remove the watermark and in doing so, would most likely degrade the quality of the image.

Block diagrams of the proposed watermark embedder and watermark retriever are shown in figures 1 and 2 respectively.

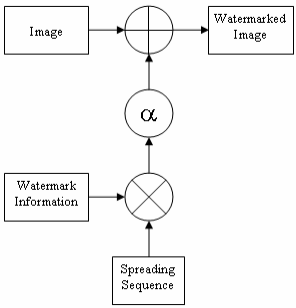


Figure 1: Watermark Embedder

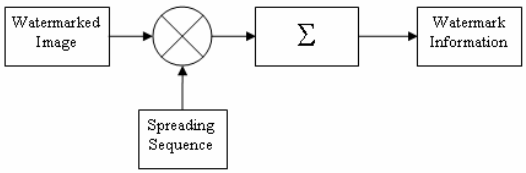


Figure 2: Watermark Retriever

The inputs to the watermark embedder are the image to be watermarked, a maximum length shift register sequence (m-sequence) having a user defined chip rate of m

chips and a user defined amplification factor, α. In the watermark embedder the watermark information is converted to a binary string and the string spread using the m-sequence. The spread spectrum binary string is then amplified according to α. The spread spectrum watermark with constant amplitude is then added directly to the image data (pixels) in a sequential manner from the top left of the image data, row by row, to the bottom right of the image data. The watermark embedder then yields the watermarked image.

The inputs to the watermark retriever are the watermarked image and the m-sequence used in the watermark embedder. In the watermark retriever, starting from the top left of the image data, the cross correlation, Σ between the first m pixels of the image data and the m-sequence is calculated. A negative cross correlation is considered a binary 0 and a positive cross correlation is considered a binary 1. The resulting binary digit is the first digit of the retrieved binary watermark string. This process is repeated using the next m pixels to retrieve the next binary digit. This process is then repeated using the next m pixels in a sequential manner, line by line, each time, until the whole binary watermark is retrieved. The retrieved binary watermark it then converted to ASCII and the watermark retriever yields the retrieved watermark information.

Once the original watermarking scheme was implemented the original watermark retriever was modified in order to improve the watermark retrieval process. A block diagram of the improved watermark retriever is shown in figure 3.

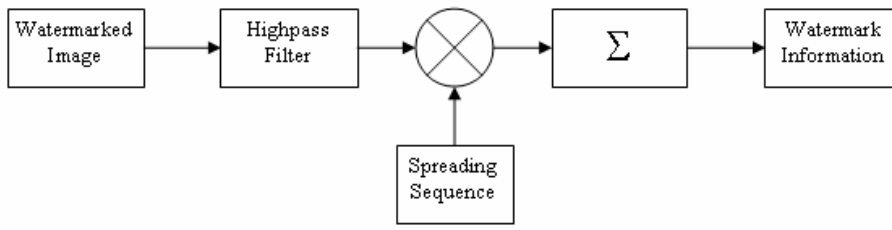


Figure 3: Improved Watermark Retriever

The improved watermark retriever is almost identical to the original but with the addition of a highpass image filter. The retrieval process is the same as in the original but the watermarked image is highpass filtered before the cross correlation is performed. Consider the watermarked image in the frequency domain, the high powered image data resides in the low frequency components of the spectrum and the low powered spread spectrum watermark covers most of the frequency spectrum, as we would expect since it is intentionally made to be a noise like signal. Remember, the watermarking process can be considered as a hidden communications channel in which the image is noise and the watermark data is the signal. SNR is the ratio of signal power to noise power and if the majority of the noise can be removed, the noise power is reduced and the signal to noise ratio increases and therefore, there is a higher chance of successful watermark retrieval. If the watermarked image is highpass filtered, the majority of the image data (noise) will be removed as well as some of the watermark data, leaving mostly the watermark data (signal).

4. Results

A 646 x 484 bitmap (BMP) image was watermarked several times using an m-sequence with 8,191 chips and 147 characters of watermark information so that the maximum number of image pixels would be modified when the watermark was embedded. Each time the image was watermarked the watermark amplitude was increased so as to assess the maximum watermark amplitude that could be used before the watermark was perceptible. It was found that the threshold of human perception was reached when α=4.

Various images were watermarked and then scaled. Attempts were made to retrieve the watermarks in order to assess the schemes robustness to image scaling. None of the attempts were successful.

An image was watermarked with α=4 (this was the highest practical value that could be used as this value was bordering on the threshold of human perception) and compressed using JPEG compression. The image was then converted back to a BMP and watermark retrieval was attempted in order to assess the schemes robustness to JPEG compression. This process was repeated for various m-sequences. None of the attempts were successful.

A 2048 x 1536 BMP image was repeatedly watermarked with 40 watermark characters using a selection of m-sequences. Each time the image was watermarked the watermark amplitude was increased by 1 and the number of successfully retrieved characters, using the original watermark retriever and the improved watermark retriever, were recorded. Figures 4, 5 and 6 show the results obtained using m-sequences with 255, 1,023 and 4,097 chips respectively. Figure 7 shows the results obtained using an m-sequence having 23,767 chips and only 30 characters of watermark information. 30 characters of watermark information were used with this m-sequence due to the limited number of pixels available in the image.

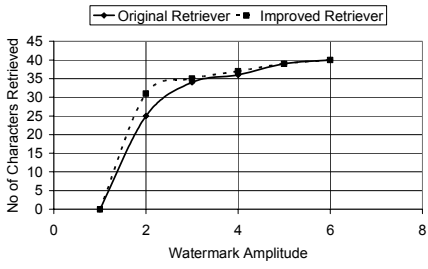


Figure 4: 255 Chips

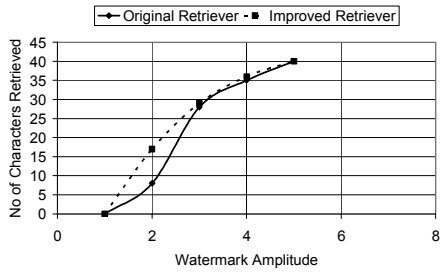


Figure 5: 1,023 Chips

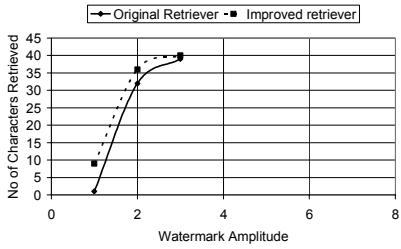


Figure 6: 4,097 Chips

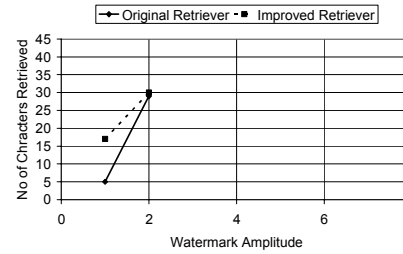


Figure 7: 23,767 Chips

5. Discussion

A functional watermarking scheme was designed and implemented in which a BMP image could be watermarked with an imperceptible watermark and the watermark successfully retrieved. The scheme was not however, robust to common image processing techniques. This limitation was due to the relatively low watermark amplitude that must be used in order for the watermark to remain imperceptible and was therefore caused by deficiencies in the watermark embedding procedure.

Whilst α=4 provided a watermark that was on the threshold of human perception for the image shown in the results, this result is not true for all images. Watermarks are more likely to be perceptible in certain areas of images, such as where there is very little variation in the pixel intensities (e.g. black background or the sky). Watermarks are less likely to be perceptible in areas where the pixel intensities are varying (e.g. grass or trees). Depending on the image content, an increase or reduction in watermark amplitude may be desirable.

From the results it is clear that the improved watermark retriever does offer watermark retrieval performance gains, as we would expect from the theory. It is also clear from the results that the improvement that is gained increases with increasing m-sequence length. This is due to the fact that using a larger m-sequence, increases the SNR after highpass filtering and therefore, the probability of bit error is reduced and there is more chance of successfully receiving the transmitted signal (i.e. the watermark information).

6. Conclusion

The watermarking scheme proposed in this paper is a blind spatial domain image watermarking scheme, loosely based on the principles of spread spectrum communications. The scheme provides an imperceptible watermark for BMP images and is intended for use in copyright protection applications. However, the scheme is somewhat weak and not robust to common image processing techniques and is therefore not suitable for copyright protection applications.

Whilst the research carried out has provided a good insight into the problem of digital image watermarking, there is scope for further development of the proposed scheme. Future research should:

1. Concentrate on investigating ways to improve the robustness of the watermark to image processing techniques.
2. Investigate HVS models and adapt the watermark embedder to adopt an adaptive watermarking approach in order to improve the robustness of the watermark.
3. Investigate the effects of embedding the watermark data in the R, G or B pixels only or perhaps embedding the first spread spectrum bit in the R pixels, the second in the G pixels and the third in the B pixels and so on.
4. Investigate methods of applying forward error correction coding to the embedded watermark in order to further improve the robustness of the watermark.

Digital image watermarking has been proposed as a solution for copyright protection of digital images and is an exciting and interesting subject area. The subject has gained large international interest over recent years and remains a very active area of research. As yet there is no flawless solution to the problem of digital image copyright protection and while image piracy continues and expands, research in this area will continue until flawless solutions are devised.

7. References

Caronni, G. (1995), "Assuring ownership rights for digital images", *Proceedings of Reliable IT Systems*, Vieweg Publishing Company, pp251-263.

Cox, I., Kilian, J., Leighton, F. T. and Shamoon, T. (1996), "Secure spread spectrum watermarking for images, audio and video", *Proceedings of the International Conference on Image Processing*, Vol. 3, pp243-246.

Cox, I., Kilian, J., Leighton, F. T. and Shamoon, T. (1997), "A secure, robust watermark for multimedia", *IEEE Transactions on Image Processing*, Vol. 6, No. 12, pp1673-1687.

Hartung, F. and Kutter, M. (1999), "Multimedia Watermarking Techniques", *Proceedings of the IEEE*, Vol. 87, No. 7, pp1079-1107.

Ruanaidh, J. J. K. O., Dowling, W. J. and Boland, F. (1996), "Watermarking Digital Images for Copyright Protection", *IEE Proceedings on Vision, Signal and Image Processing*, Vol. 143, No. 4, pp.250-256.

Interference Self-Cancellation Technique in M-QAM Modulated OFDM System

M.A.Yousuf and M.A.Abu-Rgheff

Mobile Communications Network Research, School of Computing, Communications and Electronics, University of Plymouth, United Kingdom
e-mail: mosa@plymouth.ac.uk

Abstract

The synchronisation is an important task in any digital communication system without proper synchronisation the transmitted data is not received properly. The OFDM signal waveform synchronisation can be performed either in time or frequency domain.

one of the major problem in designing orthogonal frequency division multiplexing (OFDM) systems is their inherent sensitivity to any frequency shift in the signal, a frequency offset between the local oscillators at the transmitter and receiver causes single frequency shift in a signal, while time varying channel can cause a spread of frequency shifts known as the Doppler spread., frequency shift ruin the orthogonality of OFDM sub carrier and cause inter-carrier interface (ICI), the amount of ICI for sub carriers in the middle of the OFDM spectrum is approximately twice as large as that for sub carrier at the band edges, because the sub carrier in the middle has interfacing sub carrier on both sides, so there are more interferes, within a certain frequency distance. Therefore quickly diminishing the performance of the system.

In this project, the impact of ICI has been analysed and solutions to combat ICI have been presented. The method is a self-cancellation scheme, in which redundant data is transmitted onto adjacent sub-carriers such that the ICI between adjacent sub-carriers cancels out at the receiver

Keywords

ICI self-cancellation, inter-carrier interference, multi-carrier modulation, OFDM

1. Introduction

Orthogonal frequency division multiplexing (OFDM) is a multi-carrier transmission technique to combat multi-path fading in wireless communications. The basic principle of OFDM is to split a high data stream into a number of low rate streams that are transmitted simultaneously over a number of sub-carriers. One of the major reasons to employ OFDM is because of its robustness and high spectral efficiency against the frequency selective fading or narrow band interference. In single carrier system a single fade or interference will affect the entire system, but in the case of multi-path carrier system only a small percentage of sub carriers will be affected. Error correction coding could be used to correct the few erroneous sub carriers.

OFDM system is extremely sensitive to synchronisation errors and that is occurring

due to the oscillator impairments and sample clock differences. The demodulation of the received radio signal to base band, possibly via an intermediate frequency, involves oscillators whose frequencies may not be perfectly aligned with the transmitter frequencies. This results in a carrier frequency offset. Many estimation has been approached to estimate and correct the carrier frequency offset (CFO) before demodulation. To mitigate the frequency offset problem. Currently three type of approaches have been proposed including frequency domain equalisation (Ahn and Lee, 1993; Dhahi *et al.*, 1996), time domain windowing, (Li and Stette, 1995; Muschallik, 1996), and the ICI self-cancellation scheme (Zhao and Häggman, 1996; Alam, 1997), this paper concentrates on the third method.

ICI self-cancellation scheme is a very simple way for suppressing ICI in an OFDM. The main idea is to modulate the input data symbol onto a group of sub carriers with predefined coefficients such that the generated ICI signals within that group cancel each other, hence the name self-cancellation**.** Moreover ICI is cancelled out by repeatedly modulating a symbol on two adjacent sub carriers with an 180^{o} phase difference between them. In system2, In the ICI cancellation modulation block, the same signal must be mapped onto an adjacent pairs of the sub carriers rather onto single sub carriers so sub carriers must be divided into even and odds, the same data to be transmitted must be manipulated on to a even sub carrier to get them cancel out with odd ones at ICI cancellation demodulation module. This is what I am using the simplest technique for ICI cancellation Scheme After giving work in (Zhao and Häggman, 1996; Alam, 1997), the further discussions of the ICI self-cancellation scheme are given in (Armstrong, 1999; Seaton and Armstrong, 2000), where the scheme is also called polynomial cancellation coding (PCC).

2. Analysis of inter-carrier interference

The main drawback of OFDM, however, is its vulnerability to small differences in frequency at the transmitter and receiver, normally referred to as frequency offset. This frequency offset can be caused by Doppler shift due to relative motion between the transmitter and receiver, or by differences between the frequencies of the local oscillators at the transmitter and receiver. In this project, the frequency offset is modelled as a multiplicative factor introduced in the channel, as shown in Figure 1.

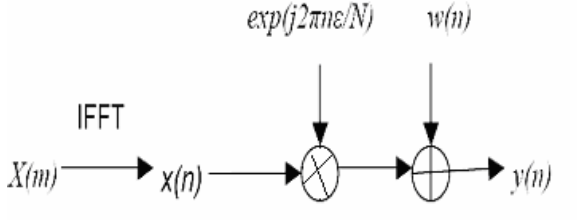


Figure 1: frequency offset model

The received signal is given by:

$$y(n) = x(n)\, e^{\frac{j2\pi n\varepsilon}{N}} + w(n) \quad \textbf{(1)}$$

Where ε is the normalised frequency offset, and is given by $\Delta f N T_s$. Δf is the frequency difference between the transmitted and received carrier frequencies and T_s is the subcarrier symbol period. w(n) is the AWGN introduced in the channel.

The effect of this frequency offset on the received symbol stream can be understood by considering the received symbol Y(k) on the k^{th} sub-carrier. From (2) we found that the received symbol Y(k) on the k^{th} sub-carrier is,

$$Y(k) = X(k)S(0) + \sum_{l=0, l\neq k}^{N-1} X(l)S(l-k) + n_k \qquad k = 0,1,.....N-1 \quad \textbf{(2)}$$

All the equations in this section has been taken from (Y. Zhao and S. Haggman, 2001). Where N is the total number of subcarriers, X(k) is the transmitted symbol (M-ary Quadrature Amplitude Modulation (M-QAM), for example) for the k^{th} subcarrier, n_k is the additive noise and S(l-k) are the complex coefficients for the ICI components in the received signal. The ICI components are the interfering signals transmitted on sub-carriers other than the k^{th} sub-carrier. The complex coefficients are given by

$$S(l-k) = \frac{\sin(\pi(l+\varepsilon-k))}{N\sin(\pi(1+\varepsilon-k)/N)} \exp(j\pi\ (1-\frac{1}{N}(l+\varepsilon-k)\,)) \quad \textbf{(3)}$$

The carrier-to-interference ratio (CIR) is the ratio of the signal power to the power in the interference components. It serves as a good indication of signal quality. The equation of CIR has given in (Zhao and Haggman, 2001)

3. ICI self-cancellation scheme

ICI self-cancellation is a scheme that has been introduced by Yuping Zhao and Sven-Gustav Haggman in 2001 in (Zhao and Haggman, 2001) to combat and suppress ICI in OFDM. Succinctly, the main idea is to modulate the input data symbol onto a group of subcarriers with predefined coefficients such that the generated ICI signals within that group cancel each other, hence the name self- cancellation.

4. ICI cancelling modulation

The ICI self-cancellation scheme requires that the transmitted signals be constrained such that

$$X(1) = -X(0), X(3) = -X(2),.....X(N-1) = -X(N-2) \quad \textbf{(4)}$$

Using equation 3, this assignment of transmitted symbols allows the received signal on subcarriers k and k + 1 to be written as

$$Y'(k) = \sum_{\substack{l=0 \\ l=even}}^{N-2} X(l)[S(l-k) - S(l+1-k) + n_k \quad \textbf{(5)}$$

$$Y'(k+1) = \sum_{\substack{l=0 \\ l=even}}^{N-2} X(l)[S(l-k-1) - S(l+1-k) + n_{k+1} \quad \textbf{(6)}$$

And the ICI coefficient S'(l-k) is denoted as

$$S'(l-k) = S(l-k) - S(l+1-k) \quad \textbf{(7)}$$

It is seen that on algorithmic scale |S' (l-k)| is much less then |S (l-k)| for most of the l-k values. Hence, the ICI components are much smaller in (7) than they are in (3). Also, the total number of interference signals is halved in (7) as opposed to (3) since only the even subcarriers are involved in the summation.

5. ICI cancelling demodulation

The received signal is redundant and this redundancy is introduced by ICI modulation since each pair of subcarriers transmit only one data symbol. This redundancy can be exploited to get better the system power performance, while it certainly decreases the bandwidth efficiency. To get benefit of this redundancy, the received signal at the $(k+1)^{th}$ $(k + 1)^{th}$ subcarrier, where k is even, is subtracted from the k^{th} subcarrier. This is expressed mathematically as:

$$Y''(k) = Y'(k) - Y'(k+1) \quad \textbf{(8)}$$
$$= \sum_{\substack{l=0 \\ l=even}}^{N-2} X(l)[-S(l-k-1) + 2S(l-k) - S(l-k+1) + n_k - n_{k+1}$$

Subsequently, the ICI coefficients for this received signal becomes

$$S''(l-k) = -S(l-k-1) + 2S(l-k) - S(l-k+1) \quad \textbf{(9)}$$

If compared both previous ICI coefficients | for the standard OFDM system i.e. S(l-k)| and for the ICI cancelling modulation i.e. |S'(l-k)|, then |S''(l-k)| has the smallest ICI coefficients, for the majority of l-k values, followed by |S'(l-k)| and |S(l-k)|. The combined modulation and demodulation method is called the ICI self-cancellation scheme.

6. Results and discussions

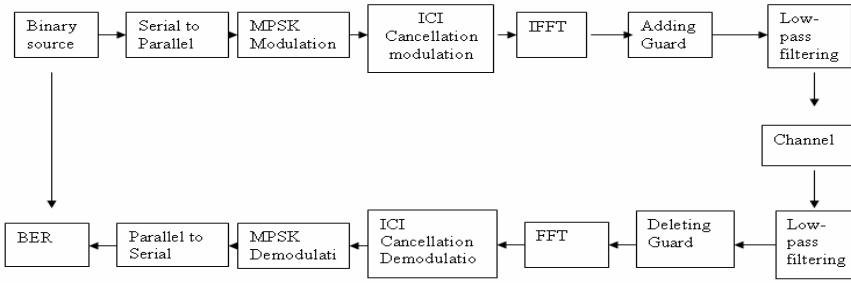


Figure 2: Structure of ICI self-cancellation system

Three types of systems have been considered for comparisons.

1. standard M-QAM modulation OFDM system without ICI self cancellation
2. M-QAM modulation OFDM system with proposed scheme
3. (Zhao and Häggman, 2001) (system3)

The block diagram of the ICI self-cancellation scheme(system2) is showing in Fig(2), while block diagram of standard M-QAM modulation OFDM system without ICI self cancellation(system1) can be obtained by simply removing the "ICI cancellation modulation" and "ICI cancellation demodulation" blocks, the bandwidth of both systems have same 1/bits/Hz/s. and the third system (system3)can be obtained by simply add differential coding block before ICI cancellation modulation block, the same values of Eb/No has been set to analyse the BER performance which gives the fair comparison between them.

By using Equation (4) in ICI cancellation modulation block same signal must be mapped onto an adjacent pairs of the subcarriers rather onto single subcarriers so subcarriers must be divided into even and odds, the same data to be transmitted must be manipulated on to a even subcarrier to get them cancel out with odd ones at ICI cancellation demodulation module. This is what I am using the simplest technique for ICI cancellation Scheme.

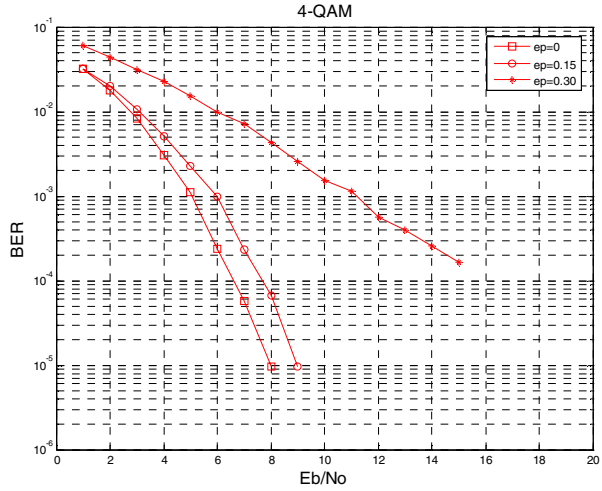


Figure 3: Effect of changing the frequency offset on Bit Error Rate for M = 4, QAM

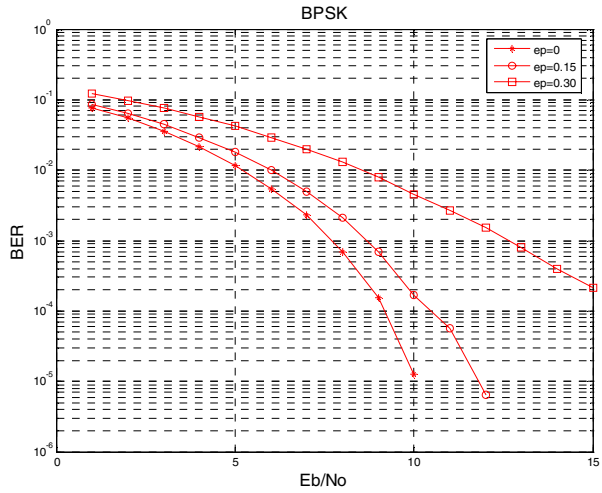


Figure 4: Effect of changing the frequency offset on Bit Error Rate for BPSK

The above result shows that increase the frequency offset hence degradation of performance, but in the case of BPSK, severe frequency offset like 0.30 does not deteriorate the performance too greatly. But for QAM with an alphabet size 2 performance get worst more quickly if the frequency offset is small, however the M-QAM has a better performance by mean of BER it has a lower BER than the BPSK, if we go higher alphabet size, but the BER of 4QAM varies more dramatically by increasing the frequency offset I.e. ε than that of BPSK as you can read from the graph if you pick 7th value of the graph at frequency offset 0.15,the Eb/No is 3dB at 0.002 of BER but at frequency offset 0.30, at same value of BER the value of Eb/No is coming out 7dB at same value so the difference is around 4dB. If we chose same

value to analyse the effect that how rapid it get worst at high frequency offset at 4QAM so we will analyse that at same value the BPSK graph at frequency offset 0.30 does not deteriorate the signal than that of 4QAM because the difference of two signals at frequency offset 0.15 and 0.30 at same point (7^{th}) is not too high (2dB difference) then that of 4QAM(4dB difference at same point), but still 4QAM has a better BER than that of BPSK, hence it is analysed that larger alphabet sizes are more sensitive to ICI. So it is evident that if ε become larger than the desired part decreases and undesired part increases I.e. |S (0)| and |S (l-k)| respectively from equation (3). So it is impossible to reduce down ICI except the ε values can not be decreases and it is only possible by increasing the sub carrier separation, but the bandwidth efficiency would be reduced as the time-domain symbol length is reduced so the cyclic prefix will take a large portion of useful signal. But by keeping small BER constant if we increase the alphabet sizes than we can have the better performance as shown in fig (5) below , at ε = 0.05 if we are increasing the alphabet size than performance get better more rapidly. As mentioned above, the bandwidth efficiency get reduce in this scheme by half. This could be overcome by transmitting signals of larger alphabet size. For example, using 4QAM modulation together with the ICI self-cancellation scheme can provide the same bandwidth efficiency as standard OFDM 1 bit/Hz/s. using the hypothetical results the improvement of the CIR must increase the power efficiency in the system and gives better results for the BER. Hence, there is a trade off between power and bandwidth in the ICI SC scheme.

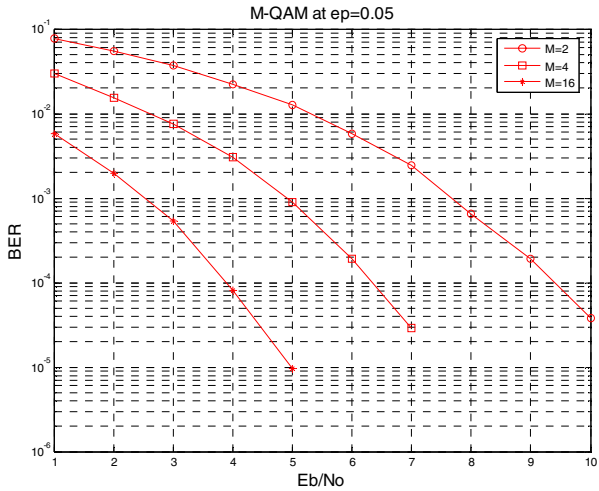


Figure 5: Effect of changing M on Bit Error Rate for ε = 0.05

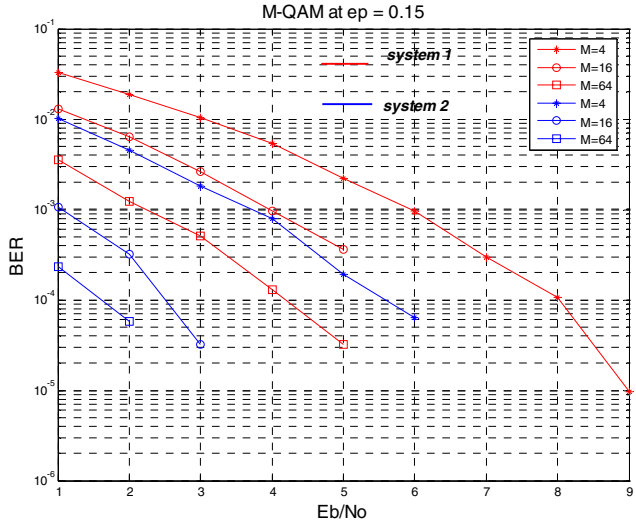


Figure 6: Comparison of Standard OFDM (system 1) with the proposed system (system2) with different values of M at ε =0.15

Method	M=4	Gain	M=16	Gain	M=64	Gain
System 1	6dB		4dB		2.2 dB	
System 2	3.8 dB	2.2 dB	1.0dB	3dB	XXX	XXX

Table 1: Required SNR and improvement of BER of 10^{-3} for M-QAM at ε =0.15

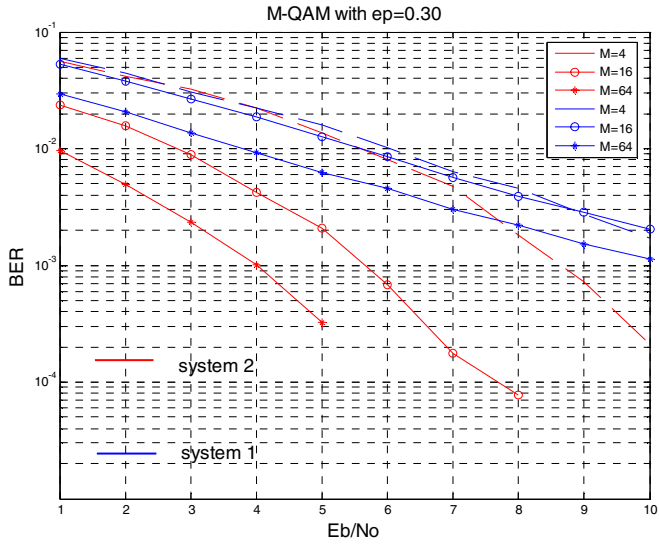


Figure 7: Comparison of Standard OFDM (system1) with the proposed system (system2) with different values of M at ε =0.30

Method	M=4	Gain	M=16	Gain	M=64	Gain
System 1	13dB		14dB		10.2dB	
System 2	8.7dB	4.3	5.7dB	8.3dB	4dB	6.2dB

Table 2: Required SNR and improvement of BER of 10^{-3} for M-QAM at ε =0.30

For the frequency offset of 0.15 at fixed BER of $10^{-3,}$ the gain improvement in Self cancellation scheme (SC) (system2) as compared to the standard OFDM(system1) is approximately 3dB as tabulated in table (1), this mean that same BER is achieved with the self cancellation scheme(system2) using a lower SNR and hence reduced the power consumption,

In table (2) the SNR of the standard system(system1) and self cancellation schemes system (system2) is again compared for higher value of frequency offset of 0.30, it is observed that at high frequency offset and high alphabet size the gain improvement is not significant,

Further it has been observed from the graph that in the presence of small frequency offset and small binary alphabet size, self cancellation scheme gives the best results. The frequency offset ε =0.30 Self cancellation scheme does not offer much increase in performance, however the self-cancellation scheme does not completely cancel the inter-carrier-interference (ICI) from the adjacent sub carrier so the effect of this residual ICI increases for larger alphabet sizes and frequency offset values.

7. Numerical comparisons between proposed system (System-2) and Zhao & Haggman system (System-3)

Eb/No	5	6	7	9	10
BER	0.04	0.03	0.01	0.002	0.0009

Table 3: System 3 at ep= 0.3

Eb/No	5	6	7	9	10
BER	0.015	0.008	0.0045	0.0007	0.0002

Table 4: System 2 at ep= 0.30

It is clear that the BER of the proposed system (system2) is better than the Zhao and Haggman system (system3), as shown in table (4) and table (3) respectively. I have proven that without differential coding at the same frequency offset, similar result even much better result can be obtained. This is what I have as a main contribution to this paper that I have proposed a system which has low complexity and produces better results than that of system3.

8. Conclusion

Self-cancellation scheme (system2) is not required any software for implementation or complex hardware. Even no channel equalisation is needed, and not differential coding is needed unlike system3, How ever it is not bandwidth efficient as there is a redundancy of two for each carrier, but by using 4-QAM the ICI self-cancellation scheme can provide the same bandwidth efficiency as standard OFDM 1 bit/Hz/s, however on low frequency offset and large alphabet size bandwidth efficiency can be achieved up to 8bits/Hz/s.

In this project simulation were performed in an AWGN channel. This model could be adapted easily by a flat-fading channel with perfect channel estimation. Further work have been done by performing simulations to investigate the performance of the ICI-self cancellation scheme in multi-path fading channels by using the effect of Doppler spread. So it has been analysed from the above graphs this scheme also works great in a multi-path radio channel with Doppler frequency spread, under the condition of large frequency offsets and same bandwidth efficiency, it is concluded that ICI-self cancellation scheme performs better than standard OFDM,

If the BER simulation results of the ICI self-cancellation scheme (system2) is compared to the result of error correction coding system (Zhao and Häggman, 2001) (system3), system2 give better performance (fig 6 and fig 7) than that of system3 , in addition if ICI-self cancellation scheme(system2) is not required error correction coding such system produce better performance and robust to both AWGN and ICI, as compared to system3 , hence we say that system2 has a low complexity and less overheads than that of system3 and system1.

The concept of self cancellation could be extended in the method of self cancellation scheme; data is mapped onto the pairs of sub-carriers. This results in cancellation of the ICI component which is due to the linear variation in weighting co-efficient over group of three adjacent co-efficient, mapping the data onto larger group of sub-carriers, higher order ICI cancellation can be achieved (Armstrong, 1999)

In summary the proposed scheme can be used as low-complexity. In practical it can be used where the lower rate and increase power consumption can be tolerated.

9. References

Ahn, J. and Lee, H.S., (1993), “Frequency domain equalisation of OFDM signal over frequency nonselective Rayleigh fading channels”, *Electron. Lett.*, vol. 29, no. 16, pp. 1476–1477, Aug.

Alam, S., (1997), “A general coding method to minimise intercarrier interference in OFDM mobile communication systems”, *Proceedings of the International Wireless and Telecommunications Symposium (IWTS’97)*, vol. 1, Malaysia, May 14–16, pp. 231–235.

Al-Dhahir N. and Cioffi, J.M., (1996), “Optimum finite-length equalisation for multicarrier transceivers”, *IEEE Transactions on Communications*, vol. 44, no. 1, pp. 56 – 64, January.

Armstrong, J., (1999), "Analysis of new and existing methods of reducing intercarrier interference due to carrier frequency offset in OFDM," *IEEE Transactions on Communications*, vol. 47, pp. 365–369, March.

Jeon, W.G., (2001), "An equalisation technique for orthogonal frequency-division multiplexing systems in time-variant multipath channels", *IEEE Transactions on Communications*, vol. 47, no. 1, pp. 27 – 32, July.

Li, R. and Stette, G., (1995), "Time-limited orthogonal multicarrier modulation schemes", *IEEE Transactions on Communications*, vol. 43, pp. 1269–1272.

Moose, P.H., (1994), "A Technique for Orthogonal Frequency Division Multiplexing Frequency Offset Correction", *IEEE Transactions on Communications*, vol. 42, no. 10, October.

Muschallik, C., (1996), "Improving an OFDM reception using an adaptive Nyquist windowing", *IEEE Transactions on Consumer Electronics*, vol. 42, no. 3, pp. 259 – 269, August.

Seaton, K.A. and Armstrong, J., (2000), "Polynomial cancellation coding and finite differences", *IEEE Trans. Inform. Theory*, vol. 46, pp. 311–313, January.

Zhao, Y. and Häggman, S., (2001), "Intercarrier interference self-cancellation scheme for OFDM mobile communication systems", *IEEE Transactions on Communications*, vol. 49, no. 7, pp. 1185 – 1191, July.

Zhao, Y. and Häggman, S.G., (1996), "Sensitivity to Doppler shift and carrier frequency errors in OFDM systems—The consequences and solutions," *Proceedings of the IEEE 46th Vehicular Technology Conference*, Atlanta, GA, Apr. 28–May 1, pp. 1564–1568.

Nonlinear Dynamical Analysis of EEG for Early Detection of Degenerative Diseases in the Brain

N.S.Soumahoro and N.J.Outram

School of Computing, Communications and Electronics, University of Plymouth, Plymouth, United Kingdom
e-mail: nicholas.outram@plymouth.ac.uk

Abstract

Alzheimer's disease (AD) is the most common degenerative brain disease characterised by mental deficits and behaviour disturbance. The electroencephalogram (EEG) has been for a long time a useful tool for diagnosis and prognosis of degenerative brain diseases. The EEG is assumed to be a nonlinear system. Nonlinear methods derived from the chaos theory are used to characterise nonlinear EEG activity in patients with Alzheimer's disease. The correlation dimension (D2) and the largest Lyapunov exponent (L1) are two common nonlinear methods which reflect the complexity of the cortical dynamics underlying EEG signals. The applicability of nonlinear methods is first tested. The presence of nonlinear dynamics is determined by comparing the time reversal asymmetry statistics of the original EEG data set with the one of the phase-randomised surrogate EEG data set constructed with the original data. The results show that EEG signals have been created by a nonlinear process. D2 and L1 are then computed to estimate the complexity of EEG patterns in patients with AD and healthy controls. D2 and L1 values from patients with AD are for most electrodes lower than those from healthy controls which indicate a decrease complexity of the dynamics underlying diseased brains. The D2 and L1 values patients with AD in early stage suggest that the occipital and temporal regions of the brain are the most affected by the disease and that the disease is more apparent in the left hemisphere of the brain. The brains of patients with AD thus show behaviours less chaotic than the one of normal controls.

Keywords

Alzheimer's disease, Chaos, EEG analysis, Correlation Dimension, Lyuapunov Exponent

1. Introduction

In 2002, about 3.6-10.3% of people aged 65 years or over in Western countries (Jeong, 2004) were affected by degenerative brain diseases that are progressive nervous diseases associated with the gradual deterioration of brain cells. Alzheimer's disease is one of the most common types of degenerative brain disorders. The main symptoms of this disease are the decrease or loss of cognitive abilities (such as memory and language functions) and motor functions (in the late stage of the disease) (Jeong, 2004; Lipsitz and Goldberger, 1992). The early detection of such diseases in order to apply effective treatments is becoming increasingly necessary (Jeong, 2004).

The main scope of this work is to apply methods derived from the Chaos Theory to the electroencephalogram (EEG) of a dozen subjects in order to distinguish people with AD from normal controls and thus define patterns of degenerative brain disease (Alzheimer's disease in particular). An aim is to independently replicate the work of Jeong *et al*. (1997 and 2001). It has been shown that these methods are an effective way to achieve this goal due to the properties of the brain and the electroencephalogram it produces. Indeed, as the brain is assumed to be a nonlinear system, the use of nonlinear methods is an appropriate way to obtain significant results (Jeong, 2004; Jeong *et al*., 1997 and 2001; Besthorn *et al*., 1995; Stam *et al*., 1995; Jelles *et al*., 1998).

The two methods used in this project to measure and compare the complexity and determine the features of the electrical activity of a diseased and normal brain are the Correlation Dimension (D2) and the maximal Lyapunov Exponent (L1). They are explained in Section 2. Section 3 presents the results of the experiments from this work and shows the difference in the values of D2 and L1 between patients with AD and normal controls. All these results are discussed in Section 4 and conclusions are given in Section 5.

2. Methodology

The brain is assumed to be a nonlinear system which means that its dynamics are governed by nonlinear paradigm such as sensitivity to initial conditions (Kantz and Schreiber, 2004). Moreover, its structure (neuronal network) and its behaviour have become increasingly complex with the evolution of the human race. Thus, to measure the complexity of brain activity, it is appropriate to use nonlinear dynamical analysis rather than simple statistics which ignore the dynamics of the signal. Applying nonlinear methods to EEG may help characterise the nature of the underlying systems that produce these observed signals. Furthermore, for any physiological system, aging and disease are linked to a loss of complexity of the system's dynamics (Lipsitz and Goldberger, 1992). Consequently, considering a brain with neurological damage (e.g. degenerative diseases), the complexity of the EEG should be far less than that of a "safe" brain.

The EEG recordings used in this project were recorded from the twenty-one scalp loci of the International 10-20 System. The study covered three different patients population: three patients with AD, one patient with AD in the early stage and seven normal controls. The data sets were recordings of 240s, each down sampled from 256Hz to 128Hz by averaging two consecutive sets of samples. The initial recordings were four minutes long and only data between 60s and 300s were used to avoid regions with electrical artefacts.

The EEG time series are first transformed into multi-dimensional embedded phase space. In that way, each state of the system, represented with a state vector x(n), corresponds to a single point in the phase space at discrete time n whose dimension is equal to the length of the state vectors. The evolution of x(n) in the phase space is known as trajectories. These trajectories follow a multi-dimensional manifold called

an attractor. The reconstruction of the attractor in phase space can be carried out through the technique of delay coordinates (Duke and Pritchard, 1992 and 1995; Kantz and Schreiber, 2004).

Let s(n) be the scalar observations: s(n) is a sequence which depends on the current state of the system. The delay reconstruction consists in taking the observed values at multiples of a fixed sampling to form a vector. A delay reconstruction in *m* dimensions is given by: $x(n) = (s(n-(m-1)\tau), s(n-(m-2)\tau), \ldots, s(n-\tau), s(n))$, with *m* the embedding dimension and τ the embedding delay. These two important parameters can be optimised when used to yield a faithful representation of the state space and recover the attractor of the system. A fairly good way to determine the optimum embedding delay τ is to use the first zero of the autocorrelation function of the signal (Kantz and Schreiber, 2004). The choice of the optimal embedding dimension is usually made by computing the studied algorithms for increasing values of *m*: the optimal *m* is the minimal one for which the result of the algorithm becomes invariant (Duke and Pritchard, 1992 and 1995; Kantz and Schreiber, 2004).

2.1 Surrogate data test

Before applying non-linear methods to the EEG, nonlinearity is tested for using the surrogate data technique. The rationale of this is that analysing linear systems using nonlinear methods do not produce results as concluding as those obtained using linear methods. The surrogate values are constructed from the original data sets in three steps: 1) a fast Fourier transform (FFT) is applied, 2) the resulting complex component is multiplied by random phase uniformly distributed in $[0, 2\pi]$, 3) an inverse FFT is applied to the result of the second step. The resulting surrogate data have the same power spectra and cross-spectra as the original data. However, the surrogate data differs from the original one in the sense that they have lost the phase and frequency relationship for the individual channels (Stam *et al*., 1995). The null hypothesis of surrogate data testing for linearity is that the data has been created by a stationary Gaussian linear process (Kantz and Schreiber, 2004). This null hypothesis is tested using the time reversal asymmetry statistic. The statistic of the surrogate data being higher than the one of original data then the EEG is assumed to be a produced by a nonlinear process which rejects the null hypothesis (Stam *et al*., 1995; Kantz and Schreiber, 2004). Nonlinear methods can be effectively applied.

2.2 Calculation of the correlation dimension

D2 measures the degrees of freedom of the attractor of the underlying system. This means that it estimates the number of independent variables necessary to characterise the system. D2 is expected to be a non integer value for a chaotic system. The higher the value of D2 the more complicated the behaviour of the system (Kantz and Schreiber, 2004). To compute D2 it is necessary to first compute the correlation sum that calculates the number of pairs of space vectors whose Euclidean distance is smaller than a given radius (Jeong *et al*., 1997 and 2001; Stam *et al*., 1995; Jelles *et al*., 1998; Duke and Pritchard, 1992 and 1995; Kantz and Schreiber, 2004). For all pairs of neighbouring state vectors x_i and x_j in a system which are closer than ε, the correlation sum is defined as (Kantz and Schreiber, 2004):

$$C(\varepsilon) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta\left(\varepsilon - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right)$$

Where Θ is the Heaviside function given by:

$$\Theta(x) = \begin{cases} 0, x \le 0 \\ 1, x > 0 \end{cases}$$

The correlation sum computes the average number of points on the reconstructed attractor within Euclidean distance of ε of each other. As the number of points N tends to infinity, and the distance ε tends to zero, the correlation sum will be proportional to ε^{D2}. Therefore, if the number of points is sufficiently large, and evenly distributed, a plot of the correlation sum versus ε on a double logarithmic scale will yield an estimate of D2:

$$D2 = \lim_{\varepsilon \to 0} \lim_{N \to \infty} d(N, \varepsilon) \qquad \text{Where} \quad d(N, \varepsilon) = \frac{\partial \ln C(\varepsilon, N)}{\partial \varepsilon}$$

To estimate the correlation dimension, one computes the slope of the linear part of the logarithmic plot (Duke and Pritchard, 1992 and 1995; Kantz and Schreiber, 2004). In practice, the estimation of D2 is not straightforward as the determination of the linear region of the double logarithmic plot is not precisely defined. For practical chaotic data sets, the formula that can be used to compute the correlation sum is the following (Kantz and Schreiber, 2004):

$$C(m, \varepsilon) = \frac{2}{(N - n_{\min})(N - n_{\min} - 1)} \sum_{i=1}^{N} \sum_{j=i+1+n_{\min}}^{N} \Theta\left(\varepsilon - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right)$$

A temporal offset n_{min} is used to exclude local neighbours from the sum.

2.3 Calculation of the maximal Lyapunov exponent

Chaotic systems demonstrate sensitive dependence on initial conditions. Small changes in initial conditions result in the trajectory of similar points in phase space rapidly diverging. This can be due to unpredictability or chaotic behaviour. For chaotic system the divergence of trajectories is exponential.

The maximal Lyapunov exponent estimates the mean exponential divergence (or convergence) of nearby trajectories of the attractor in phase space. A positive measure of L1 means that the system being investigated demonstrates a chaotic behaviour. The higher is the value of L1, the more complex is the behaviour of the nonlinear system (Kantz and Schreiber, 2004).

L1 can be computed doing the following: first estimate and plot the curves reflecting the evolution over the time of the distance between each space vector and it neighbours, and then calculate the slope of these curves (when they exhibit a robust

linear increase) to obtain the maximal Lyapunov exponent (Jeong *et al*., 1997 and 2001; Stam *et al*., 1995; Duke and Pritchard, 1992 and 1995; Kantz and Schreiber, 2004; Das *et al*., 2002).

Given a point on the attractor, the Largest Lyapunov Exponent is defined as (Das *et al*., 2002):

$$L_1 = \frac{1}{t_M - t_0} \sum_{k=1}^{M} \log_2 \frac{L(t_k)}{L(t_{k-1})}$$

Where $L(t_i)$ represents the distance between the initial point and its nearest neighbour at time t_i.

3. Results

The D2 values are computed using the researchers software (Figure 1) and verify using the TISEAN package (Figure 2). The L1 values (Figure 3) are only computed with the TISEAN software (Kantz *et al*., 2000).

The two first figures show that the correlation dimensions of EEG signals from patients with AD are lower than the ones of EEGs from healthy controls for the majority of the electrodes. The D2 and L1 values of the patient with AD in the early stage are generally lower than those of the healthy controls but slightly higher than those of the patients with AD in later stage. Figure 1 and figure 3 show that all patients with have significantly lower D2 and L1 in the occipital region than healthy controls. This result corresponds to Woyshville and Calabrese (1994) demonstration (Jeong, 2004). Moreover, the D2 and L1 values from most of the temporal regions of the patient with AD are considerably lower than those of normal controls which concurs with Jelles *et al* (1999a) results (Jeong, 2004). It can also be observe that the left hemisphere of the brain is more affected than the right one. These findings of a decreased dimensional complexity of EEGs in patients with AD support the results of a number of researches in the area of nonlinear dynamical analysis of EEG (Jeong, 2004: Besthorn *et al*., 1995; Jeong *et al*., 1997 and 2001; Stam *et al*., 1995; Jelles *et al*., 1998; Duke and Pritchard, 1992 and 1995). The decrease of D2 and L1 values of EEG of patients with AD may be associated with deficient information processing or communication in the brain injured (Jeong, 2004; Jeong *et al*., 1997 and 2001). Moreover, in the practical case of EEG analysis, L1 can be interpreted as the flexibility of information processing which means the “facility of the central nervous system to reach different states of information processing from similar initial states” (Jeong *et al*., 1997). Therefore, the brains of patients with AD exhibit a reduced flexibility of information processing.

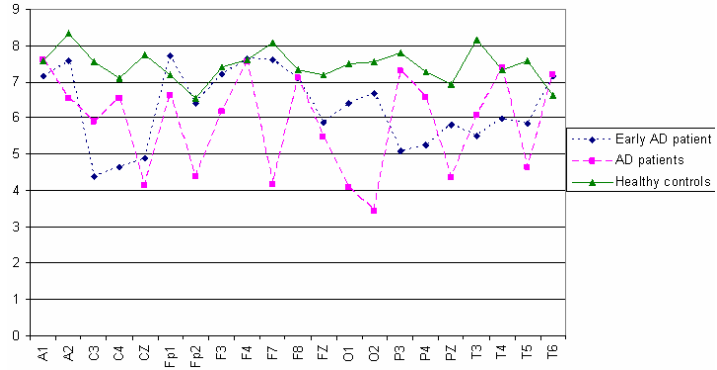


Figure 1: Correlation dimension of EEG data from patients with AD and healthy controls

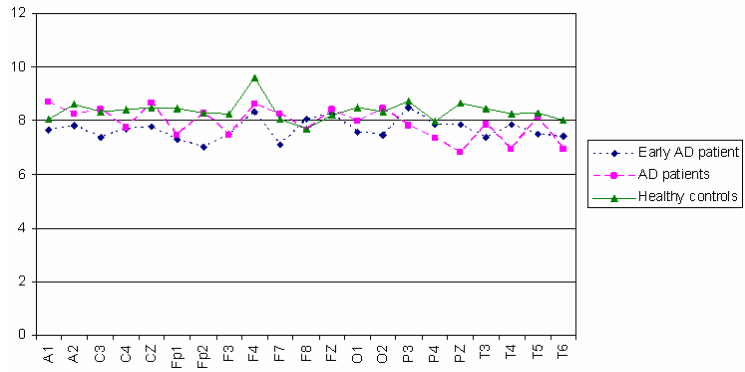


Figure 2: Correlation dimension of EEG from patients with AD and healthy controls using TISEAN

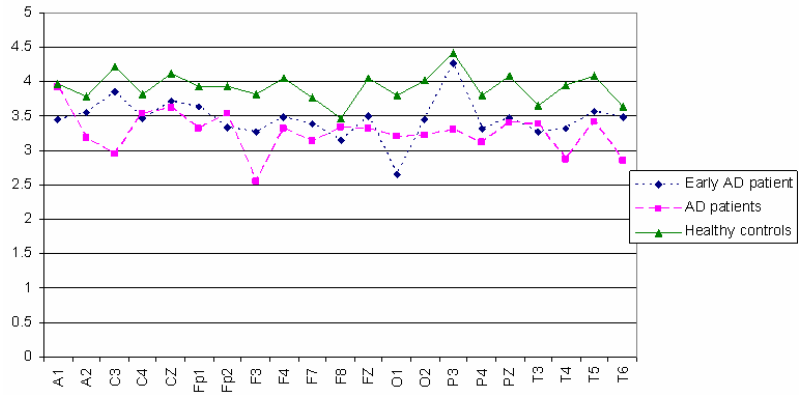


Figure 3: Maximal Lyapunov exponent of EEG from patients with AD and healthy controls using TISEAN

4. Discussion

The results obtained suggest that nonlinear dynamical analysis of EEG may be a useful tool in diagnosis AD more accurately and earlier. However these findings are limited by practical considerations: 1) The number of EEG data sets was not large enough to make a reliable generalisation of the results. Ideally, the statistical significance of these results would have been calculated. 2) The EEG data length used in this research was 30720 points (EEG recordings of 240s with a sampling rate of 128Hz). The longer the data sets to be tested the more accurate the results. However, long EEG data records reduce the chance that the brain remains a stationary system and increase the chance that the records present an artefact (Duke and Pritchard, 1992 and 1995). Moreover, the computational time of the D2 and L1 values increases significantly with the data length. 3) The determination of the optimal embedding parameters is not straight forward: if the embedding parameters are set too small, then the dimension of the true attractor will be underestimated; if they are set too large, then the structure of the attractor will be distorted (Duke and Pritchard, 1992 and 1995). 4) The pre-processing of the EEG data (filtering, sampling, A-to-D conversion) and the noise can significantly alter the estimation of the D2 and L1 values. 5) The determination of the linear scaling region is the difficult part in the estimation of D2 and L1: sometimes there is not any apparent linear region and sometimes many linear regions equally valid can be present on the double logarithmic plot of the correlation sum versus epsilon. In the latter case, the choice of the linear scaling region is very subjective. To overcome this problem, many researchers use linear regression to determine the linear scaling region (Besthorn *et al*., 1995; Jeong *et al*., 1997 and 2001; Jelles *et al*., 1998; Duke and Pritchard, 1992 and 1995). 6) The determination of the neighbourhood is complicated because of the problem temporal correlation aiming to distorted values of D2 and L1.

Besides the practical limitations, the two nonlinear methods used in this research also have some limitations. Indeed, the application of the D2 and L1 assumes the EEG signals to be stationary (i.e. their statistics such as mean, variance and probability density function do not change with time) and low dimensional chaos. These two characteristics are usually debating by researchers. Many researches demonstrate that the EEG is generated by deterministic neural processes and that the nonlinearity is due to low dimensional chaos. Conversely, diverse works show that EEG is neither a chaotic signal of low dimension, nor a stationary signal (Stam *et al*., 1995; Duke and Pritchard, 1992 and 1995; Das *et al*., 2002). Furthermore, the application of D2 implies infinite amount of data which is impossible in practice. Note that by definition, finite sets of points are zero-dimensional. In fact, to determine the dimension of the attractor, an extrapolation is made from finite length scales to infinitesimal scales sometimes resulting in failure of the dimension estimation process.

5. Conclusion

Nonlinear dynamical analysis (NDA) of EEG may be useful to predict the progression of the disease. The quantification of the advance of AD using NDA of EEG could help doctors to make more accurate and reliable prediction on the evolution of the disease and to use appropriate clinical treatments. Early detection of AD using NDA of EEG can be reached by the estimation of D2 and L1 in patient having risk factors for the disease.

Many encouraging results were found in this project. However, in the future, further research on the early stage of AD should be done.

6. References

Besthorn, C., Sattel, H., Geiger-Kabisch, C., Zerfass, R. and Forstl, H. (1995), "Parameters of EEG dimensional complexity in Alzheimer's disease", *Electroencephalography and Clinical Neurophysiology*, Vol 95, pp 84-89.

Das, A., Das, P. and Roy, A.B. (2002), "Applicability of Lyapunov Exponent in EEG data analysis", *Complexity International*, Vol 9, http://www.complexity.org.au (Accessed September 2005).

Duke, D.W. and Pritchard, W.S. (1992), "Measuring chaos in the brain: a tutorial review of nonlinear dynamical EEG analysis", *International Journal of Neuroscience*, Vol 67, pp 31-80.

Duke, D.W. and Pritchard, W.S. (1995), "Measuring chaos in the brain: a tutorial review EEG dimensional complexity", *Brain and Cognition,* Vol 27, pp 353-397.

Jelles, B., Van Birgelen, J.H., Slaets, J.P.J, Hekster, R.E.M., Jonkman, E.J. and Stam, C.J. (1998), "Decrease of non-linear structure in the EEG of Alzheimer patients compared to healthy controls", *Clinical Neurophysiology*, Vol 110, pp 1159-1167.

Jeong, J. (2004), "Invited review: EEG dynamics in patients with Alzheimer's disease", *Clinical Neurophysiology*, Vol 115, pp 1490-1505.

Jeong, J., Chae, J-H., Kim, S.Y. and Han, S-H. (2001), "Nonlinear dynamic analysis of the EEG in patients with Alzheimer's disease and vascular dementia", *Journal of Clinical Neurophysiology*, Vol 18, pp 58-67.

Jeong, J., Kim, S.Y. and Han, S-H. (1997), "Nonlinear dynamic analysis of the EEG in Alzheimer's disease with optimal embedding dimension", *Electroencephalography and Clinical Neurophysiology*, Vol 106, pp 220-228.

Kantz, H. and Schreiber, T. (2004), "Nonlinear Time Series Analysis", 2nd ed. Cambridge University Press, Cambridge, ISBN: 0-521-52902-6.

Kantz, H., Schreiber, T. and Hegger, R. (2000) "TISEAN package", http://www.mpipks-dresden.mpg.de/~tisean/TISEAN_2.1/index.html (Accessed September 2005).

Lipsitz, L.A. and Goldberger, A.L. (1992), "Loss of 'complexity' and aging", *JAMA*, Vol. 267, No 13, pp 1806-1809.

Nestor, P.J., Scheltens, P. and Hodges, J.R. (2004), "Advances in the early detection of Alzheimer's disease", Nature *Neuroscience Reviews*, http://www.nature.com/naturemedicine (Accessed September 2005).

Stam, C.J., Jelles, B., Achtereekte, H.A.M., Rombouts, S.A.R.B., Slaets, J.P.J. and Keunen, R.W.M. (1995), "Investigation of EEG non-linearity in dementia and Parkinson's disease", *Electroencephalography and Clinical Neurophysiology*, Vol 95, pp 309-317.

Non-linear analysis of the Human Electroencephalogram for the detection of Alzheimer's disease using Mutual information

B.Souchier, C.Goh and N.J.Outram

Signal Processing Group, University of Plymouth, Plymouth, United Kingdom
e-mail: {cgoh, noutram} @plymouth.ac.uk

Abstract

Alzheimer's disease (AD) is the most prevalent neurodegenerative disease in the world. There is at present no cure for this cognitive disorder, although several acetylcholinesterase inhibitors such as galanthamine and donepezil have been registered in several countries for symptomatic relief. Nonetheless, unless sufferers are diagnosed in the early stages, they cannot reap the maximum benefit of the treatments that may extend the time before significant mental decline occurs (Correy-Bloom et al, 1998). Therefore, a method of detection which can accurately detect the disease in its early stages would both increase the efficiency of the medication and decrease the cost of the disease. The electroencephalogram (EEG) has been widely used to detect AD. Mutual Information (MI) is a method which looks at the interactions between channels. As AD is concerned with the cortical functional disconnectivity caused by neuronal death, the MI of subjects with AD may be lower than that of healthy ones. Our study aims to investigate the use of MI in the detection of AD. Jeong et al (2001) have found lower MI for EEG measured on ill patient than for EEG measured on normal controls. In order to replicate this study, we have applied the MI on 11 patients suffering from AD and 8 healthy subjects. In addition, we used the Parzen estimator of the probability density function used in the computation of the MI as opposed to the conventional histogram base approach. An important relationship between power and MI has been established. Once the power is normalised, the MI of healthy subjects are found to be higher than the MI of ill subjects. These differences are more significant when we look at the MI between distant electrodes.

Keywords

Alzheimer's disease, electroencephalogram (EEG), mutual information, non-linear analysis

1. Introduction

The majority of the population meets some cognitive decline when they reach 70 years old and beyond. For some, this decline can be stronger than for others and could be due to dementia. Alzheimer's disease (AD) is the most common form of dementia accounting for 60% of all sufferers. In USA, 4.3 million people suffer from AD. In 2050, this number is expected to reach 15 million (Leifer, 2003). AD is the third most expensive disease in the USA, after cardiovascular disease and cancer. At present, $100 billion are spent on treatment for AD each year and this is expected to increase to 4 times in the next 40 years due to the escalating numbers of sufferers.

"AD is a neurodegenerative disease of the central nervous system that begins in middle to late life and results in severe dementia and ultimately death" (Thakur, 2000). The first symptoms are a deficit in short-term memory, followed by some gross impairment of the recent memory, some language skills, judgement and behaviour. This is followed by a strong cognitive decline which leads to an infantile stage and eventually to the death of the patient due to secondary infections and diseases (Jeong, 2004; Leifer, 2003, Thakur, 2000).

At present, no medication exists to cure the disease. However, a good medication can considerably slow its progression. With an appropriate medication, a patient can maintain in his/her normal lifestyle for some additional years. Nonetheless, the key element concerning the efficiency of the treatment is that the disease has to be detected very early. Using current psychological tests and interviews, general practitioners can diagnosis AD for up to 90% of cases. However, they fail to diagnose dementia in 24% to 72% of cases (Brodaty et al, 1993). Misdiagnoses lead to a delay in the treatment and decrease their efficiency, therefore there is an urgent need for alternative methods to be developed.

The EEG is a record of the electrical activities of the brain. Due to its non-invasiveness and real-time depiction of brain activities, it has become an attractive tool in clinical practices with applications including the detection of seizure and epilepsy and brain tumours, Binnie et al (4). Research have revealed that the EEG could be a potential candidate for the early detection and differential diagnosis of dementia, Ktonas (5), Rosén (6). However, the complex and nonstationary characteristics of the EEG make automating the analysis process a challenging task. Both linear and non-linear methods have been applied to the analysis of the EEG.

Linear analysis of the EEG mainly consists of dividing the time series into different frequency bands (delta, theta, alpha and beta). We can noticed marked differences between AD and healthy subjects. The EEG of an AD patient will usually show an increased of the theta and delta activities and a decrease of the alpha and beta activities (Jeong, 2004, and references within). "A good correlation is found between EEG spectral measures and cognitive deterioration scores" (Jeong, 2004). EEG can also be used to differentiate between AD and other dementia (Jeong, 2004 and references within). All the functionalities of the brain are resulting from the interaction between different cortical areas. As AD involves the loss of neurons, we can expect these interactions to be weaker for sufferers than for healthy ones. The linear method of coherence computes the linear interaction between channels. Several studies show that the coherence decreases in different bands (Adler et al, 2003, Locatelli et al, 1998).

Two methods of non-linear analysis measure the complexity of the EEG, namely, the Correlation dimension and the Lyapunov exponent. It has been found that "AD patients have reduced values of the Correlation dimension in the occipital EEG compared with those of healthy subjects" (Jeong, 2004). These findings have been confirmed by numerous other studies (References within Jeong, 2004). The Lyapunov exponent has also been used to positively detect AD (Jeong et al, 1998; Jeong, 2004). The synchronisation is an other non-linear method which calculates the linear and non-linear interactions between channels (Stam et al. 2002). Different

studies have shown a decreased synchronisation for the ill patients in the beta band (Stam et al, 2003), in the alpha and beta bands (Stam et al, 2005; Pijnenburg et al, 2004) or in the alpha, beta and theta bands (Koening et al, 2005). The mutual information (MI) performs a similar measurement. Nevertheless, Jeong et al (2001) have shown that the MI in AD EEG was lower than in healthy EEG, especially between inter-hemispheric or distant electrodes.

The focus of this paper is to apply the MI approach to the detection of AD. We will use the Parzen estimator of the probability density function (PDF) to compute the MI as this approach is more accurate than the histogram based one. The use of the Parzen estimator in computing the MI has not yet been applied to the detection of AD. The results showed an important relationship between power and MI: once the power is normalised, the MI of healthy subjects are found to be higher than the MI of ill subjects. These differences are more significant when we look at the MI between distant electrodes.

The paper is organised as follow. In section 2, we will look at the methodology used. We will first see the data used in this study. Following this, we will look at the computation of the MI. The use of the Parzen Estimator will be studied thereafter. In section 3, the settings for some parameters of the Parzen estimators and the results that we obtained are discussed. Finally, some concluding remarks and future directions are given in Section 4.

2. Methodology

2.1. Data

The EEG recordings were obtained using the traditional 10-20 system, in conjunction with a strict protocol, Jasper (18). The common average montage (using the average of all channels as the reference) was used in all recordings. For all data, the recorded sampling rate was 256 Hz, which was further reduced to 128 Hz for analysis by averaging sets of two consecutive samples (for storage reasons). The EEG recordings encompassed various states: awake, hyperventilation, drowsy and alert with periods of eyes closed and open. To prevent electrical artefacts, which commonly occur at the beginning of a recording, and to give a standard four minutes of data to analyse, only data from 60 s to 300 s from each record was used. This segment of data including artefacts was analysed and with no á priori selection of elements 'suitable for analysis'. EEGs were collected from 11 AD patients and eight age matched controls (over 65 years of age). All of the age-matched controls had normal EEGs (confirmed by a Consultant Clinical Neurophysiologist).

2.2. The Mutual Information

MI is a measurement based on information theory. It was first introduced by Shannon in the Mathematical Theory of Communication (Shannon, 1948). The MI quantifies the information content that a random variable has about one other. If the

MI is equal to zero, then the two random variables are totally independent. If the MI is maximal, then the two random variables are completely linked.

If we have a random variable $X = \{x_1,\ldots,x_N\}$. Then, the information content of the element x_i of X is:

$$\log2\ PX(xi) \tag{1}$$

Where $P_X(x_i)$ the probability that x_i is ith element of X.

We can then define the entropy, noted $H(X)$ which is the average of the information content of all the elements x_i of the variable X:

$$H(X) = -K \sum_{x_i \in X} P_X(x_i) \log P_X(x_i) \tag{2}$$

where K usually equal to 1.(Shannon, 1948)

We can also define the entropy of two variables X and Y. With $P_{XY}(x_i,y_j)$ the joint probability of the variables X and Y, the entropy $H(X,Y)$ is equal to:

$$H(X,Y) = -\sum_{x_i \in X, y_j \in Y} P_{XY}(x_i, y_j) \log_2 P_{XY}(x_i, y_j) \tag{3}$$

We can define the quantity $H(X|Y)$

$$H(Y|X) = \sum_{x_i \in X} P_X(x_i) H(Y|X = x_i) \tag{4}$$

$$= -\sum_{x_i \in X} \sum_{y_j \in Y} P_{XY}(x_i, y_j) \log_2 \frac{P_{XY}(x_i, y_j)}{P_X(x_i)} \tag{5}$$

$$= \mathrm{H(X,Y)} - \mathrm{H(X)} \tag{6}$$

$H(Y|X)$ is then the information about X and base on the knowledge in Y minus the information in X. It is then the information in Y which is found only in Y. Indeed, all the information in X has been discarded. So the common information to X and Y is not present any more.

The MI between X and Y, $I(X,Y)$, is simply the information in Y from which we discard the information in Y which is not common to X, $H(Y|X)$.

$$\mathrm{I(X,Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)} \tag{7}$$

$$I(X,Y) = \sum_{x_i \in X} \sum_{y_j \in Y} P_{XY}(x_i, y_j) \log_2 \left(\frac{P_{XY}(x_i, y_j)}{P_X(x_i) P_Y(y_j)} \right) \tag{8}$$

2.3. The Parzen Estimator

The calculation of the MI is not very complex. Nevertheless, we have to calculate the PDF and the joint PDF of the variables *X* and *Y*. This computation is rather time consuming. The PDF is a function which draws the distribution of the variables *X* or *Y*. The distribution of a variable *X* shows the spread of values that elements of *X* can take. For instance, if the values are equally distributed, we have the normal distribution.

The classical way to find the PDF is to compute the histogram. Using this technique, we simply divide the different values that can take the element of the variable *X* into groups, termed 'bins' and count how many elements of *X* falls into each group. The problem with histogram is that it is a discontinuous measure, and that depending on the width of the bins, we will have a rough approximation or a distribution with lots of fluctuations

An alternative to the histogram approach is to use the Parzen Estimator (Parzen, 1962), which is derived from the histogram to compute the PDF. The PDF $f(x_i)$ of a variable *X* is then:

$$f_n(x_i) = \frac{1}{n * h} \sum_{xj \in X} K\left(\frac{x_i - x_j}{h} \right) \tag{9}$$

K(z) is the kernel of the estimator. With it, we can adapt the estimator to the kind of distribution that we are expecting. For instance, for a uniform distribution, we might use the rectangular kernel which makes the estimator equivalent to the histogram:

$$K(y) = \begin{cases} \frac{1}{2} & if \quad |y| \le 1 \\ 0 & if \quad |y| > 1 \end{cases} \tag{10}$$

In order to obtain a true PDF, the kernel needs to have the following properties (Parzen, 1962):

$$\int K(y) dy = 1 \tag{4}$$

The parameter *h* has to satisfy:

$$\lim_{n\to+\infty} h(n) = 0$$
$$and$$
$$\lim_{n\to+\infty} nh\ (n) = \infty \tag{5}$$

The Gaussian kernal is commonly used for the Gaussian distribution.

$$Kernel(x, sigma) = \frac{1}{(\sqrt{2\pi} * sigma)^u} e^{-\frac{x*x^T}{2*sigma^2}} \tag{6}$$

where sigma maps to parameter h in equation (5) and is related to the standard deviation of the variable X.

3. Results and discussions

3.1. Function of the parameters

In the Parzen estimator of the PDF, we have to determine the value of sigma. If sigma is not optimal, then the PDF will be either too smooth or random and the MI will be erroneous. In order to obtain the optimal value for sigma, we first looked at the PDF of a part of a channel for different values of sigma. In the following figures, we can see that if sigma is too small (Figure 5), then the PDF shows too many fluctuations. On the other hand, if sigma is too high (Figure 7), we see a poor PDF approximation. Figure 6 shows the PDF computed with a suitable sigma.

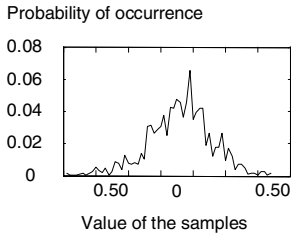


Figure 5: Distribution of an EEG for a sigma equal to 0.0008

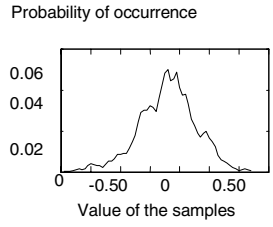


Figure 6: Distribution of an EEG for sigma equal to 0.005

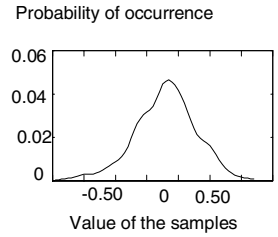


Figure 7: Distribution of an EEG for sigma equal to 0.01

The MI is calculating the common information between two channels. Due to the transmission of information through the brain, some of the particularities of the distribution due will disappear if a long part of the EEG is used in the computation.

Therefore, we need to limit the length of data that might be used in order to compute the PDF. We have then to determine what is called the window size. If we take too few samples in our PDF, we loose information as the PDF that we obtain is not really characterising the data. In order to find the optimal window size, we have looked at the entropy of a channel for different window sizes as shown in the Figure 8, as well as the variance of the entropy calculated with a same window size but on different parts of a same channel (figure 9). As the entropy as well as its variance converges from a certain value, we can find the minimum value that the window size should take. We have determined the window size to be 9secs (1152 samples).

The Parzen Estimator is a continuous function. However, as we are using a computer, we are handling discrete data. We have then to determine how many points will be used to calculate the Parzen estimate of the PDF used in the computation of the entropy. We have then to find the minimum number of point from where the results converge. Comparing the different values of the MI for different number of points, the optimal value has been decided at 64 points.

No transmission in the brain is instantaneous. Therefore, we introduce a delay between the two channels. In order to include all the information transmission that took place between the two channels, we calculated the MI for all delays until a certain value from where we estimate the information as complete. In Jeong et al. (2001) this maximum is determined to be 500ms. We have then calculated the MI for all the delays from 0 to 500 ms and average the results.

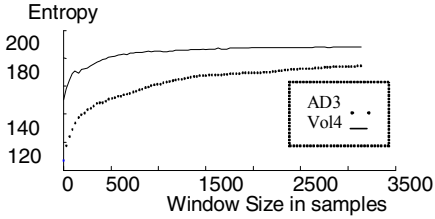


Figure 8: Entropy EEG as function of the window size

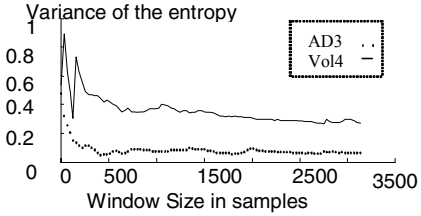


Figure 9: Variance of the entropy of EEG as a function of the window Size

3.2. Relationship between MI and power

The MI and the power may have a strong link. Indeed, when we look at the MI of two random variables X and Y with a Gaussian distribution, we find this relationship. With σ_1, σ_2 and σ_3, the standard deviation of respectively X, Y and the joint distribution of X and Y, noted z. We see in the equation (44) the appearance of the power Σx^2.

$$I(X,Y) = \frac{1}{2\pi\sigma_3^2 \log_e(2)} \left[\log_e(2) * \log_2\left(\frac{\sigma_1\sigma_2}{\sigma_3^2}\right) \sum_{x\in X}\sum_{y\in Y} e^{-\frac{z^2}{2\sigma_3}} + \sum_{x\in X}\sum_{y\in Y} \frac{x^2}{2\sigma_1} e^{-\frac{z^2}{2\sigma_3}} + \sum_{x\in X}\sum_{y\in Y} \frac{y^2}{2\sigma_2} e^{-\frac{z^2}{2\sigma_3}} - \sum_{x\in X}\sum_{y\in Y} \frac{z^2}{2\sigma_3} e^{-\frac{z^2}{2\sigma_3}} \right] \quad (44)$$

We have normalised the data by dividing each sample by the mean of the power of the sequence. The power has been calculated by adding the square of each sample. Having normalised the signal power of the dataset, the MI calculated from ill patients is found to be lower than the MI calculated from normal controls. The difference between these MI is more significant when the MI is calculated between distant electrodes than for local electrodes. These results of Jeong et al (2001). The Table 1 contains the mean and the standard deviation of the MI for the different formation of electrodes.

The high standard deviation found for the AD patients is due to some differences between two groups of measurement of EEG. The first group is composed by the 3 first EEG. The second group, composed by the 8 other abnormal EEG has been measured later. It may be interesting to look if the differences are correlated with the stage of the disease of the patients or if the differences are due to the measurement itself.

Formation	**Interhemispheric**		**Distant across the central line**		**Local**	
Patients	**AD patients**	**Normal control**	**AD patients**	**Normal Control**	**AD patients**	**Normal Control**
Mean	**2.3997**	**4.4879**	**2.331**	**4.4514**	**2.5702**	**4.2965**
Standard deviation	**0.9395**	**0.2435**	**1.0289**	**0.2866**	**0.9547**	**0.4332**

Table 1: Mean of the MI of normal and abnormal EEG

4. Conclusions

We have established a strong link between power and MI. Once power is normalised out, the MI of patient with AD is significantly lower than the MI of normal controls. The differences of MI are also greater for distant electrodes. This confirms the study of Jeong et al (2001) and the assumption that AD patients have difficulties in signal transmission in the brain. We have not proved any increase in the results with the use of the Parzen Estimator. However, it has been successfully used.

Future work includes the study of MI between different channels for different frequency bands for the early detection of AD. In addition, it is proposed that the MI should also be used to discriminate between AD and other dementia.

5. References

Adler G., Brassen, S. Jajcevic A. (2003), "EEG coherence in Alzheimer's dementia", *Journal of Neural Transmission,* n° 110, pp1051–1058.

Brodaty H., McGilchrist C., Harris L., Peters K.E. (1993), "Time until institutionalisation and death in patients with dementia: Role of caregiver training and risk factors.", *Archives of Neurology*, Vol. 50 pp 643–650.

Corey-Bloom, J., Anand, R., Veach, J., for the ENA 713 E352 study group (1998), "A randomised trial evaluating the efficacy and safety of ENA 713 (rivastigmine tartrate), a new acetylcholinesterase inhibitor, in patients with mild to moderately severe Alzheimer's disease", *International Journal of Geriatric Psychopharmacology*, n°1, pp 55-65.

Jeong J., Kim S.Y., Han S.H. (1998), "Non-linear dynamical analysis of the EEG in Alzheimer's disease with optimal embedding dimension", *Electroencephalography and Clinical Neurophysiology*, Vol. 106, pp220–228,.

Jeong J., Gore J.C., Peterson B.S. (2001), "Mutual information analysis of the EEG in patients with Alzheimer's disease", *Clinical Neurophysiology*, n° 112, pp 827–835.

Jeong, J. (2004), "EEG dynamics in patients with Alzheimer's disease", *Clinical Neurophysiology*, n° 115, pp 1490-1505

Koenig, T; Prichep, L; Dierks, T; Hubl, D; Wahlund, L. O; John, E. R; Jelic, V." (2005), "Decreased EEG synchronisation in Alzheimer's disease and mild cognitive impairment", *Neurology of Aging*, n° 26, pp 165-171.

Leifer B. P. (2003), "Early Diagnosis of Alzheimer's disease: clinical and economical benefits", *Journal of the American Geriatrics Society*, Vol. 51, n° 5, Supplement, pp281-288.

Locatelli T., Cursi M., Liberati D., Franceschi M., Comi G. (1998), '"EEG coherence in Alzheimer's disease", *Electroencephalography and Clinical Neurophysiology*, Vol. 106, pp 229–237

Parzen E. (1962), "On estimation of a probability density function and mode", *Annals of Mathematical. Statistics*, vol 33, pp 1065-1076.

Pijnenburg Y.A.L, van de Made Y., van Cappellen van Walsum A.M., Knol D.L., Scheltens Ph., Stam C.J. (2004), "EEG synchronisation likelihood in mild cognitive impairment and Alzheimer's disease during a working memory task", *Clinical Neurophysiology*, n° 115, pp 1332–1339. OK

Shannon C.E. (1948), "A Mathematical Theory of Communication", *Reprinted with corrections from The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October.

Stam C.J., van Dijk B.W. (2002), "Synchronisation likelihood: an unbiased measure of generalised synchronisation in multivariate data sets", *Physica D*, vol 163, pp 236-251.

Stam C.J, van der Made Y., Pijnenburg Y.A, Scheltens P., (2003), "EEG synchronisation in mild cognitive impairment and Alzheimer's disease", *Acta Neurologica Scandinavica*, n° 108, pp 90–6. OK 23

Stam C.J., Montez T., Jones B.F., Rombouts S.A.R.B, van der Made Y., Pijnenburg Y.A.L., Scheltens Ph. (2005), "Distributed fluctuations of resting state EEG synchronisation in Alzheimer's disease", *Clinical Neurophysiology*, vol 116, pp 708-715.

Thakur M. K., (2000) "Alzheimer's disease – A challenge in the new millennium", *Current Science*, Vol. 79, n°1, pp 29–36

Section 3

Information Systems Security & Web Technologies and Security

Social Engineering: A growing threat, with diverging directions

J.V.Chelleth[1], S.M.Furnell[1], M.Papadaki[2], G.Pinkney[2] and P.S.Dowland[1]

[1] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2] Symantec, Hines Meadow, St Cloud Way, Maidenhead, Berkshire, United Kingdom
e-mail: info@network-research-group.org

Abstract

The age old problem of social engineering is still a threat that does not receive due attention. Due to the advancements in information technology and the explosion of the Internet, attackers have many more avenues to pursue social engineering attacks. Inadequate efforts to educate employees and staff about social engineering and password management, inappropriate usage of messaging systems, poor implementation and awareness of security policies, all lead to people being exposed to potential incidents. This paper talks about social engineering and the new avenues that it has diverged into; and how social engineering plays a part in assisting other attack schemes. The paper first introduces the concept of social engineering. It then looks at different attack methods that have proliferated due to the help obtained by social engineering schemes. The paper establishes that, in addition to being a technique in its own right, social engineering can also be used to assist other types of attack, including viruses and worms, phishing, and identity theft.

Keywords

Social Engineering, Viruses, Worms, Identity theft, Phishing

1. Introduction

Typically when security is spoken of in terms of information security, it is all about having secure systems and networks; anti-virus, firewalls, Intrusion Detection Systems (IDS), etc. A lot of effort is put into implementing technical security and this creates a notion that the systems/network are not susceptible to attacks and hence exploitation. However, non-technical details are often forgotten and this gives attackers a means to slip past the otherwise heavily guarded IT security systems. Keeping this in perspective, the paper talks about attacks that arise due to social engineering; a concept that has been used often to exploit computer systems and individuals alike.

The main goals of social engineering as identified by Paradowski (2001), who calls such attackers as Cyber Cons, are fraud, network intrusion, industrial espionage, and identity theft. Of course, social engineering exploits existed long before the existence of computers - there always were individuals who deceived others into giving out valuable information. However, in the age of Information Technology, attackers found a new medium to carry out their exploits.

2. Setting the scene

Social engineering is an art of deceiving, to obtain information by pretending to be someone that s/he is not. Such information would not normally be provided because it may be personal or sensitive/protected. It is a human tendency to trust people without knowing much about the person. People believe what they hear and on the basis of how they hear it. If someone confidently tells a lie, it is believed much more than someone who tells the truth in an uncertain or loath manner.

Attackers make use of this human nature, to maximise their benefits by getting out information that will be useful in carrying out their exploits. A large organisation that is very concerned about security may have deployed a corporate anti-virus solution, a comprehensive firewall scheme, and even a strong intrusion detection system (IDS). If an attacker wants to penetrate this system and has the gift of the gab, then why would s/he waste time trying to overcome these security deployments? Instead there is a very good chance that s/he could get information from the employees by employing social engineering tactics.

Figure 1 shows a partial screen capture of the output from a popular tool called Sam Spade. It is a network-query tool like *nslookup*, *whois* and *traceroute*, but GUI-based. Querying a popular shopping site, and hitting a few buttons, we observe technical contact details. Attackers who use social engineering methods need to have some basic idea about the organisation to start their attack. Using such information, the attacker can query individuals and network their way into the organisation. To do so, they could pose as trusted vendors, new hires, contractors, electricians, a high level officer and so on. If an attacker is confident to pull off the personality of a higher official, it becomes very easy to get the information required. Employees often want to impress their seniors, possibly for selfish reasons and to get higher up the ladder; nevertheless this causes sensitive information to be given away rather easily to attackers.

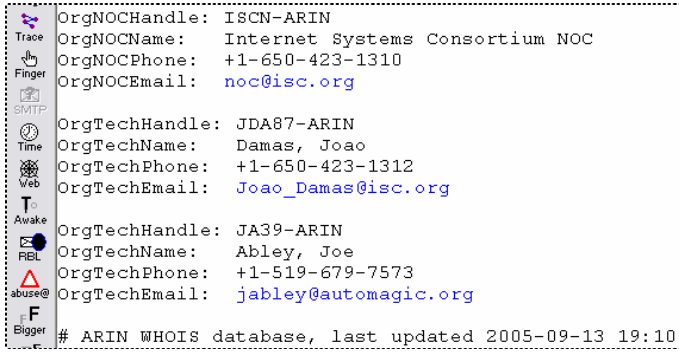


Figure 1: Partial screen capture from Sam Spade tool

That is one way to use social engineering by gathering information and then exploiting the relationship. The possibilities are endless and depend on how innovative the attacker is. If information can be easily obtained by simply querying, then why would an attacker go through the trouble of surpassing firewalls? This also

indicates that an attacker need not be highly proficient programmer and s/he does not need great technical skills; a flare to talk and confidence can do equal or more damage, at times much quicker and with far less at stake.

3. Combination of attacks

Social engineering methods will not always be used in isolation. Often, a combination of two or more attack methods is used to exploit the target, as shall be seen in this section which relates different attack methods to social engineering tactics.

3.1 Obtaining passwords

People are rightly considered as the weakest link in IT security. Even after an organisation has taken every step towards a great IT security deployment, if the system administrator gives away a password, it jeopardises the entire organisation and it falls prey to social engineering attacks; the entire security infrastructure will fail. Even employees sharing their passwords and login details cause vulnerabilities to arise in the security system. Colleagues in an organisation come to trust each other and often lend their passwords to each other for different reasons (many of which may be legitimate) but it can never be predicted when an individual might misuse the login details at his /her disposal. Additionally, when an employee gets to know more about another employee, at times it becomes easy to guess the kind of passwords that the latter would use.

It should be noted that in any organisation, there are so many systems that have to be maintained by an administrator, that is becomes very cumbersome to remember different and complicated passwords; and hence there exists a trend to use simple dictionary words, birthdates and family names, equipment names, model numbers or even keep the default passwords. It is common to find passwords like ‘cisco’ and ‘intel’.

3.2 Viruses and worms

Social engineering tactics have frequently been used as part of virus and worm attacks, as shown in Table 1. Users have been tricked into opening and running harming messages that claim to be legitimate programs or applications. The human mind is inquisitive and can get tricked into responding to such incentives, which cause their systems/network to be exploited successfully.

The success and possible damage of an unsolicited mail with a harmful payload can be judged by its rapid spread and also by the number of machines compromised. Social engineering comes to the rescue and helps to increase the levels of curiosity and want. Clearly, most individuals would have been thrilled to see a ‘*Love letter*’ in their rather boring inbox and would be interested to open it; and a good proportion of users would have clicked away without checking and thus infected their systems.

Name	Appearance	Background tasks
Christma Exec, 1987 (Virus)	Promises to draw a Christmas tree, and does draw it.	Sends out copies of itself in the users' name.
Happy99/Ska, 1999 (Worm /Trojan)	Displays fireworks on the screen.	Modified the WSOCK32.DLL file. Caused a 2nd mail to be sent with the worm to the same recipient.
Melissa, 1999 (Virus)	Promised account names and passwords of erotic sites.	Affected the document template in Microsoft Word and ran a macro that opened Outlook and sent mails to 50 recipients.
PrettyPark, 1999 (Worm)	Bears an icon of a character from a television show, South Park.	Modifies system registry. Emails itself to addresses in the Windows address book. Mails private system data and passwords to IRC Servers.
Love Letter, 2000 (Worm)	Appeared to have a Love letter text file attached to it, which was actually a VBS script.	Infected Windows and system directories. Sent out emails to addresses in Outlook and also tried to spread through IRC channels.
Anna Kournikova, 2001 (Worm)	Pretended to carry a JPEG picture of the tennis star.	If executed, it emailed copies of itself to all addresses in Outlook.
Gibe, 2002 (Worm)	Disguised as a Microsoft security bulletin and patch.	Secretly installs a backdoor onto the system.

Table 1: Social engineering methods used by viruses and worms (Chen, 2003)

3.3 Phishing

Phishing, which has now become a very serious threat, also employs clever social engineering methods. Emails are shown to be sent by banks and other financial organisations, and they get people to divulge personal information like bank account numbers, passwords, etc. The success of such an attack depends on the number of individuals who actually are duped by clicking the links within the emails. Hence, phishing mails are sent out in huge numbers to have at least a small percentage of users clicking away to their doom.

As seen in Figure 2, the user is made to believe that the email has been sent by eBay. This kind of mail is among the sophisticated phish mails, which has developed over time and has been designed to look genuine. Many users will actually click on the link given and be taken to a site that looks like ebay.com, but actually is a illegitimate site that will ask for the user's personal details. Depending on the level of complexity, malicious code could also be run on the users' machine when they are directed to the fake site.

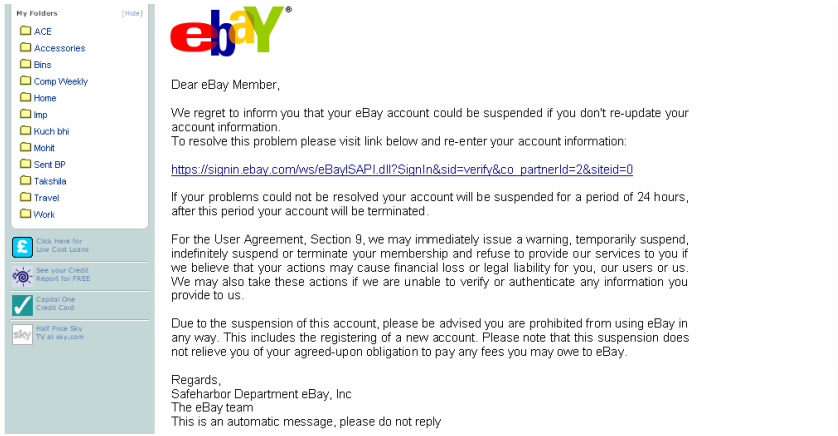


Figure 2: A phishing email, supposedly from eBay

Although, the email is cleverly disguised, such damage can easily be avoided if the mail is scrutinised a little; in this particular case the following observations can be made:

- the email comes from a spoofed id: support_num_100737@ebay.com;
- it is not a personalised email, mentions 'Dear eBay member';
- there is no sender name, and comes from a so called 'Safeharbor Deparment'.

However, the attempt may still be sufficient to fool casual or naïve users. As such, it is useful for potential victims to be educated about safe surfing habits and social engineering tactics.

3.4 Identify theft and fraud

Identity thefts are again an area that can make use of social engineering methods for fraud to maximise the attackers' benefits. Identity thieves existed long before the information technology age, but now the Internet has made it so much easier to masquerade as someone else. Anonymity on the Internet gives rise to many more attacks, because it is very difficult to track down individuals who can use various means to conceal their true identity. Different attacks methods combined with social engineering give many new avenues to commit *efraud* by stealing another individuals' identity.

Trojans slipped into an individuals' machine by means of viruses or worms can collect personal information that might be later used to steal money, or impersonate the individual. User account information gathered by simply asking, or by guessing, or using tools - can be used to log into different systems and exploit them, or log into an online shopping site and benefit the free purchases. Phishing attacks can grab personal information, and the attacker can steal from the victim without being caught until they have fled. Attackers can penetrate into an organisation and cause a variety

of damages from physical destruction of equipment to stealing valuable proprietary information.

To commit a successful efraud, the attacker must employ different levels of social engineering to gain knowledge about the target. The more information that is gained, the easier it is to crack the target. Similarly for an Identity theft, the attacker must know as much information about the individual he is impersonating; failing which his/her cover will be blown and s/he could be caught. Thus we see that social engineering is the basis for successful e-fraud.

4. Conclusion

To quote Kevin Mitnick: "Why do hackers use social engineering? It is easier than exploiting technology vulnerability. You can not go and download a Windows update for stupidity... or gullibility" (Gedda, 2005). It is a perfect explanation given by a former hacker and famous social engineering expert. Many attackers would rather just ask information from an unsuspecting employee using social engineering skills or by combining social engineering tactics along with other attack patterns.

The onus lies in the hands of the organisation to educate their employees about social engineering tactics. All factors that contribute to social engineering exploits should be considered and the employees must be made aware of such patterns. Security policies should be devised and there should be specific clauses addressing the problem of social engineering. These policies should be promoted from time to time and employees should be trained on best practices. This has to be a continuous process, as the only way to combat a non-technical issue, is to cultivate non-technical safeguards too, along with maintaining the technical levels of protection.

5. References

Chen, T. (2003), "Trends in viruses and worms", *Internet Protocol Journal*, vol. 6, September 2003, 23-33

Ernst & Young (2004), "*Global Information Security Survey*", http://www.ey.com/global/download.nsf/International/2004_Global_Information_Security_Survey/$file/2004_Global_Information_Security_Survey_2004.pdf, (Accessed: November 07, 2004)

Gedda, R. (2005), "*Hacker Mitnick preaches social engineering awareness*", http://www.computerworld.com.au/index.php/id;1016567243;fp;16;fpid;0, (Accessed: August 28, 2005)

Paradowski, C. (2001), "*The Cyber Con Game - Social Engineering*", http://www.giac.org/certified_ professionals/practicals/gsec/0971.php, (Accessed: September 02, 2005)

Issues Affecting the Extraction of Data from the Web

S.Butt and A.Phippen

Network Research Group, University of Plymouth, Plymout, United Kingdom
e-mail: info@network-research-group.org

Abstract

Much of the data found on the Web is of some value, especially data found on company websites. Therefore, the collection and storage of such data would provide a valuable resource. Due to the sheer volume of data it is impractical to expect a human being to be capable of accurately collecting it through browsing the Web. A solution to this problem is to automate the task of data extraction. Unfortunately, differing standards in the quality of documents on the Web restrict the amount of data retrieved and the accuracy of an automated process. This paper examines the types of data that may be required to be, and can be, extracted from a web page, as well as issues affecting the accuracy of data extraction. It is then suggested that the use of standard document syntax and structure, and the use of self descriptive elements, to aid the automated data extraction process may improve information flow between businesses on the Web.

Keywords

Data extraction, spidering, semantic web.

1. Introduction

The Web represents a vast repository of data. This data is presented to users of the Web in the form of web pages generated using HTML. How the data is represented within the source HTML file is generally of little or no concern to the author of the page as long as it can be interpreted correctly by a web browser and, in turn, understood by the end user. Whilst this approach to data presentation is adequate in terms of making the data human readable, it can present difficulties when attempting to perform automated data extraction involving little or no human interaction.

It can be assumed that much of the data found within a web page will be of interest to someone. In some cases this data will also have some value. Most companies nowadays have their own websites, which contain, at the very least, some information regarding the nature of their work, and contact details. Such information is of value as it relates to business – the company itself uses this data to acquire business, but it could equally be used by others (e.g. competitors, industry analysts) as it is in the public domain. The volume of data found within such sites can be great; therefore a human based approach to data extraction would prove time consuming (as well as being prone to simple human error). An automated method of efficiently extracting data from websites – providing information of value to its user – would have a value proportional to that of the data itself. Unfortunately, HTML is used to describe the layout and appearance of data for viewing through a browser and

does not strictly enforce rules regarding syntax, which could make an HTML document more meaningful to a computer. This can seriously hinder attempts to automate the extraction of data.

This paper seeks to highlight the major issues associated with automating the data extraction process based upon the results of practical research into the area. It is largely concerned with the value of data extraction to businesses and the discussion presented here focuses on this specific subset of WWW users. Following on from this discussion it presents ideas for means with which problems in this area can be overcome, measures that can be taken by companies to facilitate business to business data sharing and the value of such measures.

2. The purpose of data extraction

The perceived value of data extracted from the Web will differ depending upon the perspective from which it is viewed. One company may be interested in acquiring contact details for potential customers, another company may be interested in profiling the types of technologies used within certain websites (e.g. CGI, applets, Flash, etc). It is also possible that certain information could be inferred from data extracted from the Web. For example, information regarding the type of server being used to host websites could be used to build an independent model of server usage. What data has value cannot be stated here, as it depends largely upon the aims of those who seek to exploit the data found on the Web. It will be assumed, therefore, that all data held within a website is potentially of value to someone. The following will centre on the extraction of data in general, with no specific focus on the usage of extracted data (although, examples may be used to illustrate the potential of the methods discussed herein).

Currently, there are a number of projects concerned with the topic of data extraction from the Web (Laender et al, 2001; Chen et al, 2001; Lage et al, 2004; Baumgartner et al, 2005; Myllimaki, 2002). Many of these focus on correcting the structure of Web documents and/or the retrieval of data from within the deep Web. The deep Web is a term used to refer to those parts of the Web that are not easily accessible using simple spidering (simply retrieving interlinked, static web pages – also referred to as crawling). Generally speaking these are dynamic web pages whose content is based upon user input, the contents of a back end data store or the results of server side processing. The contents of the deep Web are outside the scope of this discussion. However, the issues discussed within this paper may provide a basis for research into the value of data extraction from the deep Web.

All of the projects looking into the topic of automated data extraction on the Web – including that which was carried out as part of the research for this paper – encounter the same problems. These problems are a result of the inconsistency, across the Web, of the presentation of data within websites.

3. An automated process

A simple way of extracting data from the Web is for a human user to navigate websites page by page, viewing the source code for each, noting down each piece of valuable data that they come across. The obvious problems with this form of data extraction are the speed with which the process can be performed by a human and simple human error. It is safe to assume that a percentage of data within each website is likely to be either missed, or incorrectly recorded – resulting in the need for an additional process that validates the extracted data. This could also be a human process, or it could be automated, either way it increases the time and resources used to perform a single task. The value of the data extracted may be outweighed by the cost of actually performing the data extraction. Automating this process would lead to a significant improvement in performance and accuracy, and would allow for more efficient use of the human user's time.

3.1 The DataChecker system

For the purposes of this research a simple data extraction system was developed that was used to study the type of data that could be retrieved from the Web and the ease with which this could be achieved. This system, called DataChecker, was designed to accept a list of URLs pointing to websites from which data was to be extracted. The system navigates through each site (to a certain depth, so as to not get lost within large sites, or stuck in a loop of interlinked pages) collecting data as it goes. It is composed of two main components: the spider component is provided with URLs, which it follows, retrieving the HTML at each location. The extractor component takes the HTML returned by the spider and extracts data from it based on a file defining what should be extracted. Any URLs relative to the website currently being spidered are passed from extractor to spider, so perpetuating a retrieval/extraction cycle that runs as long as there are pages to retrieve within the site (or DataChecker reaches a predefined page limit).

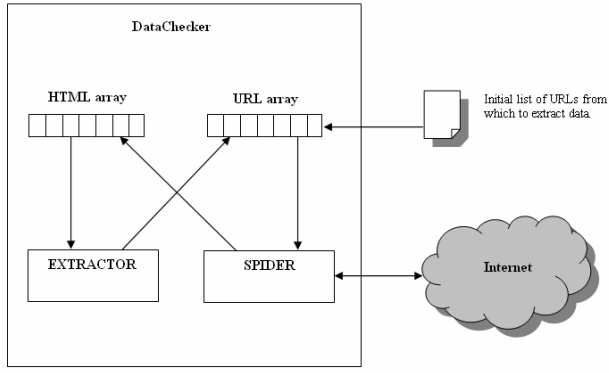


Figure 1: Simplified workings of the DataChecker system.

When provided with a list of URLs DataChecker splits them between a number of arrays. Each array is then associated with a spider and an extractor (as is an empty array that is eventually to be populated with the HTML returned by the spider). Each

spider and extractor, operating on separate threads of execution, constantly checks the status of the two arrays (the spider checking the URL array and the extractor the HTML array) to see if they have any content. As soon as content is found it is then processed in the relevant fashion (URLs used to retrieve HTML, HTML used to extract data).

The retrieval of data by the spider is a simple task – the spider is essentially dumb, leaving the majority of the processing to the extractor. In the DataChecker system data is extracted from HTML using regular expressions. These are used to extract two types of data – that which is actually present within the HTML code (e.g. Hyperlinks and email addresses), and that which can be inferred from the code, such as whether or not a scripting language is being used (the extractor does not extract the data straight from the HTML, rather a Boolean value is generated based upon its existence). As the HTML retrieved by the spider can be represented as one long string of text, regular expressions are used to find matches for certain types of data within the HTML. A regular expression can be used to verify the existence of a certain string pattern – that is, a collection of specified characters arranged in a specific fashion – within a certain string. It can also record the matched substring(s) – allowing for future analysis. In this way the regular expression can be used to extract both types of data mentioned above. To extract a certain piece of data from a website simply requires a regular expression that will correctly match every occurrence of that data on every web page that is processed. This proves much more difficult than the previous sentence makes it appear.

```
<\s*a.*(href)\s*=\s*(?:\"(?<1>[^\"]*)\"|(?<1>\S+)).*>
```

Figure 2: An example regular expression - extracts links from within the anchor (a) tag.

When developing DataChecker an alternative method to regular expressions was considered. Due to the similarity of HTML and XML (both of which are derived from SGML) it was initially decided that the code retrieved by the spider would be treated as XML. In this way it would be possible to make use of XSLT to extract data from the HTML. Unfortunately, the same problem that would also affect extraction through the use of regular expressions – which will be discussed shortly – meant that this method was not viable (without significant extra processing of the HTML).

As mentioned earlier, the rules governing the syntax of an HTML document are loose and not as strictly defined as XML. The W3 Consortium has defined exactly how to build a correctly formatted XML document (Bray et al, 2004). This standard way in which an XML document is built ensures that attempts to find a piece of data, defined by a certain tag within a well formed XML document will not be hampered by inconsistencies in the style or content of the code. As HTML does not have such strict restrictions placed upon it one website developer may present a piece of data in one way, whilst another developer presents it in a slightly different way. The difference between the two methods may not be great, but the fact that there is a difference makes extraction of that data more difficult. The less strict definition of how HTML documents are constructed can also lead to errors within the code that

may be missed by the developer, but would lead to difficulties when attempting to extract data from it using a system such as DataChecker. The sheer number of web pages on the Internet means that the probability of errors within a selection of retrieved web page is going to be high. Such errors and differences in the presentation of data create difficulties for the extractor. If attempting to treat the HTML as XML, the majority of pages will be classed as incorrectly formatted and will not get processed correctly. This problem could be overcome by pre-processing the HTML to ensure that it adheres to the XML standard (turning it from HTML into XHTML), but this would increase the processing overhead for the system. Regular expressions can be used on every page returned by the spider, but to extract the correct data the expression needs to match all variations on how that data can be presented. It also needs to be robust enough to recognise errors in the code, but still extract the correct data. Such regular expressions would be extremely complex and hard to define. A certain percentage of data within websites would not be matched due to errors in the HTML.

Website developers creating sites that adhered to the XHTML standard (W3C HTML Working Group, 2002) would go some way towards solving the problems outlined above. However, it is not practical to assume that all developers would change their ways just to accommodate those wishing to perform automated data extraction. Many companies, though, could benefit from a standard method of presenting information about themselves and their products that would allow for easy extraction and processing.

3.2 Business to business data interchange

A great many companies take advantage of some form of EDI (Electronic Data Interchange) to do business. Vast amounts of business data are exchanged electronically, often with little or no human interaction. Such methods of doing business can increase productivity, decreased administrative overheads and lead to more efficient business models. Increasingly, the Internet is being employed to facilitate this exchange of information. Already, many businesses have their own websites – just another way in which information regarding the company and its business can be relayed via the Internet. This information, whilst delivered electronically, is largely consumed by the human computer user – prone to mistakes and slow to digest the data contained within the websites. By utilising automated data extraction processes and providing company data in a standardised format, this channel of data delivery could be used to provide more information than can be seen when simply viewing a web page in a browser.

Improving the consistency of company websites through the use of a mark-up language with strict rules on how the document is structured may help in attempting to allow computers to autonomously extract data. However, this only goes some way towards solving the problems experienced in the course of this research. Fewer errors in a web page's mark-up code will improve a data extractor's ability to read data from a document, but, for some data there are a number of different ways in which they can be defined within a page. For example, a hyperlink can appear within a number of different tags (e.g. The 'A' tag, the 'AREA' tag, the 'IMG' tag).

Therefore, a data extractor needs to check for every different way in which a piece of data can be defined – adding to the processing overhead. Also, some data of value may have no standard way of being described. This data may be within the main body of text of the web page, or it may be within an image and therefore unable to be extracted by a text based system. An example of such a piece of data is the name of the website's developer. There is no standard way of including this information within a web page, but it could be of use to someone wishing to find out which websites were developed by which developer. The free flow of information between companies through their websites is hampered by problems such as this. To overcome these problems requires changes to the way in which websites are viewed (in terms of an information delivery media) and described at the level of the markup language.

4. Semantic issues

Berners-Lee et al, 2001 describes what is being hailed as the next major evolution of the Web. The Semantic Web is seen as a way for the vast amounts of data present on the Web to be made meaningful to computers, not just users. Currently, computers have very little idea about the nature of the data held on the Web, they simply retrieve data as instructed and present it to the user to interpret. In this way the search for specific data can be a frustratingly inefficient one, with the computers of the Web providing little in the way of assistance beyond acting as a data delivery service. What the Semantic Web proposes is a way in which the data held by the Web can be given meaning, which computers will understand. This should, theoretically, allow much of the work involved in the search for data to be carried out by computers rather than their users, with machines being able to make informed judgements about relevant data based upon a certain set of rules.

One of the key issues in the Semantic Web is the use of meta data – data used to describe a resource on the Web (i.e. a web page). The use of meta data to describe the contents of a web page would, in theory, allow the computer to understand, to a certain extent, those contents, as well as the user. The idea of a descriptive language used to facilitate some form of automated processing of web pages is of interest with regards to this research. A web page containing meta data, guiding the extraction of valuable data could improve the performance of automated data extractors and help to eliminate the problems encountered in the course of this research. An example of how this could be implemented would be by using additional mark-up tags to highlight data for extraction. Information regarding each individual item of data could be provided within the tags, either in the tag name or as an attribute. Such information could describe, in textual form, data that is presented in a form that is not extractable, e.g. an address presented as an image. The web page would then become a hybrid document containing both HTML specific information for describing the presentation of the document, and information provided specifically for use by applications seeking to extract data from the document. Such information would have no effect on the presentation of the web page, therefore its appearance would go unchanged and the human user would be unaware of any changes to the content of the document. Highlighting the data that can be extracted could also have

the benefit of restricting the extraction of data – the data that the author of a website does not want extracting would not be highlighted in such a way (or would be highlighted as restricted). Whether or not an extractor ignores these restrictions is entirely down to the author of the extraction software – it is simply a matter of etiquette. This is not entirely unlike the situation with the current method of restricting access to resources on the Web using a robots.txt file (Koster, 1996).

A standard would be required to ensure that the description of extractable data was consistent across the Web and all data extractors would be capable of understanding what data was of value to it. The adoption of such a standard within the business world would provide greater integration between business processes and applications, and the publicly visibly data provided by company websites.

5. Conclusion

Given the vast amount of data available on the Web, automating the data extraction process is the only way to harvest large collections of it quickly and efficiently. Of the many software solutions already available to perform such a task, all must contend with the inconsistencies and errors inherent within Web documents. Many attempt to overcome these problems by pre-processing pages before extracting data – adding to the processing overhead.

Rather than the singular approach of trying to fit the application to the resources, a more collaborative effort to improve the contents of documents on the Web could lead to a greater flow of information. A standard already exists (XHTML) that, if adopted by website developers, would increase the consistency of the structure of web page data and, therefore, help to improve the performance of data extraction software. Coupling improved document structure with descriptive elements, that provide meaningful information for applications processing the document, would create an environment in which automated data extraction applications could thrive.

It is impractical to expect all web developers to alter their development practices to accommodate the changes outlined above. However, such changes could provide benefits to businesses. By providing an easy way in which to extract data relating to a company and its business from the company's website, the company could improve its visibility and draw the attention of interested parties. The flow of information between businesses, and between business and customer, would increase as the speed and efficiency of the automated data extraction process did. Accurate repositories of company data could grow, fed by data extractors, acting as directories which could be queried in order to find information that fulfils the customer's needs. Many uses could be found for the data extracted from company websites.

As stated earlier in this paper, any data found within a website may have some value. Improving the means by which data of value can be acquired could benefit the owner of the data as well as the seeker of the data. Further research into the areas identified within this paper could facilitate such improvements and lead to improvements in the description, and clarity, of documents on the Web.

6. References

Baumgartner R., Frölich O., Gotlob G. (2005). "Web Data Extraction for Business Intelligence: the Lixto Approach" [online]. Available http://www.dbai.tuwien.ac.at/proj/lixto/WebBI.pdf [Accessed 1st September 2005].

Berners-Lee T., Hendler J., Lassila O. (2001). "The Semantic Web". *Scientific American*. May 2001.

Bray T., Paoli J., Sperberg-McQueen C.M., Maler E., Yergeau F., Cowan J. (2004). "Extensible Markup Language (XML) 1.1". *W3C Recommendation* [online]. Available http://www.w3.org/TR/2004/REC-xml11-20040204/ [Accessed 22nd August 2005].

Chen H., Chau M. Zeng D. (2001). "CI Spider: a tool for competitive intelligence on the Web". *Decision Support Systems*. Vol. 34, No. 1, pp. 1-17.

Laender A.H.F., Ribeiro-Neto B., da Silva A.S. (2002). "DEByE – Data Extraction by Example". *Data and Knowledge Engineering*. Vol. 40, No. 2, pp. 121-154.

Lage J.P., da Silva A.S., Golgher P.B., Laender A.H.F. (2004). "Automatic generation of agents for collecting hidden Web pages for data extraction". *Data and Knowledge Engineering*. Vol. 49, No. 2, pp. 177-196.

Koster M. (1996). "Evaluation of the Standard for Robots Exclusion". *The Web Robots Pages* [online]. Available http://www.robotstxt.org/wc/eval.html [Accessed 22nd August 2005].

W3C HTML Working Group (2002). "XHTML 1.0 The Extensible HyperText Markup Language (Second Edition)". *W3C Recommendation* [online]. Available http://www.w3.org/TR/xhtml1/ [Accessed 12th September 2005].

Changing Trends in Vulnerability Discovery

S.W.Tope[1], S.M.Furnell[1], M.Papadaki[2] and G.Pinkney[2]

[1] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2] Symantec, Hines Meadow, St Cloud Way, Maidenhead, Berkshire, United Kingdom
e-mail: info@network-research-group.org

Abstract

There is an increasing awareness of vulnerabilities in computer software. Vulnerabilities have to be found before an exploit can take place. Thus it is up to those wishing to seek exploits and those seeking defences or remedies to find these vulnerabilities. This research presents the results into changing trends and focus on the different operating systems towards the most widely used ones. It also examines the shift towards exploitation of vulnerabilities in other software, such as applications, as well as revealing how some of the types of vulnerabilities are declining whilst others are as frequent as ever.

Keywords

Vulnerabilities, Operating Systems, Applications, Exploits

1. Introduction

Vulnerabilities and exposures to exploits in software have become more widely known and there is an increasing awareness as to how they can be exploited. A vulnerability is considered to be a "security flaw found in a certain technology. The technology may be an operating system, an application program, a network protocol, a mathematical algorithm, or sometimes a hardware component" (The Honeynet Project, 2004). It is known that vulnerabilities exist and the growth in awareness has led to much more comprehensive recording and sharing of information of such vulnerabilities.

An exploit is where there is an attempt to take advantage of the vulnerability. This may be by use of a program or by manual methods, although it is not always easy to achieve. The rewards to some have now become apparent and increasingly popular. The view of Kaspersky is that "the use of system exploits to get a foothold in the corporate network and spread rapidly has now become commonplace as writers of malicious code have woken to the potential 'helping hand' provided by vulnerabilities in common applications and operating systems" (Kaspersky et al, 2004).

Thus the fact that the vulnerabilities exist is not in itself very useful in designing or preparing defences. It is the types of vulnerabilities that exist and their exploitation that is of additional use. This does not mean that vulnerability trends and the

knowledge of where they are being found are not of importance. If we do not know the types of software that are vulnerable and open to exploits, then there is a lack of direction as to where to place defences or how to correct the problems.

The aim of the investigation presented in this paper was to ascertain whether certain operating systems were the focus of attention more than others, and whether this focus has remained the same. Additionally, it aims to determine whether the focus of attention is on operating systems or whether it has shifted to other software such as applications. If there is such a change in focus, either between operating systems, then why has this taken place? If a change in focus for vulnerability seekers is taking place, how are the software developers coping, so as to develop more secure code, or are the same vulnerabilities still being discovered?

2. Reporting standards and databases

There are variations in the figures given by different sites and reporting authorities as to the number of vulnerabilities. In the early days, there may well have been a lack of understanding of the vulnerabilities and their importance. There was also a lack of cooperation, collection of statistics, and reporting or sharing of information. Additionally, there was not the knowledge or publicity of the ways and methods of utilising such vulnerabilities for malevolent or other purposes. Thus some databases show a more dramatic increase in the number of vulnerabilities as they have fewer recorded in the past and are more ready to accept reported vulnerabilitities now without checking on the actual vulnerability.

The de facto standard in the security industry is the Common Vulnerabilities and Exposures (CVE) started by MITRE in 1999 (Rhose, 2003). The idea came from work done by Mann & Christey (1999) where they found different names for the same vulnerabilities being used by different sites. They realised that there was a need for a common and standardised approach. This involved consistency in naming as well as free and complete sharing of information (Rhose, 2003). Prior to being recognised as CVEs vulnerabilities are referred to as CANs (Candidates). The numbering method to date has given different prefixes of ‘CVE’ and ‘CAN’, though from 19th October 2005, they will all have a CVE prefix with a status line.

Such is the importance and recognition now placed on vulnerability discovery and recording that the Department of Homeland Security in the USA has revealed that the National Vulnerability Database (NVD) will be maintained by the National Institute of Standards and Technology (NIST). There are a number of databases maintained by companies such as Secunia (secunia.com) and SecurityFocus (securityfocus.com) and these can be of equal importance. Rhose (2003) considered that an advantage of Bugtraq was the speed that information was updated and with less formality. One problem is that many reported vulnerabilities fail to be actual vulnerabilities or exposures.

3. Utilisation of data

The primary sources that was utilised for extrapolating data in this study were the CVE project and the National Vulnerability Database maintained by NIST. Additional utilisation of databases included those maintained by Secunia and SecurityFocus. The investigation also looked at other research carried out by companies such as the technology research firm Yankee Group. This information is more easily extrapolated and can be used as a basis for comparisons.

The decision to concentrate on the information provided by the CVE project is in part due to the consistency of the information provided, its history and not least the high regard with which it is held. In addition, the information provided in the National Vulnerability Database was utilised not so much for absolutes and the overall trend but rather as a means of comparing different operating systems. This was useful when comparing like for like software.

Other databases, such as Bugtraq, were not used due to the length of time they have been operating and the consistency of their data for comparison over any length of time. Initially, there were few reports but this has risen rapidly so that almost any flaw or bug is reported, including many that are not vulnerabilities. Other research and work was still examined in order to compare with the extrapolated data for comparison. The SecurityFocus newsletters were included in the research. This was in order to discover the changing focus that has taken place on Microsoft and Linux distributions as well as how the vulnerabilities were reported.

4. Results

The figures for the MITRE list show a fairly constant rate of vulnerability reports and recording since 1999. It should be remembered that some of the CANs may still be rejected, especially for the more recent years. The figures for the National Vulnerability Database show a constant rise though there was a fall in 2003. Where they both agree is the sudden and rapid rise in reports in 2005. It should be noted that the number for 2005 had already passed the figure for 2004, even though only the first six months were included (see Table 1 below). The daily average is rising and for example, on 12.7.2005 there were 35 new CANs and on 26-27.7.2005 there were 45 new CANs. These are not unusual figures and certainly reflect the trend at the time of the study.

Year	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005** (to 1.7.2005)
CVEs & CANs	1562	1202	1396	1580	1085	1704	2104
NVD (NIST)	914	1013	1672	1858	1189	2161	2222

Table 1: Vulnerabilities Recorded

Source: http://www.cve.mitre.org/ and http://nvd.nist.gov/

Whilst there are still vulnerabilities being discovered in the Windows operating systems as can be seen in Table 2, they are not as dramatic as they once were. The figures for XP Professional actually declined and though rising once again are doing so at a similar rate to the rest of vulnerabilities being discovered.

Year	2000	2001	2002	2003	2004	2005 (to 1.7.2005)
Number of Vulnerabilities	1	-	22	19	17	28
As % of Total Vulnerabilities	0%	-	1%	2%	1%	1%

Table 2: Windows XP Pro

Source: http://nvd.nist.gov/

The results for Red Hat Linux (the most common distribution of Linux) are similar for 2002 to 2004. Prior to that, Windows XP had not been released and as such cannot be compared. Whilst the figures are similar, they are for Red Hat Linux 6, 7, 8 and 9. Thus they are actually different versions, but often a user will upgrade.

Year	2000	2001	2002	2003	2004	2005 (to 1.7.2005)
Number of Vulnerabilities	47	49	17	22	10	12
As % of Total Vulnerabilities	5%	3%	1%	2%	0%	0%

Table 3: Red Hat Linux

Source: http://nvd.nist.gov/

These figures do not include vulnerabilities in all kernel revisions. This is a difficulty in comparison and Microsoft will keep their operating system going for longer (with the possible exception of the Millennium Edition) with service packs being added. Additionally, there is some conflict with the figures for the Linux Kernel as well as other Red Hat systems. Red Hat Desktop had 19 vulnerabilities in 2004 and 48 in the first six months of 2005, whilst the figures for Fedora were 20 and 47 (source: http://nvd.nist.gov/). Prior to 2004, the recorded vulnerabilities for the Linux Kernel never exceeded 21, and in 2003 they stood at 16 for the entire year. The past two years has seen a dramatic rise to 45 in 2004 and 58 in the first six months of 2005. This is continuing with 71 being reported in 2005 as at the end of August.

In examining a number of minor or more obscure operating systems, there were often no vulnerabilities recorded. Whilst much publicity has been made of mobile malware, there is only a single recorded vulnerability (CAN-2005-0681) on the Symbian operating system, four affecting specific handsets and four others relating to areas such as Nokia Electronic Documentation. Apart from the very obscure operating systems, the research did examine others such as the different forms of BSD. It was only Free BSD that has a significant number of vulnerabilities to date with some attention paid to Net BSD and Open BSD. The remainder such as 386 BSD, BSD I, BSD/OS and Eclipse BSD barely register (see Table 4).

BSD Variant	**Number of Vulnerabilities** (up to 19.5.2005)
Free BSD	265
Net BSD	88
Open BSD	77
386 BSD	0
BSD I	7
BSD / OS	9
Eclipse BSD	0

Table 4: BSD Vulnerabilities

Source: http://www.cve.mitre.org/

The same can be said for Unix variants. Whilst HP-UX, IRIX, SCO and Solaris have a significant number of vulnerabilities, others such as Sun OS, System V and TRU64 barely register (see Table 5).

Unix Variant	**Number of Vulnerabilities** (up to 19.5.2005)
HP-UX	127
IRIX	127
SCO	131
Solaris	226
SUN OS	34
System V	33
TRU64	32

Table 5: Unix Vulnerabilities

Source: http://www.cve.mitre.org/

Year	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005** (to 1.7.2005)
Number of Vulnerabilities	169	142	160	227	91	134	93
As % of Total Vulnerabilities	18%	14%	10%	12%	8%	6%	3%
Total number of Vulnerabilities	914	1013	1672	1858	1189	2159	2223

Table 6: Microsoft Products

Source: http://nvd.nist.gov/

All of the Unix variants show a decline in the number of vulnerabilities being discovered and reported. SCO had 27 Candidates in 2004 and to date (19.5.2005) has not had a single reported vulnerability in 2005. Solaris is down to 10 from 26 in 2004, IRIX to 3 compared with 11 in 2004, and HP-UX has 4 compared with 10 in 2004. This is at a time when the overall numbers are going up elsewhere. Thus attention is certainly moving away from the Unix operating systems. This may well be in part due to the maturity of the systems and the lack of new kernels, versions etc. It was observed in the statistics maintained by Secunia that the number of vulnerabilities for Red Hat Linux began with a surge and then fell away. The problem in interpreting these figures is that as a new version of the distribution was

released, so the focus of attention shifted. Still, the trend was a downward one, even prior to the release of newer versions of the 'distribution'.

The downward trend of Microsoft products in the number of vulnerabilities that are being found is supported by recent research by the Yankee Group. They found that "Microsoft flaws continue to flow – but at a significantly reduced rate" (Jaquith, 2005). Their view using NIST ICAT data (now NVD) was that the focus has shifted towards security products. This may in part be due to security companies seeking to discredit each others' products. Their study revealed that they were responsible for 26% of vulnerability discoveries involving rival security products. Another view is that the growing focus and danger is with drivers. There is a view that device drivers are the most dangerous as they are part of the kernel and the quality of software developer is not as high as that of the operating systems. In an audit of the Linux 2.6.9 kernel by the security firm Coverity, over half of the flaws were in device drivers. (Lemos, 2005). Thus it may not just be a general switch of attention to other software, but in particular types of software such drivers, networking and security software that have more access to the system as well as potential for damage. In 2004 Cisco had 75 vulnerabilities recorded (NVD) compared to 29 in 2003. The actual figures for 'drivers' do not as yet support the above theory. Mitre had only 45 entries (as at 1.7.2005), and although there was a jump in 2004, the figures are not large enough to have any marked impact on the overall statistics. Certainly, within Linux applications there are few vulnerabilities in the 'Administration' category (172 in 505 applications- of which 71 were for 'ethereal'), in the System category (233 in 1013) but alters noticeably in Networking Applications. In the DNS section of Networking there were 54 vulnerabilities in 40 applications, 'e-mail' had 201 in 163 applications, and 'firewalls' had 108 in 102 applications.

Instead of looking at the attention paid to the operating system in isolation, the operating system as a whole, with applications and drivers installed was investigated. SecurityFocus (www.securityfocus.com) has been publishing weekly newsletters since September / October 2000 for vulnerabilities in Linux and Windows (as well as a general newsletter). In June / July 2001 (a period of 9 weeks – see Figure 1) there were 129 Microsoft Vulnerabilities reported and only 28 for Linux (including applications, drivers etc.). From the beginning, the Linux newsletters included vulnerabilities in applications and other software that affected Linux. Initially the Microsoft Newsletters comprised Microsoft products only. The method of compiling the Microsoft newsletters began to change towards the end of 2001 depending upon the editor (#60, dated 12.11.2001, was the first), though the last to include only Microsoft products was #111 dated 4.11.2002 (6 vulnerabilities in Microsoft products).

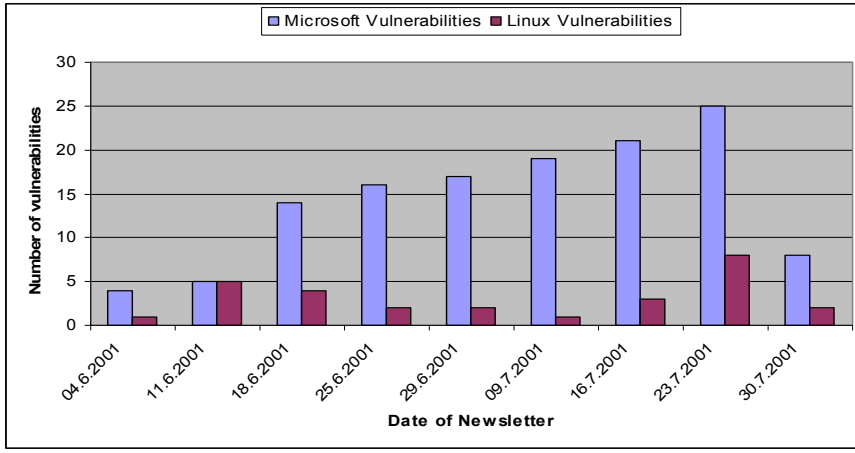


Figure 1: SecurityFocus Newsletters

Source: http://www.securityfocus.com/newsletters

After these were included in the Microsoft Newsletter, the relative difference between the two systems has significantly altered (see extrapolated quarterly figures in Figure 2). Apart from a fall in 2003 / 2004 that matches the general trend, the numbers are once again rising. This significant rise began in the last quarter of 2004 and is being sustained.

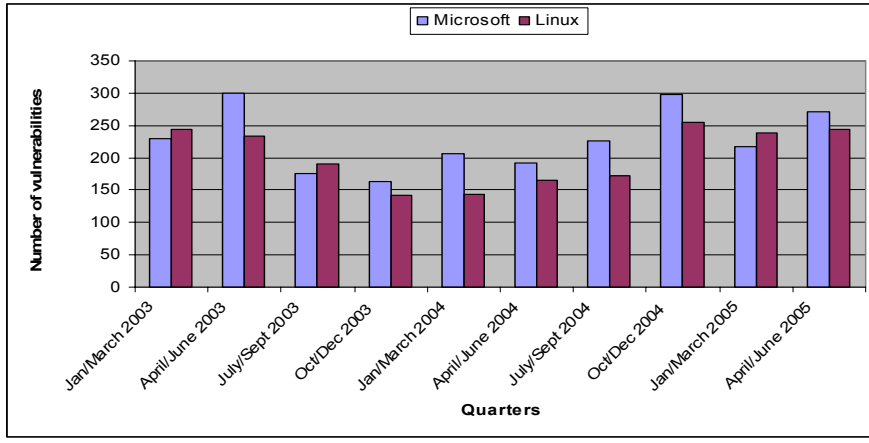


Figure 2: SecurityFocus Newsletters

Source: http://www.securityfocus.com/newsletters

It is certainly clear that there is less emphasis on Microsoft products per se and as the figures are similar from one Operating System to the other, the implication is that more emphasis is being paid to software other than the operating systems themselves. It is not so much of a question as to trends in operating systems now but rather as an overall platform and how they are affected by other interacting pieces of software. Thus the attention may well be shifting away from operating systems.

The trend of increases in applications relating to operating systems and the changes in the variety of software and operating systems available has led to SecurityFocus now reviewing their 'security mailing lists'. It is felt that the present lists are too confining as so much is contained within a finite number of lists. They now expect to have more mailing lists in the near future.

NIST lists different vulnerability types in its National Vulnerability Database. These are:

- Design Error
- Race Condition
- Configuration Error
- Environmental Error
- Exceptional Condition Error
- Access Validation Error
- Input Validation Error – (a) Buffer Overflow & (b) Boundary Condition Error

Most of the errors have remained fairly constant as a proportion to the overall numbers of vulnerabilities. This is not so for Design Error that has fallen from a peak of 29% in 2002 to 14% in the first six months of 2005. Configuration Error has fallen from 6% to 2% over the same period of time, and Buffer Overflows have fallen from 23% in 2003 to 12% in the first six months of 2005 (even though the overall percentage for 'Input Validation Errors' has remained constant). As different software usage becomes more popular so does the emphasis in different types of attack. Vulnerabilities involving SQL Injection have risen from 7 in 2001, to 42 in 2003, 108 in 2004, and 209 in the first six months of 2005. Vulnerabilities involving PHP have risen from 35 in 2003 to 120 in 2004, and 165 in the first 6 months of 2005.

5. Discussion

The figures when comparing different operating systems indicate that vulnerabilities tend to be discovered and reported principally in those operating systems that are more commonly used. When an operating system has reached a certain point then the number of vulnerabilities that are discovered and reported becomes somewhat similar. Operating systems have been the focal point in the past and the numbers of vulnerabilities in the principal operating systems has generally remained at the same numeric level though rising in 2005. Figures for minor or rarely used operating systems are certainly not rising. There has certainly been a shift away from Microsoft products as the source of vulnerabilities, possibly due to increased attention to security as well as the diminishing returns offered to vulnerability seekers. They are still a major focus of attention due to their market presence, as well as certain hostile views that are commonly held by some in the computing fraternity.

There is some evidence of a shift towards applications and other software, though specific to certain types such as networking and security products. This does not

mean that there is less emphasis upon the principal operating systems as these are maintaining their share of vulnerabilities being discovered. Rather, software other than operating systems is coming under increasing scrutiny. The evidence shows that there is convergence between Microsoft and other vendor applications. This convergence may alter, as rising trends that other vendor applications have shown may continue. The rate of vulnerabilities being discovered is rising rapidly since the end of 2004, and this rate would appear to be similar for the principal operating systems as it is for non-operating system software.

The rise and shift in focus for those seeking vulnerabilities is still developing. Some products are more mature than others and thus may be written with different levels of attention to security. The larger companies are aware of the attention that the media and public now pay to security and the discovery of vulnerabilities. Certain types of vulnerabilities have received considerable publicity in the past, principally 'buffer overflows or overruns' and these have declined considerably along with 'design errors'. Certainly the larger companies are more aware of how to ensure that code is written in a manner that to avoid these types of vulnerabilities and security audit tools are better at locating these vulnerabilities before being released. Despite this, the problem has not been eradicated and even the most well known software still suffers from these types of vulnerability. Certainly, as software such as PHP becomes more popular, so the attention of those seeking vulnerabilities becomes focused. This is not only due to the popularity of the software, but also the increase in awareness by the vulnerability hunters.

6. Conclusions

The focus of attention remains upon the software that is more commonly used. Whilst Microsoft has taken steps to produce more robust and safer software, those seeking vulnerabilities have started to look elsewhere. They will still concentrate on the most commonly used operating systems or software in which a vulnerability will possibly impact the most. It is not so much that Microsoft does not remain the focus of attention, but rather there is an increased awareness of vulnerabilities elsewhere, and as Linux distributions become more user friendly and increase in popularity so there is more attention being paid to their flaws. The focus of attention is still directed towards specific areas and is not arbitrary.

7. References

Jaquith A, (2004) "*Fear and Loathing in Las Vegas: The Hackers Turn Pro*" http://www.yankeegroup.com/ public/products/decision_note.jsp?ID=13157 Accessed 29.6.2005

Kaspersky E, Emm D, Gostev A, and Blanchard M (2004), "*Malware Trends in 2004*", http://www.viruslist.com/en/trends Accessed 15.02.2005

Lemos, R (2005) "*Device Drivers filled with Flaws, threaten security*" http://www.securityfocus.com/ print/news/11189 Accessed 29.6.2005

Mann D & Christey S (1999) "*Towards a Common Enumeration of Vulnerabilities*" http://www.cve.mitre.org/ docs/cerias.html Accessed 12.2.2005

Rhose M, (2003) "*Vulnerability naming schemes and description languages: CVE, Bugtraq, AVDL and VulnXML– A SANS GSEC practical*", http://www.sans.org/rr/whitepapers/threats/ 1058.php Accessed 12.12.2004

The Honeynet Project (various) (2004) *Know Your Enemy* 2nd Edition Addison-Wesley Professional ISBN: 0321166469

Uses and dangers of peer-to-peer and instant messaging in a business environment

T.Quaden[1], S.M.Furnell[1], M.Papadaki[2] and G.Pinkney[2]

[1] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2] Symantec, Hines Meadow, St Cloud Way, Maidenhead, Berkshire, United Kingdom
e-mail: info@network-research-group.org

Abstract

Peer-to-peer (P2P) and instant messaging (IM) applications have become very popular ways of downloading the newest media files and to chat with friends. When introduced to a business environment and not properly managed these applications present the business with new risks that might not be accounted for. This research paper focuses on the dangers of P2P and IM applications from the perspective of an organisation and discusses some measurements that can be taken to minimise these. It was found that both IM and P2P applications open up holes for all kinds of malware to enter the corporate network. Furthermore due to the nature of the majority of content downloaded from P2P networks (copyrighted files and inappropriate material) a business might be held liable for illegal or inappropriate material downloaded by its employees. Even though, when properly managed, IM software can greatly increase business performance, if unmanaged can lead to incidents such as data or even identity theft/spoofing. There are several ways to minimise or prevent the risks of using such applications. Well-defined security policies can help educate users of the risks; management software helps detect, block or manage IM & P2P applications; and some companies even offer Enterprise IM solutions.

Keywords

Insider threats, Peer-to-peer (P2P), Instant Messaging (IM)

1. Introduction

In today's world with broadband Internet access widely available, peer-to-peer (P2P) and instant messaging (IM) applications have become more popular than ever. As Figure 1 shows, by mid 2004, 90% of commercial and non-commercial enterprises in North America were using IM software (Ostermann Research, 2004) and market research firm IDC estimated that more that 506 million people world wide would be using IM products by 2008 (Leavitt, 2005).

P2P applications share a similar popularity. A study in 2004 found that about 40% of business Internet users have used P2P applications to download and share files online using their business network (Ostermann Research, 2004). One of the most popular P2P applications, Kazaa which uses the FastTrack filesharing network, usually has in excess of 2 million users online at any time (see Figure 2).

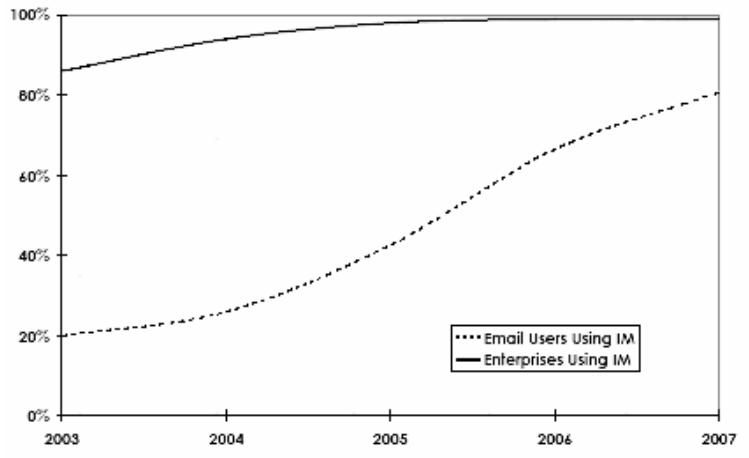


Figure 1: predicted increase in IM use (Ostermann Research, 2004)

Some of these applications also have legitimate uses in a business environment but they can also introduce new problems to the business networks they are used on, especially if not properly managed. This paper starts by identifying some legitimate uses of such applications in a business environment, discusses some of the problems associated with them and follows up with suggestions on how to avoid these problems.

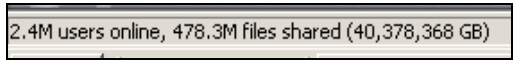


Figure 2: Number of users on Kazaa network and amount of files shared

2. Business uses of IM & P2P

If used properly, IM can help increase business performance. IM offers faster response times than e-mail, and is cheaper and less intrusive than using a telephone. The study "Measuring IM Productivity in the Enterprise" conducted in 2004 by the Radicati Group, found that the use of IM can save a company an average of 40 minutes per user every day, resulting in about $37.5 million per year in productivity savings when applied to a 5.000 employee company (Instant Messaging Planet, 2004).

P2P applications are mainly used to download and share all kinds of files, such as music, movies, and software applications. Most of these however are copyrighted and are therefore distributed illegally, which can lead to expensive law suits for which companies can be held liable. There are, however, some legitimate uses of P2P technology within a business. P2P can be used as a means to increase a business' storage capacity or to use previously unused processing power among their workstations and servers. University of Wisconsin researchers estimated that on

average most businesses only use about 25% of their available storage (Computerworld, 2001). Using P2P technologies, businesses would be able maximise the storage they use and therefore be able to save money that was previously spent on storage servers. These P2P applications, however, differ entirely from those like Kazaa, which generally have no legitimate use in a business environment and only use up valuable bandwidth.

3. Problems & Risks associated with IM & P2P

Most IM and P2P applications used in enterprises are consumer-grade clients that are not specifically made for use in business environments and in most cases have been brought into the business network by users installing their own clients without the IT department's permission or knowledge (see Figure 3). In March 2004 the mean number of different IM applications in use by employees throughout American businesses reached 2.9 (see Figure 4). The result of this is that in many businesses employees use a variety of different IM and P2P applications that are not managed in a proper way and therefore present the business network with new risks not accounted for.

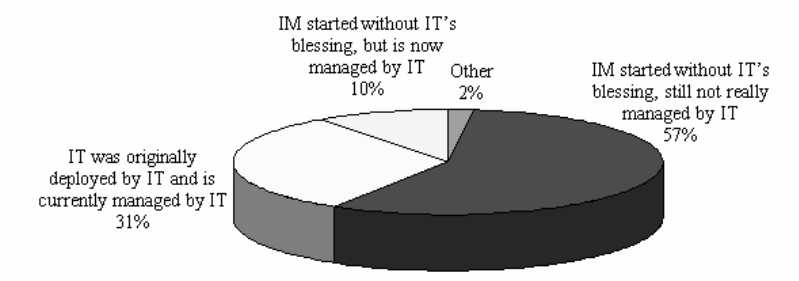


Figure 3: Methods by which IM entered the business (Ostermann Research, 2004)

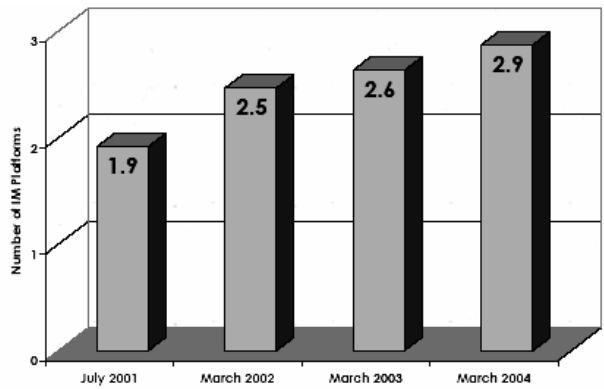


Figure 4: Mean Number of IM Platforms per Enterprise (Ostermann Research, 2004)

3.1 Copyright Violation

The most obvious danger to a business that comes to mind when users use P2P applications is copyright violations. This is due to the fact that most users use such applications to download media files (i.e. music or movies). Most of these are protected by copyrights which makes downloading such files without having purchased them illegal. If such copyrights violations are detected within a business it can lead to lawsuits that can cost the business considerable amounts in fines and in terms of reputation loss. In early 2005 the British Phonographic Industry (BPI) announced that in 23 cases illegal music uploaders had to pay up to £4500 as compensation (Leyden, 2005). More serious cases in the U.S. made illegal movie sharers pay up to $30,000 or even up to $150,000 if it was done 'wilfully' (Sherriff, 2004).

3.2 Inappropriate material

One popular use for P2P application is the downloading of inappropriate material such as pornography. Viewing of inappropriate accounted for 47% of computer abuses in the UK in 2004 (Audit Commission, 2005). Viewing of inappropriate material creates can create a hostile work environment and result in damaging the reputation of a business quite considerably. To demonstrate how inappropriate material can damage a business' reputation just consider this simple example: an employee finds an image or video clip of some inappropriate nature (e.g. pornography or some racist/sexist joke) and decides to use the company email to send it on to a bunch of friends. These (possibly working at different businesses) then send it on to others. Eventually this email could somehow end up in someone's email that might be offended by it and because the original source is still in the email header they can recognise the person and most likely also the business it originated from. This could give the entire business a bad reputation and might even result in loss of business with companies that feel offended by it.

In the case of employees viewing illegal material (e.g. child pornography) the business can possibly be held liable for it which can result in high fines and in confiscation of network hardware for investigation resulting in business disruption and bad publicity.

Additionally the downloading of inappropriate material wastes network bandwidth that could otherwise be used for legitimate business operations.

3.3 Malware

Another serious risk when using P2P and IM software is malware. There have been hundreds of viruses, Trojans and other malicious pieces of software circulating the various P2P and IM networks. The Symantec search web site (Symantec, 2005) comes up with 441 results for Kazaa, 605 results for MSN, 418 results for ICQ, 100 results for AIM and 936 results for IRC when searching for malware and vulnerabilities. IMLogic has set up a threat center to monitor IM and P2P threats worldwide. During the first week of September 2005 alone it reports 43 different malicious pieces of software that circulated P2P and IM networks, most of them

circulating the IRC network (IMLogic, 2005a). This shows the vast amount of vulnerabilities and risks of malware infection when using such applications.

Bruce Hughes, director of malicious-code research at security firm TruSecure, found that about 45% out of 4,778 files he downloaded with Kazaa contained malicious code such as viruses or Trojans (Wired News, 2004). Hughes also considered that about 85% of the malicious files can easily be detected by up-to-date anti virus software. To confirm these statistics, 30 executable files (.exe) were downloaded during this research using Kazaa while running Mc Affee VirusScan Enterprise 7.0.0. Out of the 30 files downloaded 14 were found to be infected by the Trojan 'Downloader-TS' and one file by the virus 'W32/HLLP.Hantaner.dam' (see Figure 5). Considering that 50% of the files downloaded were infected by malware, the results are similar to Hughes' research, demonstrating the high risk of malware infection when using P2P applications such as Kazaa.

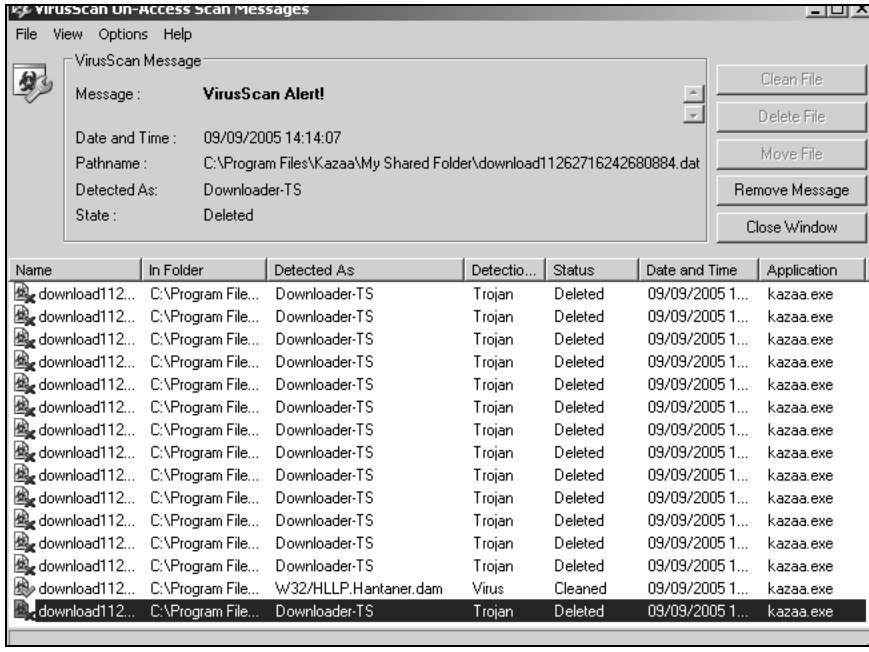


Figure 5: Trojans and viruses found downloading software from Kazaa

Many of these worms or Trojans can compromise system security by creating backdoors or lowering system security leaving systems vulnerable against more active attacks such as system intrusions. An example of this is 'Gabby.a', a worm that targets AOL's AIM and ICQ networks tries to trick users into clicking a hyperlink that leads to a webpage which then infects the user's computer with a worm that opens a backdoor into the system and stops windows services such as firewalls and antivirus software (Leavitt, 2005).

The main problem is that most IM and P2P applications are very adept at circumventing network firewalls and intrusion detection systems (IDSs) by finding

open ports to use and therefore opening up security holes in otherwise fairly secure networks. This makes it very hard if not impossible to block all P2P and IM traffic from passing through the firewall. Additionally most IM applications use different ports and protocols to communicate with their servers making it even harder for administrators to block all of them.

Even when blocked some employees will still have the desire to use them and may find ways around the block by changing the ports the applications use in the configurations or by using applications such as Hopster. Hopster acts as an anonymous proxy between the user's computer and other computer on the Internet and enables users of popular P2P and IM application to bypass censoring firewalls and proxies by making traffic look like 'innocent' HTTP requests (Hopster, 2005).

3.4 Data theft

The risk of data theft applies to both P2P and IM software. Due to the nature of P2P applications where users share files on their computer with the rest of the P2P network users it can happen, especially when used by unknowing users, that folders with private/confidential data are accidentally shared. This can obviously damage a business quite considerably. A very good example of such an incident happened start of 2005 when highly confidential documents about human traffickers belonging to the Dutch armed forces were found online. They contained at least 75 pages of phone numbers and tapped conversations, and (according to Dutch newspapers) they were first found on a P2P network in unencrypted form (Libbenga, 2005). It is assumed that an employee of the Dutch armed forces took the documents home to work on and accidentally shared his entire hard drive when using a P2P application.

Since most IM applications also allow file transfers between users this risk also exists, but in most cases the user would deliberately have to send the files to someone else. This however, is not the only way in which data can be stolen. Most consumer-grade IM clients (e.g. MSN Messenger, AIM) do not support internal network routing. This means that even if the recipient of a message is within the same internal network, the message is first sent to an external server and then back into the network. This fact combined with the fact that most clients also do not support any form of encryption means that intercepted messages can easily be read with the use of any standard packer sniffer, further increasing the risk of confidential data being stolen.

3.5 Identity theft and spoofing

Another concern with the use of consumer-grade IM applications is that users choose their own names which are not controlled by any company policies. This can lead to an outsider pretending to be a co-worker and using social engineering attacks to gain confidential data or valuable information to use in an attack against the business network. IM Authentication mechanisms also often lack sufficient encryption, which can result in account hijacking and further social engineering attacks by outsiders and competitors.

4. Recommendations

Since it is very difficult, if not impossible, to block all P2P and IM traffic, it becomes very important to have defined security policies regarding the use of such applications. SurfControl, a corporate Internet security vendor, conducted a survey of U.S. businesses that identified that even though 90% of respondents had a policy regarding Internet use but only 51% had a policy regarding the use of IM software (Leavitt, 2005). This shows that many businesses either do not seem to be fully aware of the security risks involved or do not realise how many of their employees actually use IM.

Especially in small to medium sized enterprises, where policies are easier to enforce, a well- designed policy that educates the users on the dangers of IM and P2P might help reduce the use of such applications considerably. Policies could either completely forbid the use of such applications or in cases where employees use consumer-grade IM applications for their work they could specify a specific client that all users may use in order to provide more control. If the business knows what applications are used and knows about the protocol used by these, it becomes much easier to monitor the network traffic and discover any unusual behaviour such as the propagation of malware or inappropriate material.

Some companies specialise in developing IM management software which allows administrators to detect, block and manage IM and P2P traffic and authentication. An example of such software is IMLogic's IM detector pro (IMLogic, 2005b). Some of these management tools allow administrators to set up alerts that can be triggered by certain actions such as file transfers or keywords within IM conversations in order to reduce the risk of disgruntled employees handing out confidential data to competitors (Richeson, 2003).

The best option for businesses wanting to use IM is to purchase IM software specifically developed to be secure in business environments. There are companies that developed Enterprise IM (EIM) solutions, including AOL and IBM, that support features such as encryption, proper authentication against local directory service, internal network routing and many other features that otherwise require specialised IM management software. EIM also enables administrators to control client functionality allowing different users to use different features such as file transfers, voice and video chat depending on what they require to fulfil their job. Overall EIM solutions allow for much greater flexibility and customisation while providing a business with a secure IM solution that can greatly increase business performance and result in quicker responses and money savings.

5. Conclusions

This paper clearly identifies that P2P and IM applications can present a business with serious risks. It also identifies a lack of awareness of these risks on both user's and administrator's side. While some P2P applications were found to have little to no use within a business environment, IM applications, even though they present some

risks, can help greatly in improving business performance. There are many ways to manage the use of such applications but choosing the right policies and security measures depends entirely on each business' needs and budget. Businesses will have to assess the risks and compare them to the benefits that they can gain from IM technologies. Future research could investigate into ways of calculating return on investment of IM implementation and possible tools to aid with these calculations.

6. References

Audit Commission (2005), *ICT Fraud and Abuse 2004 - An update to yourbusiness@risk*. Audit Commission Publications, UK. June 2005

Computerworld (2001), "*Potential Uses Help Brighten Future of P2P*", http://www.computerworld.com/managementtopics/ebusiness/story/0,10801,58004,00.html, (Accessed 11.09.2005)

Hopster, (2005), www.hopster.com (accessed 11.09.2005)

IMLogic (2005a), *IMLogic threat center*, http://imlogic.com/im_threat_center/index_viewall.asp?mr=top3&hr=top3 (Accessed 11.09.2005)

IMLogic (2005b), *IM detector pro*, http://www.imlogic.com/products/im_detectorpro.asp (Accessed 11.09.2005)

Instant Messaging Planet (2004), "*Study: Enterprise IM Could Reap ROI in Days*", http://www.instantmessagingplanet.com/enterprise/article.php/3448631 (Accessed 11.09.2005)

Leavitt, N. (2005), "Instant Messaging: A New Target for Hackers", *IEEE Computer Society*, http://csdl2.computer.org/comp//mags/co/2005/07/r7020.pdf (Accessed 11.09.2005)

Leyden, J., (2005) "*BPI nails 'music pirates'*", published 04.03.2005, http://www.theregister.co.uk/ 2005/03/04/bpi_fileshare_settlements/ (Accessed 11.09.2005)

Libbenga, J., (2005), "Classified Dutch military documents found on P2P site", *The Register*, published 30.01.2005, http://www.theregister.co.uk/2005/01/30/dutch_classified_info_found_on_kazaa/ (Accessed 11.09.2005)

Ostermann Research (2004), "*Managing IM and P2P Threats in the Enterprise*", http://wp.bitpipe.com/resource/org_971197299_840/Osterman.pdf (Accessed 11.09.2005)

Richeson, J. (2003), "*Finding the Right Instant Messaging Solution for Your Company*", SANS Institute, http://www.tietronix.com/pressCenter/WhitePapers/Instant%20Messaging%20and%20Security.pdf (Accessed 11.09.2005)

Sherriff,L., (2004), "*MPAA takes filesharers to court*", published 17.11.2004, http://www.theregister.co.uk/ 2004/11/17/court_mpaa_suits/ (Accessed 11.09.2005)

Symantec (2005), Symantec search, http://www.symantec.com/search/ (Accessed 11.09.2005)

Wired News (2004), “Kazaa Delivers More Than Tunes”, http://www.wired.com/news/business/0,1367,61852,00.html (Accessed 11.09.2005)

Section 4

Computing, Computer Applications, E-Commerce and Interactive Intelligent Systems

Impact of E-Commerce on International Supply Chain Management in Shanghai Custom Department

B.Xu and A.Phippen

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

After China's accession to WTO, with an increasing growth of international trade in China, the importance of international supply chain has accepted a huge attention. Undoubtedly, effective international supply chain management will contribute to improving free trade with globalising trend. In this instance, facing fierce challenges outsides, in order to ensure high-quality operation of international supply chain, the extent to which the performance of Custom Department in China is extremely important. The more effectively Chinese government performs, the more profits the international and domestic business will be generated. In addition, China is becoming a "world wide manufactory". Local government in China has to establish a good business environment for organisations. As an important port city in China, the performance of Shanghai Custom Department (SCD) will directly influence the business environment in China. Thus, the introduction of E-commerce, which is positive factor to improve the development of international supply chain, has been explored and integrated into the service of Custom Department to meet the increasing demand of export and import.

Considering all kinds of factors concerned above, the paper aims to explore the research about whether the application of E-commerce plays an important role to improve the implementation of international supply chain in SCD. This paper is combined the methods of literature review and primary research including questionnaire and telephone interview to ensure the effectiveness of research. By analysing the information gained, a result about the extent of E-commerce to affect the operation of supply chain will be found. Based on the negative issue of E-commerce, which results in a low improvement to operate supply chain in organisations, this research offers a series of recommendation to enhance the use and impact of E-commerce.

Keywords

E-commerce, International Supply Chain, Shanghai Custom Department

1. Introduction

With the development of the international trade, how to master business timely to keep customer satisfied and meanwhile to be against all competitors in the market has received amounts of focuses among all kinds of organisations. In this instance, with some absolute advantages of supply chain management, which include "*the security of products availability, inventory management and just-in distribution and delivery*" (Gattorna and Walters 1996), the supply chain management has become an extremely important arm to assist companies to do business effectively. However, in practice, despite that the importance of supply chain management has been realised, it is still difficult to find an approach complete the optimisation of supply chain

management for most of organisations. This, in turn, results in the introduction of advanced and effective technology, called E-commerce.

According to the statement by Bocil, *et al* (1999), E-commerce can be defined as "*a broad technology platform to exchange of business data between two or more organisation's or firm's platform*". In fact, as a technological tool to promote the communication among people, E-commerce has been integrated into supply chain management in the process of doing businesses, which contributes to lowering amounts of costs involved in the international transactions for all of organisations.

As a result of China accessing into the world trade organisation (WTO), with a purpose of globalisation and liberalisation to trade, broad attentions have been focused on the international trade in China. In this instance, it is necessary to establish and offer an effective and fair business environment towards all kinds of international organisations for Chinese government. Facing these challenges and opportunities concerned above, therefore, this paper puts more insight into SCD, which is which is a key window to do international business towards outsides in China. Presently, SCD, undoubtedly, has applied with the technology related to E-commerce to operate custom clearance with many companies.

Hence, combined with these factors mentioned earlier, this report will aim to explore the research about the extent to application of E-Commerce on supply chain management from the aspect of SCD. During the research, primary methods including questionnaire and telephone interviews will be conducted.

2. Background

Lardy (2002) pointed out that, after November 2001, China has been a member of the WTO, which means the evolution of the trade related investment, non-tariff barriers, very specific market access benefits in goods and services. Thus, huge opportunities have been provided to foreign investors and overwhelming possibility to do international trade has been bridged. According to the survey of statistics, Fewsmith (2001) stated that Chinese tariffs on imports of industrial goods are dropping from average of 24.6% to 9.4% between 1997 and 2005. In the future, foreign firms will be authorised to sell, distribute and market industrial goods, such as steel industry. That is to say, the Custom Department will have to face a big challenge.

Under these circumstances, Shanghai has received widespread attention in terms of the huge amount of the Foreign Direct Investment (FDI) and soaring international trade. Therefore, Shanghai government has to provide an excellent logistics service to build good business environment and meanwhile reduce the supply chain risks when doing international trade. There is no doubting that SCD, in this instance, plays an essential and inevitable role to take responsibility for establishing a high-quality logistics service. (easipass.com 2001)

News.xinhuanet.com (2002) reported that in order to build a good logistics service successfully and effectively, in 1998, China Custom Department boosted a positive plan regarding the introduction of E-commerce. In practice, several modern cities in China were selected to test the effectiveness of E-commerce. Naturally, Shanghai became one of the experimental units. After that, by a series of evaluation and test of using E-commerce, SCD and some relative research institutions eventually, implemented the E-commerce system on the basis of EDI technology in 2002.

Li (2005) presented that, by using the E-commerce technology, the performance of operating the process of custom clearance in SCD has been improved rapidly. This phenomenon can be shown in the following table 1, which shows the distinct advantage of saving time cost due to using the E-commerce technology compared with the time cost occurred with manual customer clearance.

Kinds of Customer Clearance for export	Manual Average Total Time Cost	E-commerce Average Total Time Cost
Shopping	96 hours	30hours
Air	72 hours	14hours

Table 1: Comparing Manual system and E-commerce system (Li, 2005)

3. Research Design

In order to explore the research about the impact of E-commerce on supply chain management in SCD, in the first place, a questionnaire with the form of e-mail was conducted to collect primary data. Saunders *et, al* (2003) said that the principle of questionnaire design is to get the validity and reliability of the data collection. According to these rules, a questionnaire in Chinese was designed to email to 200 companies, which can be seen as a sample of the questionnaire.

During designing questionnaire, the questions were divided into five distinct parts for each respond. The table 2 displays the different part.

Part	Question No.	Description
1	1—4	General information about the respondents
2	5—8	To examine the extent of acceptance of E-commerce in SCD
3	9—13	To identify and evaluate current situation of E-commerce in SCD
4	14—16	To find out the problems of current situation of E-commerce in SCD
5	17—19	To discuss possible development of E-commerce in SCD in the future

Table 2: Main objectives of questionnaire

After that, all of responding was collected eventually, which occupied for 21% of respondents rate, and then a further analysis was conducted to test hypotheses related to the questionnaire.

In addition, it needs to note that because of the low respond rate, apart from the first method of questionnaire to collect primary data, the telephone interviews also were conducted further. By using the telephone interviews with five interviewees in different companies, including one people working in SCD and four people working in the domestic and international companies, more detailed information has been obtained so that contributes to doing research more clearly and deeply.

4. Results and Analysis

This research selected three main objectives to do survey, which includes the acceptance, current situation and the future development of E-commerce in SCD,
First of all, considering the extent to accept E-commerce in SCD, the result showed 66.7% (exactly yes) and 21.4% (generally yes) of respondents who indicated the concept of E-commerce in SCD is accepted. That is to say, there are around 88% of respondents who voted positive opinion about the acceptance of using E-commerce in SCD and E-commerce does have impacted on the operation of customer clearance in SCD.

In order to display the reason why there is the acceptance of using E-commerce in SCD, a special question was designed and offered several reasons to lead to the use of E-commerce in SCD. As a result, the respondents have given the quite high rates for cost reduction, service improvement and competitive ability improvement, there are 4.51, 4.37 and 4.02, respectively (5 is the highest rate), which can be realised that cutting cost is the most important reason and advantage related to using E-commerce.

Secondly, considering the current situation of using E-commerce in SCD, unfortunately, there is an important issue negatively influencing the growth of E-commerce. In general, depending on the different company size, the rate to use E-commerce with SCD is dramatically different, which obstacles the dramatic progress of using E-commerce. As shown in figure 1, the percentage of the small business and middle size firms using E-commerce in SCD (20%) were rather lower than the percentage of large and supper-large companies to use (60%). That is to say, the bottleneck existing in the operation of small and middle size companies forms a issue of applying with E-commerce widely.

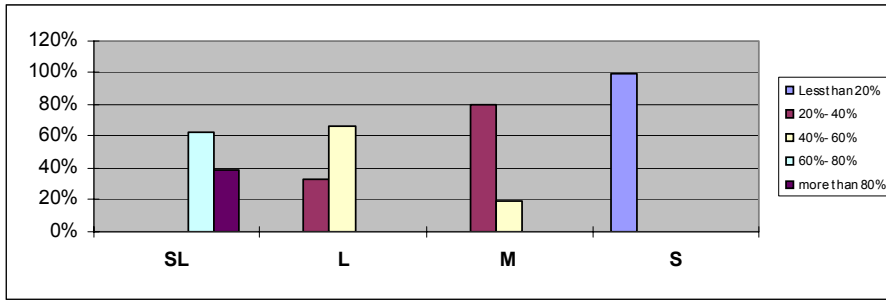


Figure 1: Different Size organisation of Using E-commerce in SCD

By using questions designed from the disadvantage of E-commerce, it could be found that there are many reasons resulting in the phenomenon that the small and middle size company unable to use E-commerce technology when doing operation with SCD. Table 3 showed these reasons and ranking these reasons with different importance to concern for companies.

	Means of Groups				*f* statistic	Compare with *f0.025≈3.30*
	S	M	L	SL		
Internet Security	4.35	4.13	4	3.95	1.18	<
High Interface Cost	4.64	4.31	3.75	3.19	6.95	>
High Maintain Cost	4.56	4.46	3.92	3.02	3.54	>
Re-Training or Recruitment Cost	4.46	4.31	3.63	3.32	3.21	<
Legal Issues	4.00	4.04	3.90	3.47	1.14	<
Network Facility Problems	3.50	3.38	3.29	3.21	0.25	>

Table 3: ANOVA of reason for rejecting of using E-commerce in SCD

As can be seen from table 3, *F*-test in ANOVA indicated that respondents in different groups have great dissidence on criteria of high maintain cost and re-training or recruitment cost. In fact, the factor of high interface cost could be considered as the most important part to affect the decision to disuse E-commerce in SCD for most of small and middle size companies (group **S** with mean 4.56 and group **M** with 4.46), whilst large and supper-large size companies have rated it at 3.92 and 3.02 respectively, which was a relative lower rate compared with that in small and middle size companies. Similarly, this phenomenon could be found in the factor about re-training or recruitment cost.

In addition to these reasons concerned above, the issue about the developing phase of E-commerce cannot be ignored. Since 1998, the E-commerce was spent just 8 years on developing by government in China. As a result of limitation of developing E-commerce in China, therefore, there is the fact that the government had to put more attentions on the development of E-commerce in the large and supper-large companies to gain huge benefits rather than that of small and middle size companies gaining relative low benefits.

In order to explore this reason in detail, the telephone interview was further conducted and thus the problem influencing small and middle size firms to use E-

commerce in SCD could be achieved. By interviews, it could be found that in SCD, because most of small and middle size firms still were not asked strictly to use the E-commerce system (EDI system) to connect with SCD for dealing with the transactions of export and import in the process of customer clearance. Hence, these kinds of firms not having E-commerce technology have to go to the SCD branch to deal with such transactions manually. Alternative, for the some of non-EDI users, they prefer to choose the agency, called KESIDA, (www.ksdinfo.com) which has been authorised by Shanghai Custom Department, to do export and import E-transactions for companies.

At last but not least, most respondents strongly believed the development of E-commerce in SCD in the future will be a great development. The following figure displayed the fact obviously.

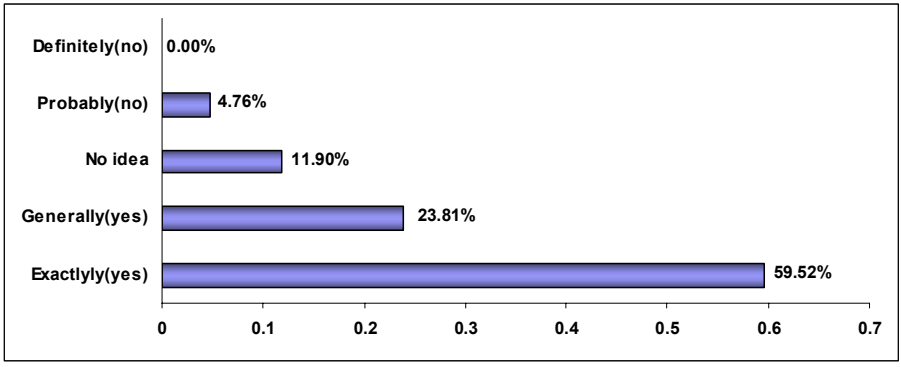


Figure 2: Percentage of Idea on Development of E-commerce in SCD

As shown in figure 2, there are 59.52% (exactly yes) and 23.81% (generally yes) of respondents who indicated the idea on development of E-commerce in SCD is accepted. That is to say, there are around 83.33% of respondents who voted positive opinion about the E-commerce in SCD would have a great development in the future.

5. Conclusion

At present, the use of E-commerce has made a great contribution in various fields and given remarkable profits for its users. According to the theory, the application of E-commerce can equip small companies to compete with large and supper-large organisations. However, by research, it cannot be denied that the development of E-commerce still exited unbalance in practical work between large and small size companies. This paper has reported that the cost of the E-commerce and some barriers (such as Government Issue) have lead to the unfair situation of E-commerce in SCD at present.

In order to solve this problem concerned above, some recommendations could be concluded. Firstly, SCD has to take urgent measures to lower the cost related to E-commerce as quick as possible, such as the interface, maintain and security cost. In

addition, currently, each E-transaction is cost $2.36, which appears to be higher cost for small size company to apply. Therefore, these kinds of firms should look some kinds of technologies to then reduce the cost. Undoubtedly, the introduction of XML could be the best measure to solve the problem. The advantage of the XML is described as a text-based format which is extensibility, universal and easy to understand.

Although there are lots of problems existed in the current E-commerce system in SCD, such as the limitation of use for small and middle companies, the system in SCD should take some measures to overcome these issues timely and so that help users, including small and large size companies, to make more benefits against the cost of E-commerce.

Summing up all the analysis, it could be safely concluded that despite that there are some E-commerce technology which have lagged behind the development of E-commerce in western countries, the positive advantages to support the progress of E-Commerce still exists among people' awareness. This, in turn, contributes to establishing a good environment to develop and promote E-commerce and then offering more opportunities to gain overwhelming profits for all kinds of organisations.
In the further research, author thinks that the sample of companies should be more. Also, there should be a big sample size for the interviews. If possible, face-to-face interview could be the best way to do depth survey and get more information through open questions. Moreover, the E-commerce in China is a developing period. Therefore, this area of research should provide more and more recommendations for development of E-commerce in China.

6. Reference

Brewer, A.M., Button, K.J. and Hensher, D.A., (2001) *Handbook of Logistics and Supply-Chain Management*, Oxford: Pregamon.

Bocij, P., Chaffey, D., Greasley, A. and Hickie, S., (1999), "*Business Information Systems: Technology, Development and Management*", London: Financial Times

easipass.com (2001), Available: http://www.easipass.com/ytzx/hyzxjs/ hyzxjs_xgjg_06_001.htm [Accessed at 18th June 2005]

Fewsmith, J. (2001), "The Political and Social Implications of China's Accession to the WTO", *The China Quarterly,* pp 573-591

Gattorna, J.L. and Walters, D.W., (1996), *Managing the supply china: a strategic perspective*, Macmillan

Lardy, N., (2002), *Integrating China into the Global Economy*, The Brookings Institution Press, Washington, DC.

Li, Y., (2005), Available: http://www.news365.com.cn/wxpd/sh/shcj/t20050113 _360525.htm [Accessed at 20th June 2005]

news.xinhuanet.com (2002), Available:http://news.xinhuanet.com/fortune/2002-03/01/content_296380.htm [Accessed 28th March 2005]

Saunders, M., Lewis, P. and Thornhill, A., (2003), *Research Methods for Business Students*, 3rd Edition. Harlow: Prentice Hall.

Using WS-Addressing To Perform Asynchronous Web Service Calls

J.Hayward and A.Phippen

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

The Simple Object Access Protocol (SOAP) specification does not define an asynchronous message exchange pattern for web services. Business Process Execution Language (BPEL), however, does provide a mechanism for using web services asynchronously, but it relies on hard-coding the end points for the request and the response. This results in tight coupling between the client and the service. WS-Addressing provides a mechanism for defining end points and relating messages with each other. By using the Microsoft Web Service Enhancements the project proposes a mechanism with which web service responses can be routed to an available application using information defined with the WS-Addressing specification. This mechanism loosens the coupling between the client and service and can improve system reliability.

Keywords

Web Services, Asynchronous, WS-Addressing, BPEL

1. Introduction

With the introduction of Business Process Execution Language (BPEL) there has been interest in composing web services into new web services. One example is that of a travel company that uses booking services of airlines, hotels, and car rental companies to provide a new service that enables a consumer to book a complete trip. This new composite service could use BPEL to create a fully automated service where the appropriate services are used and the results compared to provide the consumer with the best option.

Processes that are controlled by computer systems provide a number of benefits. They provide flexibility enabling quick responses to changes in required services. New services can be created to meet new business requirements. Through auditing of completed processes compliance can be assured and costs calculated and controlled. However a majority of processes within companies require human involvement. These human activities stop the automatic process and can take a variable amount of time to complete. For example, in a document editing process an author may take months to write a document. Once the author has finished his task the process needs to continue.

This paper researches mechanisms by which web services can be invoked asynchronously by using open standards and not relying on bespoke interfaces. The WS-Addressing specification provides information that enable routing and messaging information to be included in Simple Object Access Protocol (SOAP) messages. The SOAP standard itself does not specify an asynchronous message exchange pattern. However by including WS-Addressing information within the SOAP message headers asynchronous messaging can be supported. By including the WS-Addressing information in SOAP headers rather than in the published web service interface synchronous web services can ignore the routing information. From the information included in the response from a web service it can be determined if the web service is synchronous or asynchronous.

2. Asynchronous Calls with BPEL

BPEL provides a mechanism for asynchronous calls. The <receive> element following an <invoke> element provides an advertised entry point that the response from the remote web service can use. Juric (2004) describes a typical asynchronous callback as a "*[call to a web service] providing a port type through which the web service invokes the callback operation*". The port type used by the callback as defined by the <receive> element is publicly declared within the BPEL process Web Service Description Language (WSDL) document (Christensen *et al*, 2005). The BPEL asynchronous callback mechanism is illustrated by Figure 1, shown below:

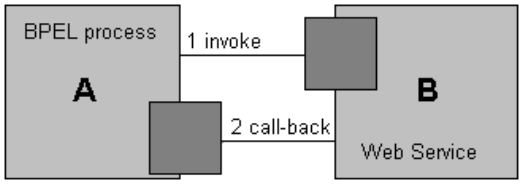


Figure 1: Example BPEL Asynchronous Call (Adapted from figure by Juric (2004))

Juric (2004) explains that in the above diagram the client (A) and web service (B) are BPEL processes. Because the client BPEL process (A) is itself advertised as a web service then a "callback" port type must also be defined within the WSDL. The BPEL process that is the invoked web-service (B) then performs an <invoke> to use the callback to the client (A). This model assumes that all the web services a client will be dealing with will themselves be a BPEL defined process as there is a "partner link" between the two processes. Indeed Andrews *et al* (2003) describe within the BPEL specification that partner links "*represent dependencies between services*". In the airline example supplied by Juric (2004) the callback is hard-coded into the third-party web service (B). Weerawarana *et al*. (2005) give an example whereby the client (A) supplies the partner-link to the web-service (B) which is then used to perform the callback. This reduces the hard-coding, but still relies on the web service (B) being a BPEL process and a tight-coupling exists whereby the web service (B) interface requires a partner-link definition.

The issue still remains of performing a callback from a web service which is either not a BPEL process or contains a hard-coded callback. As with the Weerawarana *et al* (2005) example, the interface of any web service needing to be invoked asynchronously must define a parameter for receiving at least one location of a web service where the response can be sent. This infers that the interface of the web service must be compromised to include callback information. Any client using the client must understand the format of the callback information and provide it along with other parameters the service requires.

Consider a task management system that is used by a BPEL defined processes to assign manual tasks to personnel. The BPEL definition will invoke the task management system to create a new task for a particular person. When the person completes the task within the task management system they acknowledge that the task is complete. The task management system must then perform the callback to the BPEL engine so that the process can continue. In order for the task management system to interact with the BPEL engine in this asynchronous way callback information must be supplied.

There are various categories of information that is required to be included within the callback information. These are:

- Details of where to send the reply
- A unique identifier for the message
- Details of where faults should be sent

Zdun *et al* (2003) details a framework which uses an asynchronous web service call proxy external which provides a simple API to client code. The approach of using a proxy is similar to that provided by the Microsoft .NET framework, where it is the proxy that waits for a response from the service and then performs the provided call back to the client code (Microsoft, 2003). Zdun *et al* (2003) use the proxy to either poll the service to determine if a result is available or it waits for the response from the service and then performs the required callback to the client. With the proxy approach there are a number of issues that arise when the length of time for the reply from the service could be more than a few minutes, especially when human interaction is involved. These issues include:

- Each poll of the server generates at least one network message and response. With many instances running this could unnecessarily flood the network and downgrade performance.
- It is unclear what occurs if the client crashes. Does the proxy continue, or does it crash too? If the proxy stays up what happens to the callback go?

3. Defining Callbacks with WS-Addressing

The Web Service Addressing (WS-Addressing) specification provides a framework for supplying information that would be required for a callback mechanism between remote services. WS-Addressing provides transport-neutral mechanisms to address

Web services (Gidgin *et al*, 2005). This provides the means to identify a web service endpoint and a way to use these in SOAP messages for the exchange of messages between Web Service providers and requesters (Weerawarana *et al*, 2005).

WS-Addressing contains the headers *ReplyTo* and *FaultTo* which define locations where response messages are to be sent (Weerawarana *et al*, 2005). Additionally a *MessageID* provides a unique identifier for the message so that a message can be related to a specific request by the sender. The *Source* header defines where the request came from. When callback information is supplied to the service the WS-Addressing headers can be supplied in the <header> element of the SOAP message.

Below is an example of a request SOAP message and the response that utilise the WS-Addressing specification (Adapted from example by Weerawarana *et al*, 2005).

WS-Addressing Request Example

```
<S:Envelope xmlns:S="http://www.w3.org/2003/05/soap-envelope"

xmlns:wsa="http://schemas.xmlsoap.org/ws/2004/08/addressing">
  <S:Header>
    <wsa:Source>
      <wsa:Address>http://source.com/sender</wsa:Address>
    </wsa:Source>
    <wsa:MessageID>guid:7hf8dhycd-f8djdi9-difcjdkfd
    </wsa:MessageID>
    <wsa:ReplyTo>
      <wsa:Address>http://reply-machine.com/replyreceiver
      </wsa:Address>
    </wsa:ReplyTo>
  </S:Header>
  <S:Body>
    <!--- BODY CONTENTS --->
  </S:Body>
</S:Envelope>
```

WS-Addressing Response Example

```
<S:Envelope xmlns:S="http://www.w3.org/2003/05/soap-envelope"

xmlns:wsa="http://schemas.xmlsoap.org/ws/2004/08/addressing">
  <S:Header>
    <wsa:RelatesTo>guid:7hf8dhycd-f8djdi9-difcjdkfd
    </wsa:RelatesTo>
  </S:Header>
  <S:Body>
    <!--- RESPONSE CONTENTS --->
  </S:Body>
</S:Envelope>
```

The WS-Addressing specification states that the *RelatesTo* header contains the ID passed to the service with the *MessageID* header in the request (Weerwarana, 2005). It is header that enables the application receiving the response to match the response to the process that generated the original request. By using WS-Addressing in the SOAP headers the response can be routed to an appropriate location. However, the SOAP specification (Box *et al*, 2000) does not contain an inherent asynchronous messaging exchange pattern. The web service still needs to be written to specifically read the WS-Addressing headers in order to provide asynchronous calling.

As described earlier, the sender of the request may not be available to receive the asynchronous result. This is especially true where the asynchronous call is to allow human interaction which could take months to complete. Therefore the callback address supplied in the WS-Addressing *ReplyTo* header may be an un-reliable one. One solution is for the *ReplyTo* address to point to a host that acts as a mediator or broker. The broker will receive the response and forward it onto either the source of the request or, if it is unavailable, any available host. The source of the request is supplied to Web Service as the *Source* header and so needs to be included in the SOAP message sent to the broker. Figure 2 illustrates this sequence of events:

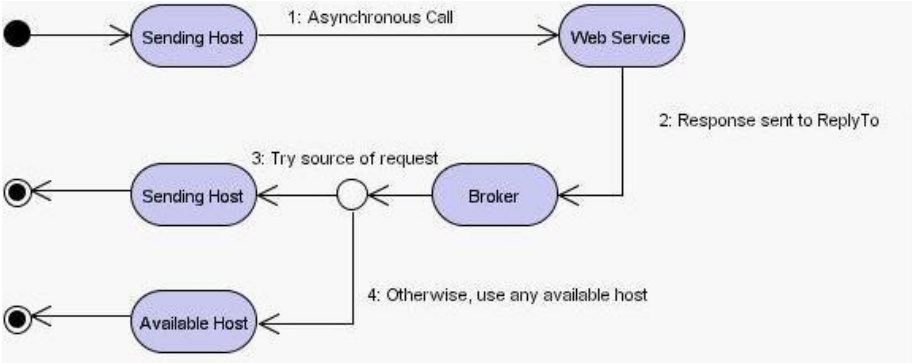


Figure 2: Using a broker to route asynchronous call responses.

In this architecture the Broker needs to contain a level of intelligence. It must interact with a directory, such as a Universal Description Discovery and Integration (UDDI) server where hosts capable of processing the response are advertised, and select an applicable host.

4. Testing Using WS-Addressing for Asynchronous Messaging

Both Microsoft .NET and Apache Axis 2 support asynchronous calls to published web services that are not specifically designed to be accessed in synchronous manner. The Visual Studio .NET (Microsoft, 2003) help topic "*Asynchronous Design Pattern Overview"* states that "*One of the innovations provided by the asynchronous pattern is that the caller decides whether a particular call should be asynchronous.*" This is confirmed for invoking web services in the help topic "*Communicating with XML Web Services Asynchronously*" which states "*note that an XML Web service does not have to be specifically written to handle asynchronous*

requests to be called asynchronously". This feature is only achievable through using the Microsoft .NET client API and not through generic SOAP calls. Also this API only allows the callback to the calling object instance. The callback cannot be routed to a different host or run-time instance. The Apache Axis 2 asynchronous support works in the similar way, in that it is achievable through a client API using a non HTTP transport (Zdun *et al* 2003).

The Microsoft Web Service Enhancements (WSE) (Microsoft, 2005) provides additional classes to the .NET framework providing support for the additional web service specifications such as WS-Security, WS-Trust, WS-Policy and WS-Addressing. To use the WSE to test the support of WS-Addressing in providing the callback information a simple web service was created using Microsoft Visual Studio .NET 2003 and deployed on host running Microsoft Advanced Server 2003. The WSE provides classes that enable the SOAP headers to be accessed from the deployed web service. This enables the service to access the values of the WS-Addressing headers.

The first step was to install the WSE on the Windows Advance Server 2003 computer where the web service is hosted. Once the WSE was installed the supplied Configuration Editor was used to enable the web service project to operate with the Web Service Extensions. The code for the Web Service project was then edited to include the *Microsoft.Web.Services2* namespace. The Microsoft.Web.Services2 namespace contains all the classes required to access the SOAP headers such as *ReplyTo* and *MessageID*. Within the web service the following code retrieves the *SoapContext* object and gets the string containing the supplied *ReplyTo* address:

```
SoapContext ctxt = RequestSoapContext.Current;
string replyTo = ctxt.Addressing.ReplyTo.Address.Value;
```

The web service is still called using the Request/Response pattern and on completion, when the *return* statement of the web service is reached, a standard SOAP response is returned. This response can be interpreted as a confirmation message that the SOAP message was successfully received. Below are the headers included in the response generated by a request to the test WSE enabled web service.

```
<soap:Header>
  <wsa:Action>http://tempuri.org/AddResponse</wsa:Action>
  <wsa:MessageID>uuid:e40fccdf-3af9-4856-b65c-f3e3db1ad6a7
  </wsa:MessageID>
  <wsa:RelatesTo>uuid:92bccf15-b71d-4a34-b05e-c967b17830bf
  </wsa:RelatesTo>
  <wsa:To>http://uop-project:13000/</wsa:To>
</soap:Header>
```

As part of the standard request/response pattern the above response is sent back to the requesting host when the web service completes. When a web service is called

that is not WSE enabled then the WS-Addressing headers are not included within the response. The *To* header in the above response matches the *ReplyTo* header as defined in the request. When using the WSE the response to the request is not routed via a new connection to the supplied *ReplyTo* header. The response, including the headers, is returned to requesting application. Therefore, the *To* header in the above example response does not contain the correct location. For the web service to reply asynchronously it generates its own SOAP request message that is sent to the location defined by the received *ReplyTo* header.

The application has to determine whether the called web service provides an asynchronous or synchronous service. The application has to determine if the response received is the result of the service being executed, or if an asynchronous response delivered in the future will contain the result of the service. The BPEL specification provides the statements <invoke> and <receive> to explicitly state that a call to a web service must be performed asynchronously. Therefore for BPEL processes it must be assumed that developer of the process can recognise whether the required service is to be accessed asynchronously or not.

Following the initial test a proof of concept for the proposed broker architecture earlier was created. The proof of concept required four separate elements. These were:

- An application that would call a web service with WS-Addressing information contained within the SOAP headers. The application would also be capable of receiving SOAP messages through an open port. When this application started it would register itself on a UDDI server advertising a URL where SOAP messages could be sent. When this application closed it would un-register itself from the UDDI server.
- A simple web service that reads the WS-Addressing headers and creates a response SOAP message. That response SOAP message is sent to the endpoint defined by the supplied *ReplyTo* header. The *ReplyTo* header contains the address of the broker, this is where the asynchronous reply is sent.
- An application acting as the broker. The broker reads the WS-Addressing headers. messages and firstly attempts to forward the whole SOAP message to an available application as advertised on a UDDI server.
- A UDDI server with publishing permissions enabled.

The proposed mechanism uses open standards to reduce the coupling between client software and an asynchronous web service. The elements described above, when used together, provide a proof of concept that an asynchronous web service can be implemented and still be de-coupled from the client. This demonstrates that by passing the routing information in the SOAP headers that the public interface of web service was not unnecessarily cluttered. Additionally, by routing the asynchronous reply via a broker, further benefits are gained. The final end point of response is not determined at the time of the original request. Any number of end points may be available if originator of the request is no longer available. This is a distinct possibility where the time between request and response could be measured in weeks or months, and would improve the reliability of the overall system. The reliability

can be further improved by the broker queuing responses and retrying where no end points are available.

5. Conclusions

With the advent of the web service standards greater collaborations between technologies is possible. Now business processes can use these distributed services and combine them to create new services. However, business processes often require human interaction. When a human becomes part of the process the process must stop and wait for a response. This could take seconds, minutes, or even months. Therefore asynchronous calls to web services are vital for including human interaction within a business process.

BPEL supports this by providing hard-coded links between services. In an example given by Juric (2004), process A uses service B, but service B must eventually callback to process A. This paper proposes that WS-Addressing is an open standard that enables routing information to be defined within SOAP messages, therefore, enabling a web service to be accessed asynchronously. By supplying the WS-Addressing information to the web service through SOAP headers, the web service interface remains unaffected, and web service need not read the headers if the information they contain is not required. This means that the application calling the web service does not need explicitly specify that the interaction should be asynchronous or not. The web service itself can provide asynchronous callbacks if required.

The mechanism proposed in this paper was tested through developing an application that uses WS-Addressing to enable a web service to be called asynchronously. Microsoft WSE was used to enable a web service to read the supplied WS-Addressing headers allowing the service to be used asynchronously. The test web service was created using .NET and published on Microsoft IIS 6.0. Although this was an entirely Microsoft environment it was used solely to test optionally including the WS-Addressing within the headers of a SOAP message and allowing the web service to read these headers and act upon them. These tests also proved that when a web service is able to perform asynchronously that the asynchronous reply can be routed to a host supplied within WS-Addressing headers that is different for the requesting host. This mechanism, therefore, offers a peer-to-peer approach to web service interaction, using open standards, rather than a traditional client/server approach. The peer-to-peer approach negates the need to poll the service for results and allows the response message to be routed to an available client or clients if required.

6. References

Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klien, J., Leymann, F., Liu, K., Roller, D., Smith, D., Thatte, S., Trickovic, I., Weerawarana, S. (2003) “Business Process Execition Language For Web Services 1.1” ftp://www6.software.ibm.com/software/developer/library/ws-bpel.pdf

Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., Mendelsohn, N., Nielsen, H. F., Thatte, S., Winer D. (2000) "*SOAP 1.1*", http://www.w3.org/TR/2000/NOTE-SOAP-20000508 W3 Consortium (accessed 3rd December 2004)

Christensen, E.; Curbera, F.; Meredith, G.; and Weerawarana, S. (2001) "*Web Services Description Language (WSDL) 1.1.*" http://www.w3.org/TR/wsdl (access 3rd December 2004)

Gidgin, M., Hadley M. (2005) "*Web Service Addressing 1.0 – Core*", http://www.w3.org/TR/ /2005/WD-ws-addr-core-20050331 (Accessed 7th July 2005).

Juric, M. B., Mathew, B., Poornachandra, S. (2004) *Business Process Execution Language for Web Services*, Packt Publishing, ISBN 1-904811-18-3, First Published October 2004.

Microsoft (2003) "*Asynchronous Design Pattern Overview*" http://msdn.microsoft.com/ library/default.asp?url=/library/enus/cpguide/html/cpconasynchronousdesignpatternoverview.asp (Accessed 25th July 2005)

Microsoft (2005) "*WSE version 2.0 SP3*" http://www.microsoft.com/downloads/details.aspx?familyid=fc5f06c5-821f-41d3-a4fe-6c7b56423841&displaylang=en (Accessed 26th July 2005)

Weerawarana, S., Curbera, F., Leymann, F., Storey, T., Ferguson, D. F. (2005) *Web Service Platform Architecture*, Prentice Hall, ISBN 0-13-148874-0

Zdun, U., Voelter, M., Kircher, M. (2003) "*Design and Implementation of an Asynchronous Invocation Framework for Web Services*" in proceedings at International Conference on Web Services Europe – 2003.

Ecommerce – Completing the Supply Chain

T.S.Moe[1] and M.Hudson-Smith[2]

[1]Faculty of Technology, University of Plymouth, Plymouth, United Kingdom
[2]University of Plymouth Business School, University of Plymouth, Plymouth, United Kingdom
e-mail: melanie.hudson-smith@plymouth.ac.uk

Abstract

This paper identifies and investigates a selection of success factors that effect delivery success rates from the perspective of SM ecommerce companies. With the help of a survey of SMEs from the Devon and Cornwall region of the UK, low order fulfilment cost and order size fitting through the letterbox are both identified as beneficial for a company's online success. Based on the analysis of success factors, the paper presents some general guidelines for SMEs considering a move into ecommerce.

Keywords

Electronic commerce, Small- to medium-sized enterprises, Delivery, Electronic supply chain management

1. Introduction

The Internet has opened up a whole new world for many businesses. It enables companies to conduct business in a completely new manner. How to successfully implement a successful ecommerce strategy is a widely covered topic and current literature present a number of factors that are important for an ecommerce company's success. Trust and security are often being identified as the most important factors (Srinivasan, 2004). As the customer may not have any physical contact with the business it is important that they feel secure and trust that the personal information they give over the Internet is treated confidential and transmitted in a secure way. The issue of online trust has been researched by several researchers. Goodwin (1996) argues that "trust is the grease that keeps the wheel turning" (Cited in: Srinivasan, 2004, p.66). Hoffman and Novak (1999) discuss the need for control and argue that "cyberconsumers feel they lack control over the access that Web merchants have to their personal information" (Cited in: Srinivasan, 2004, p.67). Jiang (2001) addresses the trust issue from a technological perspective. He argues that two methods are available to build trust; Public Key Infrastructure (PKI) and SSL.

Although trust and security are considered to be perhaps the most important factors for an online company's success, there are other factors that might be of just as great importance. Hoek (2001) argues that companies often have poor logistics and supply chain management. He identifies problems such as products not delivered on time,

products being out of stock and costly delivery as problem causers for ecommerce companies. The latter problem is in fact one of the major differences between traditional commerce and ecommerce. In traditional commerce the customers usually have to "deliver" the product themselves. In ecommerce this has to be done by the retailer (Tarn *et al*. 2003) or a third party company. Punakivi *et al*. (2001) argue that delivery logistics in ecommerce has been one of the main factors leading to large losses for companies. Research conducted on this topic focus mainly on the e grocery sector. Logistic problems arise when the customer is not at home to receive the order. KaEmaEraEinen (2001); Punakivi and Saranen (2001); Saranen and SmaEroos (2001) have done research on the standard delivery method of attended reception and with unattended reception boxes. The result of this research shows that unattended reception boxes are the most cost efficient home delivery alternative (Punakivi and Saranen, 2001). However, the initial cost associated with this delivery alternative is likely to be substantial and only affordable for larger companies. SMEs are less likely to have sufficient resources to invest in this kind of infrastructure (Chappel and Feindt, 1999). A general solution to this is to outsource the logistic functions (Delfmann *et al*. 2002). A case study by Hudson Smith & Smith (2005) describes a small UK company delivering low value, large sized products. In the study period the company changed its delivery partner three times. Neither was performing to satisfactory standards at satisfactory prices. Hudson Smith & Smith identified two key issues that were problematic for the ecommerce company; the relatively high cost of delivery and the absence of the customer when the product was being delivered to his home. Based on their analysis of the case study and past studies of ecommerce they identified five factors that will either worsen or lessen the effects of their two key issues; Cost of carriage, time of delivery, size of goods, delivery range (e.g. local or national) and frequency of purchase.

In the UK research undertaken by ecommerce software vendor Actinic (2004) revealed that only 3% of UK's SME retailers have got the facilities to take orders online. The main reason for not adopting ecommerce was unsuitable products. This could suggest that the companies reluctant to adopt ecommerce have not succeeded in developing a good ecommerce logistics strategy. If so, this can be linked to Hudson Smith & Smith's case study where the size of the goods, cost of carriage, time of delivery and delivery range proved to be a problem.

2. Rationale

Security and trust appears to be generally accepted as the most important success factors, yet Hoek (2001) argues that logistics and supply chain management are very important elements of success. Logistics problems have been identified in the grocery business and in the case study by Hudson Smith & Smith (2005). In addition research undertaken by Actinic (2004) revealed that few SME companies in the UK have got the facilities to take orders online and the main reason for not pursuing ecommerce is unsuitable products. The author believed that more research was needed addressing SMEs e logistic strategies. It was in this area the author carried out his research.

3. Research Aim

The aim of the research presented in this paper was to identify the key success factors that effect delivery success rates from the perspective of SME ecommerce companies. With the work of Hudson Smith & Smith in mind, five hypotheses were developed:

Ho: *"High cost of order fulfilment is not a hindrance for SM ecommerce company's success"*

Ho: *"Large order size is not a hindrance for a SM ecommerce company's success"*

Ho: *"Loyal customers are important for a SM ecommerce company's success"*

Ho: *"Flexible delivery services are important for a SM ecommerce company's success"*

Ho: *"Large geographical delivery range is important for a SM ecommerce company's success"*

In this paper, the author has chosen to focus on the three first hypotheses, as these produced the most interesting results.

4. Approach

The author adopted a positivistic paradigm for the study. The method used to collect the primary data for the study was a survey. The survey was distributed to a sample of around 1100 SMEs from Devon and Cornwall. The sample was supplemented by 2 of the original 1100 distributing the survey to the regional office of the Federation of Small Businesses (FSB), Digital Peninsula Network (Penzance), West Cornwall Business Network and West Cornwall Together Local Strategic Partnership. This gave a total sample of somewhere between 1500 and 2500. 301 respondents filled in parts, or entire questionnaire. However, as the research was concerned with companies with online booking facilities, delivering to members of the public, a smaller sample was used for most of the analysis.

5. Survey Design

The questionnaire was divided into three sub sections, starting with general demographic questions before moving on to more specific questions about the Internet and the respondents' ecommerce activities. The questionnaire was designed to filter out respondents selling products mainly to other businesses as the study was concerned with home delivery. It was also designed to filter out respondents without online booking facilities. The remainder of the questionnaire was designed to gather information about respondents order fulfilment costs, delivery range, customers

purchase frequency, loyalty schemes, gross margins, total cost to customer etc. All gathered to try and establish any relationships between the success factors described in the above hypotheses and online profitability.

6. Discussion of Findings

As the findings are concerned with online profitability it is worth mentioning what variable was used to measure online success. The main variable chosen to indicate an ecommerce company's success was a question asking the respondents if their online activities were more or less successful than their other activities. All the respondents had both offline and online sales, making this variable a viable indication of online success. The author assumed that if the respondents are more or equally successful with their online activities, this qualifies as being successful online.

A significant relationship was determined between the high cost of order fulfilment and the online profitability of a company. The findings suggested that respondents paying between £1-£4.99 or less for post and packaging are likely to be more profitable than respondents paying more. The relationship can be seen graphically in the clustered bar chart below. The connection between high cost of order fulfilment and online profitability can be seen as a confirmation of Hudson Smith & Smith's research which indicated that high cost of carriage was a hindrance for SM ecommerce companies. Based on these findings, the first of the above hypotheses was rejected. This was the expected outcome of the test.

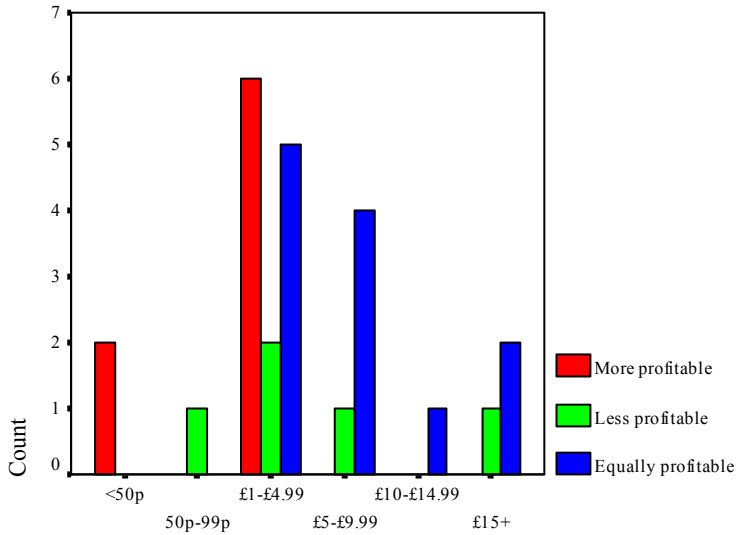


Figure 1: Packaging and Delivery cost vs. Online Profitability

A significant relationship was determined between the size of an average order and the online profitability of a company. It was discovered that respondents with an average order size that would fit through the letterbox are likely to be more successful with their online activities. The respondents with an average sized order

fitting through the letterbox were less likely to be unsuccessful with their online activities. The relationship can be seen graphically in the clustered bar chart below. Hudson Smith & Smith's research identified large orders to be problematic. These findings do point in the same direction. As soon as the order does not fit through the letterbox, extra time and resources have to be spent on getting the order to the customer if they are not at home. This extra time and resources is likely to affect the company in one way or the other. In the case study described in Hudson Smith & Smith's research the company in question incurred extra cost and grief due to the fact that the customer was not present the first time delivery was made. They ended up switching logistics partner a number of times. Based on these findings, the second of the above hypotheses was, expectedly, rejected. When testing this hypothesis it was also discovered that 80 % of the respondents with a high gross margin (over 60 %) on their orders, typically delivered letterbox sized orders, indicating a higher profitability on this type of order.

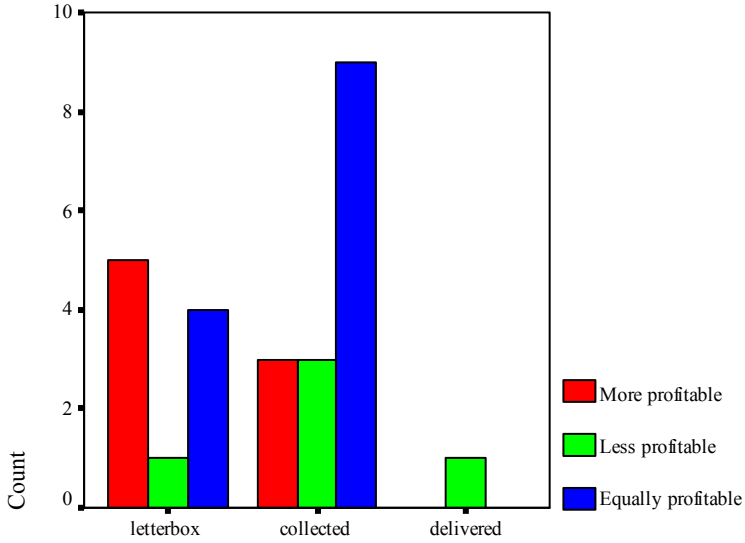


Figure 2: Order Size vs. Online Profitability

No significant relationship was found proving that loyal customers are important for SM ecommerce company's success. This analysis was carried out by analysing the relationship between purchase frequency and online success. No conclusive relationship could be found. Based on these findings, the third of the above hypotheses was rejected. This was not expected and it was surprising to learn that only 17.2 % of the respondents had loyalty promoting schemes. It would be natural to assume that giving customers and incentive to stay with a particular company is very important online, were the switching cost is very low. However, the respondents with loyalty schemes were not more profitable online than the respondents that did not.

7. Implications of Findings

Having analysed and discussed a number of significant findings leading to the rejection of the above hypotheses, it is now time to take a step back and consider the implications of the findings for an SME. It was hoped that the results of this research could be used to develop some general guidelines for SME companies considering a move into ecommerce. In this section some pointers for such a company will be suggested:

- *Keep order fulfilment costs low*. A company wishing to move into ecommerce should calculate their estimated order fulfilment costs before starting up. As can be seen above, respondents with an average order fulfilment cost of less than £ 5 were likely to be more successful online. An SME should aim to keep their estimated cost at the same level. This is by no means exact science and should merely serve as a general guideline. Factors such as e.g. product price and customer type are likely to affect the relationship between order fulfilment cost and online success. In addition the total value of the order must be considered. In the study, 64.3 % of the respondents with an average order fulfilment cost of less than £ 5 had a total cost to the customer between £ 10 and £ 59.

- *Consider order size*. It would be beneficial if the company could do some research and calculate their estimated average order size. The important thing to consider with order size is that as soon as the average order size do not fit through the letterbox it is likely to have implications for the company. This does not mean that companies with larger average order size will not be successful online, it merely suggests that more consideration is needed as the order size increase. Especially if the company have to rely on a third party for delivery as at least 78.4 % of the respondents in the study did. Depending on what kind of third party company that is chosen a few questions should be answered before a contract is signed. The price, delivery options and re-delivery policy are always useful to know. E.g. if the company use Royal Mail to deliver their goods, the orders can be redelivered or collected at a local depot at no extra cost to the company. This would be a good and reasonably cheap option for companies with an average order size accepted by Royal Mail. However, if the company have larger average order size, or customers requiring very prompt delivery etc., Royal Mail might not be the best option. Courier could be a better alternative. Courier companies usually offer fast delivery from door to door. However, as can be seen in the case study described in Hudson Smith & Smith's research couriers do not always fulfil their promises. The company in the case study had, during the study period, three different courier partners. They found it very difficult "balancing price viability with even an acceptable level of delivery service" (Hudson Smith & Smith, p2). It is therefore recommended that several companies are considered before making a decision on courier partner. If possible, contact SMEs already trading online and ask about their experience with different courier companies.

- *Consider potential loyalty schemes*. Regardless of the industry and amount of competition a company find itself surrounded by, loyal customers are likely to be beneficial. Even though the survey results did not indicate any positive relationship between online success and purchase frequency, or respondents with loyalty schemes, it could very well prove to be advantageous. Especially if the SME finds itself in an industry with considerable competition. Research undertaken by the Millard Group and ActiveMedia argue that sites with loyal customers are likely to be more successful than those without (King, 2001). Obviously, if the company manage to compete on price and be the cheapest available (and the customer is aware of it) then they could expect customers to shop at their website. However, as is often the case, prices on different companies website's are usually on a similar level and the company need to find other ways of standing out from the crowd. Offering discounts on quantity, give special rates to frequent purchasers etc. could prove to yield good results. It should at least be considered before a company invest in online facilities.

The implications of this research are likely to be more relevant for companies already operating offline. All of the relevant survey respondents had both offline and online activities making the results more applicable for offline companies considering a move into ecommerce.

8. Limitation of Findings

The Devon and Cornwall regional sample used for this study was skewed and likely to not be completely representative of SME's in the UK in general. As an example the number of respondents having online booking facilities was 35.4 %. In comparison, Actinic's (2004) research revealed that only 3 % of UK's SME retailers had the facilities to take orders online. In addition some of the finding's validity may have been reduced due to the small number of respondents answering particular questions. However, all the tests carried out in this research still give good, indicative results.

9. Conclusions

In this research paper order fulfilment cost, order size and customer loyalty have been analysed in relation to a company's online success. Low order fulfilment cost and order size fitting through the letterbox were both identified as beneficial for a company's online success. The analysis concerning customer loyalty was inconclusive and did not generate a conclusive link to online success. Some general pointers for SMEs considering making a move into ecommerce have been suggested. These are not strict rules, merely guidelines to some issues that should be addressed before the company decide to go through with the online expansion. Instead of only focusing on traditional issues such as designing a secure, trustworthy, appealing website, companies will hopefully realise that the actual fulfilment of the sale needs to be carefully considered. Whether companies will find the authors' guidelines

helpful or not remains to be seen, however they will at least point out potential order fulfilment problems.

The research presented in this paper will be a good starting point for future investigations of factors affecting ecommerce company's delivery success rates. With the exception of the e grocery market, not much literature has been written on this topic. The study identifies the importance of low order fulfilment cost and small order size for online success. These findings can be seen as a confirmation of parts of Hudson Smith & Smith's proposed theoretical framework. It also encourages further research of the relationship between customer loyalty and online success, which was not established in this research, but is natural to assume exists.

In a wider context, e logistics is an area commonly undermined when analysing factors affecting the success of an ecommerce company. Current literature tends to favour analysis of other factors such as trust and security. This research has hopefully demonstrated the importance of e logistics awareness for SME companies considering making a move into ecommerce and will encourage further research into this area.

10. References

Actinic (2004), "2004 Actinic Ecommerce Report", Available at: http://www.actinic.co.uk/docs/ 2004_Actinic_Ecommerce_Report.pdf [Accessed 21 June 2005].

Chappel, C. and Feindt, S. (1999), "Analysis of E-commerce Practise in SMEs", Europe's Information Society, available at: europa.eu.int/ISPO/ecommerce/ g8/documents/kitebestpractice.doc [Accessed 2 September 2005].

Delfmann, W., Albers, S. and Gehring, M. (2002), "The impact of electronic commerce on logistics service providers", *International Journal of Physical Distribution & Logistics Management*, Vol. 32, No. 3, pp.203-222.

Hoek, V.R (2001), "E-Supply Chains-Virtually Non-Existing", *Supply Chain Management: An International Journal*, Vol. 6, No. 1, pp.21-28.

Hudson Smith, M and Smith, D. (2005) "I'm Sorry We Can not Deliver...Fulfilling the E-Business Promise", Proceedings of the 12th International EurOMA Conference, Budapest, Hungary, pp849-853.

Jiang, S. (2001), "WebALPS Implementation and Performance Analysis: Using Trusted Co-servers to Enhance Privacy and Security of Web Interactions", Technical Report TR2001-399, Dartmouth College, Hanover, NH.

KaÈmaÈraÈinen, V. (2001), "The Reception Box Impact on Home Delivery Efficiency in the E-Grocery Business", *International Journal of Physical Distribution& Logistics Management*, Vol. 31. No. 6, pp.414-426.

King, C. (2001), "Loyal Customers Drive Online Success", Internetnews, available at: http://www.internetnews.com/ec-news/article.php/572131 [Accessed 4 September 2005].

Punakivi, M. and Saranen, J. (2001), "Identifying the success factors in e-grocery home delivery", *International Journal of Retail & Distribution Management*, Vol. 29, No. 4, pp.156-163.

Punakivi, M., Yrjølå, H and Holmstrom, J. (2001), "Solving the Last Mile Issue: Reception Box or Delivery Box?", *International Journal of Physical Distribution & Logistics Management*, Vol. 31, No. 6, pp.427-439.

Saranen, J. and SmaÊros, J. (2001), "An analytical model for home delivery in the new economy", Working paper, Available at: http://www.tuta.hut.fi/ecomlog/ [Accessed 3 December 2004].

Srinivasan, S. (2004), "Role of trust in E-business success", *Information Management & Computer Security*, Vol. 12, No. 1, pp.66-72.

Tarn, J., Razi, M., Wen, H. and Perez, A. (2003), "E-fulfilment: the strategy and operational requirements", *Logistics Information Management*, Vol.16, No.5, pp.350-363.

The State of Elderly in ICT Adoption at Rural Area

S.Y.Lee and A.Phippen

Network Research Group, University of Plymouth, Plymouth, United Kingdom
e-mail: info@network-research-group.org

Abstract

Despite the rapid growth of Internet, elderly often being left out while their needs and opinions seldom being heard and represent. Given the number of elderly population started to rise over the years, their commercial potential have draw attention of government and private party which show interest of the neglected market segment. This paper refers to study 87 sample respondents from Plymouth, United Kingdom. Behind the sense, many encouragements and Internet facilities had been built at rural area to enhance the development. This paper discusses the state of readiness for elderly to accept ICT having the current condition, measuring their ability, emotional constraints and physical barriers. Additionally, it would looks into effectiveness of encouragements used to promote ICT and connection of social impact to elderly's learning attitude. The result suggests different approaches of sustaining local area elderly's interest in ICT adoption and government effort in monitoring existing scheme.

Keywords

Elderly, Information Communication Technology (ICT), social, barriers, Internet

1. Introduction

Elderly citizen has become the most rapidly expanded population, over 16% of the population of UK is now 65 and over, majority of them are baby boomers born after the period of Second World War (1946-1964) who become older during first half of this century (ONS, 2004). Many of the elderly have move to reside at rural area after retirement. In fact they are looking forward for free lifestyle of 10-20 years with current advanced of medical technology, elderly is expect to have higher life expectancy. The fact that ageing process bring limitations to elderly, reason of finding out what are their needs in IT and what are best Information Communication Technology (ICT) make used by them are essentials. At the current condition, Internet and ICT could offer assistance in personal communication, providing alternative method for financial matters, information provider and online purchasing which could help to prevent isolation, loneliness, depression and social exclusion for elderly.

For this paper purposes, the elderly is defined as anyone age 60s and above to accommodate the public acceptance on elderly both based on qualification to received pensioner scheme and commonly noted in literature review. This study would starts discussing the issue from previous web literature, journals, and annual reports. Then this paper describes the survey among Plymouth respondents that was conducted to measure their ability in physical, preventing ICT barriers and awareness

of training at local area. Finally, the state of elderly readiness is presented after the evaluation of encouragements being done.

2. Issue on elderly and the use of Internet

It is a debatable question since the raise of Internet on the measurement of acceptance from elderly group to use Internet and the ability of the environment to provide suitable learning atmosphere. The difference existed between rural and urban area is due to vast development, which could not be aligned either by magnitude or speed of development. This could terribly affect the competitiveness of rural areas, by reducing the interest to attract business investments and more importantly, reducing the people's level of exposure who are living there. Although many efforts have been made by the government to provide broadband access to rural area, it has not been achieving to fully successful. By autumn 2003, about 16% of rural villages and 4% of remote rural areas had access to affordable broadband Internet connections, compared to 99% of the urban populations and 80% of UK population (Countryside Agency, 2004).

Elderly's participation in ICT has always been reluctant by barriers either from personal problems or technical problems. Personal problems derived mostly about an elderly learning perception, the ability of external environment including confident, location, and income. A quarter of older people live in private household in rural area are in low income (Social Exclusion Unit, 2005). Additionally, physical problems would result into functional restriction in requirements on ICT product devices including elderly's sight problem in seeing fine detail on the computer monitor and hearing abilities which eventually not a big problem for operating a computer but create problem during communication with tutor at training session. On the other hand, technical problems vary from software, personal computer and mode of connection. At current stage significant of these barriers have yet to be truly defined for rural area, as it would be difficult to determine the estimation of barriers for each individual.

There were researches on the acceptance of Internet among elderly to social life of themselves and household happiness, focusing a major reason on communication with their own children particularly. Ito and colleagues (Swindell, 2002; cited by Ito *et al*. 1999) study shown people who are regular users of SeniorNet in the USA revealed positive effect on medium social interaction and individual empowerment. ICT could potentially improve communication with younger relative. Additionally, elderly who utilise Internet saw it as a tool to strengthen the social bonds (Trocchia and Janda, 2000). In contrast, Kraut and colleagues (Swindell, 2002; cited by Kraut *et al.* 1998) reviewed that limited access to Internet by elderly usually because of loneliness and significant research on adults Internet users shown obvious declines in everyday household communications, declines in the size of social circles, and increases in depression.

Elderly is likely to have pessimistic attitude finding learning as unavailable and unappealing at older age. Their perception indicating life experiences as sufficient

and stable would make the learning process become stunted, embedded as strong perceived values and taken granted as normal behavior among society. This would be the major reason, which result ICT encouragement among elderly to be slow. This culture is particularly difficult to change in depth when rooted in custom and attempts of modification would be resisted. They may have felt, or been made to feel, that learning is not for them (NAGCEL, 1998). Many of them might suffered from social exclusions, blinded by unawareness, neglecting pleasure and achievements that could be deserved cause by lacking of self confidence and opportunities.

3. The Survey

The study is conducted with 87 respondents from University Third Age (U3A) at Plymouth which is member of Third Age Trust, an independent association throughout the UK that consist of elderly member mostly in retirement stage or early retirement. The distributed survey consisted of 12 questions in 2 pages with further interview with 6 elderly communicate through telephone conversation and email. The whole survey and interview session were carried out within a month.

4. The Barriers for Elderly

The study revealed that majority of elderly had used personal computer with total of 78%. However, only 58% of them were equipped with personal computer and Internet connection at home while 33% having none of them. The elderly expressed their barrier by income constraints usually worrying on their ability of not earning anymore and spending on income received through pensioner scheme. The costs engaged with buying a new personal computer and its maintenance cost were major worry. 70% of elderly is using broadband while 30% remains subscribing to dial up. Some elderly felt that the cost of subscribing to broadband is rather high and felt the usability of dial up sufficient for their own personal use.

In another separate question asking about barriers, it was found top three barriers were confidence, no guidance/tutoring and follow by no personal computer. (Refer figure 1) Although income came fourth, it was thought that this barrier has indirectly spread to affect the ability of owning personal computer and landline telephone connection. Elderly were even worried about their confident and how all these ICT equipments works. It alerts the alarm bell on the urgent present of tutor to build the confident among elderly. Many elderly are afraid to be expose to Internet partly because they felt at their age; they would not be able to cope with new technology since they have been ‘left behind too far’. There is also little evidence found through elderly interview on the help their seek from contemporaries who are computer educated. At the same time, those who are computer educated would not initiate to provide guidance.

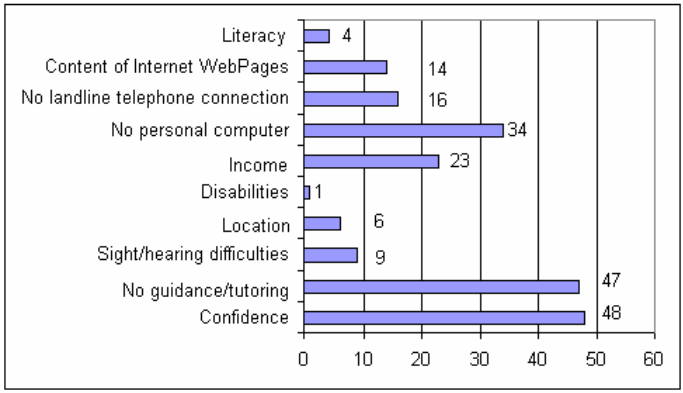


Figure 1: Barriers Preventing Internet for Elderly

The study also revealed that the elderly that felt for extremely interested and very interested attitude to learn Internet, were those who basically used a computer before which eventually explain their high preference to Internet. Perhaps unsurprisingly, those who have not used a computer before show moderate interest to no interest at all. It revealed that elderly that has not experienced with computer and Internet before would not have the curiosity about learning them. Internet simply does not exist as attractive learning tool for them to improve the life that they are having right now.

The current atmosphere sees encouragement of ICT being promoted by government and private institution where many online training centres were setup. However, elderly still preferred to get online at home supported by 93% of the respondents. A rather disappointing result illustrating only 4% of respondents would go to online centres to get online. It certainly marks the failure of public online centres, which claimed to make remarkable statistic in these few years providing online centre to national coverage. Elderly found traveling problem did contribute to the access of online centre. Many of them did not own transportation and felt traveling by public transports simply kills the desire to reach the destination. This would be even worst for elderly that have disabilities physically.

Some elderly felt reluctant on their pace of learning if being put into training at learning centre. The public online centre, typically, public libraries did not target specifically in elderly but member of the public. Therefore, it is likely everyone in the public online centre would be including learners in various age. It would easily create illusionary exclusion for them for their lower response rate and possibly higher attention being put on younger people. However, the survey indicates no request being rise on the need for learners to be at the similar age as the elderly. Some identified they rather preferred to have training at home, however, most of this training service are make available by private institutions which in deed required consuming of income.

The sample respondents showed little problem of sight and hearing difficulties with relatively small groups of them were engaged with these barriers. It was found that some elderly have problem seeing fine details in monitor screen while complaint on

glare and flicker were among the sight problem. In term of hearing, it was not prove with the sample respondents as much of the time Internet communication with user is based on text although multimedia started to involve in as one of the media; majority of the WebPages content is still pure text. It might be alternative method of communication for elderly that have hearing difficulties, ICT could do more than normal conversation through telephone but with text based, chatting and could allows the impossible to become reality.

Elderly that is inexperienced with computer and Internet is usually lack of motivation to allows themselves to be expose to ICT and advanced of technology. Generally they have accepted the change and admits new technology simply did things differently compare to older days. Nevertheless, they themselves are not emotionally prepared or interested in participation in term of learning by themselves. The need for changes, which do not stand out strongly to urge them on the reason for adapting ICT in their life. Many of them are regular user of reading newspapers, books and watching television that are the major source of information they felt comfortable and reliable of continue practice.

5. Impact of ICT to Elderly's Social Life

When looking matter at different perspectives by putting aside of ICT advantages to elderly, majority of elderly that have not adapting to ICT might be suffering from loneliness, depression and isolation. Over the years, elderly in social exclusion is not a new area or topic that catches society attention. It is there but the seriousness of it is unpredictable with many national statistic continues to revealed the increasing number. It is also at the same time, ICT failed to show evidence on improvements to current elderly's social life. In most of the cases, elderly felt that ICT is destructing the society. They felt many people would be just sitting in front of computer communicating to the never known other person. Despite the other person communicating at the opposite side is someone known in real life, it would doubtful whether communication through the web could provide the level intimacy and stimulation they expected.

In contrast, many active elderly that gathered with their contemporaries regularly shown positive social life helping them from loneliness, isolation and showed tangible benefits such as bringing them happiness and the interest of continuing their participation in future activities. These include language classes, group exercises, playing games together, exchanging information on cooking, gardening and pets. It drew the human contact closer and with real life face-to-face communication. Elderly very often stresses the need to communicate with real people where relationship is established through communication. However, many of them do not realise that they could be communicating with someone they known through ICT. The idea of ICT does not appear to be an alternative solution in many of elderly daily communication life.

On the other hand, ICT has not been showing the strong impact to improve current social life for elderly. Although there is possibly many things to be done through the

Internet to help bringing communication closer such as email replacing letter, online chatting replacing telephone conversation. Online chatting is currently available in different approach not only limited to text based but also handwriting functionalities with assist of web camera and voice chat are among the changes that could possibly suite different elderly with or without disabilities. Elderly might or might not know the existence of all these, they might just limit their thinking into saying chatting is complete lack of human contact but at the end of the day, we could now chat and see someone through monitor screen despite limited by geographical area in real time. There is a lack of acknowledge for elderly to know, they might simply subjective to certain thinking presumably not to be change in a short period of time. There could be possibility that they not knowing what ICT could provide to them.

The survey describes that email is second highest activities did online. (Refer figure 2) Elderly expressed that email functionality of capable attaching photos has help them to exchange photos with their children that lives abroad. They would not be worrying on mailing letters anymore where now they trusted the electronic version of mail could perform better and it helps to save money too. It was agreeable that email has help better communication between elderly parents with their working children. In the same question, it also revealed that chatting as the lowest activities performed online. It proves the prediction that many elderly seldom uses online chatting service even for those elderly who are regular users of Internet. It could be possibly relate that not knowing exactly what chatting service could offer and perception in believing chatting as complete waste of time has limit the benefit that elderly could have.

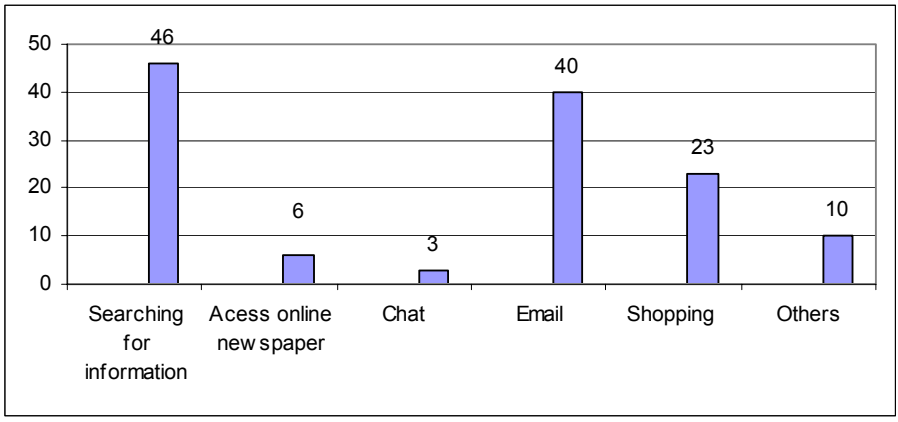


Figure 2: Activities Done Online by Elderly

6. The Failure of ICT Attractiveness

The failure of ICT to attract participants typically from elderly group could constitute a wide variety of reasons. It could be from the state of elderly themselves not keen into learning ICT. There is simply no reason, need or purpose for elderly to learn personal computer and Internet when the alternative method is not prove to them of being helpful and to be able to run the computer system and Internet requires more

effort and time dedicated in learning process more than completing the task. There is a need for comparison to be done by elderly themselves by letting them to feel in reality the benefits of ICT could provide to them. Spreading the word is not enough to make elderly felt the same as touching the devices themselves, operating with guidance and working on it independently. It is absolutely crucial to let the satisfaction override the hesitation before elderly could engage in ICT. The perception of learning plays an important role in motivating elderly considering their age to start learning again. It certainly requires a push of encouragements from many responsible parties, effectively from family member for this momentum to start and keep moving.

Elderly population is usually in retirement period where they are particularly careful in spending, considering the fixed cost of their life and inability of being employ anymore. Income consuming to adapt into ICT could be high for them especially those who are small pensioners or elderly living alone. They might not be willing to buy the personal computer and spend on subscribing Internet connection for something that they have not been used to and something definitely invested to learn. They would not know whether the learning process would result in positive manner or otherwise. The uncertainty leads them to worries and resistance. Additionally, this investigation indicates income problem would spread into many other problems which could not provide them in their preference learning ambience, their home. Limitation of income would not be able to provide elderly with assistive technology such as hardware or devices specialise for elderly with disability. These devices certainly not produce in mass production, difficult to access and come with high price. The ill equipped environment certainly would not be able to teach elderly or suitable for elderly to learn as well.

The rural area is usually difficult to reach where attention and encouragement are among the slowest being facilitated. In order for the elderly people to have changes in their life, they require the tools to make changes. These are basic requirements including personal computer, Internet connection, the availability of tutor or guidance and the existence of training or online centre convenient for elderly to access. It is obviously the basic requirements for an elderly to start stepping forward to new technology. The need for online centre is not only for the benefit of having a person there to teach but it also provide opportunity for elderly that could not afford buying personal computer and subscribing Internet connection to try how is it to function the equipments and what is all about WebPages on the web. The effort to provide all these requires participation of government and other private institutions as this is not only concern about elderly, it involves rural area development, public facilities, infrastructure, human resource and business collaborations.

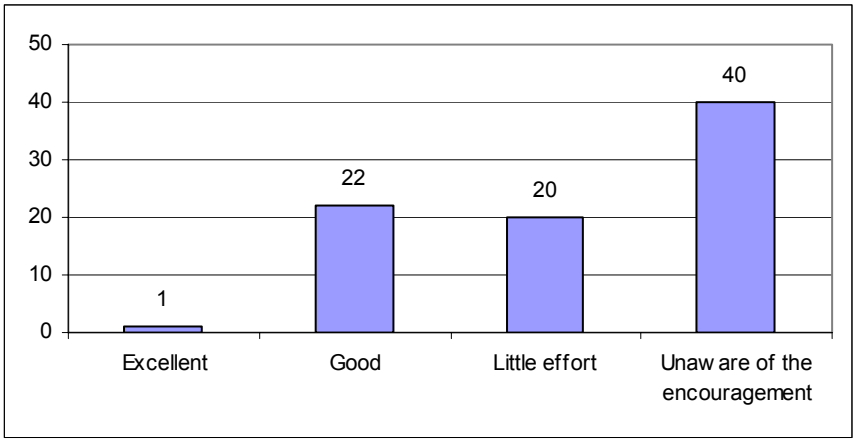


Figure 3: Elderly Rating on Government IT Encouragement for over 50s

The current IT encouragement done by government is targeted to member of the public rather specifically to elderly. In this study it also revealed that 48% of the respondent stated that they were unaware of government encouragement directly reflect the failure of government encouragements. (Refer figure 3) Elderly has to compete for a place to learn Internet when most of the time, they are only capable of reaching Internet access at public libraries and taster course at private institution because majority of them would not prefer to pay and learn in a course usually run by private institutions. Government effort in elderly has been seen little in ICT but life long learning programme which target to learning at older age has been promote since 1998. IT has been part of the learning area but not in attention and receiving low responses. Regardless of learning programme or ICT encouragements, it definitely fails to catch the attention of elderly in the method of addressing the problem or accessing the elderly at rural area.

The major fall of existing encouragements organised by either government or private organisations is due to lack of commitment for continuous development. When problems are there, the responsible bodies are aware of it, this is where all the activities, training, encouragements programmes took place catching the eye of everyone including elderly. It was a complete success for many programmes collaborated by private organisations but it soon dies out over period of time. There is no progress reporting, it is treated as something occurred in the past. There is lack of responsibility for further communication with the elderly, local organiser and trainers. There is an urgent need of local training officer to coordinate the plan, constantly communicate to elderly and reporting the condition to organisers so that the promotion would suite the local elderly needs. There is also the problem with lack of coordination of trainers in public access points. This eventually limits the number of training programmes a location could offer to the elderly. Trainers should be trained in term of technical knowledge and the way they communicate during teaching session with elderly because some of them might have disabilities.

7. Conclusion

ICT could do more for elderly if only elderly is aware of it and is given the opportunity in learning. Although elderly might felt many barriers prevented their participation, the government, ICT industry and society could do more to help in term of solving the problems and making the basic requirements ready. It would be an effort by everyone in order to encourage elderly participation. Learning attitude must be promoted by life long learning programmes, which should accommodate more ICT courses and organisations to take part. There is alternative way of enhancing social life with ICT through virtual communities created for a local area and encouragement to learn through group of elderly so that they would not feel neglected in the process of learning. It is important that before carrying out any promotion catered for elderly, the effort was spend on understanding the local elderly needs, follow by spreading the ICT culture step by step then into the depth courses.

8. References

Countryside Agency (2004) '*The State of Countryside 2004'*, West Yorkshire, Countryside Agency Publication

Office for National Statistic (2004), "*Population Ageing*", www.statistics.gov.uk/cci/nugget.asp?id=949, (Accessed 11 June 2005)

Social Exclusion Unit (2005) '*Excluded Older People Social Exclusion Unit Interim Report'*, London, Office of the Deputy Prime Minister

Swindell, R., (2002), "*Technology and Over 65s? -Get A Life*", http://www3.griffith.edu.au /03/u3a/includes/linked_pages/file_downloader.php?id=317&prop=5&save=1, (Accessed 31 December 2004)

The National Advisory Group for Continuing Education and Lifelong Learning (1998), "*Creating Learning Cultures: Next Steps in Achieving the Learning Age*", www.lifelonglearning.co.uk/nagcell2/nagc-30.htm, (Accessed 10 August 2005)

Trocchia, P.J & Janda, S. (2000), "A Phenomenological Investigation of Internet usage among Older Individuals, *Journal of Consumer Marketing,* Vol. 17, pp605-616

Language acquisition in Epigenetic Robotics

E.Hourdakis and A.Cangelosi

Adaptive Behaviour & Cognition Research Group, School of Computing, Communications and Electronics, University of Plymouth, Plymouth, United Kingdom
e-mail: acangelosi@plymouth.ac.uk, ehourdakis@yahoo.com

Abstract

Recently cognitive Models have been employed for robotic linguistic studies. Among the popular approaches is the Grounded Adaptive Agent methodology, which specifies that language emerges from the embodiment interaction of the agents with their environment. Previously, epigenetic robotics was used to study the grounding transfer mechanisms of simulated agents (Cangelosi & Riga, 2004). We extend this model, by including further modalities on the cognitive system of the agent, and a learning protocol used for the grounding of perceptual stimuli.

Keywords

Symbol grounding, epigenetic robotics, human-robot interaction, embodied cognition, language evolution, imitation, grounding transfer

1. Introduction

Communication has been an undeniable vantage for the evolution of the human specie. For robotic agents however the sharing and manipulation of name labels, implies the use of language (Harnad, 1996). The programming of such system is subject to the Symbol Grounding Problem (Harnad, 1990), or how are these symbols connected to their meaning. Among the most effective methodologies employed to confront the issue, is the "Grounded Adaptive Agent Approach" (Parisi & Cangelosi, 2002; Cangelosi, 2004), where linguistic communication emerges from the interaction of the robot with its environment.

Linguistic communication however, is not an isolated capability, inherent to humans. We come to realise and comprehend language, through a series of semantic interpretations of symbols and meanings within our world. Furthermore, these symbols do not exist as arbitrary representations of some notion, but are intrinsically connected to behavioral or cognitive abilities, based on the properties of the reference system they belong to. This task of connecting the arbitrary symbols used in internal reasoning with external physical stimuli is known as the "Symbol Grounding problem" (Harnad, 1990).

Ideally reference systems are composed of objects and their associations. Peirce (1978) defined three types of association constituents, namely icon, index and

symbol. Iconic representations are connected to objects, based on their "conventional similarity" and stimulus generalisation (Deacon, 1997). Indexical associations are found in animal communication systems, while symbols are definitions of logical or combinatorial relationships with other symbols or objects. These definitions are not innate to humans, but are grounded through a series of stimuli and cognitive interaction with the world, embodied as words within a human communication system. Barsalou (1999) further states that the brain implements these basic symbolic operations by predicating conceptual properties of individuals and categories.

One issue emerging from this proposition is how grounded symbols acquire their meaning. Simply, by referring them to further symbols and so on. However, such resolving mechanism is subject to the infinite regression problem or at which point these meaningless symbolic representations actually acquire their meaning. To solve this, a model needs to ensure, that its reference mechanisms will eventually lead to one or more objects, or cognitive representations (e.g. categories) within the world.

More importantly, the great evolutionary benefit granted in competent symbolic reference systems e.g. the one of apes (Savege-Rumbaugh & Rumbaugh, 1978), lies in the ability of their symbols to facilitate double referential relationships. Thus symbols can be described using either indexical references of objects or logical definitions of relationships with other symbols. Such property allows the forming of groups of symbols, not only based on their differential stimuli properties, but also on their combinatorial capacities with other notions. A common example in human language is the one of verbs, which is used to describe how one or more objects can be manipulated.

Acquiring the ability to discriminate between diverge stimuli however, does not imply that robots will fully perceive what the object is. To identify an entity, the invariant features, which denote whether it is of a specific kind, must be detected. Categorical representations (Harnad, 1987) are the most abstract of these feature assemblages, and can be acquired by filtering the feature space of the available stimuli. In cognitive systems, this is mainly achieved by compressing the within-category differences and expanding the between-category distances, until a sufficient boundary is reached. This compression/expansion effect is called "categorical perception" (Harnad, 1987), and occurs in humans (Goldstone 1994; Pevtzow & Harnad 1997) as well as Neural Networks (Tijsseling & Harnad 1997). Later these clearly defined categories can be used to form propositions, and thus become symbolic representations themselves. Furthermore, this abstractedness of a representation acts as a measure of its quality (Cangelosi & Parisi, 2004).

The strength of symbolic communication systems, including human language, lies in the fact that they allow manipulation of grounded entry-level symbols, for the sake of acquiring further categories. Thus, in these systems, higher-level categorisation, the forming of extended cognitive representations based on existing elementary symbols, can be achieved in two diverse methods. The first, sensorimotor toil, is more straightforward as it employs the same methodologies used by an agent during the grounding of entry level symbols. Acquiring objects by "toil" implies that the agent attains any new symbolic tokens through direct sensorimotor interaction with its parts, under the guidance of corrective feedback (Cangelosi, Greco, Harnad, 2000).

This method, also known as "supervised learning", resembles our own acquisition mechanism during the infant stage, where we come to perceive and name elementary symbols and categories based on their iconic representations. Among the most important drawbacks of the method is the amount of training required for an agent to complete learning. Furthermore, categories acquired this way lack of any symbolic representations, thus restricting themselves only to iconic and categorical ones.

In contrast to sensorimotor toil, symbolic theft (Cangelosi & Harnad, 1998) is based purely on describing categories by forming propositions of previously grounded objects. These newly formed classes consist as a mere description of their symbolic representations in the form of propositional expressions, endorsed by the reference system (e.g. in a human communication system, as groups of words). It is thus permitted, to transfer knowledge between two agents that use the same reference system, simply by transferring symbolic propositions of formerly acquired terms. This mechanism, also known as theft, has been pertained as a victimless crime (Cangelosi & Harnad, 1990), as the intellectual provider of the knowledge is not degraded in any way.

2. Grounded Adaptive Agent Approach

Our model is an extension of a previous research (Cangelosi & Riga, 2004) in which robotic agents were taught a lexicon using direct instructions by humans. In these new simulations, robots are able to comprehend action names and produce linguistic instructions. We accomplish this by using the learning protocol previously suggested by Cangelosi et al (2000), for the grounding of perceptual stimuli. This consists of a three stage process, in which agents first learn to perform a basic behavior, associate the action with a linguistic signal, and latterly manipulate these symbols to form higher order propositions.

2.1 The model

Our model consists of two simulated agents, embedded within a virtual simulated environment that supports physics, using the Open Dynamics Engine. Robots are comprised of two 3-segment arms, with 12 degrees of freedom, and a torso attached to four wheels, for navigating in the environment. Their embodiment is realised from their endeavor to perform several actions against the physical constraints of the environment.

The first agent, the teacher, is pre-programmed to perform a variety of basic actions, each associated with a linguistic signal. These are demonstrated to the second robot, the imitator, which attempts to reproduce the behavior by mimicking. To accomplish this we used the same algorithm as in the preceding research (Cangelosi & Riga, 2004) for postural approximation. The imitator robot is further endowed with a Neural Network that governs all of its perceptual, cognitive and motor abilities. Firstly the agent acquires basic behaviors and their names by observing the teacher, and latterly employs the linguistic symbols that are grounded to its cognitive system, to describe further propositions of new actions. It achieves this using grounding

transfer, an effect that has been also illustrated in similar connectionist simulations (Cangelosi et al, 2001; Cangelosi & Riga, 2004).

Our network consists of a three layer feedforward MLP, taught using back-propagation. There are three modalities integrated in the network, namely vision, motors and language. Vision consists of a 25 cell representation of the artificial depiction of each object. Three objects were made available, cube, plane and bar, each associated with two different views. Each basic behavior was associated with one of these six views, with the corresponding object being placed in front of the robot on every initiation of an action. The linguistic layer consists of 12 nodes, 6 referring to the signals of the basic actions, 3 reserved for the higher order actions and three remaining nodes used to designate the name of the activated object. Motors consist of a 6 node layer, which encodes the forces that are being applied on the body of the agent. These are elaborated from the network, using the imitation algorithm. The values of these motors in the output are propagated as input in the next iteration. Such recurrrency allows the agent to have a complete perception of its own position on every time step, and additionally permits the robot to execute a behavior on the absence of the demonstrator.

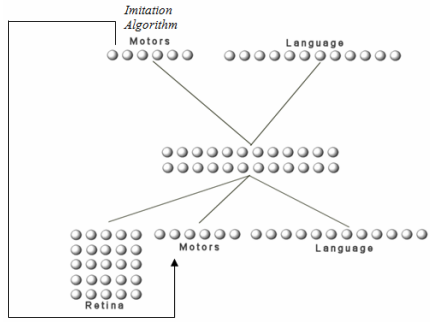


Figure 1: The network employed for the experiments

2.2 The training procedure

We attain the grounding transfer, using a 3 stage training process, in which the imitator first learns to correctly perform an action, name all basic actions and respond to their signals, and latterly use these grounded symbols to form propositions of further symbols. Weights are conveyed from one stage to the other, with inputs of the network being activated or deactivated correspondingly.

During the first stage, Basic Grounding (BG), the agent learns to execute all six basic actions. No linguistic elements are present in the network, while the imitation algorithm is used for the corrective adjustments of the motor output.

The second stage, Entry Level Naming, was concerned with associating the previously acquired behaviors to linguistic signals. It featured three sequential activation cycles, in which the agents were first taught to respond to a linguistic

signal, and latterly name the action that was being performed. The first cycle, Production, taught the imitator how to name all 6 primitive behaviors. Motor and visual information were still made available, while the network was additionally taught the six linguistic nodes corresponding to the basic behaviors. Following production was the Comprehension activation cycle, where we taught the imitator to correctly respond to a linguistic signal, without having the ability to perceive the object associated to the action. To accomplish this retinal information of the network were disabled, while we provided the word corresponding to the action name in the input. During the final cycle, imitation, all associated inputs were activated, including retinal information and the language signals of the 6 basic behaviors in both input and output of the network.

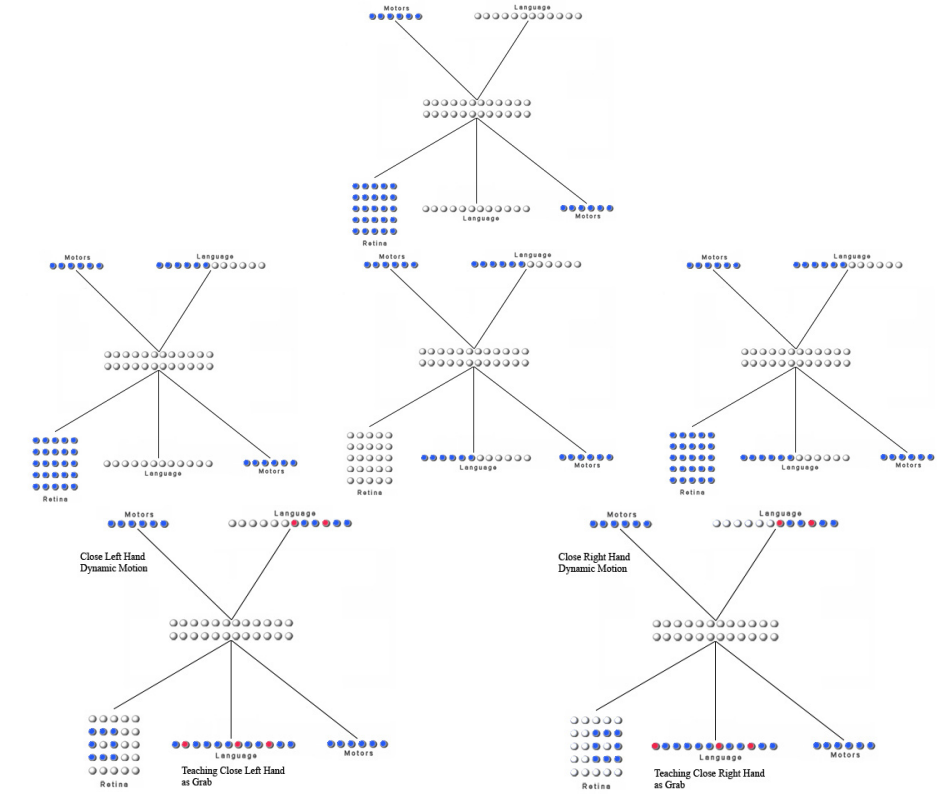


Figure 2: All 3 stages of training. Top: Basic Grounding, Middle: Entry Level Naming, Bottom: Higher Order Naming

The final stage of learning, Higher-Level, taught the imitator to perform composite actions that were derived from propositional descriptions formed from previously grounded symbols. The stage had all network inputs activated, including the 3 additional linguistic units that were reserved at the beginning of the experiment, as well as the ones responsible for naming the objects. To endorse the concept of true imitation and autonomy of the system, the imitator, in contrast to the Basic Grounding stage, was taught to its own motion information, instead of a learning

algorithm. To accomplish that, the requisite motor forces for producing an action were recorded to a file straight after the end of the first stage. Composite actions were trained twice, once for each arbitrary action using every time the retina associated with the initial object:

During this final stage, the system had already some grounded symbols, from the previous stages. Therefore our main objective was to assess how these intrinsic representations respond to different definitions of propositions. To achieve this we elaborated two diverge patterns of activation. The first was concerned with teaching the agent only the names of the higher order actions, while the second, employed the higher level signals in conjunction to the basic level name.

2.2 Results

All stages were trained successfully, with very low errors on the final epoch. Results were evaluated on the ability of the imitator to perform the actions, as well as appropriately manipulate the linguistic signals. The first stage, basic grounding, featured the simplest elaboration of motor dynamics, as only the six basic behaviours were taught. The network required 1000 epochs to converge, with a final average error below 0.004. At the end of the stage, the imitator was able to execute all actions flawlessly, when being presented with an object. The addition of the three activating cycles during the second stage, introduced to the network supplementary information which had to assimilate. Due to this, training required 3000 epochs to complete, with 300 subsequent steps being used for each action. The error on the final epoch was 0.03. This was a constituent of the errors produced by the three activation cycles. The first, production output an error of 0.05, comprehension an error of 0.04, while the error of the imitation cycle at the end of the final epoch was 0.002. It is evident from the examination of the errors for each cycle that imitation produced the best performance of the network. This is due to the linguistic input units that the cycle featured, which provided a further measure for the categorisation of the basic actions.

Finally, the two alternatives considered for the higher order stage, were equally successful, requiring an average of 1500 epochs to be completed, with 0.4 learning rate and 0.6 momentum as the network parameters. The final error for both training trials was 0.002.

After the completion of all the experiments, the agents were assessed against their abilities to transfer the grounded symbolic representations in their cognitive system, to new ones, using propositions. Our grounding transfer test aims at evaluating the aptitude of the imitator agent, to perform a new composite action with any of the objects previously associated, in the absence of the linguistic descriptions of the basic actions. Thus the agent is requested to respond solely on the signal of the composite action (e.g. Grab) and selectively one of the two retinas that were associated with the objects. In addition, while the imitator was taught only the motion of the dissected action for each composite behavior, testing evaluated the performance of the composite action, a behavior never seen before. The stage was comprised of two basic phases, one for testing the left view of the object and one for the right. All inputs were propagated through the network with no training occurring.

In addition, two testing phases were run, one for each alternative trained during the previous stage.

For the first training option, the average error of the network was 0.021 for the left view and 0.015 for the right. The second tests were concerned with assessing the alternative of Higher-Order training that included the linguistic signal of the basic actions. The agent was again requested to perform a new composite action, only on the basis of the higher order signal. The table below summarises the testing results, for both evaluation modules:

	Close Both Arms	Close Both Elbows	Lower Both Shoulders	Average
Left View	0.023	0.027	0.013	0.021
Right View	0.024	0.01	0.0079	0.015
Left View (Alt)	0.048	0.077	0.04	0.05
Right View (Alt)	0.05	0.029	0.019	0.03

Table 1: Results of the grounding transfer test

Such low errors confirm our hypothesis that previously grounded symbols are transferred to the new behaviors. Alternatively, the common error an agent produced when performing an unknown action was lying in the range of 1.0 – 2.0.

3. Discussion and Conclusion

The experiments were concerned with the grounding of external stimuli, to the cognitive system of a simulated agent, by means of imitation. This type of research can provide a solid basis upon which future epigenetic robotic projects could be expanded. The model provides a well established method, using socio-psychological principles, and practices strongly resembling the human nature. The main strength of the project is the multi-modality of the cognitive system, which enabled the imitator to have a more complete representation of the external environment.

Another important aspect that was addressed was the use of reference systems. Such hierarchical structures are vital for language research as they provide a technique for associating objects to their meanings. The common practice up to date was to associate such notions by providing a direct link, from the object to its representation. As this was created artificially, it had no meaning for the robot itself. Thus the agent was aware of the existence of the association, but unable to discriminate what each end was representing, sort of a meaningless association. In our experiments however, Neural Networks were employed, to allow the construction of this link based on guided learning and the agent's own autonomy. The result of this approach is that the robots have a complete representation of the object-meaning association, and can thus use each one accordingly without requiring any sort of human intervention.

Behaviors are initially grounded to the cognitive system of the imitator, during the Prototype Sorting stage. These newly symbolic representations, are associated with signals, so that they can be used to describe further behaviors on the higher order stage, using a process named grounding transfer. It has been pointed out by researchers that for a system to be able to achieve this effect it must be fully autonomous, and allowed to build its own representations on the environment (Cangelosi & Harnad, 2000) (Cangelosi & Parisi, 2004). This essentially occurs during a series of perception categorisation effects, upon which the agents first learn to discern the diverge stimuli input from the environment and latterly categorise it to similarity groups. Results confirm the grounding transfer, as agents are able to perform flawlessly behaviors that have never seen before. They do this by merging the previous grounded actions to one on the basis of the linguistic propositions they are provided.

As an extension to the preceding model, this research gradually introduces more complex linguistic information, such as associating an action with more than one object. Thus the agents are allowed to associate explicit propositions with specific external stimuli, as a proto-form of syntactical associations. Similar experiments have illustrated that there is an association between syntactical categories, and the regions of activation within the neural network (Cangelosi, 2001). More specifically, it has been demonstrated that verbs tend to activate regions of the weight space that are associated with the motor categories. In contrast, nouns, activate nodes that are associated with physical properties of an object, such as shape. Such hypothesis can be employed in order to further extend our model as to co-relate special linguistic representations to the motor categories, which can latterly emerge as verbal instructions, or even full grammatical propositions.

4. References

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Cangelosi, A., Harnad, S. (1998) The adaptive advantage of symbolic theft over sensorimotor Toil: Grounding language in perceptual categories. *Presented at the 2nd International Conference on the Evolution of Language, London, April 1998. Submitted to the journal Evolution of Communication.*

Cangelosi A., Greco A. & Harnad S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, 12(2), 143-162

Cangelosi A. & Riga T. (2004). An epigenetic robotic model for the sensorimotor grounding of language. *Artificial Intelligence*

Cangelosi A., Riga T, Greco A., (2001) Symbol Grounding Transfer with Hybrid Self-Organising/Supervised Neural Networks

Cangelosi A., Parisi D. (Eds.) (2002). *Simulating the evolution of language*. London: Springer-Verlag

Cangelosi, A., & Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, 89(2), 401-408.

Cangelosi A. (2004). The sensorimotor bases of linguistic structure: Experiments with grounded adaptive agents. In S. Schaal et al. (Eds.), *Proceedings of the Eighth International Conference on the Simulation of Adaptive Behaviour: From Animals to Animats 8*, Cambridge MA, MIT Press, pp. 487-496

Deacon T.W. (1997) The Symbolic Species:The coevolution of language and human brain, London: Penguin.

Goldstone, R. (1994). Influences of categorisation of perceptual discrimination. *Journal of Experimental Psychology: General*, 123, 178-200

Harnad S. (Ed.) (1987) *Categorical Perception: The groundwork of cognition*. New York: Cambridge University Press

Harnad S. (1990) The Symbol Grounding Problem. *Physica D*

Harnad, S. (1996) The origin of words: A psychophysical hypothesis. In Velichkovsky B & Rumbaugh, D. (Eds.) *Communicating Meaning: Evolution and Development of Language*. NJ: Erlbaum: pp. 27-44.

Peirce C.S. (1978). *Collected papers. Vol. II: Element of logic*, C. Hartshorne & P. Weiss (Eds.), Cambridge, MA: Belknap.

Pevtzow, R. & Harnad, S. (1997) Warping Similarity Space in Category Learning by Human Subjects: The Role of Task Difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.) *Proceedings of SimCat 1997:Interdisciplinary Workshop on Similarity and Categorisation*. Department of Artificial Intelligence, Edinburgh University, pp. 189-195.

Savage-Rumbaugh, S. & Rumbaugh, D.M. (1978). Symbolisation, language, *and Chimpanzees: A theoretical reevaluation on Initial language acquisition processes in four Young Pan troglodytes*. Brain and Language, 6: 265-300.

Tijsseling A. & Harnad S. (1997). Warping Similarity Space in Category Learning by Backprop Nets. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.) *Proceedings of SimCat 1997:Interdisciplinary Workshop on Similarity and Categorisation*. Department of Artificial Intelligence, Edinburgh University, pp. 263- 269.

How not to evolve a neural network for robot football players

M.E.Ellen and A.Cangelosi

School of Computing, Communications, and Electronics,
University of Plymouth, United Kingdom
e-mail: mojo-the-waffle@excite.com

Abstract

This paper examines the results of research conducted to investigate the possibility of using genetic algorithms to evolve a neural network to control the behaviour of robot football players. The first step was to try to evolve networks to control just the motion of the robots, to move towards the ball, and then hopefully to avoid opponents. If that step had been successful, the next step would have been to try to evolve grounded communications between team mates. The results of evolution of neural networks showed that there was something wrong with the set up of the experiment. The errors made will be examined in this paper, so that hopefully they will not be made again.

Keywords

Artificial Neural Networks, Genetic Algorithms, Robot Football

1. Introduction

Robot soccer is a growing sport. There are many variations sanctioned by the governing body FIRA (www.fira.net), encompassing different types of robot, including but not limited to: Humanoid robots, Khepera, and two wheeled 7.5cm^3 robots with vision capabilities (Mirosot robots). The most popular robots are the Mirosot robots.

The rules for robot football vary from type to type. For the Mirosot league there are three categories (at the moment): three a side, five a side and seven a side. Each one has a different sized pitch. There have to be areas on each robot to place the team colours on (this is decided by the referee before the match). There are various rules governing the flow of play, including what data maybe transmitted to the robots and when, and how the winning team is decided. There are even rules governing fouls. It is quite clear that this game is as rich and full as real football.

The impetus for developing soccer playing robots varies. For some it is the pursuit of a team of robots that can play football well enough to beat a human team. For others it is all about pushing the boundaries of robotics and AI technology.

Robinson et al. (2002) discuss the use of developing robotic football players in order to teach undergraduates. The technology for the robots used there takes a step back

from the most up to date, in order to curb costs and make modifying the designs easier. The major problem with the development of the robots seems to be within the motion control system, as it seems to be difficult to make a robot move in a continuous straight line.

Robinson et al. (2004) go into detail about the technological difficulties involved with designing and building the Mirosot robots. Of particular interest are the sections on AI and multi agent real time control.

The paper wishes to address the problems of agent control. If it is so hard to make robots move towards a ball and slow down quickly enough, and account for friction, and all the other factors that a mirobot designer must consider why do it? Why not let mother nature do it for you?

Obviously mother nature is a metaphor for using techniques that have their inspiration routed in nature. There are two specific tricks that nature uses that this paper is interested in. Genetics and neural networks.

With respect to genetics, scientists have come up with a computational analogy called evolutionary computing, and more specifically genetic algorithms (GA). The type of problem that GAs can solve is varied. Hofmeyr and Forrest (2000) discuss the use of genetic algorithms to develop an artificial immune system that would be used as part of a computer's security system. They argue that a computer network is an ideal environment for the cultivation of an adaptive security system, as the network is always changing. An important point they make is that the system should be evolved in situ, i.e. as part of a network security system, rather than evolved in isolation, so that it has a the context for its operation.

Another example is Hong, Huang and Lin (2001), who describe a system that is used to decide on the best next move in a two player game. Their basic premise is to improve the overall traversal of the game search tree, by using genetic algorithms to find the optimal path of local partial search trees.

Nature's other problem solver, the neural network, has its own analogy in computer science. Neural networks can be use for such things as pattern recognition, for example: Huang (2005) discusses a neural network for extrapolating high resolution images from low resolution images. The results of using this technique show that a combination of hand-coded algorithms and the multi-layer perceptron (MLP), i.e. the Artificial Neural Network (ANN), is superior to either by themselves. The paper gives conclusive evidence as a series of zoomed in pictures, where resolution is increased in a variety of ways.

Also Artificial Neural Networks can be used to simulate biological neural network, for example: Pantic, et al. (2002) use an ANN to study the affects of transitioning from one memory state to another. (i.e. recognising one pattern then another.). Their paper describe an experiment to examine the affect of pre-synaptic activity on memory state transitioning. Their model reasons that the activity of returning some of the spent activation from a neuron firing back to the originating neuron acts as reinforcement for remembering a certain pattern. Their overall conclusion is that the

synaptic dynamics can increase the speed of memory state transitioning (in their model) by making any memory state less stable and thus more susceptible to noise.

With respect to this study, a combination of a genetic algorithm and neural networks were used to control the motion of computer simulations of robots. Two experiments took place. The first was to evolve a network that was able to move towards a moving object, the second was to evolve a network to move towards a stationary object. If this sounds like a backward step, that is because it is. The reasons for such a happening are discussed later on.

2. Method

All the software used in the experiments was written by the author, and is available from him.

Experiment 1 Network Design

For the first experiment the ANN is made up as follows: the input layer is made up of four neurons, they signify the angle the robot is from the ball, the distance the robot is from the ball, the x velocity of the ball and the y velocity of the ball. The hidden layer is made up of seven neurons. The output layer is four neurons, two for the left and right motor speeds, and two for the left and right motor rotation direction. This can be seen in figure 1.

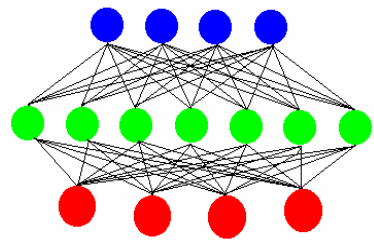


Figure 1: Layout of the network for the first experiment.

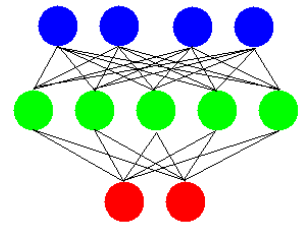


Figure 2: Layout of the network for the second experiment.

Experiment 2 Network Design

The design is of two input neurons and four output neurons, with a layer of five hidden neurons between, this can be seen in figure 2.

The input neurons receive the distance the ball is from the particular robot, and the angle between the direction of the robot and the ball. The output neurons contain left and right motor speeds, and whether each motor is to be going in reverse or forwards.

Genetic Algorithm Design

The algorithm is very simple. The robot moves for 299 steps, the fitness of the robots is judged on how many of those 300 steps move the robot closer to the ball. There are 32 robots in a generation, the top four are chosen, and their neural networks are cloned with a 0.001 probability that each neuron's threshold may mutate, and a 0.001 probability that connection weights will mutate.

Simulation Software

For evolving the networks the software used was basically an implementation of the evolver class, which evolves one generation of networks then ends, run until it was stopped or it reached a predefined fitness limit.

There was no interface to change the the GA parameters on the fly, so new parameters meant that the programme needed to be recompiled.

To view the effectiveness of a neural network there was the robot football simulator. This had a very simple interface – start, stop and quit buttons and a viewing area where the game was drawn. The main class behind this was the pitch class, it also used the team class too. The results section has examples of output from the simulator.

Procedure for experiment 1

10000 generations were run of this network. The 10000 generations were run three times, each time the network was initialised using a different random seed. At the start of each game the robot is placed in the centre of the pitch, and the ball is placed in a random position.

Procedure for experiment 2

This experiment allowed the network to evolve until the total fitness for a generation was greater than 8000 (the maximum fitness for a generation is 9568 (299 movements towards the ball x 32 robots in a generation)). At the start of each game the robot is placed in the centre of the pitch, and the ball is placed in a random position. The network used the three random seed procedure.

3. Results

Experiment 1

The results displayed in figures 3a, b and c show a similar trend of erratic evolution, without getting close to a convergence.

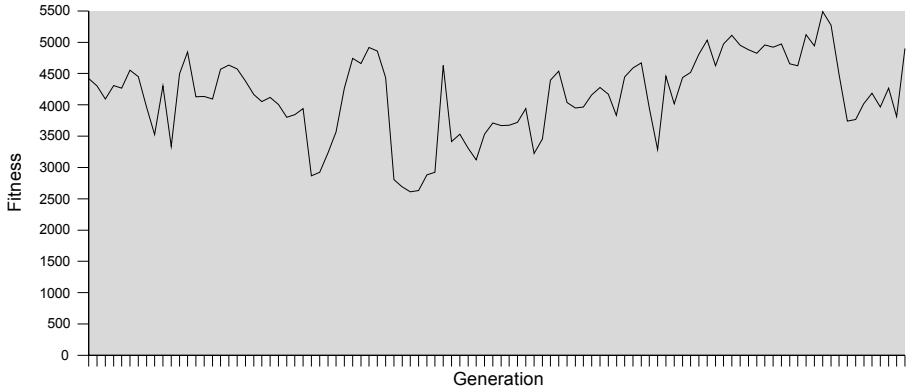


Figure 3a: Average fitness every 100 generations. Seed 1.

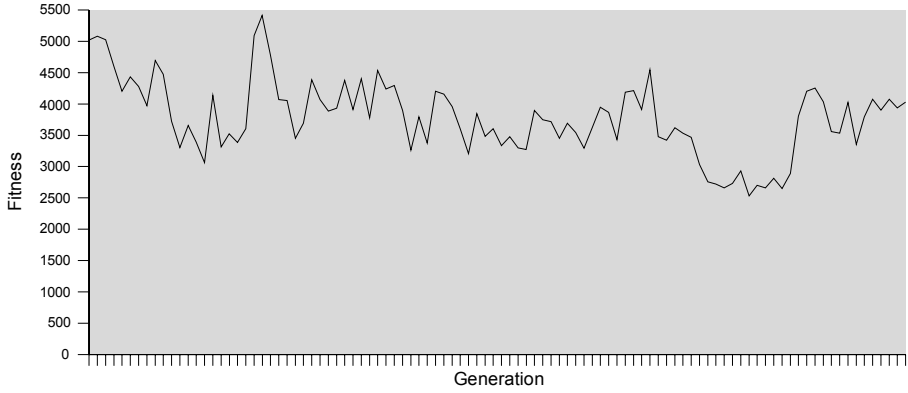


Figure 3b: Average fitness every 100 generations. Seed 2.

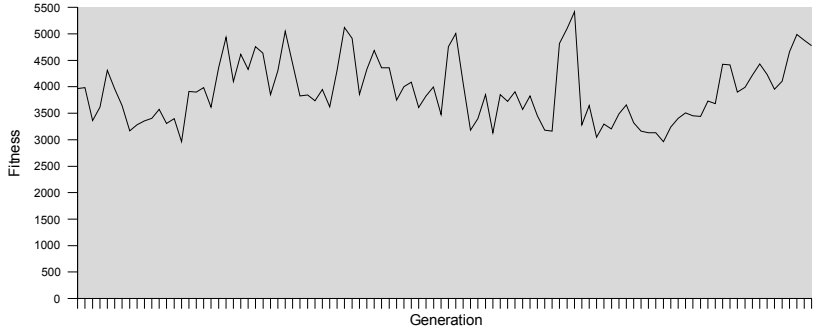


Figure 3c: Average fitness every 100 generations. Seed 3.

Experiment 2

The data shown in figures 4a, b and c are the last 10000 generations for each seed before the target fitness was reached.

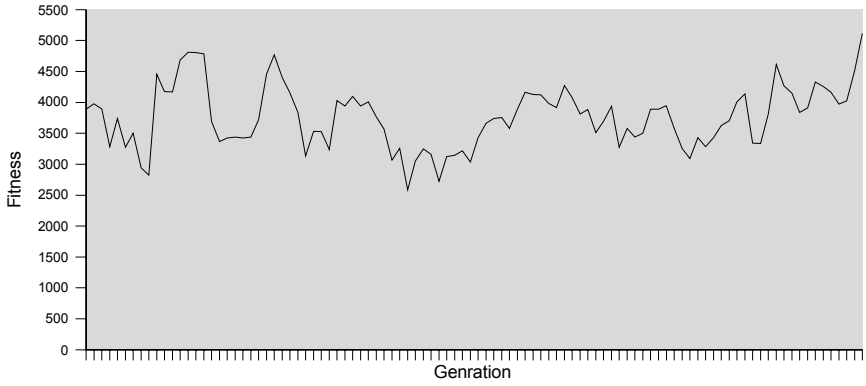


Figure 4a: Average fitness for every 100 generations. Seed 1.

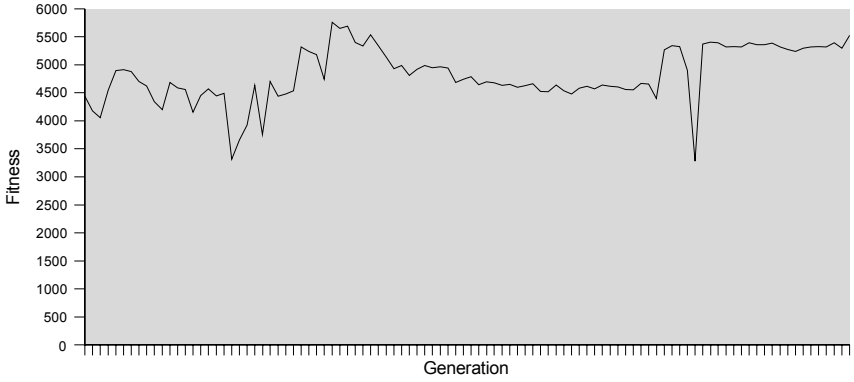


Figure 4b: Average fitness for every 100 generations. Seed 2.

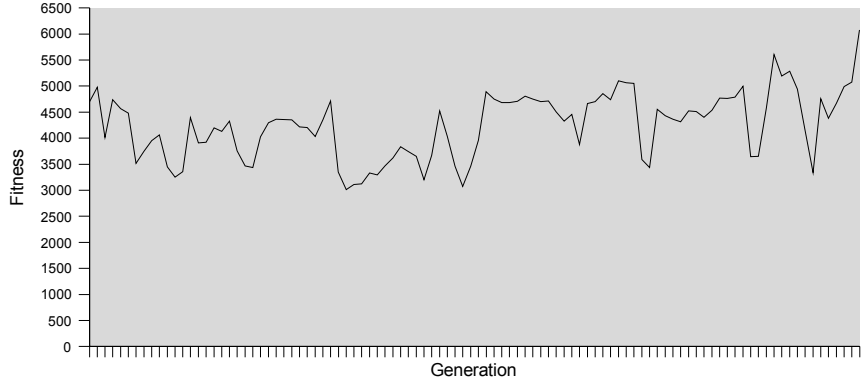


Figure 4c: Average fitness for every 100 generations. Seed 3.

Figures 4a and 4c do not seem to be showing signs of convergence, however figure 4b is showing definite signs, as it plateaus and then improves, and plateaus again.

4. Discussion

The results for both experiments show that there is a lack of ability to evolve a neural network that can move to a stationary point or move towards an object in motion. Figure 4b shows an exception to the norm, and perhaps hope that eventually the system would evolve working networks.

The fact that the lack of convergence comes as a shock is down to the triviality of the tack at hand. Nolfi and Floreano (2000) in their Evorobot literature show conclusive evidence that a neural network can be evolved to moved towards a fixed point quite successfully, and in far fewer generations than in this instance.

Since the neural network is so simple it is unlikely to be the cause of this failure to evolve. The mutation rate for the GA was set so low, it cannot have been the overarching defect here either. It would seem that the cause for the lack of convergence lies in the simplicity of the GA as a whole.

Because the GA picks the best four from 32 then convergence to one network configuration should happen quickly, and comparing the networks with a file comparing utility, it is. However due to the small size of the neural network it is quite likely that any mutation, to the threshold of a neuron or the weight of a connection, will cause the network to cease being fit. Also the likely hood that the top four are there by chance is quite high, considering that evolving the ability to move toward one spot on the pitch is something event this GA can do in a short space of time.

Given these factors, and the fact that the results tend to fluctuate around half the maximum fitness, it is safe to say that the GA is at fault.

Improvements that should be made to the GA, firstly the size of the population should be increased to perhaps 48, or maybe even as high as 60. Then the top eight networks should be chosen, to give a better selection of the population. The next important change should be to go to a crossover style of GA, rather than a cloning one, to give the better genes a better chance of survival.

5. Conclusion

The attempt to evolve a neural network to control constructs for robot football players did not succeed. It came close, but by the time it was known what was wrong with the system it was too late to redesign it and run more experiments. The major factor in the poor design of the GA was the lack of reading done to investigate the pros and cons of GA designs before the implementation phase. During the analysis phase a lot of reading was undertaken to better understand how GAs work, and this brought to light the inherent flaw in the design due to its simplicity.

Admittedly the design of the neural network was not perfect from the beginning, however a lot more was already known about neural networks from the authors past experience to be able to evaluate the situation more successfully to that respect.

Because there was the beginnings of convergence towards the ends of the experiment, this shines a glimmer of hope on the premise the study was undertaken. Hopefully if future iterations of the experiment take place, by this or any other scientist, their understanding of the nature of GAs will prevent them from falling into the same pit as this one.

6. References

Hong T-P., Huang K-Y. and Lin W-Y. (2001), Adversarial Search by Evolutionary Computation. *Evolutionary Computation,* volume 9, issue 3, pp. 371-385.

Huang, Y-L. (2005), Wavelet-based image interpolation using multilayer perceptrons. *Neural Computing & Applications,* Volume 14, Issue 1, pp. 1-10.

Hofmeyr S. A., & Forrest S. (2000), Architecture for an Artificial Immune System. *Evolutionary Computation,* volume 8, issue 4, pp. 443-473.

Nolfi, S. & Floreano, D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organising Machines.* London: The MIT Press.

Pantic L., Torres J.J., Kappen H.J., Gielen S.C.A.M. (2002) Associative Memory with Dynamic Synapses. *Neural Computation,* Volume 14, Issue 12, pp. 2903-2923.

Robinson, P., Bugmann, G., Kyriacou, T., Culverhouse, P., Norman, M., & Simpson, A. J.(2002). Mirosot as a Teaching & learning tool. *Proc. of 2002 FIRA World Congress,* (ed. Ju-Juangn Lee), pp. 309-314.

Robinson, P., Hall, P., Wolf, J., Phillips, R., Peck, C., Culverhouse, P., Bray, R., & Simpson, A. J. (2004).

The Technology and Challenges of Mirosot Robot Football. http://www.tech.plym.ac.uk/robofoot/publications/paulrobinson/Korea%20KAIST%20paper%202003.pdf, (Accessed 2 December 2004)

Author Index

Distributor:
Network Research Group
University of Plymouth
Drake Circus
Plymouth
PL4 8AA
United Kingdom

ISBN 978-1-84102-179-9
90000 >
9 781841 021799

http://www.network-research-group.org