ITB Journal

Issue Number 5, May 2002

Contents

Papers for submission to the next ITB Journal should be sent to the editor at the address below. Alternatively, papers can be submitted in MS-Word format via email to brian.nolan@itb.ie

Brian Nolan
Editor
ITB Journal
Institute of Technology Blanchardstown
Blanchardstown Road North
Blanchardstown
Dublin 15

Editorial

I am delighted to introduce the fifth edition of the ITB Journal, the academic journal of the Institute of Technology Blanchardstown. The aim and purpose of the journal is to provide a forum whereby the members of ITB, visitors and guest contributors from other third level colleges can publish an article on their research in a multidisciplinary journal. The hope is that by offering the chance to bring their work out of their specialised area into a wider forum, they will share their work with the broader community at ITB and other academic institutions.

In this issue, we have again a diverse selection of topics. Dr. Mohamad Saleh of ITB applies fresh thinking to third level education for the modern marketplace. Michael O'Donnell of ITB explores the inner workings of identification protocols in cryptography. Róisín Faherty of the Institute of Technology, Tallaght raises many important issues in her paper on the effects of mobile computing on teleworking. A joint paper by Fionnuala Darby of ITB and Geraldine Lavin explores the role of knowledge and value development within management consulting. and contend that impression management, aided by positive perception is an important motivational force. Moving to the field of linguistics, Brian Nolan of ITB presents an analysis of morphological causation in Irish. At the intersection of computer science and online learning, Matt Smith (ITB) has an interesting paper on pitch circles in which music theory is mapped into a computer-based learning tool writen in the Java programming language. Finally, an interesting paper by Anthony Keane of ITB looks at the physics of cormic radation.

We hope that you enjoy the papers in this issue of the ITB Journal.

Brian Nolan
Editor
ITB Journal
Institute of Technology Blanchardstown
Blanchardstown Road North
Blanchardstown
Dublin 15

Common Sense Approach To Third Level Education For The Modern Market

Dr. Mohamad Saleh
School of Informatic and Engineering
The Institute of Technology Blanchardstown
Road North, Balanchardstown, Dublin 15

Mohamad.saleh@itb.ie

Abstract

A healthy industry is of crucial importance to the well being of a nation. Recently, new technology has caused tangible changes in the modern market by creating continuous needs and demands for skills and technologies in various fields. The educational system is a major component in the provision of the skilful workforce in a society. However, the recent unprecedented rate of industrial expansion has resulted in new industries experiencing recruitment difficulties. It is felt that these issues can be addressed by establishing reliable collaboration between industry and third level educational systems to fulfill the needs using less resources.

This paper is an attempt to examine the mechanism of probable future industrial needs for skills and technologies with some focus on the response of the third level education towards these needs.

1. Introduction

In recent years, there has been a manufacturing quality revolution, which began with Taylor around 1920 and the division of labour. Then Schewhart developed the Control Chart. They were the dominant manufacturing force in the world and concentrated on the "product out" rather than the "market in" situation. The Japanese then embraced their ideas and ironically with Deming and Juran (both American) and home grown talent (Ishikawa and Taguchi et. a!) they developed today's Quality Concept which are based on Total Quality Management (TQM), and "market-in". Due to these concepts the manufacturing industry has gone from strength to strength.

Total Quality Management (TQM) is a philosophy of never-ending improvement achievable only by people. This has grown from the view that quality cannot be "inspected in" to a

product or service. The essential feature of TQM is the improvement of quality, which depends on the attitude of the workforce. In this context, the quality improvement in any organisation must be the responsibility of every member of the organisation. Thus, Total Quality Management is inseparatible from general management practice.

Manufacturing process can be the act of providing something, which somebody wants. Therefore, the educational system is no different from a manufacturing process. However, this system is at present falling behind the manufacturing system with regard to quality within its industry. Thus, in order to progress, it is felt that the educational system should adapt the concept of TQM, similar to that used by the manufacturing system, to respond to the new development and indeed to survive in the modern market place.

At present we are living in a world of manufacturing goods where information technology-based manufacturing is often described as the second industrial revolution. The first industrial revolution brought about the invention of mechanical machines that delivered unparalleled power. Today's industrial revolution, however, is all about processing machines that deliver brain power, and the precision in manufacturing industry. In fact, this has been associated with significant changes in the market for the buying and selling of industrial and consumer products across the globe. Consequently, the relationship between the manufacturer and supplier is now faced with dynamic changes. Therefore, special qualifications and skills are needed for these new developments in the industry. As a result, this influences industrial long-term requirements for the education and training of its employees. Opinion about the impact of the present and the future developments in manufacturing industry diverge widely. In this regard, predicting the exact nature of the future of manufacturing education with any certainty is difficult [1]. However, the future of manufacturing will most likely be a smaller workforce with a higher-order of multi-disciplinary critical skills in management and labour processes to respond radically to the opportunities and constraints that will arise in the market [2].

Competence in the optimal use of information and communication technologies, supporting a global co-operation of enterprises, will be the future key to industrial countries remaining competitive, in both the race to bring new products to the market and sustaining a profitable presence in that market [3]. The multimedia communication services and electronic document systems have great influence in the management of technical documentation within the modern factory [4]. The development in computer-integrated-manufacturing (CIM), and the subsequent appearance of new methodologies and concepts such as World Class

Manufacturing (WCM), Total Quality Management (TQM); Just-in-Time (JIT), Material Requirement Planning (MRP) and Concurrent Engineering (CE) in the automated factory have caused great emphasises to be placed on the integration of the physical and intellectual capabilities of employees [5]. In addition, research and development could be conducted in control of monitoring discrete event systems like Flexible Manufacturing System (FMS) in conjunction with neural networking [6]. Micro-technology might be the determining factor of advanced manufacturing technology in the immediate future. The development of new types of miniaturised and micro-robots with human-like capabilities will play an important role in different applications and tasks [7]. These issues, in recent years, have had a significant impact upon engineering and engineering education in terms of the knowledge and skills required to develop quality market orientated products from a wide range and variety of complex engineering systems. Thus, the interface between industry and education and training is becoming increasingly complex [8]. The growing awareness of the critical importance of Lifelong Learning is likely to change professions and its whole approach to 'qualification' [9]. It has been pointed out that the developments in science and technology have created new interfacial disciplines. These disciplines could be the areas where Lifelong Learning could be in greatest demand [10]. This paper discusses a practical module of industrial training and education at third level in relation to the requirements of modern technology.

2. Third Level Education and Industry

The old style education systems were evolved along three direct traditional lines:
Humboltian, Napoleonic and Anglo-Saxon. These traditions are centuries old . The difference between them lies in where the power resides. In the Humboltian tradition, found in much of Europe, the faculty is very strong, the central administration is weak and there is little government interference. In the Napoleonic tradition, found in France, Poland and Russia, the government has powerful influence and the institutions and faculties are subservient to it. In the Anglo-Saxon tradition, found in the United Kingdom and the former British Colonies, the University's Central Administration has responsibility for the institution, it has control over the faculties and operates quite independently of government. In the latter half of the Twentieth Century, new institutions of higher education have been developed. These have a less scholarly focus and are directed towards the employment market. These institutions have developed within the academic tradition of the nation. Therefore, the power repository in these new institutions is similar to that of the old traditional universities.

In the later half of the 20th Century, new institutions of higher education were developed. These have a less scholarly focus and are directed towards the employment market [11]. A frightening statistic is that in 'some fields', 20% of an engineer's knowledge becomes obsolete every year [12]. This suggests that the globalisation of engineering education requires further characteristics for the modern degree programme. To meet the challenges of the 21st century, universities must become learning organisations which are skilled at creating, acquiring and transferring knowledge, modifying their behaviour to reflect new knowledge and insight [13]. It may be that as access increases into undergraduate courses, postgraduate schools will become the place for higher education [14].

Our society has many groups with legitimate interests in third level education. Thus, the quality of third level educational systems has a considerable influence on the economic well being of a society. Accordingly, the ability of a country to attract industrial and commercial investment is dependent on many factors; the availability of an educated workforce is among them. Therefore, the level of investment and the range of activities, which international organisations bring to a country, are very dependent on the number and quality of the country's graduates. Countries with weak third level educational systems attract labour-intensive, low-knowledge industries; whereas those with strong third level educational systems attract high-value, knowledge-based industries. So, the relationship between the third level educational

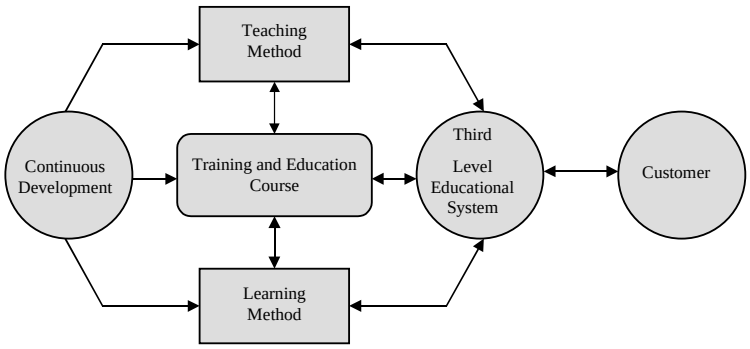


Figure 1.

system and society can be defined based on the customer-satisfaction pattern as shown in Figure 1. This suggests that the third level educational system should forecast/evaluate the needs of the customer and accordingly provide what is required to fulfil them. This can be achieved through appropriate course and methods of delivery with a Lifelong Learning process for lecturers/staff members and students. In this respect, the third level educational

system plays an effective role in knowledge transfer and provision of the required skills in a society.

3. Industrial Training and Education

The technological, economical and political changes around the globe have brought about a new scheme of industrial training and education. As a result, eemployees must now be able to learn as quickly and effectively as possible, using the available resources as efficiently as possible. However, the level of employee developments varies according to the regional economic situation in the world. During 1980's, many books and articles lauded Japanese employee development as a key factor in Japan's economic advantage over the United Sates and other countries. Accordingly, firms around the world adapted Japanese management practices, heeding dire warnings that companies that failed to do so would fall quickly behind in the competitive global economy [15]. In Asia, the current crises and the attendant business down-turn caused a hard stance within the business and financial communities to cut waste and create greater efficiencies. According to Korea's Human Resource Development Magazine 70% of survey respondents had cut their training investments 12.5% by June 1998. More than 16% responded that their training investments remained the same for 1998, and 11.4% said their training investments had increased from previous year [16]. Due to the economic growth in Ireland, the Government established the business, education and training partnership in late 1997 to develop national strategies to tackle the issue of skill needs, manpower needs estimation and training for business [17]. Subsequently, there was a novel pilot project set up between the IT (Institute of Technology) sector in third level education and the national/multinational industry in Ireland for training and education.

Employee learning can be planned/formal or unplanned/informal. Formal learning is referred to as training in the classroom-based, instructor-led training. Informal learning occurs in a company during such activities as team and customer interactions, meetings, cross-training and shift changes. Both formal and informal employee learning could be either under or outside control of the employer. Estimates from several recent reports suggest that most of what the U.S. employees learn occurs informally. The Bureau of Labour Statistics reports that in 1995, employees who worked in firms with 50+ workers received an average of 44.5 hours of formal and informal training during a six-month period. Of that total 70% occurred formally; the remaining was formally outside the control of the employer. On the other hand, the best evidence suggests that employees of Japanese firms also obtained much of their learning informally, but the proportion is substantially smaller than the US figure of 70%. It is estimated that the international percentage of employees receiving training is 75.5% in

Europe, 68.9% in Canada, and 67.3% in Japan respectively. It has been found that employees value training, they will be more loyal to a company that provides it – a fact that attracts the budget-focused eyes in any organisation. Also this enables the company to create the kind of employees it wants [18].

4. Model for Industrial Training and Education

As technology is changing so fast, it does not take long for existing skills to become obsolete. As a result, this makes the market a very competitive labour place. Therefore, it is imperative that third level education, in conjunction with the industry, provides all the necessary tools to attract candidates/employees to keep abreast of the new changes in technology to enable them to build on their skills. This can be achieved by means of strategic collaboration between third level education and industry through research and development and Lifelong Learning as shown in Figure 2. The following approach is based on our recent experience in training and education towards addressing the technical engineering skills. This model can be adjusted to a great many situations for running training and education like a business.

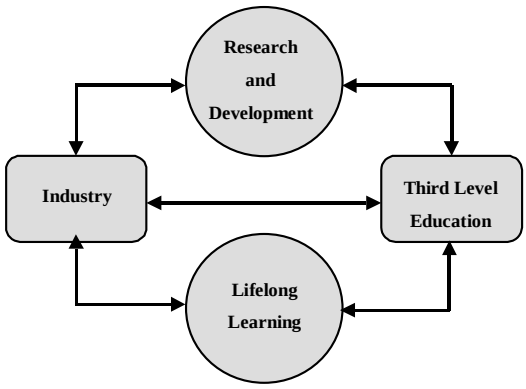


Figure 2.

5. Course Strategy

The main steps in designing a training and educational course for industry are:

(i) Define the present and the future requirements based on the nature of the industry.

(ii) Organize resources available for the course.

(iii) Set scheme for entry and assessment based on the quality assurance benchmark.

(iv) Establish/adopt the teaching method and course contents based on immediate requirements and the focus technique .

(v) Provide a measure for the overall delivery cost.

(vi) Provide a reliable database of information based on give and take information.

(vii) Establish a course management.

The Human Resource Department in the relevant industry should be approached by the third level educational institute to discuss and agree the terms and conditions.

6. Course Costs for Delivery

Costs form the very core of most business people's concept of efficiency. Yet few businesses are even aware what their true training costs are because, in most companies' budgets, many training costs are in effect hidden from the line item covering training expenditure. In fact, only a fraction of the true costs of training are visible in most organizations [19]. The average international expenditure per employee rises to US$960 in Europe, US$531 in Canada and US$579 in Japan respectively [15]. This training expenditure is taken as a percentage of the payroll of the employees.

5. Conclusion

Overall, a successful industrial training and educational program must be flexible, with confident collaboration and reliable give and take information needed for all involved.

Third level education should reflect the rapid change in manufacturing technology. This can be accomplished by the 'market in' approach with strategic links with industry. Course delivery and the learning process should be based on the extensive use of computer technologies and on multimedia communication. This should be based on the teamwork concurrent delivery method rather that sequential method. In this sense, the educational model at third level education will be more like the medical school model where lecturers practice, coach learning and do research. However, it is hoped that the new trend in this education will be balanced between academia and industry.

REFERENCES

1. Mohamad Saleh,"Probable future trends in manufacturing engineering at third level education", Proc. Int. Conf., AMPT'99, Dublin, Ireland 2-6 August 1999, pp. 1949-1957.
2. Marinescu, I.D, Lavelle, J.P, " Innovation and globalization of manufacturing 2000", Int.J. Ind. Eng. Appli. Pract. Vol. 5, no. 3, 1998 Sept., pp. 244-248
3. Hirsch, B.E, Thoben, L.D, Hobeisel, J." Requirements upon human competencies in globally distributed manufacturing", J.Comput. Ind. Vol. 36, no.1-2, 1998 Apr.,pp. 49-54

4. Papandreou, C.A, Adamopoulos, DX., "Architecture of a multimedia communication system for technical documentation in a modern factory", J. Comput. Ind. Vol. 36 no. 1-2, 1998 apr., pp. 83-93.
5. McEwan, A.M., Sackett, P.," The human factor in CIM systems", J. Comput. Ind. Vol. 36 no. 1-2, 1998 Apr., pp. 39-47.
6. Zamai, E. Chaillet-Subia, A., Combacan, M., "An architecture for control and monitoring of discrete events system", J. Comput. Ind., vol. 36, no. 1-2, 1998 Apr., pp. 95-100.
7. Fatikow, S. Benz, M.," A micro-robot-based automated micro-manipulation station for assembly of micro-systems", J. Comput. Ind., vol. 36, no. 1-2, 1998 Apr., pp. 155-162.
8. Graham R. Mackenzie, " Industrial pressure for change in UK education and training", J.Eng. Sci. & Education, vol. 8, no 6, 1999 Dec., pp. 268-270.
9. John E. Midwinter, 2 The challenge of lifelong learning", J. Engineering Science & Education, vol.8, no.6, 1999 Dec., pp. 271-280.
10. Lee, E. A. and Messerschmitt, D.C., Engineering an education for the future", IEEE Comput. Mag., 1999 Jan., pp.77-85.
11. Mohamad Saleh, et. al " A revolutionary style at third level education towards TQM", Proc. Int. Conf. , AMPT'99, Dublin, Ireland 2-6 August 1999, pp. 1949-1957.
12. L. Otata, "Studying for the future", ASEE PRISM, 1993 Oct., pp. 22-29.
13. J.L. Melsa, "Trends in engineering education", J.Eng. Sci. & Education, vol. 6, no. 6, 1997 Dec., pp.239-244.
14. George Brown, " Higher education: an international perspective", Colloquium, University Teaching & Learning: policy & practice, Proceedings (IUTN), Dublin, Ireland, 1998 Dec.
15. Daniel P. McMurrer, Mark E. Van Buren, " The Japanese training scene", Training & Development Mag., The American Society for Training & Development Mag. (ASTD), 1999 Aug., pp. 43-46.
16. Yonjoo Cho, Hye-Young Park and Stancey Wager "Training in Changing Korea" Training & Development Mag., ASTD, 1999 May., pp. 98-99
17. First report of Expert Group on Future Skills Needs, Forfás, Ireland, 1998.
18. Margaret Olesen, " What makes employees stay", Training & Development Mag., ASTD, 1999 Oct.., pp. 48-52
19. David van Adelsberg and Edward A. Trolley, " Running training like business", Training & Development Mag., ASTD, 1999 Oct., pp. 56-61.
20. Mohamad Saleh "A Module of Industrial Training and Education at Third Level Education", Proc. Int. Conf.MIT, Beijing, China, 200.
21. Mohamad Saleh, "Industrial Integrated Approach to Mellennium Education ", Int. Conf. Mech. & Materials in Design, Orlando, USA, 2000.

Identification Protocols in Cryptography

Michael O'Donnell

School of Informatics and Engineering, ITB Blanchardstown, Dublin, Ireland

Abstract

In this paper we examine the role of Identification Protocols in the field of Cryptography. Firstly, the rationale behind the need for Identification Protocols is discussed. Secondly, we examine, in detail, challenge-response protocols, based upon zero-knowledge proofs, that form a subset of Identification Protocols in general. Thirdly, the mathematical tools necessary for the understanding of how these protocols work is given. Finally, we discuss four main Identification Protocols: Fiat-Shamir, Feige-Fiat-Shamir, Schnorr and Guillou-Quisquater. This discussion includes the theory, practical examples and the security aspects of each protocol.

1. Introduction

"If you think cryptography is the answer to your problem, then you don't know what your problem is."

Peter G. Neumann
Quoted in the *New York Times*

[12]

At the moment there are a plethora of terms, definitions and even some disagreement about what constitutes message authentication, data origin authentication, transaction authentication, key authentication and entity authentication or identification.

For consistency and clarity, we will use the definitions as outlined in the "Handbook of Applied Cryptography" by Menezes *et al*. [1]

2. Authentication Methods

Authentication is any process by which an entity (which could be a person or indeed another computer) establishes its identity to another entity. Authentication can generally be divided into two main areas:

1. Entity authentication or identification
2. Data-origin authentication

Authentication in the broadest sense encompasses not only these two areas, but also protection from all **active attacks**. Active attacks involve some modification of the communication between source and destination or indeed the substitution of an authentic communication by another communication entirely. This contrasts with encryption, which only provides protection from **passive attacks**. Passive attacks involve eavesdropping or monitoring of a communication between source and destination. [2]

2.1 Data origin authentication (message authentication)

Data origin authentication is a type of authentication whereby a party is corroborated as being the original source of specified data created at some (typically unspecified) time in the past. By definition, data origin authentication includes **data integrity** - which guarantees that data has not been altered in an unauthorised manner since the time it was created, transmitted or stored by an authorised source.

Message authentication is a term that is similar to data origin authentication. It provides data origin authentication and data integrity but does **not** provide **uniqueness** or **timeliness**.

Uniqueness guarantees that the source of the data can be identified with an established degree of confidence and that the source we are dealing with is who they purport to be. Timeliness guarantees that the data was sent by the source at a time which can be verified by both the source and destination.

Two common methods of providing message authentication include:

1. Message authentication codes (MACs)
2. Digital signature schemes

While MACs and digital signature schemes may be used to establish that a specified party generated data at some time in the past, they establish no uniqueness or timeliness guarantees. These methods alone cannot detect for example **message replay** (where a message is passively captured and is subsequently retransmitted to produce an unauthorised effect).

2.2 Transaction authentication

Transaction authentication not only provides message authentication but also uniqueness and timeliness of data, thus preventing undetectable message replay. The uniqueness and timeliness guarantees are provided by random numbers in **challenge-response protocols**, sequence numbers and timestamps. [1]

A typical scenario, involving the use of challenge-response protocols in a private-key cryptosystem such as DES (Data Encryption Standard), consists of the following three steps: [3]

1. Bob chooses a challenge, *x*, which is a random 64-bit string. Bob sends *x* to Alice.

2. Alice computes
$$y = E_k(x)$$
and sends it to Bob.

3. Bob computes
$$y' = E_k(x)$$
and verifies that
$$y' = y$$

2.3 Entity authentication (identification)

Entity authentication is the process whereby one party is assured of the identity of a second party involved in a protocol and that the second party has actually participated in the execution of that protocol. These protocols, which are more commonly known as **identification protocols,** use **asymmetric** techniques, but do not rely on digital signatures or public-key encryption. Neither do they use sequence numbers or timestamps. Instead they use random numbers, both as a challenge and a commitment, based upon interactive proof systems and zero-knowledge proofs. [1]. Challenge-response protocols based upon zero-knowledge proofs are the main topic of investigation in this paper.

2.4 Key authentication

Key authentication is the property whereby one party is assured that no other party aside from a specifically identified second party (and possibly additional identified trusted parties) may gain access to a particular secret key. This requires both the secrecy of the key and identification of those parties with access to it. The identification requirement here differs in one very important respect from that of entity authentication outlined above. Here the requirement is knowledge of the identity of parties that may gain access to the key, rather than corroboration that actual communication has been established with such parties.

Kerberos is a key transport protocol that is based on **symmetric** techniques. In symmetric techniques, the same key is shared between the two parties, Alice and Bob, with the proviso that the key is kept secret from an adversary such as Oscar. It may, however, be shared with a trusted third party or a key distribution centre (KDC). Kerberos is an example of a challenge-response scheme providing both entity authentication and key establishment based on symmetric encryption.

Public-key techniques may also be used for challenge-response based identification providing both entity authentication and key establishment.

The table summarises the properties already defined.

	Property		
Type of authentication	**Identification of source**	**Data integrity**	**Timeliness or Uniqueness**
Message authentication	Yes	Yes	----
Transaction authentication	Yes	Yes	Yes
Entity authentication	Yes	----	Yes
Key authentication	Yes	Yes	Desirable

3. Identification Objectives

Identification is the process through which one ascertains the identity of another person or entity. In our daily lives, we identify family members, friends and coworkers by their physical properties, such as voice, face or other characteristics. These characteristics, called biometrics, can be used on computer networks with special hardware. Entities on a network may also identify one another using cryptographic methods. An **identification scheme (or**

protocol) allows Alice to identify herself to Bob in such a way that someone listening in cannot pose as Alice later. One example of an identification scheme is a **zero-knowledge (identification) protocol.** Zero-knowledge protocols allow a person (or a server, website, etc.) demonstrate they have a certain piece of information without giving it away to the person (or entity) they are convincing.

Suppose Alice knows how to solve the Rubik's cube and wants to convince Bob she can do so without giving away the solution. They could proceed as follows. Alice gives Bob a Rubik's cube which he thoroughly messes up and then gives it back to Alice. Alice turns away from Bob, solves the puzzle and hands it back to Bob. So Bob is convinced that Alice solved the puzzle without giving away the solution.

This idea may be adapted to an identification protocol if each person (or entity) is given a "puzzle" and its answer or indeed each entity makes up a "puzzle" that only they themselves can solve. The security of the system relies on the difficulty of solving the "puzzle". In the case above, if Alice was the only person who could solve a Rubik's cube, then this could be **her** puzzle.

The idea is to associate each person with something unique; something that only that person can reproduce. This, in effect, takes the place of a face or voice, which are unique factors allowing people to identify one another in the physical world. [6]

In general terms, an identification protocol involves a **claimant** *A* and a **verifier** *B*. The verifier is presented with, or presumes beforehand, the purported identity of the claimant. The goal is to corroborate that the identity of the claimant is indeed *A*, i.e. *A* provides entity authentication.

3.1 The Objectives of Identification Schemes

From the point of view of the verifier, the outcome of an identification protocol is either acceptance of the claimant's identity as authentic or rejection of the claimant's identity. More specifically, the objectives of an identification protocol include the following:

1. In the case of honest parties *A* and *B*, *A* is able to successfully authenticate itself to *B*; i.e. *B* will complete the protocol having accepted *A*'s identity.

2. *B* cannot reuse an identification exchange with *A* so as to successfully impersonate *A* to a third party.

3. The probability is negligible that any party *C* distinct from *A*, carrying out the protocol and playing the role of *A*, can cause *B* to complete and accept *A*'s identity.

4. The previous points are true even if a large number of previous authentications between *A* and *B* have been observed or indeed *C* has participated in previous protocol executions with either or both *A* and *B*. [1]

3.2 Applications

There are many common everyday situations where it is necessary to prove "electronically" one's identity.

Typical scenarios include:

- In consumer payment transactions, a token or a card may be presented, bearing an identifier (a code), and a sample signature or an encoded PIN for automated authentication processes.
- To log in remotely to a computer over a network, it suffices to know a valid user name and the corresponding password.
- In establishing a new relationship between a customer and a financial institution, we may need to provide a set of "tokens" such as a passport, driver's licence and/or utility bills bearing the applicant's home address.

It is important to note that in all identification schemes or protocols the phrase "proof of identity" should be avoided in favour of the term "evidence of identity". [22]

3.3 Basis of Identification

Entity authentication techniques may be divided into three main categories, depending on which of the following the security is based:

1. *What you know*.
   Examples include standard passwords, passphrases, PINs and secret or private keys whose knowledge is demonstrated in challenge-response protocols.

2. *What you have.*
   This is normally a physical accessory like a passport or a magnetic-striped card like a credit card. It also includes token-based systems, such as smartcards that have an embedded microprocessor, or password generators that provide time-variant

passwords. Providing something you know – like a password, is usually augmented by the possession of some object, to prove one's identity.

For example, challenge-response systems can include a hand-held token, typically the size of a small calculator, with a keypad. To access a system using challenge-response technology, a user activates the token using a PIN and then enters into the token a series of codes that are challenged by the system. This technology has some limitations however, including the number of user steps, the longer time required to authenticate the identity of a user and the size and complexity of the token.

3. *What you are.*
   This category includes methods that make use of human physical characteristics or involuntary response patterns known as biometrics. These include fingerprints, voice, retinal patterns, hand geometry and face or body profile. [21]

3.4 The Quality of Identification Schemes

In each of the three methods outlined above, there is always the need to strike a balance arising from:

- False rejection (Type I error)
- False acceptance (Type II error)

In a false rejection type error, an identity is accepted that should have been rejected, resulting in the acceptance of imposters and mistaken identity. In a false acceptance type error, an identity is not accepted that should have been, resulting in the rejection of correct matches.

Sources of poor quality identification include:

- Accidental mistakes
- Intentional mistakes, which include masquerading or spoofing (pretending to be a different entity) and the avoidance of identification.

Where quality shortfalls occur, additional considerations come into play which include:

- The **repudiability** of an assertion – for example the question of how an entity contests information stored by another party.
- The **onus of proof** – for example the need to establish on which entity the responsibility lies to establish that data is or is not of appropriate quality. [7]

This balance depends on the extent of protections against abuse and hence if it can be repudiated by the entity concerned. Hence we have the need for different **levels of authentication** that depend on many factors such as our resources, the level of protection that we require, the nature of the communication between the entities involved and the nature of the network communication channels between source and destination.

4. Levels of authentication

There are three main levels of authentication, consisting of:

1. Weak authentication
2. Moderate authentication
3. Strong (zero-knowledge based) authentication

4.1 Weak authentication

Weak authentication may be sub-divided into **one-s**tage authentication and **two-stage** authentication.

4.1.1 One-stage authentication

In the case of one-stage authentication, end users need only one item of information to verify the username they enter when they log on. This one item is usually a password, which often remains the same for a significant amount of time. Furthermore, end users tend to use passwords that are short and easy to remember – usually a family name or some word from a standard dictionary.

Even if users keep their password secret, they maybe captured by protocol analysers or maybe the subject of a **dictionary attack.** Although this brute method approach of finding a particular user's password is not very successful, they are successful at finding passwords in most systems, given the predictable nature of most people's passwords. This success rate is largely due to the **birthday paradox**. The birthday paradox states the counterintuitive result that the probability that two people share the same birthday in a group of 23 people is greater than 0.5. [2]

A further drawback of this method is that it provides no guarantee against privileged insiders or disgruntled employees gaining access to the stored passwords. One way to combat this drawback is to store the passwords in an encrypted file. In this scenario, to verify a user-

entered password, the system computes a **one-way function** of the entered password and compares this to the stored entry for the stated user-id.

Let $f(x)$ be a one-way function. This means it should be easy to compute $f(x)$ but it should be **computationally infeasible** for an opponent to calculate x given the value of y. [4] Generally a problem that cannot be solved in polynomial time is considered infeasible or intractable. Polynomial time in Big-O notation means that the running time of an algorithm with an input of size n is $O(n^t)$ for some constant t. [2]

Another method of making dictionary attacks less effective is that each password, upon initial entry, is randomly padded with an additional 12-bit salt. The salt is a 12-bit random number that is appended to the password resulting in 4,096 possible encryptions of the one password. These 12 bits are then used to modify the function $f(x)$. The result is stored in the password file, along with the user's name and the values of the 12-bit salt. When the user enters the password x, the computer finds the value of the salt for this user in the file, which it then uses in the computation of the modified $f(x)$. This is then compared to the value stored in the file. [5]

Another major weakness of schemes using fixed, reusable passwords is that passwords are transmitted in cleartext over the communications link between the user and the system. An eavesdropper may record this data, thus allowing subsequent impersonation.

4.1.2 Two-stage authentication

Token-based authentication is an example of two-stage authentication. In contrast to password authentication, which relies solely on the use of a single password, two-stage authentication incorporates a PIN in addition to a hardware or software device. Smart cards, which are preprogrammed with unique passwords, are an example of hardware tokens. Another scenario uses a physical device called a token, which generates a token code. The token displays a new code every 60 seconds; therefore each token code is used only once. [14]

Software tokens are generated by software installed on a computer. After being activated by a user upon entering a PIN, a software token provides a unique password for authentication. [15] Hardware tokens that generate One Time passwords seem like a great idea but they are expensive. Additional costs include replacing lost tokens, and administering and updating the authentication server's database. For this reason they have not been widely adopted, even though it's a much more secure system than the use of passwords alone. [16]

4.2 Moderate authentication

A natural progression from fixed password schemes (weak authentication) to challenge-response identification (strong authentication) may be observed by considering **one-time password schemes**. [1] As mentioned earlier, a major security concern of fixed password schemes is eavesdropping and subsequent replay of the password. A partial solution is the use of one-time passwords, where each password is used only once. Such schemes are safe from passive adversaries who eavesdrop and later attempt impersonation.

One variation of this one-time password scheme is the Lamport's scheme, which is based upon one-way functions. Lamport proposed an elegant scheme for one-time passwords requiring no specialised hardware. A general-purpose computer can be used to generate responses based on a secret reusable password. In the scheme, the user's secret password is never sent over an insecure link, where it could be captured by an eavesdropper. Say the secret password is w. We then use a one-way function H. H is used to define a sequence of passwords, w, $H(w)$, $H(H(w))$, etc. A fixed number of identifications are initially agreed to and each password in the sequence is then verified by the intended destination. If each password in the sequence is accepted, then the destination accepts the source as authentic. [10]

This scheme however remains vulnerable to an active adversary who intercepts and traps an as-yet-unused one-time password, for the purposes of subsequent impersonation.

4.3 Strong authentication

As outlined earlier, the idea behind strong authentication techniques or challenge-response protocols is that one entity (the claimant) "proves" its identity to another entity (the verifier) by demonstrating knowledge of a secret only known to itself, without revealing the secret to the verifier during the protocol. This forms the basis of all **zero-knowledge** techniques. The **Zero-Knowledge Identification Protocol (ZKIP)** is part of an **interactive proof system**, which uses zero-knowledge techniques in order to achieve identification.

Informally, an interactive proof is a protocol between two parties in which one party, called the claimant, tries to prove a certain fact to the other party, called the verifier. An interactive proof usually takes the form of a ZKIP protocol (or challenge-response protocol), in which the claimant and the verifier exchange messages and the verifier either "accepts" or "rejects" the fact that the claimant tries to prove. [6]

More formally, the claimant's objective is to prove to the verifier the truth of an assertion, e.g. claimed knowledge of a secret. The verifier either accepts or rejects the proof. However, in an interactive proof system such as this, the traditional mathematical notion of proof where proofs are absolute, needs to be discarded. Instead, interactive proofs are probabilistic rather than absolute and a proof in this context needs to be correct only with bounded probability, albeit possibly arbitrarily close to 1.

Interactive proofs, used in cryptographic applications, have three essential properties:

1. *Completeness*

   The verifier accepts the proof if the claimant knows the fact with an overwhelming probability, and both the claimant and the verifier follow the protocol. The definition of overwhelming probability depends on the application, but it generally implies that the probability of failure is not of practical significance.

2. *Soundness*

   The verifier always rejects the proof, if the claimant does not know the fact, as long as the verifier follows the protocol.

3. *Zero-knowledge*

   The verifier learns nothing about the fact being proved (except that it is correct) from the claimant that he/she could not already learned without the claimant. In a zero-knowledge scheme, the verifier cannot even later prove the fact to anyone else. The essential point here is that only a single bit of information need to be conveyed – namely, that the claimant actually **does know** the fact that it wishes to prove. [8]

5. Zero-Knowledge Identification Protocols (ZKIPs)

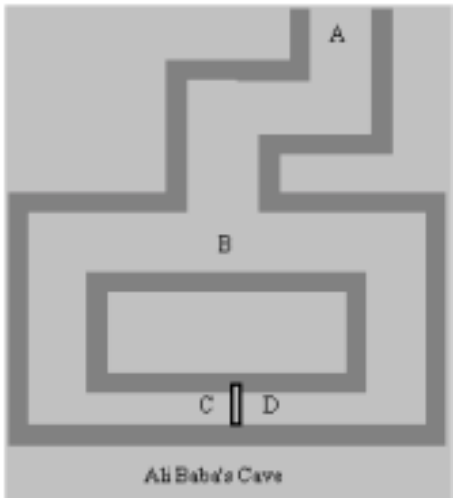
Some years ago, it was reported that some thieves set up a fake ATM machine in a shopping centre. When a person inserted a bankcard and typed in their PIN identification, the machine recorded the information but responded with the message that it could not accept the card. The thieves then made counterfeit bank cards and then went to legitimate ATMs and withdrew cash, using the PIN numbers they had obtained.

How can this be avoided? There are several situations where someone reveals a secret identification number or password in order to complete a transaction. Anyone who obtains

the secret information can masquerade as this person. What is needed is a way to use the secret information without giving away any of this secret information that can be reused by an eavesdropper. This is where zero-knowledge techniques come in.

Quisquater *et al*, [9], explain zero-knowledge protocols by retelling the legendary story of *Ali Baba and the Forty Thieves.*

Alice wants to prove to Bob that she knows the secret words that will open the door at CD in the cave, but she does not wish to reveal the secret to Bob. In this scenario, Alice's commitment is to go to C or D. A typical round in the proof proceeds as follows:



Bob goes to A and waits there while Alice goes to C or D. Bob then goes to B and shouts to Alice to appear from either the right side or the left side of the tunnel. If Alice does not know the secret words "Open Sesame", there is only a 50% chance that she will come out of the side of the tunnel requested by Bob. Bob can repeat this challenge as many times as he desires, until he is certain that Alice knows the secret words. In each round, of course, Alice randomly chooses which side of the tunnel she will go down and Bob randomly chooses which side he will request. Therefore, if Alice comes out the correct side of the tunnel for each of 10 consecutive repetitions, say, there is only one chance in $2^{10} = 1024$ that Alice doesn't know how to go through the door CD.

No matter how many times the proof repeats, Bob will never learn the secret words. Suppose Oscar is watching the proceedings on a video monitor set up at B. He will not be able to use anything he sees to convince Bob or anyone else that he, too, can go through the door. Moreover, he might not even be convinced that Alice can go through the door. After all,

Alice and Bob could have planned the sequence of rights and lefts in advance. By this reasoning, there is no useful information that Bob obtains that can be transmitted to anyone.

Note that there is never a proof, in the strict mathematical sense, that Alice can go through the door. But there is overwhelming evidence (the overwhelming probability referred to earlier) that she can, obtained through a series of challenges and responses. This is, in essence, the nature of zero-knowledge "proofs".

6. Mathematics of Zero-Knowledge Protocols

Zero-knowledge Identification Protocols (ZKIP) are based on **Euler's totient function** and **discrete logarithms** over the **subgroup *Z*/*nZ***. [11]

6.1 Euler's Totient Function

Euler's Totient Function

Euler's Totient function is written as $\phi(n)$, where $\phi(n)$ is the number of positive integers less than and **relatively prime** to *n*.

Definition

The integers *a* and *b* are relatively prime if they have no prime factors in common, that is, if their only common factor is 1. This is equivalent to saying that *a* and *b* are relatively prime if gcd (*a*,*b*) = 1 where gcd (*a*, *b*) stands for the Greatest Common Divisor of *a* and *b*.

Example:

$\phi(21) = 12$ where the 12 integers are {1, 2, 4, 5, 8, 10, 11, 13, 16, 17, 19, 20}

It follows from the above that for a prime number *p*

$$\phi(p) = p - 1$$

Now suppose that we have two prime numbers *p* and *q*. Then for *n* = *pq*

$$\phi(pq) = (p-1)(q-1)$$

Example:

21 = (3)(7)

So $\phi(21) = (3 - 1)(7 - 1) = 12$ and the 12 integers are listed above.

6.2 Congruences

Definition

Let *a*, *b*, *n* be integers.

We say that $a \equiv b \mod n$

i.e. *a* is congruent to *b* mod *n* if *a* – *b* is a multiple of *n*.

6.3 The Extended Euclidean Algorithm

Theorem

$b \in Z_n$ has a multiplicative inverse if and only if gcd (*b*, *n*) = 1

The set of residues modulo *n* that are relatively prime to *n* is denoted by Z_n^*. Any element in Z_n^* will have a multiplicative inverse b^{-1}, which is also in Z_n^*.

The extended Euclidean algorithm is an efficient way to compute b^{-1}. [24]

Example:

Compute $28^{-1} \mod 75$

Solution:

$75 = 2 \times 28 + 19$

$73 \times 28 \mod 75 = 19$

$28 = 1 \times 19 + 9$

$3 \times 28 \mod 75 = 9$

$19 = 2 \times 9 + 1$

$67 \times 28 \mod 75 = 1$

$9 = 9 \times 1$

Hence, $28^{-1} \mod 75 = 67$

Proposition

Suppose gcd (*a*, *n*) = 1.

Let *s* and *t* be integers such that *as* + *nt* = 1. Integers *s* and *t* can be found using the Extended Euclidean algorithm.

Then $as \equiv 1 \mod n$, so *s* is the multiplicative inverse for *a* (mod *n*).

6.3.1 Solving $ax \equiv c \mod n$ when gcd (a, n) = 1

1. Use the Extended Euclidean algorithm to find integers s and t such that $as + nt = 1$.
2. The solution is $x \equiv cs \mod n$.

Example:

Solve $11111x \equiv 4 \mod 12345$

Solution:

Using the Extended Euclidean algorithm, we find that

$$11111.2471 + 12345.\ t = 1$$

Hence $11111.2471 \equiv 1 \mod 12345$

Multiplying each side by 2471 yields

$$x \equiv 9884 \mod 12345.$$

In practice, this means that if we are working mod 12345 and we meet the fraction $\frac{4}{11111}$, we can replace it with 9884.

6.4 Chinese Remainder Theorem

Suppose gcd (m, n) = 1. Given a and b, there exists exactly one solution x (mod mn) to the simultaneous congruences $x \equiv a \mod m$ and $x \equiv b \mod n$

Example:

Solve $x \equiv 3 \mod 7$ and $x \equiv 5 \mod 15$

Solution:

$x \equiv 80 \mod 105$ since 105 = 7.15

Since $80 \equiv 3 \mod 7$ and $80 \equiv 5 \mod 15$, 80 is a solution. [23]

6.5 Modular Exponentiation

We now consider numbers in the form $x^a \mod n$.

Example:

Suppose we want to compute $2^{1234} \mod 789$.

Solution:

We could perform each multiplication and then work out the remainder. This method is too slow to be of any practical value, so instead we begin with $2^2 \equiv 4 \bmod 789$ and repeatedly square both sides as follows:

$$2^4 \equiv 4^2 \equiv 16$$
$$2^8 \equiv 16^2 \equiv 256$$
$$\vdots$$
$$\vdots$$
$$\vdots$$
$$2^{1024} \equiv 286$$

Since 1234 = 1024 + 128 + 64 + 16 + 2

we have $2^{1234} \equiv 286.559.367.49.4 \equiv 481 \bmod 789$.

6.6 Fermat's Little Theorem

Fermat's Little Theorem

If *p* is a prime and *p* does not divide *a*, then $a^{p-1} \equiv 1 \bmod p$

Example:

Compute $2^{43210} \bmod 101$

Solution:

From Fermat's Theorem, we know that $2^{100} \equiv 1 \bmod 101$.

Therefore, $2^{43210} \equiv (2^{100})^{432} 2^{10} \equiv 1^{432} 2^{10} \equiv 1024 \equiv 14 \bmod 101$.

6.7 Euler's Theorem

Euler'sTheorem

For every *a* and *n* that are relatively prime, $a^{\phi(n)} \equiv 1 \bmod n$

Example:

If $a = 3$ and $n = 10$,

$\phi(10) = 4$ and $3^4 = 81 \equiv 1 \bmod 10$

Definition

Z_n^* denotes the set of numbers *i*, $0 \le i \le n$, which are relatively prime to *n*.

Example:

$Z_9^* = \{1, 2, 4, 5, 7, 8\}$

Multiplication Table

* mod 9	1	2	4	5	7	8
1	1	2	4	5	7	8
2	2	4	8	1	5	7
4	4	8	7	2	1	5
5	5	1	2	7	8	4
7	7	5	1	8	4	2
8	8	7	5	4	2	1

Theorem

Z_n^* forms a group under modulo *n* multiplication. The identity element is *e* = 1.

Definition

The order of an element $a \in (G,\circ)$ is the smallest positive integer such that $a \circ a \circ ..... \circ a = a^0 = 1$ where $\circ$ is some binary operation defined on the group *G*.

So, for example, from the multiplication table above, the order of *a* = 2 in $(Z_9^*,\times)$ is 6 because $2^6 = 1$, i.e. ord (2) = 6.

In fact, we know that if *p* is prime, then Z_p^* is a group of order *p* – 1.

Definition

A group $(G,\circ)$ which contains elements *a* with maximum order ord (*a*) = | *G* | is said to be **cyclic**. Elements with maximum order are called **generators** or **primitive elements (roots)** of the group $(G,\circ)$.

Note: |*G*| is the number of elements in the group *G*. [23]

This, in effect, implies that if *p* is prime, then the group Z_p^* is in fact cyclic: there exists an element $a \in Z_p^*$ having order equal to *p* – 1.

Theorem

If *p* is prime, then Z_p^* is a cyclic group.

Example:

Suppose $p = 13$.

Then 2 is a primitive element modulo 13 because

$2^i \bmod 13 = 2, 4, 8, 3, 6, 12, 11, 9, 5, 10, 7, 1$ where $1 \le i \le 12$.

i.e. 2 generates all 12 elements of Z_{13}^*.

The element 2^i is primitive if and only if gcd (i, 12) = 1, i.e., $i = 1, 5, 7$ or 11.

Hence, the primitive elements modulo 13 are 2, 6, 7 and 11.

More generally, we can say that the highest possible exponent to which a number can belong (mod *n*) is $\phi(n)$. If a number is of this order, it is referred to as a primitive element of *n*,

i.e. $a, a^2, ..., a^{\phi(n)}$

are distinct (mod *n*) and are all relatively prime to *n*.

This implies that $a^{\phi(n)} \equiv 1 \mod n$……. Euler's Theorem

6.8 Square Roots Mod *n*

How do we find the solution or solutions to $x^2 \equiv 71 \mod 77$? Or more generally, consider the problem of finding all the solutions of $x^2 \equiv b \mod n$, where $n = pq$ is the product of two primes. It can be shown that this can be done quite easily once the factorisation of *n* is known. Conversely, if we know all the solutions, then it is easy to factor *n*.

Proposition

Let $p \equiv 3 \mod 4$ be prime and let *y* be an integer. Let $x \equiv y^{(p+1)/4} \mod p$

1. If *y* has a square root mod *p*; then the square roots of *y* mod *p* are $\pm x$.
2. If *y* has no square root mod *p*, then $-y$ has a square root mod *p*, and the square roots of $-y$ are $\pm x$.

Example 1:

Solve the equation $x^2 \equiv 5 \mod 11$

Solution:

Since $(p + 1)/4 = 3$,

we compute $x \equiv 5^3 \equiv 4 \mod 11$.

Since $4^2 \equiv 5 \mod 11$,

$x = \pm 4$.

Example 2:

Solve the equation $x^2 \equiv 71 \mod 77$

Solution:

$x^2 \equiv 71 \equiv 1 \mod 7$ and $x^2 \equiv 71 \equiv 5 \mod 11$.

Therefore, $x \equiv \pm 1 \mod 7$ and $x \equiv \pm 4 \mod 11$.

Using the Chinese Remainder Theorem, we can combine a congruence mod 7 and a congruence mod 11 into a congruence mod 77.

Here, we recombine in four ways to get the solutions $x \equiv \pm 15, \pm 29 \mod 77$

It follows from the above that if $a \equiv b \mod p$ and $a \equiv -b \mod q$, gcd $(a - b, n) = p$ and we have found a nontrivial factor of $n = pq$. In the example above, we know that $15^2 \equiv 29^2 \equiv 71 \mod 77$. Therefore, gcd (15 – 29, 77) = 7 gives a nontrivial factor of 77.

We can now state in summary an important result:

> Suppose $n = pq$ is the product of two primes congruent to 3 mod 4, and suppose y is a number relatively prime to n which has a square root mod n.
>
> Then finding the four solutions $x \equiv \pm a, \pm b$ to $x^2 \equiv y \mod n$ is computationally equivalent to factoring n.

6.9 Finite Fields

Loosely speaking, a set that has the operations of addition, multiplication, subtraction and division by nonzero elements is called a field.

Examples of fields include real numbers, complex numbers and the integers mod a prime number, i.e. Z_p. But the set of all integers is **not** a field because we sometimes cannot divide and obtain an answer in the set, e.g. 4/3 is not an integer. [25]

6.10 The Discrete Logarithm Problem

We begin by describing the problem in the setting of a finite field Z^p, where p is prime. The problem is considered to be difficult if p is carefully chosen.

In particular, there is no known polynomial-time algorithm for the Discrete Logarithm problem. To thwart known attacks, *p* should be at least 150 digits and *p* – 1 should have at least one "large" prime factor. [3]

The main advantage of the Discrete Logarithm problem in challenge-response protocols (and indeed in cryptography generally) is that finding discrete logs is difficult, but the inverse operation of exponentiation can be computed efficiently. Stated another way, exponentiation modulo *p* is a one-way function for suitable primes *p*, i.e. it is computationally infeasible.

The Discrete Logarithm problem can be formally stated as follows:

We assume that *p* is prime and α is a primitive element modulo *p*. We take *p* and α to be fixed. From the foregoing discussion, we know that the powers of α from 1 through to (*p* – 1) produce each integer from 1 through (*p* – 1) exactly once.

It follows, therefore, that given $\beta \in Z_p^*$, we can find the unique exponent *a*, $0 \le a \le p-1$, such that $\alpha^a \equiv \beta (\mathrm{mod}\ p)$. In other words, taking logs, we have $a = \log_\alpha(\beta)$

The problem of finding ***a*** is called the Discrete Logarithm problem. Note that if we dispensed with the requirement that α be a primitive root, then the discrete logarithm will not be defined for certain values of β.

7. Identification Protocols

7.1 General structure of Zero-Knowledge Protocols

The Fiat-Shamir protocol illustrates the general structure of a large class of **three-move** zero-knowledge protocols:

$A \rightarrow B$: witness
$B \rightarrow A$: challenge
$A \rightarrow B$: response

The design of these protocols ensures that only the legitimate party *A*, with knowledge of *A*'s secret, is truly capable of answering all the questions. Furthermore, the answer to any of these questions provides no information about *A*'s long-term secret.

A responds to at most one challenge (question) for a given witness, and should not reuse any witness. [1]

7.2 Fiat-Shamir Identification Protocol

What follows is a basic version of the Fiat-Shamir Zero-Knowledge protocol. The objective is for Alice to identify herself by proving knowledge of a secret *s* to any verifier such as Bob or a trusted centre called Tom, without revealing any information about *s* not known or computable by Bob or Tom prior to execution of the protocol. The security of the method relies on the difficulty of extracting square roots, modulo large composite integers *n* of unknown factorisation, which is equivalent to factoring *n*.

The protocol involves the following steps:

1. *One-time setup*

   (i) A trusted authority – call it Tom – selects two large prime numbers *p* and *q*, calculates $n = pq$, publishes *n*, and keeps the primes secret.

   (ii) Alice, a user of the centre, selects a number *s* in the range 1 to $n - 1$ that is relatively prime to *n*, calculates $v = s^2 \bmod n$, registers *v* as her public key with Tom, keeps her password *s* secret and carries both *s* and *v* with her.

   (iii) Alice and Bob agree on a maximum number of rounds *t* of the identification protocol that will be carried out at login.

2. *Protocol messages*

   Each of *t* rounds has three messages as follows:

   Alice $\rightarrow$ Bob: $x = r^2 \bmod n$

   Bob $\rightarrow$ Alice: $e \in \{0, 1\}$

   Alice $\rightarrow$ Bob: $y = r.s^e \bmod n$

3. *Protocol actions*

   The following steps are iterated *t* times – sequentially and independently. Bob accepts the proof if all *t* rounds succeed.

   (i) Alice first generates a random number *r* in the range 1 to $n - 1$. This is called her **commitment.** She calculates $x = r^2 \bmod n$ - called her **witness** – and sends *x* to Bob.

   (ii) On receiving *x*, Bob selects a bit *e*, either 0 or 1, at random – called a challenge and sends *e* to Alice.

(iii) On receiving *e*, Alice calculates $y = r.s^e \bmod n$. If $e = 0$ then $y = r$ and if $e = 1$, $y = r.s \bmod n$.

y is called the **response.** She sends *y* to Bob.

(iv) Receiving y, Bob calculates

$$z = y^2 \bmod n$$

and

$$z' = x.v^e \bmod n$$

If $z \neq z'$, he refuses the login from Alice; if $z = z'$, he accepts the round of the protocol. [1]

In the later case, if fewer than *t* rounds have been carried out, Alice starts a new round by picking another number at random and if *t* rounds are all successful, Alice's identification is complete and she is logged in.

The following congruences show that Alice will be able to prove her identity to Bob:

$$\begin{aligned} xv^e &\equiv r^2 v^e \bmod n \\ &\equiv r^2 s^{2e} \bmod n \\ &\equiv (rs^e)(rs^e) \bmod n \\ &\equiv y^2 \bmod n \end{aligned}$$

So Bob will accept Alice's proof of identity.

7.3 Example to illustrate the Fiat-Shamir Protocol

1. *One-time setup*

(i) Suppose $p = 37$ and $q = 101$; then $n = 37.101 = 3737$.

(ii) If Alice selects $s = 113$ (which is relatively prime to *n*), then her public key is $v = s^2 \bmod 3737 = 113^2 \bmod 3737 = 1558$.

(iii) Bob and Alice now agree on $t = 8$ rounds of the protocol. This completes the setup.

2. *Protocol Actions*

(i) Now Alice wants to login from a remote location. She chooses a number at random: $r = 3284$ (her commitment).

Then her witness is $x = 3284^2 \bmod 3737 = 3411$.

She sends her public key 1558 and her witness 3411 to Bob.

(ii) He receives these values and flips a coin to determine the challenge

e = 1. He sends 1 to Alice.

(iii) On receiving e = 1,

Alice calculates $y = 3284.113^1 = 1129 \pmod{3737}$.

She sends this back to Bob (the response).

(iv) He then verifies $z = y^2 \bmod 3737 = 1129^2 \bmod 3737 = 324$

and $z' = x.v^e \bmod 3737 = 3411.1558 \bmod 3737 = 324$.

Because these two values are the same, Bob accepts the first round of the protocol. Say z and z' agree over 8 rounds of the protocol. Bob estimates that the probability this person is **not** Alice is

$$\left(\frac{1}{2}\right)^8 \approx 0.0039 .$$

From Bob's perspective, the probability that it **is** Alice is 1 – 0.0039 = 0.9961.

7.4 Security of Fiat-Shamir Protocol

An adversary, Oscar, could try guessing Bob's challenge for each round of the protocol. So if Oscar can guess Bob's challenge correctly for each of t rounds of the protocol, he can fool Bob into believing that Oscar is in fact Alice. The chances of Oscar completing t rounds of the protocol successfully is

$$\frac{1}{2}.\frac{1}{2}.\frac{1}{2}.....\frac{1}{2}(t \text{ times}) = \left(\frac{1}{2}\right)^t .$$

With t = 20 rounds, the probability of Oscar succeeding in impersonating Alice is 0.000976563 – an extremely unlikely event! Now suppose that Oscar listens in on the protocol between Alice and Bob. Can he infer from it the value of s? At each round, he will only see v, e and y.

In round 1, in our example, because e = 1, Oscar will know that $r.s$ = 1129 (mod 3737) and r^2 = 3411 (mod 3737).

He could calculate the square root of 3411 mod 3737 by trail and error, find r and solve for s. With such a small value of n as in this example, this would pose little difficulty. But in real implementations of the protocol, n will be of the order of 200 decimal digits and then the square root problem becomes **intractable**. So the effectiveness of the protocol depends on the purported intractability of the square root problem modulo pq (= n).

7.5 Feige-Fiat-Shamir (FSS) Identification Protocol

The Fiat-Shamir protocol described above can be generalised and the FSS protocol is an example of one such generalisation.

In summary, Alice has secret numbers $s_1, s_2, ..., s_k$.

Let $v_i \equiv s_i^{-2} \pmod{n}$, where we assume gcd (s_i , n) = 1.

The numbers v_i are sent to Bob. Bob will try to verify that Alice knows the numbers

$$s_1, s_2, ..., s_k .$$

The protocol involves the following steps:

1. *One-time setup*:
   (i) A trusted authority – call it Tom – selects two large prime numbers *p* and *q*, congruent to 3 mod 4, calculates $n = pq$, publishes *n*, and keeps the primes secret. Tom also publishes the parameters *k*, the key size and *t*, the number of protocol iterations.
   (ii) Alice, a user of the centre, selects *k* secret random numbers $s_1, s_2, ..., s_k$ in the range $1 \le s_i \le n-1$ where gcd (s_i , n) = 1.
   This ensures that *n* cannot be factored easily.
   (ii) Alice computes $v_i \equiv s_i^{-2} \pmod{n}$ and registers her public key $(v_1, ..., v_k; n)$, while only Alice knows her private key ($s_1, s_2, ..., s_k$) and *n*.
2. *Protocol messages*
   Each of *t* rounds has 3 messages as follows:
   Alice $\rightarrow$ Bob : $x \equiv r^2 \pmod{n}$
   Bob $\rightarrow$ Alice : $(e_1, ..., e_k)$, $e_i \in \{0, 1\}$
   Alice $\rightarrow$ Bob : $y \equiv r.s_1^{e_1} s_2^{e_2} .....s_k^{e_k} \pmod{n}$
3. *Protocol actions*
   (i) Alice chooses a random number *r* (the commitment), computes
   $x \equiv r^2 \pmod{n}$ (the witness) and sends *x* to Bob.
   (ii) Bob chooses numbers $(e_1, ..., e_k)$, $e_i \in \{0, 1\}$. He sends these to Alice.
   (iii) Alice computes $y \equiv r.s_1^{e_1} s_2^{e_2} .....s_k^{e_k} \pmod{n}$ and sends *y* (the response) to Bob.
   (iv) Bob checks that $x \equiv y^2 v_1^{e_1} v_2^{e_2} .....v_k^{e_k} \pmod{n}$
   (v) Steps (i) through (iv) are repeated *t* times.

The following congruences show that Alice will be able to prove her identity to Bob:

$$x \equiv y^2 v_1^{e_1} v_2^{e_2} \ldots\ldots\ldots v_k^{e_k} \mod n$$
$$\equiv (r^2 s_1^{2e_1} \ldots\ldots\ldots s_k^{2e_k})(v_1^{e_1} \ldots\ldots\ldots v_k^{e_k}) \mod n$$
$$\equiv (r^2 s_1^{2e_1} \ldots\ldots\ldots s_k^{2e_k})(s_1^{-2e_1} \ldots\ldots\ldots s_k^{-2e_k}) \mod n$$
$$\equiv r^2 \mod n$$

So Bob will accept Alice's proof of identity.

7.6 Example to illustrate the FFS Protocol

1. *One-time setup*

(i) Tom (the trusted centre) selects the primes p = 683, q = 811 and publishes n = pq = 553913. Integers k = 3 and t = 1 are defined as security parameters.

(ii) Alice selects 3 random numbers s_1 = 157, s_2 = 43215, s_3 = 4646.

(iii) Alice computes $v_1 \equiv s_1^{-2} \pmod n$ = 441845, v_2 =338402 and v_3 = 124423.

Alice's public key is (441845, 338402, 124423; 553913) and her private key is (157,43215, 4646).

2. *Protocol actions*

(i) Alice chooses r = 1279 (the commitment), computes $x = r^2 \mod n$ = 25898 and sends x to Bob.

(ii) Bob sends to Alice (the challenge) the 3-bit vector (0, 0, 1).

(iii) Alice computes $y \equiv r.s_1^{e_1} s_2^{e_2} \ldots\ldots\ldots s_k^{e_k} = r.s_3^1 \mod n$ = 403104 and sends y (the response) to Bob.

(iv) (iv) Bob computes $z \equiv y^2 v_1^{e_1} v_2^{e_2} \ldots\ldots v_k^{e_k} \pmod n = y^2.v_3^1 \mod n$ = 25898 and accepts Alice's identity since $z = x$. [1]

7.7 Security of Feige-Fiat-Shamir (FFS) Protocol

The security of the FFS protocol relies on the difficulty of extracting square roots modulo n. This is equivalent to factoring n. The best attack, using a chosen message, has a probability 2^{-kt} of successful impersonation.

An adversary, Oscar, who doesn't know the numbers $s_1, s_2, ..., s_k$, could try and guess the string of bits $(e_1, ..., e_k)$ that Bob will send. He lets *y* be a random number and declares $x \equiv y^2 v_1^{e_1} v_2^{e_2} .....v_k^{e_k} \pmod{n}$. When Bob sends the string of bits, Oscar sends back the value of *y*.

For example, suppose Oscar can guess the correct response when $e_1 = 1$, e_2 = 1, e_4 = 1 and all other e_i = 0. However, suppose Bob sends $e_1 = 1$, e_3 = 1 and all other e_i = 0. Then Oscar, posing as Alice, will be ready to supply a square root of $xv_1v_2v_4$, but will be asked to supply a square root of xv_1v_3. This, combined with what he already knows, is equivalent to knowing a square root of $v_2^{-1}v_3v_4^{-1}$, which he is **not** able to compute.

In effect, there are 2^k possible strings of bits that Bob can send to fool Oscar. In one iteration of the protocol, the chances are only one in 2^k that Bob will be fooled. If the procedure is repeated *t* times, the chances are one in 2^{kt} that Oscar will be fooled.

Recommended values are *k* = 5 and *t* = 4, which gives the same probability as 20 iterations of the previous Fiat-Shamir scheme or 1 in a million chance of impersonation. So the FFS protocol is more efficient in terms of communication between Alice and Bob.

FFS is a pretty simple, and effective zero-knowledge proof. There is, however, an important security tradeoff that needs to be addressed. If you set *t* = 1, computation and communications can be reduced. Also, while holding *kt* constant, and incrementing *k*, while decrementing *t*, will result in the protocol no longer being able to hold the concept of *proof of knowledge*. This means that as *t* approaches 1, the protocol become less and less *sound*. [17]

7.8 FFS as an Identity-Based Scheme

Let *S* be a string that includes Alice's name, address and date of birth. Let *f* be a one-way function (a public **hash** function, for example). A trusted authority, Tom (could be a bank, for example), chooses *n* = *pq* as before and then computes Alice's public values

$$v_i = f(S, i),\ 1 \le i \le k.$$

Now, Tom, knowing the factorisation of *n*, computes a square root s_i for each v_i, and gives these to Alice. Tom can now discard $s_1, s_2, ..., s_k$ and the values of *p* and *q*. This adds to

the security of the scheme since someone who breaks into Tom's computer cannot compromise Alice's security.

Say Alice goes to an ATM for example. The ATM reads *S* from Alice's card. It downloads $(v_1,.....,v_k;n)$ from a database. The FFS protocol is then performed to verify that Alice knows $s_1, s_2,....., s_k$. After a few iterations, the ATM is convinced that the person is in fact Alice and allows her to withdraw cash.

To avoid a lot of typing on Alice's part, a better implementation would be to use chips embedded in the card and store the data in such a way that it cannot be extracted. [5]

7.9 Schnorr Identification Protocol

1. *Setup*

   A trusted authority – Tom – chooses the following parameters:

   (i) p is a large prime such that the discrete logarithm problem in Z_p^* is intractable.

   (ii) q is a large prime divisor of $p - 1$.

   (iii) $\alpha \in Z_p^*$ has order q, i.e. if β is a primitive element mod p, then $\alpha = \beta^{(p-1)/q} \bmod p$.

   (iv) A security parameter t such that $q > 2^t$. For most applications, t = 40 provides adequate security.

   (v) Tom also establishes a secure **signature scheme** with a secret signing algorithm sig_{Tom} and a public verification algorithm ver_{Tom}.

   (vi) A secure **hash function** is specified. The hash function is used to hash the message before it is signed.

2. *Issuing a certificate to Alice*

   (i) Tom establishes Alice's identity by means of conventional forms of identification such as birth certificate or passport. Then Tom forms a string ID (Alice) which contains her identity information.

   (ii) Alice secretly chooses a random exponent a, where $0 \le a \le q-1$. She then computes $v = \alpha^{-a} \bmod p$ and gives v to Tom.

   (iii) Tom generates a signature $s = sig_{Tom}(\mathrm{ID}(\mathrm{Alice}), v)$

   and gives the certificate $cert_{Alice}$ = (ID(Alice), v, s) to Alice.

3. *Protocol messages*

Alice $\rightarrow$ Bob : $cert_{Alice}$ and $x = \alpha^r \bmod p$

Bob $\rightarrow$ Alice: $e, 1 \le e \le 2^t$

Alice $\rightarrow$ Bob: $y = r + ae \bmod q$

4. *Protocol actions*

   (i) Alice chooses a random number r (the commitment), where $0 \le r \le q-1$. She then calculates (the witness) $x = \alpha^r \bmod p$ and sends her certificate $cert_{Alice}$ = (ID(Alice), v, s) and x to Bob.

   (ii) Bob verifies the signature of Tom by checking that ver_{Tom} (ID(Alice), v, s) = true

   (iii) Bob chooses a random number e (the challenge), $1 \le e \le 2^t$ and sends it to Alice.

   (iv) Alice computes $y = r + ae \bmod q$ and sends y to Bob.

   (v) Bob verifies that $z \equiv \alpha^y v^e \bmod p$ and accepts Alice's identity if $z = x$.

For discrete logarithms to be computationally infeasible (or intractable), we require that $p \ge 2^{1024}$ and $q \ge 2^{160}$.

There are two things happening in this identification protocol: Firstly, the signature s proves the validity of Alice's certificate. So Bob verifies the signature of Tom on Alice's certificate to convince himself that the certificate itself is authentic. Secondly, the value a functions like a PIN in that it convinces Bob that the person carrying out the protocol is indeed Alice.

The following congruences show that Alice will be able to prove her identity to Bob:

$$\begin{aligned} \alpha^y v^e &\equiv \alpha^{r+ae} v^e \bmod p \\ &\equiv \alpha^{r+ae} \alpha^{-ae} \bmod p \\ &\equiv \alpha^r \bmod p \\ &\equiv x \bmod p \end{aligned}$$

So Bob will accept Alice's proof of identity.

The Schnorr scheme is designed to be very fast and efficient, both from a computational viewpoint and the amount of information that needs to be exchanged in the protocol. It is also designed to minimise the amount of computation done by Alice. This makes it quite attractive in applications where Alice can use a smart card and where Bob needs to perform more

complex computations. [3] To illustrate the point above, assume that ID (Alice) is a 512-bit string, *v* is also 512 bits and *s* will be 320 bits if the Digital Signature Standard (DSS) is used as a signature scheme. The total size of the certificate which needs to be stored on Alice's card is then 1344 bits. Now we can calculate the number of bits that are communicated during the protocol. Recall the 3 steps of the protocol.

Alice $\rightarrow$ Bob : $cert_{Alice}$ and $x = \alpha^r \bmod p$

Bob $\rightarrow$ Alice: $e, 1 \leq e \leq 2^t$

Alice $\rightarrow$ Bob: $y = r + ae \bmod q$

Alice sends to Bob 1344 + 512 = 1856 bits of information in step 1.
Bob sends Alice 40 bits in step 2.
Alice sends 140 bits in step 3.
So the communication requirements are quite modest. [3]

The computations performed by Alice require the modular exponentiation $x = \alpha^r \bmod p$, which, although computationally intensive, can be performed offline. The computation of $y = r + ae \bmod q$ comprises one modular addition and one modular multiplication, which is not computationally intensive.

On the other hand, Bob's calculations are computationally intensive, since he has to verify Tom's signature on Alice's certificate and also verify that $z \equiv \alpha^y v^e \bmod p \equiv x$. A hash function is a one-way function that produces a message digest of the entire message. The message digest is combined with Alice's secret key to produce Alice's digital signature. [18]

7.10 Example of Schnorr Identification Protocol

1. *Setup*
   (i) Suppose p = 88667, q = 1031 and $p - 1$ is divisible by the prime q.
   The element α = 70322 has order q in Z_p^*, where $\alpha = \beta^{(p-1)/q} \bmod p$ and β is a primitive element of Z_p^*.
   (ii) Suppose Alice's secret exponent is a = 755 then
   $$v = \alpha^{-a} \bmod p$$
   $$= 70322^{1031-755} \bmod 88667$$
   $$= 13136$$
2. *Protocol actions*

(i) Now Alice chooses $r = 543$. She then computes:

$$x = \alpha^r \bmod p$$

$$= 70322^{543} \bmod 88667$$

$$= 84109.$$

She sends x to Bob.

(ii) Bob sends the challenge $e = 1000$ to Alice.

(iii) Now Alice computes $y = r + ae \bmod q$

$$=543 + 755.1000 \bmod 1031$$

$$= 851$$

and sends y to Bob.

(iv) Bob then verifies that

$$x \equiv z \equiv \alpha^y v^e \bmod p$$

i.e. $84109 \equiv 70322^{851}13136^{1000} \bmod 88667.$

and accepts Alice's identity if $z \equiv x$.

7.11 Security of Schnorr Protocol

While it is our hope that an adversary, Oscar, will not gain any information about a when Alice proves her identity (zero-knowledge property), the Schnorr identification protocol has not been proven **secure**. The protocol is not secure for large e, because through interaction, Bob obtains the solution (x, y, e) to the equation $\alpha^y v^e \equiv x \bmod p$, which Bob himself might not be able to compute.

But a modification to the Schnorr scheme was designed by Okamoto, which can be proven to be secure. The main difference between the two schemes is that instead of Tom choosing

$$\alpha \in Z_p^* \text{ of order } q$$

as in the Schnorr scheme, Tom instead chooses two elements

$$\alpha_1, \alpha_2 \in Z_p^*, \text{ both of order } q.$$

Tom keeps the value $c = \log_{\alpha_1} \alpha_2$ secret from all participants including Alice, which we assume is infeasible for any adversary to compute. [3]

Although an adversary, Oscar, could gain access to Alice's correct certificate (since the information on a certificate is revealed each time the identification is run), he will not be able to impersonate Alice unless he knows the value of a. Oscar would have to compute y for each round, but y is a function of a. The computation of a from v involves a discrete logarithm problem, which we assume is intractable.

7.12 Application of Schnorr Identification Scheme

Stefan Brand [18] has developed a system of **digital cash**, which uses the Schnorr identification protocol twice. Digital cash refers to electronic records or messages which serve as money and can be authenticated by the institution granting the digital cash. Essentially, it is a payment message bearing a digital signature that functions as a medium of exchange. [19]

When a customer (Alice) withdraws a "coin" from a bank, the bank binds the user's identity to the "coin", but sends along additional information that allows the customer to **blind** the signed "coin" as seen by the bank. Blinding is the process by which a bank cannot identify the person who withdrew the "coin". While this maintains the customer's identity, it poses the problem for the bank of identifying double-spenders.

Alice challenges the bank to prove knowledge of its secret key. This verifies that the bank has provided a valid signature on the "coin". When Alice spends the blinded "coin", the merchant challenges her to provide knowledge of her secret key. The merchant records the challenge and response and gives this to the bank as part of the "coin" deposit protocol. If the bank receives the same "coin" twice, the challenge and response will reveal Alice's identity if Alice was responsible for double-spending. [20]

7.13 Guillou-Quisquater (GQ) Identification Protocol

The GQ protocol is an extension to the Fiat-Shamir protocol. It allows a reduction in both the number of messages exchanged and memory requirements for user secrets. Furthermore, like the Fiat-Shamir scheme, it is suitable for applications in which the claimant has limited power and memory.

Alice proves her identity to Bob in a 3-pass protocol.

1. *Setup*
   (i) A trusted authority – call it Tom – selects two large prime numbers p and q, and calculates $n = pq$.
   (ii) Tom defines a large prime integer b that functions as a security parameter.
   (iii) Alice secretly chooses an integer u, where $0 \le u \le n-1$. Alice computes $v = u^{-b} \bmod n$ and gives v to Tom.

   The trusted authority, Tom, establishes Alice's identity (as in the Schnorr scheme) and issues the identification string ID (Alice).

The certificate $cert_{Alice}$ = (ID(Alice), *v*, *s*) is given to Alice where $s = sig_{Tom}$(ID(Alice),*v*).

2. *Protocol messages:*

The protocol involves 3 messages:

Alice $\rightarrow$ Bob : $cert_{Alice}$ and $x = r^b \bmod n$

Bob $\rightarrow$ Alice:*e*, $1 \le e \le b-1$

Alice $\rightarrow$ Bob: $y = ru^e \bmod n$

3. *Protocol actions:*

(i) Alice chooses a random *r* (the commitment), $0 \le r \le n-1$ and computes

(ii) $x = r^b \bmod n$ and sends her certificate $cert_{Alice}$ = (ID(Alice), *v*, *s*) and *x* to Bob.

(iii) Bob verifies the signature of Tom by checking that

ver_{Tom} (ID(Alice), *v*, *s*) = true

(iv) Bob chooses a random number *e*, $1 \le e \le b-1$ and sends it to Alice.

(v) Alice computes $y = r\,u^e \bmod n$ and sends it to Bob.

(vi) Bob verifies that $x \equiv v^e y^b \bmod n$.

The following congruences show that Alice will be able to prove her identity to Bob:

$$\begin{aligned} v^e y^b &\equiv u^{-be}(ru^e)^b \bmod n \\ &\equiv u^{-be} r^b u^{eb} \bmod n \\ &\equiv r^b \bmod n \\ &\equiv x \bmod n \end{aligned}$$

So Bob will accept Alice's proof of identity.

7.14 Example of Guillou-Quisquater (GQ) Identification Protocol

1. *Setup*

(i) The trusted authority, Tom, selects primes *p* = 467, *q* = 479 and computes

(ii) *n* = *pq* = 223693.

(iii) Suppose also that *b* = 503 and Alice's secret integer *u* = 101576.

Alice computes

$v = u^{-b} \bmod n$

$= 101576^{-503} \mod 223693$

$= 89888$

and gives this value to Tom.

2. *Protocol actions*

(i) Alice selects $r = 187485$ and computes

$x = r^b \mod n$

$= 187485^{503} \mod 223693$

$= 24412.$

(ii) Alice now sends $cert_{Alice}$ and $x = 24412$ to Bob.

(iii) Bob verifies the signature of Tom on Alice's certificate by checking that ver_{Tom} (ID(Alice), v, s) = true.

(iv) Now Bob sends a random challenge $e = 375$.

(vii) Alice replies with $y = r.u^e \mod n$

$= 187485.101576^{375} \mod 223693$

$= 93725$

and sends it to Bob.

(viii) Bob then verifies that

$x \equiv v^e y^b \mod n$

i.e. $24412 \equiv 89888^{375}93725^{503} \mod 223693.$

Hence Bob accepts Alice's proof of identity.

7.15 Security of Guillou-Quisquater Protocol

Extracting b^{th} roots modulo the composite integer n is necessary to defeat the protocol; this is no harder than factoring n, which we already know to be computationally intractable.

8. Comparison of Fiat-Shamir, Schnorr and Guillou-Quisquater Protocols

Each of these protocols provides solutions to the identification problem. Each has relative advantages and disadvantages with respect to various performance criteria and for specific applications.Each protocol can be compared under the following criteria:

1. *Computational efficiency*

Fiat-Shamir requires from 11 to about 30 modular multiplications or steps by the claimant (Alice) with $kt = 20$ and n is 512 bits while GQ requires about 60 steps.

2. Offline computations

The Schnorr scheme has the advantage of requiring only a single online modular multiplication by the claimant. This assumes (as outlined earlier) that the exponentiation is done beforehand. However, significant computations is required by the verifier (Bob) compared to the Fiat-Shamir or GQ scheme.

3. Security assumptions

All the protocols require the assumptions that the following problems are intractable:
For a composite integer *n*:

Fiat-Shamir – extracting square roots mod *n*

Schnorr – computing discrete logs mod a prime number *p*.

Guillou-Quisquater – extracting b^{th} roots mod *n*

9. Zero-Knowledge Identification from a Geometric Perspective

In their paper, "Identification by Angle Trisection", Burmester *et al*, [13], describe an elegant zero-knowledge scheme based on the impossibility of trisecting an angle using a ruler and compass only.

1. Setup

Alice publishes a copy of an angle Y_A, which is constructed by Alice as the triple of an angle X_A, that she has constructed at random. Because trisection of an angle is impossible using a ruler and compass only, Alice is confident that she is the only one who knows X_A.

2. Protocol actions

It follows the standard form of an iterated 3-round protocol:

(ii) Alice gives Bob a copy of an angle *R*, which she has constructed as the triple of an angle *K* that she has selected at random.

(iii) Bob flips a coin, and tells Alice the result.

(iv) If Bob says "heads", Alice gives Bob a copy of the angle *K* and Bob checks that $3 * K = R$. If Bob says "tails", Alice gives Bob a copy of the angle

$$L = K + X_A,$$

and Bob checks that

$$3 * L = R + Y_A.$$

The three steps are repeated *t*-times independently. Bob accepts Alice's proof of identity only if all *t* checks are successful.

10. Converting Identification to Signature Schemes

An identification scheme involving a witness-challenge-response sequence can be converted to a signature scheme as follows:

Replace the random challenge *e* of the claimant by the one-way public hash function *h* of the witness *x* and the message *m* to be signed. This, in effect, converts an interactive identification scheme to a non-interactive signature scheme. The challenge *e* must typically be increased to avoid off-line attacks on the hash function.

11. Conclusion

In this paper we have discussed the need for identification protocols that ensure that two parties prove their identity to each other before the actual transfer of information takes place between them. We focussed our attention upon identification schemes that are based on challenge-response protocols. In particular, we examined in detail those protocols that are based upon zero-knowledge proofs where the claimant can prove their identity to another entity in real-time without revealing any meaningful information other than the claim of being that particular entity.

We touched briefly on modern day applications of these identification protocols. As the Internet grows and becomes an essential part of our lives, e-commerce has grown and with it the need for entities to identify themselves by revealing more and more sensitive information about themselves. The Internet provides a vast array of ways in which people's privacy can be and is being intruded upon, and adds new dimensions to existing problems. It necessitates the negotiation of a whole new set of balances among the various interests. Identification protocols that can prove an entity's identity without the need for that entity to reveal any information about itself are to be welcomed. As a result, we anticipate that identification protocols will grow in importance as we seek new ways to negotiate the conflicting needs of privacy on the one hand and proving our identity on the other.

12. References

1. Menezes, P, Van Oorschot, P and Vanstone, S (1997), *Handbook of Applied Cryptography*, CRC Press Inc
2. Stallings, William (1998), *Cryptography and Network Security*, Prentice Hall
3. Stinson, Douglas R, (1995), *Cryptography – Theory and Practice*, CRC Press Inc.
4. Barr, Thomas, (2002), *Invitation to Cryptography*, Prentice Hall
4. Trappe, W and Washington, L, (2001), *Introduction to Cryptography with Coding Theory*,Prentice Hall
6. http://www.crypto.nkfust.edu.tw/infosec/faq/html/14.html

7. http://www.anu.edu.au/people/Roger.Clarke/EC/AuthModel.html011019.html
8. http://www.iks-jena.de/mitarb/lutz/security/cryptfaq/q107.html
9. Quisquater, J and Guillou, L., *How to explain zero-knowledge protocols to your children*, Advances in Cryptology – CRYPTO '89
10. Lamport, L, (1981), *Password Authentication with Insecure Communication*, *Communications of the ACM*, 24 (11)
11. http://www.acm.org/sigcomm/ccr/archive/2000/july00/nyang.pdf
12. http://www.anu.edu.au/people/Roger.Clarke/EC/AuthModel.html
13. Burmester, M, Rivest, R and Shamir, A, (1997), *Geometric Cryptography: Identification by Angle Trisection*, Department of Computer Science, Israel
14. http://www.win2000mag.com/Articles/Index.cfm?ArticleID=526
15. http://developer.novell.com
16. http://www.commweb.com/article/COM20010328S0010/1
17. Grabbe, J, *Cryptography and Number Theory for Digital Cash*, http://www.aci.net/kalliste/cryptnum.htm
18. Brands, S, *Untraceable Offline Cash in Wallets with Observers*, Advances in Cryptography – Crypto '93
19. Grabbe, J, *The Mathematical Ideas Behind Digital Cash*, http://www.aci.net/kalliste/cryptnum.htm
20. Miller, J, *Digital Cash Mini-FAQ*, http://ntrg.cs.tcd.ie/mepeirce/project/mlists/minifaq.html
21. Davies, D.W, and Price, W.L., 1992, *Security for Computer Networks*, Wiley and Son
22. http://www.anu.edu.au/people/Roger.Clarke/DV/UIPP99EA.html
23. Gallian, J, (1994), **Contemporary Abstract Algebra**, Heath
24. Fraleigh, J, (1994), **A First Course in Abstract Algebra**, Addison Wesley
25. Walace, D, (1998), **Groups, Rings and Fields**, Springer
26. Schneier, B, (1996), **Applied Cryptography**, Second Edition, John Wiley & Sons

The Effects Of Mobile Computing on Teleworking.

Roisin Faherty,

Department Of Computing,

Institute of Technology Tallaght.

Abstract

This paper examines Mobile Computing Technology with a particular focus on the effect Mobile Computing is having on teleworking. Mobile computing as it is defined for this paper is described. The enablers of this technology as well as the inhibitors to this technology are discussed. Future possible trends in the area of mobile computing are also explored. Teleworking is reviewed in terms of the advantages and disadvantages it offers to the organization. Also outlined is the use Information Technology (IT) in teleworking. This examination of teleworking leads the paper into the next step up from teleworking i.e. The virtual organization. The issues around this type of structure are outlined including, Strategic Change Issues, Virtual teams, Integration of Virtual teams, Trust Issues and Cultural Issues. The paper then examines the implications involved in managing a virtual organization. Having explored the issues and possible complications surrounding a virtual organization it is important to highlight any strategic advantage that a move to this type of organizational set up would give and organization.

1. Introduction

The aim of this paper is to give the reader an overview of mobile computing technology and to highlight the effect this technology is having on teleworking. The research for this paper was conducted by referencing a number of acknowledged text books, academic journal articles and web publications. This paper is aimed at those interested in the evolutionary changes that are becoming apparent in organizations today. Mobile computing technology is providing the basis for this revolutionary fundamental change in the way organizations are structured. However even though the technological infrastructure is in place there are a number of issues facing an organization that wants to use these technological advancements to their full potential. If these issues are not addressed by the organization the realization of a purely virtual organization may not be possible.

The paper is divided into four main sections. Section 2 considers the technological side to mobile computing, including the enablers and the inhibitors of this technology. There is also

an overview of possible future trends in this area. Second 3 discusses the effects of mobile computing on the organization. Teleworking issues, the virtual organization, management of a virtual organization and possible strategic advantages of the virtual organization are considered here. Section 4 is the conclusion of the paper.

2. Mobile computing

Mobile computing also known as Wireless Computing has begun moving out of its gestation period (Gartner Group, 1998). Today's workforce is becoming increasingly mobile and companies are keeping remote users up to date and in touch by providing notebook computers and remote access capabilities to their employees. Mobile computing seems to be the latest 'Buzz Word' in the technology industry. It can have different meanings for different people. The next section explains the term mobile computing, as it is understood for this paper.

2.1 What is mobile computing

There are many diverse definitions of what exactly mobile computing is, some example include:
"Mobile Wireless – the use of wireless devices or systems on board motorized moving vechicles, e.g. Mobile phone and Personal *Communications devices (PCD's)" (www.whatis.com)"*

This limits the term mobile to use only in relation to moving vehicles, today mobile computing has a much broader meaning. Another definition refers to mobile computing as telecommunication in which electromagnetic waves, rather than some form of wire, carry the signal over part or all of the communications path (www.whatis.com). Therefore mobile computing would be the ability to connect computers or computing devices using electronic waves rather then wires. However for the purposes of this paper mobile computing will be defined as *"Any computing device which facilitates work outside the organization."*Today the term mobile is used to refer to the many wireless technologies available, these include:

- Cellular Phones and Pagers
- Global Positioning Systems
- Cordless PC Peripherals e.g. cordless mouse
- Home entertainment-systems controls e.g. television remote control.
- Wireless Applications Protocol (WAP)
- General Packet Radio Service (GPRS)
- Universal Mobile Telecommunications System (UMTS)
- Video Conferencing

- E-mail
- Voice-mail.

These are but a few of the systems to arise out of the mobile revolution. Some of which are not particularly applicable to this paper. Mobile technologies can be divided into several forms including direct-dial and virtual private networks (VPN). Direct-dial allows users to dial in or out from a corporate LAN, or to establish LAN-to-LAN connections using traditional analogue or high speed ISDN lines (Saab, 1999). A secure VPN is a network tunnel created for encrypted data transmission between two or more authenticated parties over a shared data network. Through a VPN employees can simply dial into a corporate network or into any ISP and have direct access to the corporations data and decision-making tools regardless of their location (Saab, 1999). Using VPN technology, intranets can be created to keep off-site sales forces informed of new products, product enhancements and new price structures. Extranets implemented using VPN technology can also establish secure communication channels between corporations and their business partners, permitting proprietary information to be shared in confidential secure environments. These two complimentary technologies, direct-dial and VPN, can help maximize mobile computing while minimizing underlying costs (Saab, 1999).

2.2 Enablers of Mobile Computing

"Major driving and enabling technologies are now in place to ensure that mobile data becomes pervasive over the next five years: Internet, intranets, successful Personal Digital Assistants (PDA's), browser enabled phones and finally higher speed mobile WAN services" (WAN Wireless, Gartner Group, 1998).

Mobile computing is one of the fastest growing segments of the communications industry (Krichevsky, I., 1999). The international data corporation expects there will be one billion wireless phones by 2003. Applications for mobile computing are also growing at rapid speeds. In August and November 1999 there were 14 significant news announcements related to mobile computing (Krichevsky, I., 1999). These announcements ranged from wireless access to the Internet to the recently announced "Wireless MD", which claims to be the first two-way wireless communications product for physicians. In terms of the organization some of the key motivators for investment in mobile computing, and particularly remote access, are increased productivity, increased accuracy and improved customer service and satisfaction (Gartner Group, 1998).

According to the Gartner Group, major developments in networking, including the advent of universal wireless LAN and WAN access as well as further evolution in mobile, home-based and wearable devices are triggering the evolutionary change into the mobile era.

Trends in mobile computing e.g. Web-enabled GSM phones with larger screens, integrated web browsers, e-mail agents and personal information management applications and the growth of GSM-enabled PDA and portable PC's, has resulted in individuals and groups within the enterprise, seeking to implement mobile data connectivity initiatives in the absence of any corporate program. This means that the corporation will be obliged to support multiple applications and gateways, so even if the corporate decision to move into mobile computing is not taken it seems that the push from individuals and groups within the organization will be so strong the corporation is going to have to move in this direction (Gartner Group, 1998). Another enabler of mobile computing has been in introduction of broadband wireless services called local multipoint services (LMDS). These give new competitors an opportunity to enable infrastructure flexibility and time to provision of services. Organizations can also use LMDS to eliminate common single point of network access failures by incorporating this technology into a diverse access route strategy.

As can been seen from this section there are a number of technologies driving the growth in mobile computing, however there are some areas that would be considered inhibitors to the successful continued growth of mobile computing. The next section considers these.

2.3. Inhibitors to Mobile Computing

The inhibitors to this technology that are considered in this section are:

- Cost
- Networks
- Standards
- Complexity
- Lack of Talent

2.3.1 Cost

"Through 2003, organizations that fail to proactively manage mobile cost reduction will overpay as much as 40%" *(Gartner Group, 1998).*

With an new and emerging technology one of the major inhibitors tends to be the cost and mobile computing is no exception. Most ongoing costs for mobile projects dwarf any initial investments, making many endeavors more costly then expected (Gartner Group, 1998). According to the Gartner Group, (1998), as European organizations add up the costs of

mobile services many are surprised to see that it represents as much as 50% of their overall communications costs. The reason cited for this is a lack of competitive bidding, devolved procurement and a dearth of pan-European suppliers. However the Gartner Group does give some hope to the organizations suggesting the use of other WAN networking solutions in an attempt to control the spiraling costs of mobile communications. The other possibility is to bypass the middleman. Few organizations are aware of how much the fixed line operators charge for delivering a call to a mobile. Organizations can eliminate terrestrial network surcharges by simply linking office PBX systems with the nearest mobile network point of presence (Gartner Group, 1998).

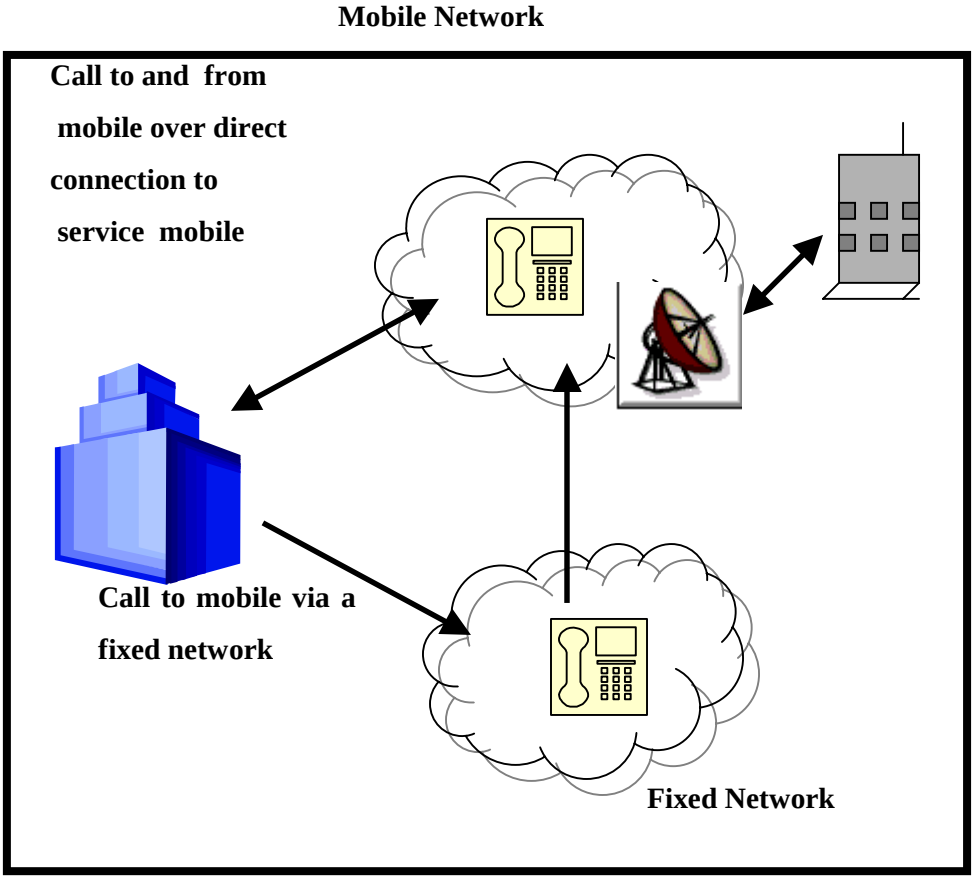


Figure 1: Gartner Group, WAN Wireless, 1998.

Few organisations are aware of how much the fixed line operators change for delivering a call to a mobile phone. This figure 1 represents the calls from an office to a mobile phone. In the upper bubble organizations could save money by connecting to mobile over a service mobile switch. This eliminates the middleman thus saving money. The lower bubble represents the fixed line call to a mobile, incurring the costs of the fixed line operators.

In order to have a complete perspective on cost it is important to include setup and maintenance costs, for organizations wishing to implement mobile computing, which can be

very high. According to Cascio (2000), for individual employees the cost to equip a mobile office varies from $3,000 to $5,000 plus another $1,00 every year in upgrades. In addition to this a virtual office requires online materials and database products. Also needed are well-indexed, automated, central files that are accessible from remote locations and a way of tracking any other mobile workers.The final cost consideration with relation to mobile computing and virtual organizations is the effect of the loss of efficiencies. When expensive equipment is concentrated in one location, multiple users can access it. However if the same equipment needs to be distributed across locations cost efficiencies may be lost (Cascio, 2000).

2.3.2 Networks

At present there is a lack of a fast reliable and affordable network infrastructure to support mobile computing (Dhawan, C., 1997). On the road, most wireless networks do not yet have the ubiquitous coverage that many mobile applications require. As for reliability the wireless networks are more prone to errors then fixed networks, with temporary disconnects still quiet common (Dhawan, C., 1997).

2.3.3 Standards

There is a distinct absence of standards in the mobile computing industry. As a result most solutions are proprietary based and therefore not interoperable. While some standards such as the IEEE 802.11 for wireless LANS, TCP/IP for transport and CDPD for the network are emerging, here is still a lack of more pervasive standards (Dhawan, C., 1997).

2.3.4 Complexity

The move to a mobile computing environment is very complex. Systems integration of a large mobile computing project for enterprise-wide deployment is a non-trivial task. The number of component involved, application modifications required, end-to-end integration and the emerging nature of the technology all contribute to its complexity (Dhawan, C., Computer Dealer News,1997).

2.3.5 Lack of Talent

According to the Gartner Group, organizations are facing a similar challenge to that which was seen in the mid- 1990's around web development. Just as with web-enabled applications, organizations that fail to proactively grow the requisite programming talent will be forced to delay or outsource projects. Today, most organizations are turning to outside resources, such as application service providers, for their wireless applications development talents. While this strategy is sufficient for the present organizations need to realize that mobile

technologies will become so embedded in their computing infrastructure that wireless access will become a common feature needed to be supported. The long term need to build customized business applications, as well as a need to link mobile applications to existing systems will force many organizations to assemble their own wireless application teams.

The next section considers the trends in the area of mobile computing. Will the inhibitors to this technology stunt its growth? Or are the enablers such that continued growth will be sustained?

2.4 Trends

Not since the Mainframe era has technology reached a level of relative stability before other discontinuous architectures entered the IT environment (Gartner Group, 1998).

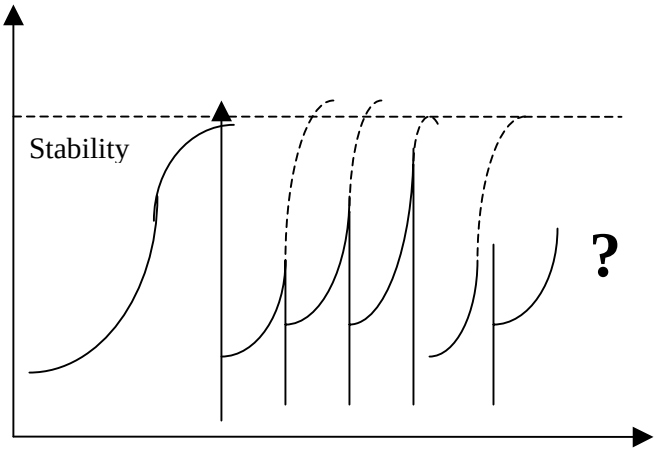


Figure 2: Gartner Group , 1998.

As can be seen from figure 2 above, very early in the life cycle of each new emerging technology over the past 10-15 years, a newer technology has superceded the previous one. With PC's, LAN's , Client/Server and now network computing, IT planners have barely had time to come to terms with new management, security support and total cost of ownership before the new paradigm arrives.

According to the Gartner Group a key element in predicting future trends in the IT industry, is to identify the driving forces of the present that will influence the future. Some of these driving forces are predetermined e.g. the number of computer science graduates entering the market in 2000 will depend largely on the number of students enrolled on these courses. Global Business Network, one of the pioneer organizations involved in planning, characterizes a subset of the forces as "Critical uncertainties", that are the key to the focal

issue. This organization represents these critical uncertainties as axes on a matrix, e.g. determining the future of the office environment in 2008 (figure 3) the two axes would be :

1. Whether the dominant equipment will still be general purpose PC's with massive memories and storage V's targeted, task specific computers,
2. Whether the devices will be wireless on dependant on a wired network solution.

Looking at the four corners created by combining these two axes leads to radically different scenarios each posing a different possible future trend.

a) **Deskbound**: If the trend stays with general purpose PC's on the wired side, then employees will be deskbound.
b) **Device Proliferation**: If the trend was towards task-specific devices on the wired side, then there would have to be rapid growth or change in the devices being used.
c) **Wearables**: If the trend were to be general purpose PC's on the wireless side, devices could be reduced in size a carried around i.e. they would be mobile, thus employees could work from anywhere, home, office, traveling.
d) **Ubiquitous Computing**: If the trend was to move both towards task-specific devices and a wireless world we would certainly be looking at computing everywhere and at once. Devices would be mobile and specific to the task required, e.g. a salespersons device would differ from a H.R. managers device, the only similarity being that all are mobile.

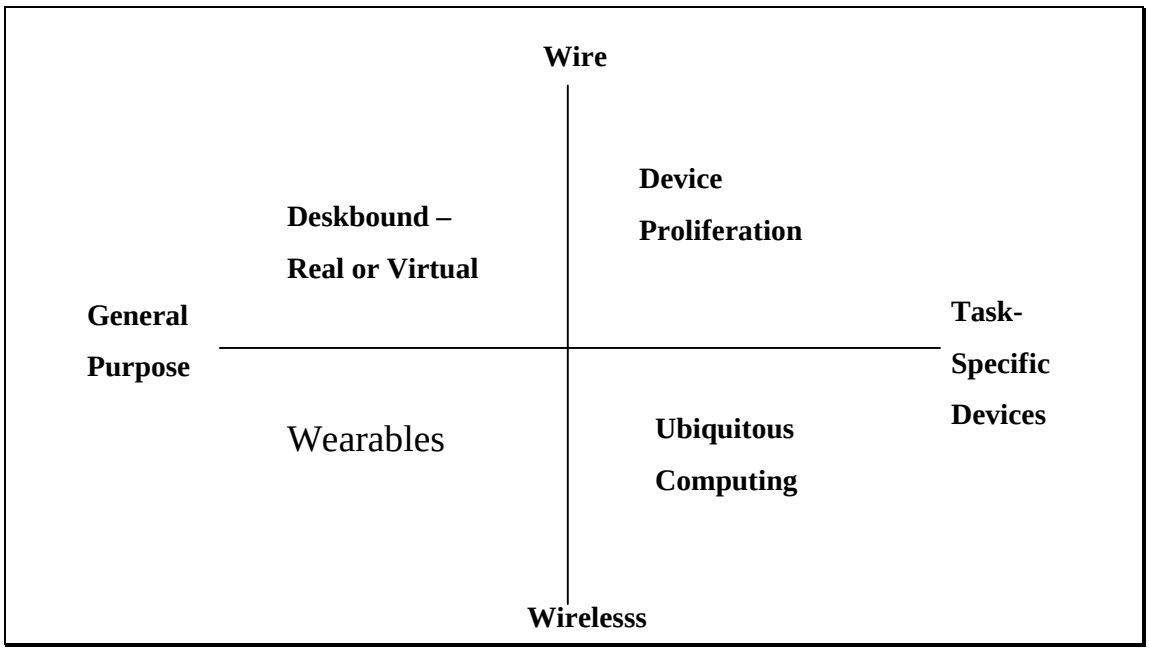


Figure 3: Gartner Group, 1998.

With regard to mobile computing the indicators seem to sow a distinct move in this direction, the decision facing organization seems to be more a question of when is best to invest? If an

organization launches into a new technology too soon, it runs the risk of suffering the painful and expensive lessons associates with new technologies. It also runs the risk of a newer technology superceding it as in Figure 3. On the other hand if the organization delays too long it runs an even greater risk, that of being left behind by its competitors.

The possible trends in this area are impossible to predict conclusively at this stage however the following have a 0.8 probability of happening according to the Gartner Group:

- The adoption of wireless data services will be driven by individual and organizational messaging and personal information management through 2002.
- Third generation wireless systems will offer higher data rates and increased functionality by 2002 in Europe.
- Through 2003, satellite systems will play a niche role in providing access in regions where terrestrial alternatives are limited or where rapid provisioning is higher priority then cost.
- Through 2003, the cost of rapid growth in enterprise use of mobile communications will not be offset by tariff reductions alone.
- Organizations that do not put in place strong procedures, policies and guidelines will loose control of mobile costs before 2003.

Having discussed what mobile computing is and the future trends in this area, the paper will now move on to look at the effects that this technology is having on the organization.

3. Effects of mobile computing on teleworking.

Study after Study asserts that mobile computing and telecommuting are going to be key factors in the next phase of the information revolution (Bartlett, J., 1997). While Mobile computing is becoming more and more widespread it is necessary to understand the effect this revolution is going to have on organizations. Will Mobile computing leave us with purely virtual organizations? The effects this new technology is having on organizations can already be seen by the increase in teleworking. Teleworking is an evolutionary step in the integration of mobile computing into the organization. The next section deals with teleworking within organizations.

3.1 Teleworking

Teleworking is an evolutionary step in the move towards the virtual organization (Huws *et al.*, 1993). Mobile computing has enabled teleworking (Huws *et al.*, 1993). Teleworking is when an employee carries out their work in a location remote form the central offices or

production facilities. The worker would have little or no personal contact with other employees, but is able to communicate with them via electronic means. Many of the issues faced with teleworking are also those faced by the virtual organization. There was a 60 percent increase in telecommuting between 1996 and 1998, to almost 16 million workers spending at least one day per week outside the traditional office environment, according to the International Telework Association and Council. The Gartner Group has estimated that this will grow to over 30 million by the year 2000.

Teleworking will now be considered under the following headings:

- Why teleworking was introduced
- Advantages of teleworking
- Disadvantages of teleworking
- Use of IT in teleworking

3.1.1 Why teleworking was introduced

In order for teleworking to be successfully introduced to an organization there needed to be an involvement from two groups of people:

- The Employer
- The Employee

Teleworking can only take place if there is sufficient convergence of interest for both parties to agree to it (Huws *et al.*, 1993).

The Employer – There is a wide range of reasons organizations might have for considering the introduction of teleworking, many of which are outlined in section 3.1.2. Advantages of Teleworking. Gil Gordon, a telework consultant who has done research on this area in the United States, outlines eight common reasons for the introduction of teleworking:

- Improved Recruitment.
- Improved retention of Staff.
- Curiosity to experiment.
- Space Savings.
- Hiring a disabled employee.
- Increased productivity.
- Response to employee demand.
- Need to improve customer services.

However the reasons for adoption of teleworking can be better analyzed based on different work sectors. For example, the software industry adopted it mainly as a way of retaining staff, whereas some companies decentralized typing and word processing task in order to minimize cost including expense of office space. And lastly consultancies companies introduced teleworking as a way to meet customers and business needs in the organization.

Having looked at management reasons for adopting teleworking, it is necessary now to look at the reasons why teleworkers themselves have chosen to work in this way. The main reasons cited by employees for wanting to telework are:

- Autonomy of work
- Balance between work and home life
- Flexibility of working hours
- Reduction in commuting

The benefits of these reasons given by employees are further outlined in section 3.1.2 Advantages of Teleworking. Teleworking must be agreed between employer and employee, in most cases the type of job being undertaken as well as the person involved are considered. The job needs to be one that can be conducted without major face-to-face communication. The individual involved needs to be self-motivated and disciplined in order to meet the work requirements from home. Having discussed how teleworking came about it is necessary to discuss the advantages teleworking offers to both the employers and the employees.

3.1.2 Advantages of Teleworking

Some of the reasons why firms might want to shift some of their workload to a virtual office environment include:

- For those in a rapid-growth mode, physical office space may be a problem, and remote working could provide an inexpensive alternative to leasing additional real estate (Faulkner & Gray, Inc., Accounting Today, 1999).
- The dwindling and ever more expensive pool of skilled professionals makes competition for the best talent tougher than ever. Offering creative alternatives to traditional work environments may help to recruit or retain those who would like to work part-time, on flexible schedules, or those with special physical or family needs.
- Finally, several studies have shown that workers can be both happier and more productive when freed from long commutes, parking expenses and formal office routines. (See Kavanagh (2000), example below).

The technologies that make this possible are available today, and are evolving and improving rapidly. These improvements in technology mean that teleworking is a very viable solution. The example cited by Kavanagh (2000) is based on BT, they have 3500 people teleworking. So far the policy has lead to a 20% growth in productivity and can increase morale for those employees finding a new balance between work and home life. BT say that peoples communication skills have also increased.

Even though teleworking offers many viable advantages to the organization it also has some disadvantages that need to be considered.

3.1.3 Disadvantages of Teleworking

As increased numbers of employees work from home or at distant peripheral offices, they can quickly and easily become disconnected from the central office (Cascio, 2000). They may begin to feel demotivated and loose self-discipline. A good tool for making teleworking more effective is the Internet. It can provide the forum for a chat, a means of communication and the technology to manage projects (Gurton, 2000).

Despite all the proposed benefits and advantages of teleworking there is now some literature to support the fact that teleworking may not be working! Despite these and other trends, the telework initiative isn't taking hold (Dvorak, Computer Shopper, 2001). This situation was recently outlined in a Wall Street Journal report, offering example after example of people specifically looking for telework and finding nothing, even at companies literally begging for employees. Those who did find telework ended up being isolated from the rest of the company, excluded from meetings, and viewed as expendable (Dvorak, Computer Shopper, 2001). Ironically, it's the newer, younger executives who are killing the telework movement. This trend will continue until they realize the old model won't work in the new century. It's too inefficient and old- fashioned (Dvorak, 2001). Two young entrepreneurs flat out said they will hire no telecommuters. Their rationale was simple-and shallow: People need to be in the office so they can be part of a team (Dvorak, 2001). The Internet Age is moving so fast that the need for teamwork on the spur of the moment is critical, they say. With cell phones, virtual private networking, e-mail, pagers, and desktop teleconferencing, you'd think virtual gatherings would be easy (Dvorak, 2001).

Teleworking relies on IT in order to be successful therefore the next section looks at the usage of IT in teleworking.

3.1.4 Use of I.T in Teleworking

The best way to outline the usage of I.T for teleworking is to use a table. The following is a tables are taken from Huws *et al.* (1993).

TABLE 1 – Teleworker's use of electronic hard and software

Proportion of working time spent using the equipment					
EQUIPMENT	**NONE**	**VERY LOW**	**<= One Third**	**Half**	**>= One Third**
Terminal	63.0	6.7	8.4	3.4	18.4
PC	45.4	19.3	15.1	6.7	13.4
Elec. Typewriter	95.0	4.2		0.8	
Modem	52.9	21.0	7.6	2.5	15.9
Wp S/W	54.6	23.5	8.4	8.4	5.0
Vidoetex	95.8	3.4	0.8		
Other	74.8	6.7	5.0	5.0	8.4

From this table it can be seen that IT equipment is used by teleworkers both for the performance of their work and for communication with other workers, managers and clients. Terminals are used by under a third of the teleworkers in the sample taken by Huws *et al.* (1993), for the majority in conjunction with a modem for connection to an external mainframe. Despite widespread availability of electronic typewriters reported by management, these were only being used by 5% of teleworkers in the sample. Word Processors were being used extensively by under 15% of the sample.

Communication is particularly important to teleworkers therefore Huws *et al.* (1993), asked not only about IT based methods but also more traditional communication methods:

As can be seen form the table Teleworkers still rely a lot on traditional form of communication e.g. the phone, and postal services, however e-mail and modems are also a well used means of communication for Teleworkers.

With the continued improvements in technology teleworking can now be taken a step further to the virtual organization. The next section looks at the possibility of a virtual organization and the issues involved in setting up a virtual organization.

TABLE 2 – Teleworkers use of Communications media

Equipment	No Use	Usage per Week <1 (Usage per Week)	1,2 (Usage per Week)	3-6 (Usage per Week)
Modem	61.3	5.9	5.9	9.2
Teletex	98.3	1.7		
Videotex	99.2	0.8		
E-mail	84.0	2.5	4.2	5.0
Disk	70.6	11.8	7.6	2.5
Texts	19.3	10.9	31.9	5.9
Tapes	88.2	2.5	1.7	1.7
Post	29.4	6.7	26.1	10.9
Courier	68.9	21.0	4.2	1.7
Meetings	25.2	45.4	19.3	2.5
Other(phone)	46.2	2.5	10.1	7.6

3.2 The Virtual Organization

The definition of a virtual organization is *"A group of skilled people who form a company but are geographically separated and communicate mostly electronically."* (Chutchian-Ferranti, J., 1999).

The diagram following in figure 3 is a representation of the virtual organization as the author sees it. The traditional organization is represented inside the office structure with each department having it's own area within the building. The virtual organization is different to this in that there is no fixed central location for the organization. Each department can be geographically dispersed around anywhere in the world. The virtual organization does not need to have all it's sales, H.R, Marketing or other functions in one area, the sales department can be made up of different people from anywhere in the world. The key difference between a virtual organsiation and a global organsiation is that it is possible for the virtual organization to have no central offices at all, whereas a global organization, while being distributed around the world, would have central offices in each location.

The key to making a virtual organization work is communication and effective technology, in a virtual organization the technology architecture is the organizational structure (Chutchian-Ferranti, J., 1999). The main advantage of a virtual organization is there are no geographical restraints. From a hiring perspective it's a major advantage. With today's difficulties in getting staff you have a wider range to choose from, not having to worry about relocation issues.

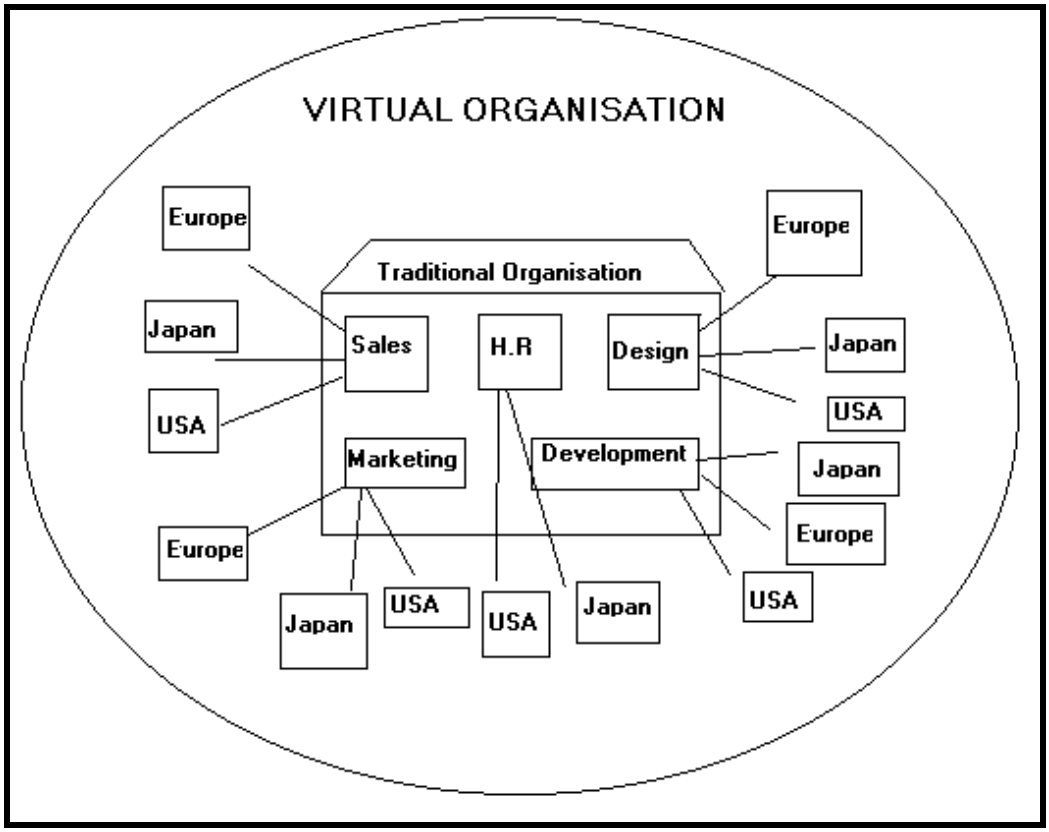


Figure 3 - Roisin Agnew January 2001

Having researched many publications, (Chutchian-Ferranti, J., 1999, Handy, C., 1995, Hsiao&Omerod, 1998, Jackson, P. J., 1999, Lipnack & Stamps, 1997, Warren, L., 2000), on the subject I have come up with the following five main issues surrounding the implementation of a purely virtual organization:

- Strategic Change Issues
- Virtual Teams
- Integration of Virtual Teams
- Trust Issues
- Cultural Issues

The following sub-sections examine these issues beginning with the strategic change issues.

3.1.1 Strategic Change Issues

A wealth of information can now be found dealing with "IT enabled" strategic change. Mobile computing is causing strategic change within organizations (Jackson, P., 1999). Hsiao and Omerod (1998) discuss different change archetypes related to IT-enabled strategic change, the one of interest to this paper is called "IT transformation". This describes the situation where IT is used to support innovation in processes, job roles and organizational culture, resulting in the overhaul of business structures. Mobile computing is effecting these

changes in organizations and these changes require understanding and analysis within a strategic business and IT context (Jackson, P, 1999). In order to ensure this, consideration needs to be given to the business model an enterprise must adopt in order to respond to the competitive conditions of its market place. In understanding and managing virtual teams, it is important to know what type of team we are concerned with and what role it plays alongside other elements of business structures and processes (Jackson, P., 1999).

Many strategic developments point towards the growth in virtual teams. These include the move towards a more horizontal organizational structure, a growth in teamwork generally and the increased use of mobile computing techniques to connect people dispersed in time and space. As Townsend *et al*. (1998) point out, changes in teams may have to transcend both organizational and national boundaries. This is because the expertise needed for many work processes is unlikely to be located in the same office or organization. For Townsend *et al*. (1998), flat structures, interorganisational co-operation and globalization make the move from face-to-face team to virtual teams an imperative.

Where virtual teams are concerned a central role is played by advanced technologies such as mobile computing, video conferencing and e-mail. This adds a layer of complexity to what is already a very dynamic situation. The redesign of the business processes in order to fit in with the changes brought about by mobile computing organizations will have to consider significant social redesign (Badham *et al*., 1997). Because these changes are likely to be highly complex with a high degree of uncertainty Badham *et al*. (1997) suggests that a lot of time be spent ensuring effective communication to encourage flexibility, address perceptions and generate involvement.

Having looked at the strategic change issues involved, the next stage is to consider virtual teams in the organization.

3.2.2 Virtual Teams

"Virtual Team developments certainly suggest a host of new opportunities for individuals and businesses alike. They point to new applications of the latest communications technologies. But they also bring with them a host of new questions and challenges" *(Jackson, P., 1999).*

It is these new questions and challenges facing the organization that this section will be considering. Organizations are having to deal with whole sets of challenges and transformation. Not only are technologies changing rapidly so too are the business environments. It is due to these changes that organizations need to look at redesigning their

structure. There needs to be a shift from hierarchies based on management control, to a horizontal form built around teamwork and employee empowerment (Jackson, P., 1999). Organizations need to encompass moves towards strategic alliances and virtual organizations (Jackson, P., 1999). These changes cannot happen overnight, it is very important that attention is paid to the management of change (Jackson, P., 1999). An organizational change into the running of virtual teams is likely to produce social, cultural and political ripples across the organization. For an organization the important thing is to understand the nature of these ripples in order to see how they can be dealt with and managed.Setting up of virtual teams can be difficult and time consuming but the greatest challenge is the integration of these teams into the organization as a whole.

3.2.3 Integration of Virtual Teams

When dealing with the integration of virtual teams into the work place there is a tendency to focus on the degree to which IT can substitute for face-to-face communications (Townsend *et al*. 1998). The general consensus is that the richer the medium the less need for face-to-face encounters. The problem with this is that verbal communication is not explicit, it often involves voice intonation and body gestures, which even the richest of media cannot convey. In order to be able to fully decide on whether IT can be a good substitute for face-to-face communications it is important to look at the boundaries of the teams involved.

Given the fact that mobile computing is eroding temporal and spatial boundaries, it would be tempting to presume boundaries no longer mattered (Jackson, P., 1999). In virtual organizations and teams, boundaries are seen as barriers, restrictions that frustrate the free flow of ideas, information and expertise. But Yan and Louis (1999) show boundaries remain an important aspect of organizational life, virtual or otherwise. They note that the absence of structures and well-defined sets of roles and responsibilities has increasing left the individual to deal with demands on time and energy. So even if teams are virtual it is important to lay out well defined structures for them to be able to work to, otherwise there could be a lot of unnecessary time wasted by individuals trying to sort this out themselves (Cascio, 2000).

In addition to considering the social and political integration of virtual teams it is important to focus on the process of knowledge creation (Jackson, P., 1999). This is particularly important in cross-functional and product development teams, where new ideas that come are intended to secure competitive success (Nonaka and Takeuchi, 1995). Team members are generally viewed as bringing their skills, knowledge and strategies to bear on the team processes. The challenge in creating teams is to combine these in order to ensure effective

and efficient creation of knowledge. Here the challenge for virtual teams becomes apparent, much of what is known is not held or communicated explicitly, it is acquired and passed on in the realm of action. It is knowledge learned by doing. When setting up virtual teams, the organization cannot presume that the members will share common metal models, figurative or symbolic language or particular routines and practices. Given the strong reliance on learning by doing for these to develop, the lack of face-to-face interaction in virtual working could prove problematic.

Once virtual teams have been set up and integrated, the organization can begin to function. Howeve other issues that need to be considered iteratively throughout the life of the virtual organization these include trust and cultural issues.

3.2.4 Trust Issues

Another common issue around virtual organizations is the idea of trust. Some organizations just presume that trust will develop simply because it is so important. There is little discussion about how trust might be achieved and what form it might take. There are some different schools of though on this matter, Lipnack and Stamps (1997) state that:

"As important as positive relationships and high trust are in all teams, they are even more important in virtual ones. The lack of face-to-face time, offering opportunities to quickly clear things up, can heighten misunderstandings. For many virtual teams trust has to substitute for hierarchical and bureaucratic controls"

In contrast, Handy (1995) says that trust cannot be presumed but must be managed. However he continues to tell us little about how trust can be enforced or the ways in which the organization can intervene in order to shapetrusting relationships. According to Cascio (August, 2000), a key ingredient in the success of a virtual organization is trust that one's coworkers will fulfill their obligations and behave predictably. A lack of trust in a virtual organization can undermine every other precaution taken to ensure success. Just as trust needs to be managed and revisited regularly throughout the life of the organization so too do the cultural issues that arise from the geographical dispersement of employees.

3.2.5 Cultural Issues

With the increased use and somewhat dependence of organizations on mobile computing, virtual teams can be seen as somewhat of a panacea for resourcing dispersed and global organizations. It is important to remember that in certain circumstances more effective human resource solutions may be possible (Jackson, P., I.S.J.,1999). Where Organizations use mobile computing to operate globally, a broad cultural perspective, as well as particular cross

cultural expertise, may be needed. In some circumstances the nature of the cultural knowledge may mean that learning it demands sustained presence in the culture in question. Having to transfer business policies and cultures to work with dispersed business teams across collaborating organizations, geography and cultures can lead to potential clashes of business and national cultures. According to Cascio (August,2000), if the members of a virtual organization are notempowered to make decisions, the technology that enables their collaboration will add little value and the competitive advantage associated with rapid responses to demands in the market place will be lost.

A virtual organization is fundamentally different to the traditional organizational setup. Due to this the management of such an organization would need to be different. Managers need to have certain characteristics and need to learn new skill. The issues around management of a virtual organization are considered next.

3.3 Managing The Virtual Organization

There are sound business reasons for establishing Virtual Organizations, but theiradvantages may be offset by setup and maintenance costs, loss of cost efficiencies, cultural clashes, isolation and lack of trust (Cascio, August 2000). According to Cascio (2000), in order to successfully set up a virtual organization managers need to ensure two things :

- Shift from a focus on time to a focus on results
- Recognize that virtual organizations, instead of needed fewer managers, require even better supervisory skills among the existing managers.

Virtual organizations are multisite, multiorganisational and dynamic, it consists of a group of people that have joined in an alliance to exploit complementary skills in perusing common strategic objectives. This grouping represents a dramatic change in the way work is completed and this presents new challenges for managers. The challenges stem from the physical separation of workers and managers wrought by such information-age changes such as mobile computing and teleworking.

It is important to note that not all managers are suited to managing employees with virtual-work arrangements. Cascio (2000) has identified the following characteristics that the virtual manager requires:

- An open, positive attitude that focuses on solutions.
- A results-oriented management style. Those who need structure and control are not likely to be effective managers in a virtual situation.

- Effective communications skills, both formal and informal, using both mobile computing methods as well as more traditional methods e.g. the phone.
- An ability to delegate effectively and to follow up to ensure work is completed successfully.

Even if a manager seems suitable for the role of virtual manager it is important to ensure sufficient training is supplied in the following areas:

- How best to us the mobile computing software to enhance team performance. Guidelines on social protocol for virtual organizations
- How best to manage the virtual environment
- How to provide feedback, this is particularly important since the traditional cues of social interaction i.e. body language and gestures, may not be available.

Communication is a major challenge for managers implementing a virtual work environment. Many managers have to learn new communication skills in order to prevent team members from feeling isolated and not part of any larger group. It is important that managers do not rely solely on e-mail. Managers need to ensure they make the most of all that the mobile computing technology can offer them. Managers should conduct effective audio meetings, use voice-mail and video conferencing. Mangers should regularly schedule virtual meetings, communicate with all team members and produce regular updates and status reports for the whole team(Cascio,2000)

According to Cascio (2000), the biggest challenge in managing the virtual organization is the management of performance. It is vital that mangers define, facilitate and encourage performance.

- Define performance – In a virtual Organization it is important that everyoneunderstands their responsibilities. A manager trying to define performance may ask thefollowing questions to clarify responsibilities:

  What are the objectives?
  Which responsibilities are shared?
  Will the teams elect their own leaders?
  What are the responsibilities of the team leader?
  How will teams make decisions?
  What decisions can be make by which teams?

The next step is to develop specific, challenging goals. To be useful these measures should be linked to the organizations strategic direction. In defining performance regular assessment of progress towards goals focuses the attention and efforts of employees.

- Facilitate Performance – Managers that are committed to managing remote workerseffectively have two major responsibilities:
    1. To eliminate roadblocks to successful performance
    2. To provide adequate resources to get a job done right and on time.

Obstacles that can inhibit maximum performance include, outdated technology, delays in receiving critical information, and inefficient design of work processes. Adequate capital resources, material resourses and human resourses arenecessary if the remote workers are to reach the goals laid out by managers.

- Encourage performance – It is important to provide sufficient rewards that employees really value in a timely and fair manner.

For a organization to set up as a virtual organization or for an organization to move to being a virtual organization is a huge undertaking with several key issues to be considered. So this begs the question:

- What are the strategic advantages that an organization will gain from becoming a virtual organization?

3.4 Strategic Advantage

In a virtual organization the employees are dispersed and their primary interaction is through some combination of electronic communication systems. The employees may never even meet in the traditional sense. This type of organizational structureoffers several advantages:

- It can save time, travel expenses and can provide easier access to experts.
- Teams can be organized even if members are not in proximity to each other.
- Organizations could use outside consultants without incurring cost for travel, logging and downtime.
- Virtual teams allows organizations to expand their potential labour markets, enabling them to hire and retain the best people regardless of their physical location.
- Employees can easily accommodate both their personal and their professional lives.
- Dynamic team membership allows people to move from one project to another.
- Employees can be assigned to multiple, concurrent teams.
- Team communications and work reports are available online to facilitate swift responses to the demands of a global market.

According to Warren (2000), there are also the following benefits to having a virtual Organization:

- **Staff Retention** –In a case study done in Nortel after introducing teleworking for employees the job satisfaction survey conducted showed that staff retention was 16% better among teleworkers rather then office based staff, with job satisfaction level 11% higher and teleworkers being 17% more productive.
- **Reduced costs** – It is estimated that real estate is the second largest overhead of an organization after salaries, therefore moving people out of offices would realize a huge saving.
- **Staff mobility** – The increased mobility of staff means that staff can now work productively in any place at any time.

These are just some of the possible advantages of a virtual organization. This area is still in its infancy and therefore is relatively unexplored. It may be years before these benefits are actually realized by an organization, and it may also be years before any unforeseen benefits arise.

4. Conclusion

This paper has given a technological background to mobile computing before moving into the main body of the paper, discussing the effect mobile computing is having on the organization. As can be seen from this paper the effects of mobile computing are being seen in organizations already in the form of teleworking. The use of e-mail, voice-mail, video conferencing, and modems have enabled organist ions to allow certain people to work from home. These people rely on modern technological advancements as their main form of communications. Teleworking can be considered an evolutionary step towards the Virtual Organization.

A Virtual Organization is very different in every way from the traditional organization and therefore there are a number of issues involved in setting a virtual Organization up:

- Strategic change issues.
- Virtual Teams.
- Integration of Virtual Teams.
- Trust Issues.
- Cultural Issues.

As well as these issues the management of the virtual organization would be very different to that of managing a traditional organization. As indicated in the paper not everyone is suited

to managing a virtual organization and therefore managers must be carefully chosen and well trained.

In order to successfully set up a virtual organization all of the issues addressed in this paper must be considered fully, failure to identify with this issues and plan for them could mean the failure of the virtual organization. The issues addressed are difficult and time consuming so it is important that the organization is aware of the possible strategic advantage that this organization can offer including, staff retention, reduced costs, staff mobility, organizations can expand their potential labour markets, employees can easier accommodate home and work life, reduction in travel times and it can provide easier access to experts.

5. Bibliography

Badham, R, Couchman, P. & Mcloughlin, 1997, Implementing vunerable socio-technical change projects in: innovation, Organistional Change, and Technology, ITB Press London.

Bartlett, J. September 1997, Not as Mobile as we would like to be, LAN Times.

Cascio, W., August 2000, Managing the Virtual Workplace, The Academy of Management Executive.

Chutchian-Ferranti, J., September, 1999, Virtual Corporation, ComputerWorld.

ComputerWorld, Spetember 1999, Business, Virtual Style.

Dhawan, C. January 1997, Analyst bets mobile PC future on better speed, business tools, Computer Dealer News.

Fenton, C., Nigeon, B., Harris, J., July 2000, Wireless Access, BT Technology Journal.

Gartner Group, Advanced Technologies Scenario, Conference Presentation, 1998.

Gartner Group, WAN Wireless, Conference Presentation 1998.

Guardini, I., D'Urso, P., Fasano, P., November 2000, The Role of Internet Technology in future Mobile Data Systems, IEEE Communications Magazine.

Gurton, A. November 2000, Are you alright? You seem do …remote, Computer Weekly.

Handy, C., 1995, Trust and the Virtual Organization, Harvard Business Review May/June.

Hartung, F., Ramme, F., Ericsson Research, November 2000, Digital Rights management and watermarking of multimedia content for M- commerce applications, IEEE Communications Magazine.

Hsiao, R.L., Omerod, R.J., 1998, A new perspective on the dynamics of information technology enabled strategic change, Information Systems Journal.

Huws, U., Korte, W., Robinson, S., 1993, Telework- Towards the elusive office, John Wiley & Sons

Jackson, P. J., 1999, Organizational change and Virtual Teams: strategic and operational integration, Information Systems Journal.

Jones J. P., 1999, Plugging into Wireless Technology, Computer Dealer News.

Kavangh , J., 1999, Teleworking solves the skills shortage, Computer Weekly.

Krichevsky, I., 1999, Wireless still at early stages of potential, Computing Canada.

Lipnack, J., Stamps, J., 1997, Virtual Teams, Wiley, New York.

Merrett, R.P., Beastall, P.V.E., Buttery, S.J., Wireless Local Loop, BT Technology Journal, October 1998.

Nonaka, I., Takeuchi, H., 1995, The knowledge creating company, Oxford university press , NY

Saab, H., 1999, Teleworking bring business, morale advantages, Computing Canada.

Seeley, M., 2000, What the papers say, Computer Weekly.

Towsend, A.M., DeMarie, S.M., Hendrickson, A.R., 1998, Virtual Teams: technology and the workplace of the future, The Academy of Management Executive.

Warren, L., 2000, The benefits of emptying the office, Computer Weekly.

www.whatis.com

Yan, A., Louis, M.R., 1999, The migration of organization functions to the work unit level: buffering, spanning and bringing up boundries, Human Relations.

Knowledge and Value Development in Management Consulting

Management Consulting from the Client´s Perspective

Developing an Understanding of the Dynamics of the Client-Consultant Relationship

Ms. Fionnuala Darby & Ms. Geraldine Lavin

Fionnuala.Darby@itb.ie

ABSTRACT

The relationship between management consultants and their clients plays a key part in the success of consulting firms. To develop an understanding of the dynamics of the client-consultant relationship, the authors reviewed literature in the areas of intangible professional services, impression management, perception and the interaction process. A model is proposed and the authors use a case study to emphasise areas highlighted by the literature.

The authors contend that impression management, aided by positive perception and the development of the client-consultant relationship, is an important motivational force in securing consulting projects. It upholds the model for understanding the client-consultant relationship.

INTRODUCTION

Management consulting is no different to any other profession in that promotion is on merit. You move forward as fast as your performance warrants. One should always differentiate in management between "what should be", i.e. normative models promoted by consultants and top managers and "what is". i.e. shop-floor arrangements and regulations lived by employees and that can only be documented by careful and prolonged empirical research.

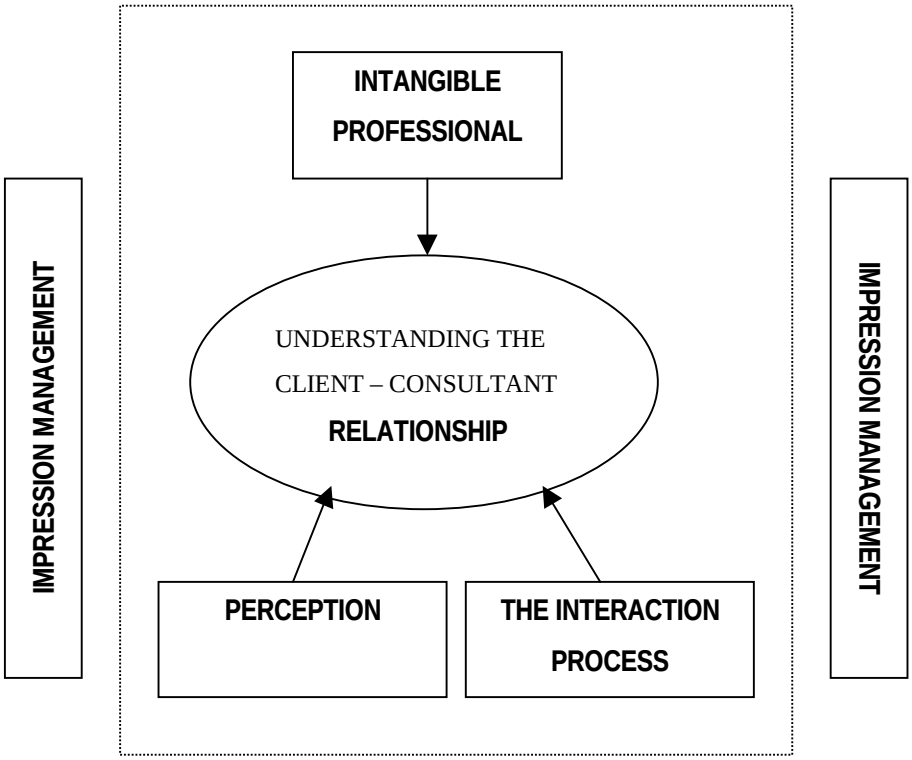
Fads are created and sold and implemented by consultants, reengineering is a good example. In summer 1990 there was an article by Hammer in the Harvard Business Review on reengineering. By March 1993 Hammer and Champy had published a book on the topic. In a positive sense fads are seen as a source of energy and dynamism to break away from the status quo. Negatively, however they maybe interpreted as a sign of panic, a regression to a "cure-all" magical time of thinking at the expense of distance and analysis.

The true measure of a consulting firm is its ability to help clients solve difficult problems. Consulting is unique in the field of business because it is driven by ideas. This paper will review the concept on impression management and the part that it plays in securing the client-consultant relationship. Examined with this idea are the interaction process, the role perception plays in selecting a consultant and the marketing and selling of an intangible professional service like management consulting.

The corner stone of this paper will develop an understanding of the dynamics of the client consultant relationship.

OVERVIEW

The kernel issue of the research is an attempt to understand the client – consultant relationship. Central to this is comprehending the concepts of intangible professional services, perception and the interaction process. Surrounding these themes is the theme of impression management which permeates the entire model. Diagrammatically the themes can be presented as follows:



LITERATURE REVIEW

Intangible Professional Services

When comparing products and services, the absence of a physical product is the primary difference. The intangible nature of services has been identified as one of the key differences between products and services, the others being inseparability, heterogeneity and perishability (Pride & Ferrell 2000; Zeithaml, Parasuraman & Berry 1985). The characteristics of a professional service, such as management consulting, include intangibility and the status of the personnel involved (Hall, Leidecker, Mills & Margulies 1983). The service industry relies on the experience achieved by the consumer, resulting in the prohibition of potential customers inspecting or trying out a service before purchase. This results in a greater reliance on the communication of the benefits on offer being accurate, attractive and credible. Accurate and current information should be available to the potential consumer, ensuring the build up of an impression of the service (Pride & Ferrell 2000). Further to the marketing effort, Kolter (1994) argues that it is the responsibility of the service provider to "mange the evidence", to "tangibilize the intangible" and utilise the project proposal to make the actions of the service provider specific.

Traditionally, there is also a difficulty in producing accurate comparisons, between two similar services, due to the lack of unbiased material available. This is being resolved in some part through the freedom pertained from the Internet. There is significantly more information available, ensuring increased comparisons and decreasing the inability of traditional price comparisons.

Proposition 1: Consumers about to engage the services of a consulting company seek independent information on their prospective service provider.

The literature also highlights the importance of repeat business from existing clients in service companies and emphasises the benefits of retaining these clients compared to securing new clients (Grönroos 1990; Liswood 1989; Reichheld & Sasser 1990; Sellers 1989). With prospective consumers, the service organisation must expend time, financial and marketing resources on engaging and securing a new client. Whereas with existing clients, provided the service organisation has built up a good relationship with the client, the process consumes less resources. This area has been examined in depth through literature on relationship marketing. Maintaining and developing good client relationships through fulfilling promises to customers is crucial to service organisations (Grönroos 1990). Indeed,

this focus on the value of existing clients convinced Levitt (1986) to deduce that the aim of business is not to make a profit but to win and retain clients.

Proposition 2: Consulting companies who focus on retaining existing clients and developing the client-consultant relationship will secure further contracts as a result.

Impression Management

Consultancy, as impression management, is an attempt to convince clients of their value and quality. What is a consultant but a professional helper. According to Clark (1995) "new lenses" are needed to see that a core feature of consulting work is the art of impression management. The processes by which individuals attempt to control the impression others form is impression management.

It should be noted from the outset that impression management theory does not imply that the impressions created by consultants in this case, are necessarily false. This paper recognises that impression management exists in securing the client consultant contract and attempts to examine the nature of impression management in understanding the client-consultant relationship.

Bolino (1999) posits that impression management appears to have a lot in common with citizenship behaviours. For the purposes of this research there is a dichotomy between citizenship behaviour and impression management. Citizenship behaviours that employees engage in, in the workplace affect the impression that an individual makes on a supervisor or co-worker. These behaviours are therefore internal to the organisation. Their main purpose being an important attempt to accomplish one´s goals i.e. usually career advancement. Impression management on the other hand is what the consulting firm engages in to influence the image others have of them. Consultants use this behaviour because what they are trying to sell is an intangible professional service to the prospective client. Until it is produced it only has potential, up to that point it remains just a promise. This behaviour is external as it occurs in the external-task environment of the company when they seek help, or wish to maximise an opportunity and so draw on a consultant´s supposed expertise.

Theorists in impression management advocate that there exists a primary human motive, both inside and outside of organisations, to be seen by others in a favourable light and to avoid negative connotations. (Rosenfeld, 1995). Drawing on the literature on the topic there appears to be three salient factors determining the motivation to manage impressions:

Goal importance and impressions

This is the consulting firms motivation to make the prospective client perceive that the consulting firm´s proposal is instrumental to the client´s success, (Leary & Kowalski, 1990). The "goal" has two sides – the client firm´s goal is economic success after project implementation. The consultant´s goal is securing the contract after presentation of the proposal.

Proposition 3: Consultants will be more likely to engage in impression management when they can identify with the importance of the client´s goals.

Value of image enhancement

Bolino (1999) describes this factor as obvious in political climates in an organisation, e.g. approaching performance appraisal deadlines. In the consulting industry value enhancement is used where there is a lack of objective criteria for assessing performance or success. The greater the ambiguity surrounding an organisation´s future, the greater the scope for the consultant to engage in successful impression management behaviours.

Proposition 4: The greater a client organisation´s environmental uncertainty the more likely a consultant will engage in impression management.

The difference between desired and current images

This is the discrepancy between the consultant´s desired image and the image they believe the client has of them. This discrepancy can arise from reputation, previous experience and former successful or unsuccessful contracts. The greater the discrepancy between the two states the more likely the consultant will be to engage in impression management. The consultant´s aim being corrective action for prior negative impressions the client may have of them. Or simply to reinforce the consultant´s current positive reputation and track record.

Proposition 5: Consultants will be more likely to engage in impression management when their image has suffered in the past or due to poor performance.

Not everyone is concerned with impression management and it is no surprise to hear that the low self-monitor does not feel the need to mould their appearance and behaviour to fit each situation unlike the high self-monitor. The consulting profession undoubtedly, engages in impression management. Impression management techniques include conformity, acclamation, flattery and association (Robbins 1999). Crosier (1997), refers to "corporate

reputation" in his article on Bromley´s *Social Psychology of Reputation* as an "unarguable credential" in analysing one consultant´s bid over another.

It is important to keep in mind that impression management does not imply that the impressions consultants convey are necessarily false. The impression manager for the consulting firm must be cautious not to be perceived as misrepresenting the situation. However it is situations of high uncertainty and ambiguity that tend to be characterised by misrepresentation and these situations provide little information for challenging a fraudulent claim (Robbins 1999) (See proposition 2).

Interpreting impression management could be described as a stand-alone theory. Like reading a novel each reader will have a different interpretation of events. Similarly because impression management involves interpreting behaviours, each client-consultant situation is likely to be different. The above propositions are to be seen in a general light and will undoubtedly vary from client to consultant to interpreter and so on. For the purposes of this research they are to be seen as highly exploratory in nature.

Perception

While impression management focuses on attempts by the consultant to be seen in a favourable light by the client, perception in this paper is the client´s view of these attempts by the consultant to secure the contract. Perception is the process that the client uses to make sense out of the consultant´s proposal. It is the means by which the client selects, organises and interprets the information put forward by the consultant firm. There exists individual differences in what people perceive and how they organise and interpret it. Therefore perceptions vary among people. Recognising the difference between what is perceived and what is real is a key element in diagnosing a situation.

According to Daft (2000) the perceiver, the client in this case, has six characteristics:

- Needs and motivation – What is the consultant´s attempt to satisfy the client´s needs by finding a solution to a problem or exploitation of a pending opportunity.
- Values and beliefs – How aligned are the beliefs and values of the consulting firm with those of the client?
- Personality – How stable is the consultant´s behaviour pattern in response to idea generation, problem-solving and the environment?
- Learning – Has the consultant made certain factors important enough for the client to pay attention to them.
- Primacy – People pay greater attention near the beginning of a presentation.

- Recency – People pay greater attention toward the end of a presentation.

It can be concluded that it is the job of the consultant to select and organise stimuli to provide meaningful experiences for the perceiver. Hentschel, Smith and Draguns (1986), stipulate that this represents the psychological process whereby people take information from the environment and try to make sense of it. In this case the client takes information from the consulting firm´s proposal/presentation that will be of relevance to their situation be it a problem or an opportunity.

In the consulting industry the client or prospective client engages in perceptual selectivity. They screen out the various objects and stimuli that vie for their attention. Certain stimuli catch their attention and others do not. It is the intention that this exploratory piece of research examines what exactly catches their attention and what does not, all the time taking into account the role impression management is playing in catching the client´s attention.

Proposition 6: On what basis does the client perceive an understanding of the consultant´s efforts in attracting their attention?

The Interaction Process

Customers are increasingly aware of the alternatives on offer and also of the rising standards of service. Consumers are demanding higher standards and better customer service, both of which businesses must provide in order to remain competitive (Lewis 1995). The changing pace of business has increased the need for more meaningful interaction between client and consultant.

The nature of the delivery of a professional service requires a high degree of client-supplier organisation interaction. Levitt (1972) identified this personal involvement in delivery and maintained that "services are invariably and undeviatingly personal….something performed by individuals for individuals". The interaction between the consulting company and the client is particularly important in the initial stages of securing the consulting project. Gummerson (1979), following interviews with 50 professionals, concluded that:

> *A professional service can only be purchased meaningfully from someone who is capable of rendering the service. What is needed is the professional who sells, not the professional salesman.*

Further research by Gummerson (1987) stresses the emphasis on person-to-person interaction between the service provider and the customer. The interaction between client and consultant is not static, it develops over a number of phases. Grönroos (1980) developed a Three Stage

Model to describe the development of the supplier organisation-client interaction process. The three stages are:

Stage 1	Interest by the consumer in the supplier organisation and the services it offers as a possible means of satisfying the consumer´s needs
Stage 2	Purchase of the services required as a possible means of satisfying the consumer´s needs
Stage 3	Repeat purchase of the same or similar services which are provided by the supplier organisation as needed.

Of particular interest to the scope of this paper is Stage 1, where it is identified that the supplier company, such as a management consulting company, is trying to generate consumer interest in the company and its services and to generate a high level of perceived quality. Stage 2 involves turning the consumer´s general interest into a sale. During Stage 3 the supplier organisation attempts to guarantee repeat sales by activities engaged in during the supply of the service to the consumer when there is frequent and interactive contact between the two parties. Grönroos (1980) advocates continuous adaptation of the supplier organisation´s operations to meet consumer´s present, expected and perceived needs.

A framework of four phases, drawing on the work of Peplau (1969) and Barber (1997) has been extended by Barber and Mulligan (1998). The four phases discussed are orientation, identification, exploration and resolution and can be summarised in the following table (Barber & Mulligan 1998):

Phases	**Issues**	**Nature**
Orientation	Forming a working allegiance Adapting to each other´s world Developing trust Negotiating rules of engagement	Client-centred
Identification	Clarifying problems Raising strategies Understanding the contextual frame Bonding in partnership	Problem-centred
Exploration	Implementing chosen strategies Modifying and experimenting Enacting mutually supportive roles Deepening understanding	Strategy-centred
Resolution	Evaluating outcomes Completing consultation Reviewing follow-up Debriefing for insight	Quality-centred

Proposition 7: To ensure successful interaction, the consulting company must concentrate on the development of a personal relationship with the key project members in the client company particularly during the preliminary stages of the project.

RESEARCH METHODOLOGY

The research conducted involved the collection of both primary and secondary data. Primary data is information collected from the original sources specifically for the task at hand (Emory & Cooper, 1991). Secondary data is gathered and recorded by someone else, prior to and for purposes other than the current needs of the researcher (Zikmund, 1991).

Research Question

To develop an understanding of the client-consultant relationship.

More specifically the paper strives to understand the part played by impression management, perception, intangible professional services and the interaction process as interpreted by the consultant, client and potential client.

Propositions

1. Consumers about to engage the services of a consulting company seek independent information on their prospective service provider.
2. Consulting companies who focus on retaining existing clients and developing the client-consultant relationship will secure further contracts as a result.
3. Consultants will be more likely to engage in impression management when they can identify with the importance of the client´s goals.
4. The greater an organisation´s environmental uncertainty the more likely a consultant will engage in impression management.
5. Consultants will be more likely to engage in impression management when their image has suffered in the past or due to poor performance.
6. On what basis does the client perceive an understanding of the consultant´s efforts in attracting their attention?
7. To ensure successful interaction, the consulting company must concentrate on the development of a personal relationship with the key project members in the client company, particularly during the preliminary stages of the project.

Research Design

The information gathered in qualitative research goes beyond what, which and when and into the realms of how and why. It consequently allows the researcher to gain deep and rich insight into people´s attitudes, needs and behaviours.

What is distinctive about qualitative research is that it is holistic in the sense that qualitative research attempts to understand the overall context of the problem. It does not follow a ready made plan and the research is open to discovery. Qualitative methods are particularly oriented toward exploration, discovery and inductive logic.

According to Cooper and Emory (1995) exploration is particularly useful when the researcher lacks a clear idea of the problems they will meet during the study.

Sample

The sample size used was a single case study. The participants were made aware that their expert training and experience in their professions would give the researchers valuable insight into the research topic. The case study examines the relationship and interaction process between a management consulting organisation and their client. The consulting project required the management consulting organisation to complete a 5 year strategic plan for the client. The time frame for the consulting project was one year. The project was completed in summer 2000.

Data Collection

Within the realms of direct qualitative methods, the researcher must choose between in-depth research methods or focus group interviews. The nature of the questioning in this study rendered inapplicable the use of focus groups. The confidential nature of the information required creates reasonable doubt that sufficiently deep responses could be elicited in a group setting. Therefore in-depth interviews with the key project members from the consulting company and the client organisation were held in October and November 2000.

Data Analysis

Qualitative methods will be used in this research due to its suitability as a means of conducting exploratory research. The qualitative findings are presented in the following section along with a discussion.

Research Limitations

The results will be limited primarily by the use of a single case study. However the authors feel that the primary benefit will be the review of literature in this area and the case study serves to enhance this review.

RESULTS & DISCUSSION

The cornerstone of this research is to understand the dynamics of the client-consultant relationship. More specifically one of the over-riding factors that emerged from the literature review and subsequently the data collection was the key role of impression management in securing the initial consulting project and developing the relationship between the two parties.

Proposition 1: Consumers about to engage the services of a consulting company seek independent information on their prospective service provider.

Proposition 2: Consulting companies who focus on retaining existing clients and developing the client-consultant relationship will secure further contracts as a result.

Proposition 3: Consultants will be more likely to engage in impression management when they can identify with the importance of the client´s goals.

Having analysed the data bearing Proposition Three in mind, the following results emerged:

- The consultant company in this research has a relatively recent formation. Due to previous work the company had done on similar projects they were invited to tender for the project.
- A tender document containing terms of reference were submitted prior to a presentation to the client company.
- The presentation by the consultant company was interactive and was made to a steering group.

The consultant company was of the impression that the potential client was looking to work with someone who could maximise a highly interactive and consultative environment to elicit information. In this regard the interactive and open presentation worked to their advantage as the format of this presentation was essentially a questions and answers style forum “around a table”. It was in this environment that the consultant company could display highly interactive and communicative skills which turned out to be a forte for them in securing the

contract. So it could not be directly said that they engaged in impression management but rather that the open nature of the presentation worked to their advantage.

Proposition 4: The greater an organisation´s environmental uncertainty the more likely a consultant will engage in impression management.

A key success factor for the consulting company was their exact interpretation of the environment that the client wished them to work in (an highly interactive and consultative environment). To this end their initial strategy was focused in displaying a confidence at the presentation of their ability to work in a highly interactive environment. The environment could be said to be certain as there was a clear strategic aim (to write a 5 year strategic plan) but the outcome of the environmental analysis was uncertain and emerged during completion of the project.

The aim of the project was the development of a project plan for a suburban area. The way in which the consultant company gathered information on the project after securing the contract was through a long and rigorous series of focus groups and interviews with key personnel in all segments of society pertaining to the development plan.

As Bolino (1990) points out where there is a lack of objective criteria for assessing performance or success then the greater the scope for the consultant to engage in successful impression management behaviours. In this scenario the client was uncertain as to how the objectives were to be achieved thereby creating uncertainty. The consultant adopted this uncertainty by reducing it to comfortable terms for the client and thereby securing the contract.

Proposition 5: Consultants will be more likely to engage in impression management when their image has suffered in the past or due to poor performance.

For the consultant this was one of the first major projects they had undertaken since their formation. The client was aware of them from previous work the consultant company had completed in the project area. So with regard to Proposition Five, it does not apply in this case since the consultant company had neither poor performance nor a dented corporate image to defend. However it is worth noting under discussion of this proposition that because the consultant company was a in an early start-up phase, it was very eager to make a lasting impression having neither an extensive portfolio of previous projects to give additional weight to their initial tender.

The consulting company would like to think that the impression they left with the client company was that they endeavoured to pursue the client´s objectives in a capable, competent and complete fashion, so much so that the client company would use their professional services in the future and would also recommend them to others.

Proposition 6: On what basis does the client perceive an understanding of the consultant´s efforts in attracting their attention?

The project in question in the case study was the development of a 5 year strategic plan for the development of a suburban region. It should be noted that the providers of information to the consultant company did so on a purely voluntary basis. Interviews were conducted with local government departments, residents and social partnerships in the area.

It was through networking that the client shortlisted three consulting companies to bid for the project. On the basis of initial proposals, the consulting companies made presentations to the client. The quality of the tender proposal along with the consulting company´s subsequent presentation to the steering group were the deciding factors in securing the contract.

What was the client company looking for?

- **Personnel Involved** - a concern for excellence and success and the ability to work closely with others on the project.
- **Body Language** – the use of body language to create a confortable setting which would be used by the consultant later in eliciting responses. This was particularly important due to the highly participative nature of the information gathering process.
- **Rapport** – in general terms the client company were seeking an affiliation with the consultant company and their strategic aims and objectives for the project. One of the initial companies invited to tender failed to secure the contract because they gave the impression to the client that they were unapproachable and viewed the project as “just another piece of business”.

The client´s initial impression of the consultant company was favourable. During the project this impression escalated. The client perceived that the consultant was always looking ahead. The consultant company had no affiliation to the suburban area but the client perceived that nonetheless they “embraced the environment” they found themselves in and took ownership of the project, something that looked unlikely from another competitor in the tender process.

The client has used the consultant company again in a training capacity and has recommended them to other organisations. The client company revealed the following about the consultants:

- Found the consultant company to be highly professional
- Committed to the project
- Made a lasting first impression

Proposition 7: To ensure successful interaction, the consulting company must concentrate on the development of a personal relationship with the key project members in the client company, particularly during the preliminary stages of the project.

CONCLUSION

On the basis of impression management and perception there appears to be little differences between the impression the consultant company feels they left with the client and the impression and perception that the client company has of the consultant. The authors contend therefore that impression management, aided by positive perception is an important motivational force in securing the bid for a consulting project. It upholds the model aforementioned for understanding the client-consultant relationship and enforces how one´s image may drive a project in the initial stages.

If the propositions in this article are found to be true in further research then they present important implications for practising managers seeking a consultant and consultant companies trying to secure contracts. The client should be careful in assessing the impression of the consultant firm so as to reduce the discrepancy between desired and current images. What is special about this piece of research is that the consultant did not have a reputation or much previous experience to rely on and won the contract solely on the content and delivery of their presentation. The authors would also make the assumption that the client was not unduly concerned about impression management swaying their decision in the selection of a consultant.

BIBLIOGRAPHY

Barber, P. & J. Mulligan (1998): The Client-Consultant Relationship in Management Consultancy – A Handbook for Best Practice, Ed by P. Sadler, *Kogan Page*

Barber, P. (1997): The Client Therapist Relationship: An Action Research Approach *British Gestalt Journal* v6 n1

Bolino, M., (1990): Citizenship and Impression Management: Good Soldiers or Good Actors?, *Academy of Management Review*, Jan 1990 v24 n1

Clark, T., (1995): Managing Consultants: Consultancy as the Management of Impressions, *Open University Press*

Cooper, D. & Emory, C., (1995): Business Research Methods, *Irwin Publications*

Crosier, K., (1997): Reputation, Image and Impression Management, *European Journal of Marketing*, May-June 1997 v31 n5-6

Daft, R., (2000): Management, 5th ed., *Dryden Press*

Grönroos, C. (1980): Designing a Long Range Marketing Strategy for Services *Long Range Planning* April 1980

Grönroos, C. (1990): Relationship Approach to Marketing in Service Contexts: Marketing and Organisational Behaviour Interface, *Journal of Business Research* v20

Gummerson, E. (1979): The Marketing of Professional Services *European Journal of Marketing* May 1979

Gummerson, E. (1987): The New Marketing – Developing Long-Term Interactive Relationships, *Long Range Planning* v20 n4

Henschel, U., G. Smith & J. Draguns (1986): The Roots of Perception (eds.), *Amsterdam*

Kotler, P. (1994): Marketing Management: Analysis, Planning, Implementation and Control, 8th Ed, *Prentice-Hall, Englewood Cliffs, NJ*

Leary, M.R., & Kowalski, R.M., (1990): Impression Management; A Literature Review and Two-Component Models, *Psychological Bulletin*, 107, p34-47

Levitt, T. (1972): Production-Line Approach to Services, *Harvard Business Review* Sept/Oct 1975

Levitt, T. (1986): The Marketing Imagination, *The Free Press, New York*

Lewis, B.R. (1995): Chapter on Customer Care in Services in Understanding Services Management, Ed. by W.J. Glynn & J.G. Barnes, *Wiley, Oak Tree Press (Ireland)*

Liswood, L.A. (1989): A New System for Rating Service Quality, *The Journal of Business Strategy*, v10 n4

Hall, J.L., J.K. Leidecker, P.K. Mills,, & N. Margulies (1983): Flexiform: A Model for Professional Service Organisations *Academy of Management Review* 1983 v5

Peplau, H.E. (1960): Psychotherapeutic Strategies *Perspectives in Psychiatric Care* v6

Pride, W.M. & O.C. Ferrell (2000): Marketing: 2000 Edition *Houghton Mifflin*

Reichheld, F.F. & W.E. Sasser Jr. (1990): Zero Defections: Quality Comes to Services, *Harvard Business Review* Sept/Oct 1990

Robbins, S., (1999): Organisational Behaviour 9th ed., *Prentice Hall*

Rosenfeld, P.R., Giacalone, R. A., & Riordan, C.A., (1995): Impression Management in the organisation, *Hillsdale, NJ: Lawrence Erlbaum Associates*

Sellers, P. (1989): Getting Customers to Love You, *Fortune*, 13 March 1989

Zeithaml, V.A., A. Parasuraman & L. L. Berry (1985): Problems and Strategies in Services Marketing, *Journal of Marketing* Spring 1985 v49

Zikmund, W.G., (1991): Business Research Methods, 3rd ed., *Dryden Press*

A Brief Characterisation of Morphological Causation in Irish

Brian Nolan

School of Informatics and Engineering

Institute of Technology Blanchardstown

brian.nolan@itb.ie

1 Introduction

In this paper we attempt to characterise some elements of morphological causation as expressed in modern Irish. Three types of causation may be identified: lexical, periphrastic and morphological. In terms of the relative weightings of each type, the morphological causative is the least productive. Its use appears to be highly constrained to two very specific domains and it is signalled by particular morphological affixes. Lexical causatives are more productive than the morphological causative. By contrast, periphrastic or analytical causatives are highly productive and wide-ranging in their deployment. We concentrate in this analysis on some data on morphological causation.

2 Morphological Causatives

2.1 Change Marked on Verb to Signal Volition and Agency

Morphological causation requires that, when expressed, some level of change be recorded on the verb. This type of causation is the least common type found in Irish. The examples in this study relate to two specific verbs, the first of which is prefixed by *dún*, as in *dún+marú* 'murder', and the second of which is prefixed by *for*, as in *for+eigniú* 'force'. These prefixes modify just two causative verbs. A characteristic of both of these base verbs is that, without the causative morphological prefix, they are lexical causatives with a resultant change of state of the undergoer participant. Neither *dún* nor *for* have any independent existence other than as a prefix. Though now opaque, it is possible that historically these may have represented compounds, though I am not aware of any supporting evidence to elaborate on this possibility. A nominal with a morphological shape similar to *dún* does exist with the meaning of 'castle/fort/haven'. A verb *dún* 'close' also exists. These do not form a compound with a

verb in any usage. The causative prefix *dún* serves to signal a very strong agency with the highest motivated intent. The actor whose agency is marked with the causative morphological prefix on the verb is always animate and human. The causative prefix *dún* has a highly restricted distribution. Its domain of use is limited to marking the strongest agency with full control, intent, motivation and deliberation on the part of the human actor on the base causative verb *maraigh* 'kill'. In contrast, the causative prefix *for* functions as a morphological marker in causative constructions signalling intensified force.

To motivate this analysis we examine three different 'cause to die' contexts. These range from the case where: a) the causation may be accidental but a time lag exists possibly between the cause and its effect; b) the causation may be accidental but no time lag exists between the cause and its effect; to c), non-accidental motivated causation on the part of the actor with an immediate consequence for the undergoer. We illustrate this as a cline with the following table.

(1)

Verb	***Bhásaigh***	***Mharaigh***	***Dhúnmaraigh***
Actor	**±** Accidental Causation	**±** Accidental Causation	**-** Accidental Causation
Undergoer	**±** Immediate Effect	**+** Immediate Effect	**+** Immediate Effect
Time	**±** Time lag	**-** Time lag	**-** Time lag

2.1.1 Time Distance between Cause and Effect

The causative occurrence of death on the undergoer can be expressed in yet a different way by a speaker, particularly when that speaker wishes to convey information as to whether the means of causing death is direct, or not. Use of this particular verb *bhásaigh* 'kill' reflects that the death may have taken place some time after the causative act was undertaken, or that the means of the killing was somehow indirect but still no less intended. Accidental causation is possible here and its effects are not necessarily immediate for the undergoer. A typical example of this is given next with a transitive clause coding for two participants in its logical structure. The clause is causative with an animate and human actor and an animate undergoer.

(2) *Bhásaigh Bran an luch.*
Kill:V-PAST Bran:N the:DET mouse:N
LIT: 'Bran died the mouse'.
Bran caused the mouse to die.

do'(Bran,0) CAUSE BECOME [**básaigh'**(an luch)]

The clause reflects [**±** accidental causation, **±** immediate effects, **±** time lag].

2.1.2 Possible Accidental Causation

In this section we examine accidental causation where, while the result of the causation is not in question, the agency, control and motivated intent of the actor may be in doubt. We can see an example of this in (3) of a causative accomplishment where there may not be strong control or motivated intention by the actor, indicating accidental causation with the possibility of immediate effects for the undergoer.

(O Baoill 1996: 23)

(3) *Muirfidh sé é féin ag obair.*
kill:V-FUT he:PN him:PN self:PART at:PP working:VN
He will kill himself working.

ag'(obair', [**do'**($sé_1$) CAUSE BECOME [**maraigh'**(**é'**($féin_1$))]])

(Foinse, 7 October 2001: 15)

(4) *Mharaigh siad an t-iriseoir.*
Kill:V-PAST they:PN the:DET reporter:N
They killed the reporter.

do'(siad) CAUSE BECOME [**maraigh'**(an t-iriseoir)]

Example (4) contains a transitive clause coding for two arguments and is causative. The expression of causativity in this example allows for accidental or unintentional causation where the intention of actor may be in doubt. There is, however, unquestionably a very strong implication of result, that is, death for the undergoer.

(Míceál O Cíosóg. Annagael: 114)

(5) "*Och, chuala tú faoin bheirt leads óga a* ***mharaigh iad féin*** *toisc nach raibh siad in ann obair a fháil. Bhuel . . . tá Cóilín agus cara leis i ndiaidh an cleas céanna d'imirt.* ***Mharaigh siad iad féin - phlúch siad iad féin*** *istigh i ngaráiste agus gás ag teacht ó inneall gluaisteáin. Maidin inniu.*"

"… have you heard about the two young lads that killed themselves because they could not find any work. Well … Cóilín and a friend have played the same trick. They killed themselves – they suffocated themselves inside a garage with the fumes coming from a car engine. This morning."

Another example is given above and, in this, we can see that the effects of the caused act unfolded over a very brief time period, but is essentially immediate. The intent here appears to be quite intentional on behalf of the actors.

These clauses reflect [**±** accidental causation, **+** immediate effects, **-** time lag].

2.1.3 Strong Agency via Morphological Marking

In the logical structure representation (6) of the verb with the causative prefix *dún*, we use the predicate **DO'**[…] to indicate a very highly motivated actor who is a prototypical agent in the causation, after Dowty (1979) and Van Valin & LaPolla (1997). The causation reflected here is not accidental and the effects on the undergoer are immediate. The verbal noun form of the verb is used in example (7) and (8).

(6) *Dhúnmharaigh an gadaí an cailín.*
Murder:V-PAST the:DET thief:N the:DET girl:N
The thief murdered the girl.

DO'[**do**'(an gadaí, [**dhúnmharaigh'**(an gadaí, an cailín)])] CAUSE
BECOME [**be'**(**dúnmharaithe'**(an cailín)]

(Ní Dhuibhne, Éilís: Dúnmharú sa Daingean: 139)

7) *Deineadh <u>dúnmharú</u> den saghas céanna i dTrá Lí mí ó shin.*

Deineadh *dúnmharú*
Make:V-IMPERS-PASS-PAST murder:VN
den *saghas* *céanna*
of:PP+the:DET same:ADJ kind:N
i *dTrá Lí* *mí* *ó* *shin*
in:PP Tralee:N month:N from:PP then:DET
Lit: '(Someone) made a murder of the same kind in Tralee a month before that'
Someone committed the same kind of murder in Tralee a month before.

(Ó Baille, Ruaidhrí: Dúnmharú ar an Dart: 52)

(8) *Chomh maith leis sin, bhí sé freagrach as ar a laghad sé <u>dhúnmharú</u> le dhá bhliain anuas.*

Chomh *maith* *leis* *sin,* *bhí* *sé* *freagrach*
As:ADV well:ADJ with:PP that:DET be:SUBV he:PN answerable:ADJ
as *ar* *a* *laghad* *sé* *dhúnmharú*
from:PP on:PP that:REL reduce::V-PAST he:PN murder:VN
le *dhá* *bhliain* *anuas.*
with:PP two:NUM years:N up:ADV
As well as that, he was answerable because he committed murder for the last two years.

Because of the constraint of its fixed distribution on the verb coding murder as distinct from killing, this example of direct causation has the strongest agency, intention and control by the actor, and with the strongest implication of the result of the action on the undergoer. The result is so strong that it is not in question. The intensified agency with the causation is morphologically marked with the prefix *dún*. Quite possibly, it is significant that the use of this prefix is reserved for describing the most serious of criminal acts, that of murder. It is this fact which bounds its distribution, presumably as a function of the means by which society recognises and communicates the height of its abhorrence for the taking of human life. These examples of *dún* constructions in clauses reflect [**-** accidental causation, **+** immediate effects, **-** time lag].

The deployment of the causative prefix *dún* is an example of morphological marking on the base verb. The base verb in this instance is, however, already causative and the marking in question indicates a heightened agency, volition, control, intent and motivation. The morphological prefix *dún* is therefore a marker of intense agency. The morphological prefix *dún* signals the highly agentive action that has immediate consequences for the undergoer. We motivated this analysis with an account of three cause-to-die verbs and found that a cline exists with attributes that range over actor and undergoer.

The iconicity pyramid of Haiman (1983) diagrams the vertical cline between less direct to direct causation. The cline we find here in this analysis augments the Haiman pyramid horizontally for morphological causation with respect to accidental/non-accidental causation with immediate/delayed effects on the undergoer.

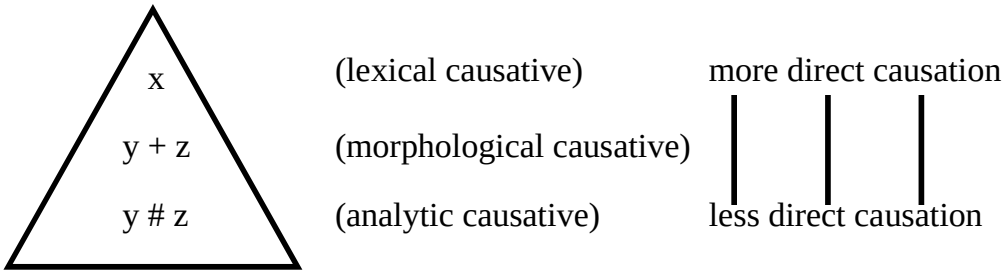


Figure 1. Haiman's Iconicity Pyramid

2.2 Change Marked on Verb to Signal Intensification of Action

The use of the morphological prefix *for* indicates an intensification of the causative force in the act of violence encoded within the base predicate in the clause. The difference between the use of the prefix, and its absence, can be captured using the examples of the verb *éignigh* 'violence' and the verb *foréignigh* 'extreme violence', within the following table.

(9)

	éighnigh	*foréignigh*
Actor	**-** Accidental Causation	**±** Accidental Causation
Undergoer	**+** Immediate Effect	**+** Immediate Effect

We illustrate this in relation to the verb *éignigh* 'force/violence/ravish', which lexically records an act of violence of some kind undertaken by a human and animate actor against a human and animate undergoer. The actor and undergoer roles can be elaborated by instances

of singular or plural count nominals. The verb *éignigh* codes for non-accidental causation with immediate consequences for the undergoer.

(Ó Cíosóg: Annagael: 184)

(10) *Fógraíodh freisin go mbeadh pionós an bháis ann i gcás marú Garda, marú saighdiúra, drugaí a scaipeadh nó a dhíol go mídhleathach, agus i gcás éigniú ban nó seandaoine a ionsaí.*

Fógraíodh *freisin* *go*
Announced:V-PAST also:PART go:PP
mbeadh *pionós* *an* *bháis*
be:SUBV:fut penalty:N the: DEF death:N
ann *i* *gcás* *marú* *Garda,* *marú* *saighdiúra,*
there:ADV in:PN case:N killing:VN Police-Officer:N killing:VN soldier:N
drugaí *a* *scaipeadh* *nó* *a* *dhíol* *go* *mídhleathach,*
drugs:N to:REL spread:VN or: CONN to:REL sell:VN to:PN unfortunates:N
agus, *i* *gcás* *éigniú* *ban*
and:CONJ in:PN case:N rape:VN women:N
nó *seandaoine* *a* *ionsaí.*
or: CONJ old:ADJ+people:N to:REL attack:VN

Lit: 'Also announced was that there will be the death penalty in the case of killing of a police officer, killing a soldier, selling drugs to addicts and in the case of rape of women or attacks on old people'.

Also announced was that the death penalty will apply in the case of killing a police officer, killing a soldier, selling drugs to addicts and in the case of rape of women or attacks on old people.

When the prefix *for* is added to the verb *foréignigh*, the act of violence is signalled as intensified with the same immediate consequences for the undergoer. In addition, as in example (11) and (12), use of the prefix appears to soften the commitment to the agency, control and motivated intent of the actor.

(Cuisle: March 1999: 11)

(11) *Bhí daoine eile sa phobal a mhaígh go mbeadh ar an pbobal roganna eile neamhfhoréigeanacha a aimsiú chun plé leis an choiriúlacht shóisialta, agus í dtús na sosanna cogaidg i 1994 díríodh aird ar chuardach na roganna eile seo.*

There were other people in the community who claimed that the people had other non-violent choices, (and) that aimed for discussion with the social parties, and in the beginning of the truces in the fighting in 1994 focused high on the search of these other choices.

(Cuisle: March 1999: 11)

(12) *Toradh ar chuardach na modanna neamhfhoréigeanacha chun plé le coiriúlachta atá sa tuairisc ar restorative justice. Is iad inclusion agus mediation na príomheilimintí sa chóras.*

The fruits of the searching for non-violent methods for discussion with the parties are in the report on restorative justice. They are inclusion and mediation, the primary elements of the system.

The constructions have the logical structure schema indicated in (13) for *éignigh* and (14) for *foréignigh*, which thereby identifies for each the situation type and its participants. The use of this prefix is indicative of increased intensity of the violent action by the actor. Used with the verb *éignigh*, the prefix is indicative of extremity of force. With respect to example (13), the undergoer participant, the NP *an ban* 'the woman' and the verb *éignigh* 'violate' can form compound into a new verb *banéignigh* 'rape of a woman'.

(13) *Éignigh an saighiúir an ban.*
Force:V-PAST the:DET soldier:N the:DET woman:N
The soldier violated the woman.

do'(an saighiúir, 0) CAUSE BECOME **éignígh'**(an ban)

(14) *Foréignigh na saighiúirí an baile.*
Force:V-PAST the:DET-pl soldiers:N the:DET town:N
The soldiers raised the town.

do'(an saighiúirí) CAUSE BECOME **foréignígh'**(an baile)

We have discussed two morphological prefixes. The prefix *dún*+V morphologically marks the verb as strongly agentive, with all the attributes that that implies. Its distribution is highly restricted. The prefix *for*+V also deploys as a morphological marking on the base verb, but codes for an increased intensity of the caused action of the verb, with a possible weakening of the agency.

3 Summary of Morphological Causation

The cline that represents the verbs of 'cause to die', within which the rightmost *dún* prefix operates, is indicated in the following table. We can see that it codes for non-accidental causation on the part of the actor and immediate causative effects for the undergoer.

(15)

Verb	***Bhásaigh***	***Mharaigh***	***Dhúnmaraigh***
Actor	**±** Accidental Causation	**±** Accidental Causation	**-** Accidental Causation
Undergoer	**±** Immediate Effect	**+** Immediate Effect	**+** Immediate Effect
Time	**±** Time lag	**-** Time lag	**-** Time lag

The cline for the two verbs of violent action is indicated below. In this, we have at one pole non-accidental causation by an actor participant while at the other pole we have the possibility of either accidental or volitional causation by the actor. With both poles we have immediate effects for the undergoer.

(16)

	Éighnigh	***Foréignigh***
Actor	**-** Accidental Causation	**±** Accidental Causation
Undergoer	**+** Immediate Effect	**+** Immediate Effect

In this paper we examined some elements of morphological causatives as found in modern Irish. Two specific instances of morphological markings, ranging over a highly constrained and restricted distribution of verbs pertaining to extreme violence against the person were explained in terms of a cline. We found that the morphological markings appear to be restricted to indicating causation with the highest level of agency, in the case of the *dún* prefix, and to intensified violent action in the case of *for*. With these two prefixes we found that they played a role in indicating a position on a cline for accidental/deliberate causation

on the part of an actor and, for the undergoer, whether the causative effects were immediate or not.

4 References

Ball, J. Martin, and Fife, James. (1993). *The Celtic Languages*. Routledge, London and New York.

Borsley, R. D. and Roberts. I. (1996). *The Syntax of the Celtic Languages: a comparative perspective*. Cambridge University Press. Cambridge, England.

Christian Brothers. (1997). *New Irish Grammar*. C.J. Fallon, Mount Salus Press, Dublin, Ireland.

DeLancey, Scott. (1983). Agentivity and Causation. *Berekely Linguistics Society*, 9: 54-63.

DeLancey, Scott. (1984). Notes on agentivity and causation. *SIL*,8: 181-213.

Dowty, David R. (1979). *Word Meaning and Montague Grammar*. Reidel, Dordrecht.

Gregor, D.B. (1980). *Celtic. A Comparative Study of the Six Celtic Languages: Irish, Gaelic, Manx, Welsh, Cornish, Breton.* The Oleander Press, Cambridge, England and New York, USA.

Haiman, John. (1983). Iconic and Economic Motivations. *Language* 59. 781-819.

Haspelmath, Martin. (1993). More on the typology of inchoative/causative verb Alterations. In Comrie. B. and Polinsky, M. (eds). *Causatives and Transitivity*. John Benjamin Publishing Company, Amsterdam and Philadelphia.

Ní Dhuibhne, Éilís. (2000). *Dúnmharú sa Daingean.* CoisLife, Baile Átha Cliath.

Ó Baille, Ruaidhrí. (1989). *Dúnmharú ar an Dart*. Cló Lar-Chonnachta, Indreabhán, Gaillimhe, Éire.

O Baoill, Donall, P. (1996). *An Teanga Beo: Gaeilge Uladh*. Instidiúid Teangeolaíochta Éireann, Baile Átha Cliath.

O Cíosóg, Míceál. (1997). *Annagael.* Coiscéim. Baile Átha Cliath.

Payne, Thomas. E. (1997). *Describing Morphosyntax: A guide for field linguists*. Cambridge University Press, Cambridge.

Shopen, T. (Ed.). (1985a). *Language Typology & Syntactic Description* Volume 1: Clause Structure. Cambridge University Press, Cambridge MA.

Shopen, T. (Ed.). (1985b). *Language Typology & Syntactic Description* Volume 3: Grammatical Categories and the Lexicon. Cambridge University Press, Cambridge MA.

Song, Jae Jong. (1996). *Causatives and Causation: A Universal-Typological Perspective*. Longman, London and New York.

Stenson, Nancy. (1981) Studies in Irish Syntax. Narr, Tubingen.

Talmy, Leonard. (1976). Semantic Causative Types in *Syntax and Semantics No. 6*. Academic Press, New York.

Talmy, Leonard. (1978). Figure and Ground in Complex Sentences. In J. H. Greenberg (Ed.). *Universals of Human Language iv: Syntax.* Stanford University Press. Stanford.

Talmy, Leonard. (1985). Lexicalisation patterns: Semantic Structure in Lexical Forms in T. Shopen (Ed.), *Language Typology & Syntactic Description 3*: Grammatical Categories and the Lexicon. Cambridge University Press, Cambridge MA.

Talmy, Leonard. (1988). Force Dynamics, in language and cognition, *Cognitive Science* 12:49-100.

Talmy, Leonard. (1996a). Windowing of attention in language in *Grammatical Constructions, their form and meaning* by Shibatani & Thompson. Clarendon Press, Oxford.

Talmy, L. (1996b). Fictive motion in Language and "Ception": The Emanation Type, in P. Bloom et al (Eds.), *Language and Space*. MIT Press, Cambridge MA.

Talmy, L. (2000). *Towards a Conceptual Semantics*, Volume 1 and 2. MIT Press, Cambridge, MA.

Van Valin, Robert D. and LaPolla, Randy J. (1997). *Syntax: structure, meaning, and Function.* Cambridge textbooks in linguistics. Cambridge University Press. Cambridge.

Whaley, Lindsay. J. (1997). *Introduction to Typology: The unity and diversity of language*. Sage Publications, London.

Pitch Circles – From Music Theory To Computer-Based Learning Tool

Matt Smith

School of Informatics and Engineering

Institute of Technology Blanchardstown

Abstract

This paper describes how a music theory with explanatory power for expression of relationships between pitch classes, chords and tonal regions can be exploited as the foundations for a computer-based tool, called 'Pitch Circles', to support musical novices learn about and manipulate such musical concepts and relationships . The paper introduces this research with a brief review of the 'direct manipulation' principles for computer interaction design on which the computer-based learning tool has been based, and of the features of tonal theories which led to our choice of a particular theory, 'Pitch Spaces', as the basis for this work.

Introduction

Lerdahl and Jackendoff's (1983) Generative Theory of Tonal Music (GTTM) was, and remains, an important and ambitious grammatical approach to modelling the structure of Western Tonal Music. In part to address the lack of detail for the 'stability conditions' of the GTTM, another of Lerdahl's contributions to music theory was his proposal for a model of 'Pitch Spaces' (Lerdahl 1988). The Pitch Space model treats pitches, chords and regions in a single framework, overcoming limitations of previous theories. In this paper we will both summarise Lerdahl's arguments for the relative strengths of the model over previous cognitive models of harmonic knowledge, and we will present an extension to the model facilitating the design and implementation of a computer-based tool, called 'Pitch Circles', for harmonic analysis and composition for musical novices. We then describe the current implementation of the computer music tool and make suggestions about how it may be used to support learning of aspects of tonal harmony by musical novices.

Effective tools for learning with computers

'Direct Manipulation' (Shneiderman, 1982) is an approach to computer interface design so users feel as through they are directly performing a task – i.e. they effectively do not notice

they are using a computer but have an experience whereby the artefacts presented by the computer (for example through the monitor or virtual reality headset) respond to their controls as if they were real world objects. Important aspects of direct manipulation interfaces are real-time response, and continual visual/audio communication of the 'state' of the system. Just as when driving a motor car the current state of the car is communicated through the angle of the steering wheel, the position of the gear lever, the pressure and position of the break pedal, and what the driver sees through the windows, direct manipulation computer interfaces likewise present continuous presentation of the state of the virtual device and ways to change that state. Direct manipulation computer systems are appropriate for novice learners (novices to computers or to the domain) since they are easier to learn, and to understand – generally there are no complex 'hidden' aspects to the state of the interactive tool or simulation, and there are no complex symbolic languages to learn to communicate with the system. The 'drag-and-drop' desktop metaphor of modern graphical user interfaces is based on direct manipulation principles, so for example when a folder has been dragged and dropped into the wastebasket the folder disappears and the wastebasket shows that it is no longer empty with a bulging icon, all without the user having to learn or remember commands such as 'delete' or the path to the item being deleted.

When designing tools to aid musical novices learn and manipulate core concepts of pitch classes, scales (regions) and chords, we have taken the approach of searching for a sound theoretical basis on which to design and develop a direct manipulation tool. Thus the relationship between two chords, or a chord and a region, or a pitch class and a chord are to be presented visually (and with audio) in a form that can be manipulated through actions that make sense to the presented representation (concentric circles that can be rotated – as shall be presented later ni this paper).

A computer tool to support learning built upon sound and sufficiently rich theoretical foundations can offer the following benefits:

- easy to use
- easy to learn to use
- a consistent framework for the progressive introduction from simple to more sophisticated concepts and manipulations (i.e. start with few components / controls and add they as needed)

Tonal framework models and motivation for Lerdahl's model

Models and pitch space frameworks

Models are an attempt to present the important characteristics of a concept or class of objects, while ignoring the unnecessary details. Those engaged in modelling attempt to simplify to the point where superfluous details have been removed, and where any further simplification would render the model too simple to be useful. Modelling takes place in the analysis and design activities of many disciplines, for example scale models of proposed building or conceptual models of the relationships between data for database design. Tonal frameworks attempt to present a model of a tonal system – so these tonal models attempt to present simple yet powerful framework for the analysis and modelling of musical concepts. Models arranging concepts in multi-dimensional spaces attempt to create a model to aid explanations of the proximity and distance of pitch classes and tonal regions – pitches and tonal regions arranged a multi-dimensional way so that distance in the space means something. For example, if two items are nearby in the space then the two items are perceptually proximate. Since many musical space models can be presented visually they can form the basis for interactive computer tools for analysis and composition of music. Figure 1 illustrates a screen shot of Holland's (1989) Harmony Space – it is a good example of a direct manipulation tool for learning about music based on Longuet-Higgins' (1962) topological model.

Several different tonal framework models have been proposed, some over 100 years ago, including different models of pitch classes by Balzano (1982) and Shepard (1982), and models of tonal regions such as Weber (1830/1851) and Schoenberg (1911/1978). For example Shepard proposes a three-dimensional model in the form of a double helix, where a chromatic circle forms the X-Z dimensions. The circle is arranged so that pitches on opposite sides of the circle are perfect fifths, and vertically above any pitch is an intersection with the double helix of a pitch one or more octaves higher. Many of these frameworks have the weakness that each only models a single level of pitch description (e.g. pitch classes or tonal regions). One tonal framework that does model multiple levels of pitch description is Longuet-Higgins (1962) – a framework that derives tonal regions from pitch classes. However, Lerdahl cites as a weakness of Longuet-Higgins' model that it cannot be straightforwardly used to describe pitch proximity.

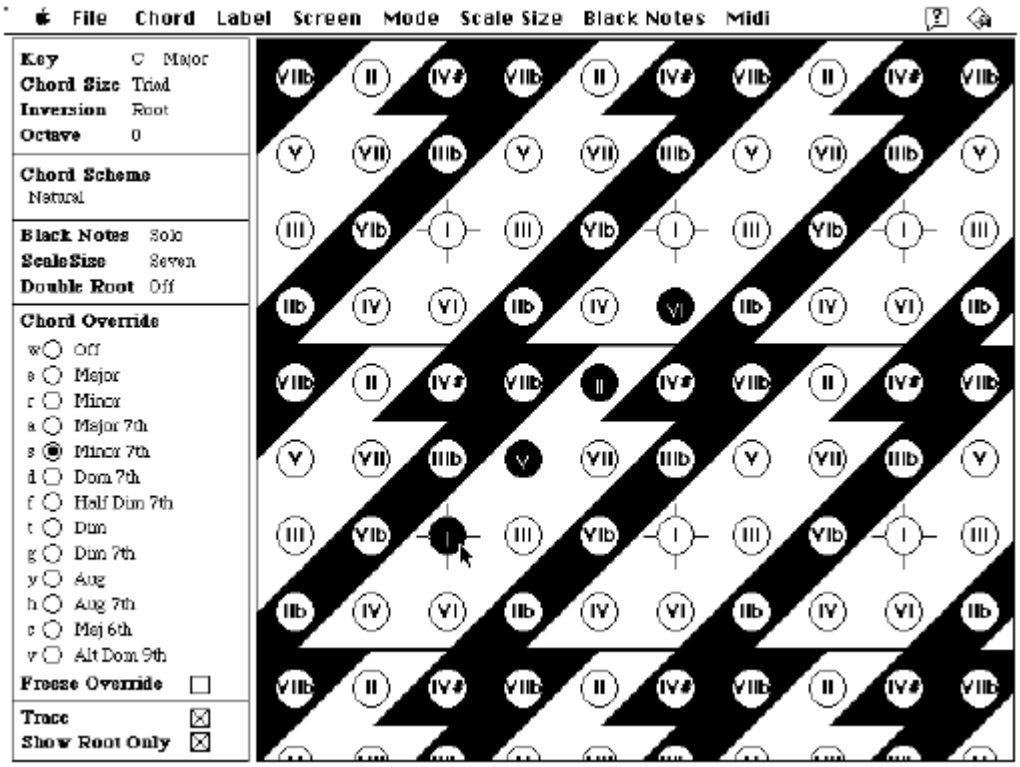


Figure 1: Holland's 'Harmony Space' direct manipulation learning tool.

In addition to the number of levels of pitch description, another important feature of a tonal framework model is whether the model is symmetrical or asymmetrical. While symmetrical models can be simpler and more general for multiple tonal systems, those models that are symmetric can be viewed as misrepresenting the non-symmetrical aspects of any particular tonal system, such as the non-symmetrical aspects of the diatonic system of Western Tonal Music.

Taking as an example the Diatonic scale for the region C major:

C (C#) D (D#) E F (F#) G (G#) A (B^b) B

clearly there is asymmetry since some intervals between scale members are 2 semitones (e.g. from C to C# to D) and between others only a semitone (from E to F). Listeners of many pieces of music from a particular tonal culture will 'overlearn' (Deutsch & Feroe, 1981) pitch, chord and region relationships. Such overlearning is one of the ways the human brain manages to process and make sense of complex perceptual input from the senses. To take a computing science example, in the same way a simulated neural network for voice recognition is trained with representative sets of data (e.g. samples of a person speaking known words) all the Western Tonal Music a person has heard through their life has been tonal training for their perception of diatonic tonal music. Lerdahl makes use of the particular

hierarchical overlearning of chromatic, diatonic and triadic spaces identified by Deutsch & Feroe to form the basis for his proposed pitch space levels.

Motivation and background to the Pitch Space framework

Lerdahl's Pitch Space framework was constructed to combine the strengths and overcome the weaknesses of earlier tonal frameworks. Most importantly the Pitch Spaces framework is able to model, in a single framework, the same asymmetry as the chords and keys (regions) it is modelling in the diatonic system at multiple levels of pitch description. Lerdahl cites several key studies as the background to the development of his model, including descriptions of pitch classes, chord spaces and tonal regions from Krumhansl (1979 & 1983), Krumhansl, Bharucha & Kessler (1982) and Krumhansl & Kessler (1982). Lerdahl's Pitch Spaces have been successfully used to predict phrase structures in a empirical investigation of listener's perceptions of tension in music (Smith & Cuddy, 1997).

Description of Lerdahl's pitch space framework

Building upon the chromatic, diatonic and triadic overlearning ideas previously mentioned, Lerdahl's framework is a hierarchy of five spaces. The hierarchy is such that each level is made up of a strict subset of those pitch classes from the level immediately below. The five spaces are shown in Table 1.

Level	Name
a	Octave space
b	Open fifth space
c	Triadic space
d	Diatonic space
e	Chromatic space

Table 1: Names of pitch spaces

Important points Lerdahl makes about the spaces are as follows:

- except for the chromatic space, the spaces describe the asymmetric patterns appropriate for diatonic music
- the diatonic space is directly represented in the framework (unlike the symmetrical frameworks mentioned earlier)
- the pitch space framework allows unified treatment of pitch class, chord and regional proximity

Lerdahl uses the Roman-numeral notation of chord / region – in this paper we shall further clarify the references to chords and regions by presenting the region numeral in parentheses. For example. I/(I) is the pitch space for the tonic chord (say C major), in the region of the tonic (the C major diatonic region). The choice of the tonic as C is arbitrary, and could be any other pitch class. A general numeric form of reference to pitch classes is adopted for many of the tables of pitch spaces in this paper, the numbers are calculated from semitone intervals from the root. So for a root pitch class of C, the numerals are:

C = 0 C# = 1 D = 2 … G# = 8 A = 9 B = a B = b

Both the numeric and alphabetic forms for the pitch space I/(I) are shown in Table 2.

Space	Pitch Class when C/(C) is chosen as I/(I)											
a	C											
b	C							G				
c	C				E			G				
d	C		D		E	F		G		A		B
e	C	C#	D	D#	E	F	F#	G	G#	A	B^b	B
ed	**0**	**4**	**3**	**4**	**2**	**3**	**4**	**1**	**4**	**3**	**4**	**3**

Space	Pitch Class											
a	0											
b	0							7				
c	0				4			7				
d	0		2		4	5		7		9		b
e	0	1	2	3	4	5	6	7	8	9	a	b
ed	**0**	**4**	**3**	**4**	**2**	**3**	**4**	**1**	**4**	**3**	**4**	**3**

Table 2: Pitch space for I/(I) in numeric and alphabetic formats.

The pitch spaces can be thought of as being repeated in both directions, or alternatively thought as wrapped around by a modulus arithmetic. So, for example, travelling along level d to the right (0 2 4 5 7 9 b), having reached b one would go back to meet 0 2 4 5 and so on.

The bottom (shaded) row in Table 2, ‘ed’, is the 'embedding distance' – this is a measure of how far from the octave space a given pitch class is for a given pitch space. This distance shifts for a given chord and region. The shallower the embedding (the closer a pitch class is to space ‘a’ at the top) the more important the pitch class harmonically for a given space. In the space for I/(I) pitch classes 0 and 7 have the shallowest embedding distances, and are therefore the two most important pitch classes for this space (see Table 2).

Lerdahl explains this vertical embedding distance measure in terms of 'skip' and 'step':

> *"In traditional usage a step occurs between adjacent members of the chromatic or diatonic scales (a chromatic or diatonic step), and an arpeggiation takes place*

between adjacent members of a triad. It is more illuminating, however, to thing of an arpeggiation as stepwise motion in triadic space [space c]. A leap of two octaves, on the other hand, is a skip in octave space [space a]. In sum, a step is adjacent motion along any level of the hierarchy, and a skip is non-adjacent motion – two or more steps – along any level."

[underline emphasis has been added to the quotation] (Lerdahl, 1988, pp. 321-322)

Pitch-class proximity

Using Lerdahl's definition of step and skip, the proximity of two pitches in a given pitch space (e.g. I/(I)) can be measured as a 'step distance' by the number of steps left or right at a given level to get from one pitch to another – e.g. in I/(I) from p0 to p4 is one step in triadic space, two steps in diatonic space and four steps in chromatic space.

Chord proxomity

Triadic chords are found in space c (triadic space), with the root defined by the pitch class in space a (octave space) – for example we can see the Cmaj chord in Table 2 constructed from the root C plus pitch classes E and G (also as 0 4 7 in the more general numeric form). Chord proximity can be calculated using two factors: the diatonic circle of fifths and the number of common tones between the two chords. Lerdahl describes how each of these factors can be modelled via his pitch spaces. He presents the 'chord circle rule', defined as instruction to *"move the pcs [pitch classes] at levels a-c four diatonic steps to the right or left (mod 12) on level d*" (p. 322). Thus there is no change to the diatonic or chromatic spaces when modelling the chord circle. The circle of fifths (See Figure 2) appears as a sequence of pitch spaces when the chord circle rule is successively applied as shown from V/(I) to ii/(I) in Tables 3 and 4.

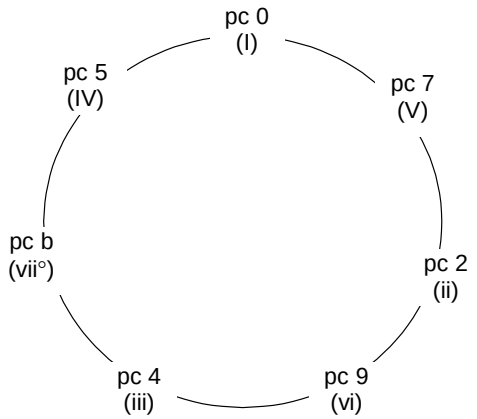


Figure 2: The circle of fifths

Note how movement of levels a-c as steps along level d (the diatonic level) naturally results in the major and minor (and diminished) triadic chords for the key (assuming I/(I) is chord Cmaj in region Cmaj):

- in Table 1 we start with a major chord – I/(I) – chord Cmaj / (region Cmaj)
- after one application (to the right) of the chord circle rule we see in Table 3 – V/(I) – chord Gmaj / (region Cmaj)
- after a second application of the chord circle rule we see in Table 4 – ii/(I) – chord Dmin / (region Cmaj)
- and so on, until:
- after a sixth application of the chord circle rule we arrive at vii°/(I) – chord Bdim / (region Cmaj)

Space	Pitch Class											
a							7					
b			2				7					
c	0		2				7					b
d	0		2		4	5	7		9			b
e	0	1	2	3	4	5	6	7	8	9	a	b

Table 3: Pitch space for V/(I)

Space	Pitch Class											
a			2									
b			2							9		
c			2			5				9		
d	0		2		4	5		7		9		b
e	0	1	2	3	4	5	6	7	8	9	a	b

Table 4: Pitch space for ii/(I)

Lerdahl proposes a straightforward metric for the chord proximity of two chords in the same region – a combination of the shortest number of steps between the chords on the circle of fifths and the number of common/different tones between the chords. The chord circle rule can be used to find the shortest number of steps (in either direction) between chords on the circle of fifths, calculations of this measured for chords from I/(I) are:

- 1 step to V / (I)
- 2 steps to ii / (I)
- 3 steps to vi / (I)
- 3 steps to iii / (I)
- 2 steps to vii° / (I)
- 1 steps to IV / (I)

When measuring the number of common/different tones Lerdahl proposes a measure of the number of different tones in the second chord at all hierarchical levels. For measures of chord proximity in the same region there will be no difference in the chromatic or diatonic levels (levels d and e), therefore the number of different tones in levels a, b and c are counted. Chord I/(I) can be though of as: 0 0 0 4 7 7, a 0 from level a, 0 and 7 from level b and 0 4 7 from level c.

We can therefore calculate the number of distinctively different tones in chords from I/(I) as follows:

- chord V/(I) has 2 2 7 7 7 b so distinctive new pitch classes are 2 2 7 b = 4
- chord ii/(I) has 2 2 2 5 9 9 so all 6 pitch classes are new = 6
- chord vi/(I) has 9 9 9 0 4 4 so distinctive new pitch classes are 9 9 9 4 = 4
- chord iii/(I) has 4 4 4 7 b b so distinctive new pitch classes are 4 4 b b = 4
- chord vii°/(I) has b b b 2 5 5 so all 6 pitch classes are new = 6
- chord IV/(I) has 5 5 5 9 0 0 so distinctive new pitch classes are 5 5 5 9 = 4

A single measure for the proximity (chord distance 'd') of two chords in the same region can be calculated by adding the two values (Lerdahl, p324):

> d (chord) = shortest number of steps on circle of fifths + number of distinctive pitch classes

Applying this formula for each chord with a root in region (I) we get the following:

- chord V/(I) = 1 step + 4 distinct pitch classes = proximity distance of 5
- chord ii/(I) = 2 steps + 6 distinct pitch classes = proximity distance of 8
- chord vi/(I) = 3 steps + 4 distinct pitch classes = proximity distance of 7
- chord iii/(I) = 3 steps + 4 distinct pitch classes = proximity distance of 7
- chord vii°/(I) = 2 steps + 6 distinct pitch classes = proximity distance of 8
- chord IV/(I) = 1 steps + 4 distinct pitch classes = proximity distance of 5

The addition of these two values have been carefully chosen to correspond to cognitive importance, so that proximity on the circle of fifths is not enough, nor just the number of common pitch classes but their hierarchical importance. Thus while vii° is closed than iii on the circle of fifths, their chord proximity calculations (8 and 7 respectively) show than Lerdahl's metric calculates iii as closer to I – this corresponds to our musical perceptions in tonal music. Likewise, while vi shares two common pitch classes with the root (0 4) and V only shares a single pitch class (7) their measure of distinct pitch classes at all hierarchical levels is the same (since 7 is the root of V and reduces the overall number of different pitch classes with the root) therefore, overall, V is found to be perceptually closer to the root than vi due to Vs proximity on the circle of fifths (1 step against vi's 3 steps).

Lerdahl goes further, and defines a measure of chord proximity across regions using a rule that gives a chromatic circle of fifths, and a measure of region proximity. Likewise he suggests how other chord levels could be introduced for modelling particular styles of music for which such chords form a fundamental base. Lerdahl (p320) suggests the introduction of a

seventh-chord level for modelling such musical styles as jazz, Debussy and Ravel, and proposes how seventh and minor chords can be modelled in the framework. One of the claimed strengths for the Pitch Space model is how a single tonal framework can support metrics for pitch-class, chord and region proximity. In this section we have summarised how Pitch Spaces can be used to measure Pitch proximity and Chord proximity with a region. Further details of the Pitch Space theory, including measure of Chord Proximity across Regions and Region Proximity, can be found in Lerdahl's publications, and we plan to demonstration their reification in the tool in our own publications in the future.

The hierarchical and numerical nature of the pitch spaces, and the simplicity of the measurement of pitch, chord and region proximity suggest the use of this framework for computational modelling. The pitch space formalism has strong explanatory power, and as Lerdahl goes on to discuss, appears to correlate with experimental results investigating pitch class stability (see Krumhansl 1979, and Krumhansl & Shepard 1979), multi-dimensional scaling of diatonic triads (Krumhansl, Bharucha & Kessler, 1982) and abstract region spaces (Krumhansl & Kessler, 1982).

Our extended pitch space model – pitch circles

The contribution we make to Lerdahl's pitch spaces is to make the model a circular model, rather than a tabular, linear arrangement. This new model is a non-repeating, two-dimensional model in the form of concentric circles – with chromatic space as the innermost circle, and octave space as the outer one. The result is that steps and skips are seen as motions around the circles to move from one space to another, and rules such as the circle-space rule are seen as rotations of the circles themselves.

An example of the model for I/(I) is presented in Figure 3 (See Figures 11 and 12 for clearer versions of this screen in numeric (0..b) and note-letter (C..B) notations).

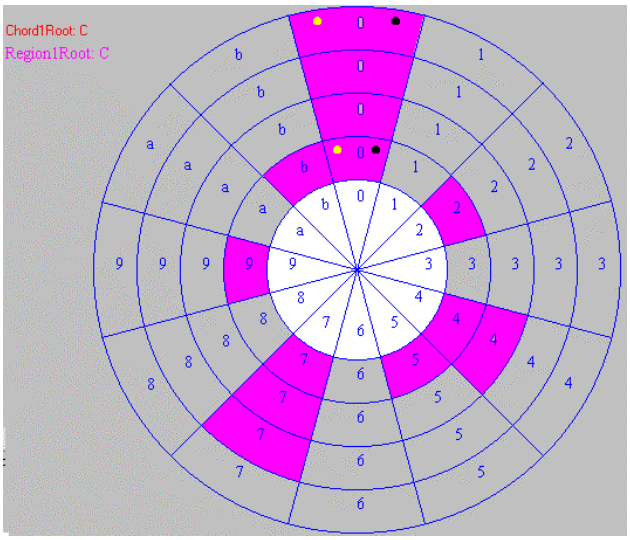


Figure 3: I/(**I**)

The modulus arithmetic required for calculating pitch classes now becomes natural and intuitive with the circular organisation — if it is necessary to go beyond pitch class zero (e.g. C) or pitch class 11 (e.g. B) one just continues to count around the circle in the appropriate direction. This straightforward extension to Lerdahl's original work makes the model much more suitable for computer-based implementation as a tool for music analysis and generation by musical novices.

The current prototype

A direct manipulation implementation of our interactive, circular model of Lerdahl's pitch spaces has been developed (as a Java application). Figures 2 and 3 show the current prototype running with I/(**I**) and V/(**I**) (shown as chords Cmaj and Gmaj in region Cmaj).

Note that at the time of implementation a flat symbol was not straightforward to display, so that all notes are presented in sharp notation (e.g. A# instead of B ^{b}) – the author is aware of this notational inaccuracy and it will be addressed in the next version of the computer tool.

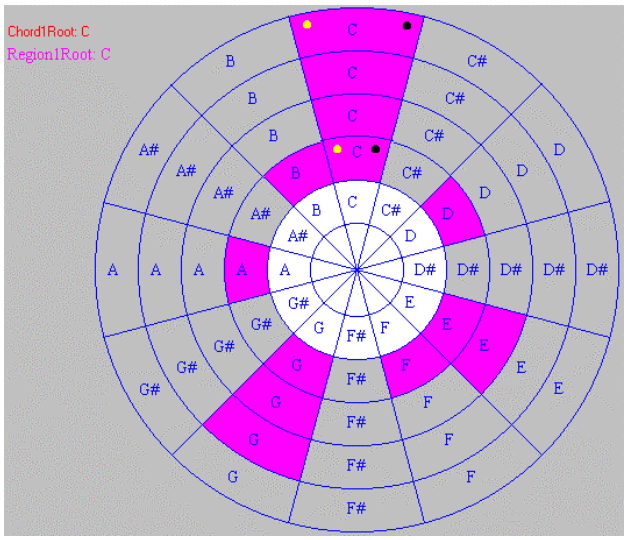


Figure 4: I/(**I**) as C/(C)

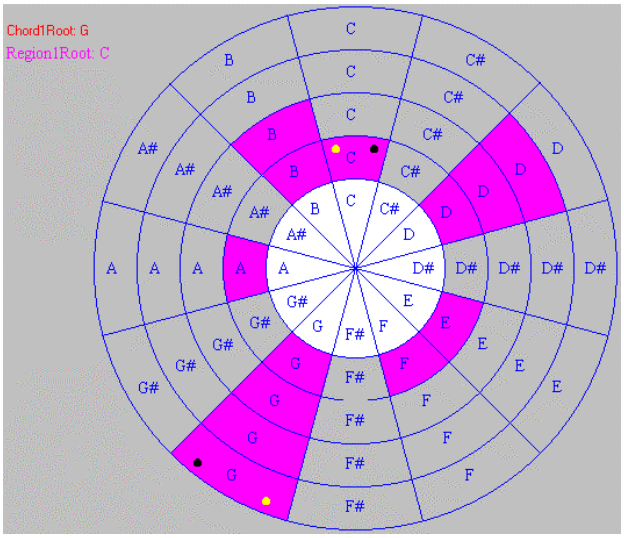


Figure 5: V/(**I**) as G/(C)

As can be seen from the difference between Figure 3 and Figure 4 (and full versions in the web browser in Figures 11 and 12) the prototype allows the user to choose whether to see pitch class numbers (0..b or 0..11) or the note letters assuming pitch class 0 is C.

The user currently has simply facilities such as stepping (rotating) the major chord circles (levels a, b and c) chromatically, and/or stepping the diatonic region circle (level d) chromatically. The chromatic circle is left unchanged. With each move the three notes at triadic level are played as a MIDI chord (in root inversion).

Proposed extensions to the implementation

We are in the process of extending our prototype for use in experiments with novice musicians for the support of simple harmonic analysis and composition tasks. If Lerdahl's claims are correct, and the asymmetry of his model provides a good cognitive fit with human harmonic problem solving, we expect to find positive results from our trials, in comparisons with alternatives such as Holland's (1989) 'Harmony Space' computer tool.

In the next few sections we shall present some examples of the kinds of questions musical novices might be asked, and how they might be able to answer them using the tool. A key feature of the tool being direct manipulation is that a student can be set tasks that are straightforward to understand and undertake, but through which they can discover concepts and rules of the subject domain, i.e. about pitch classes, chords and regions, since tasks on this tool are actions to change the state of a theoretically grounded representation of the relationships of pitch classes, chords and regions.

Chord shapes in a region

The musical novice might be presented with the system set up with Cmaj chord in the region Cmaj (as in Figure 4). In the current implementation any rotations of chord will be in chromatic steps (i.e. the chord shape is fixed as a major chord). A task that could be set to a student with the system constrained this way is:

> *In addition to Cmaj what are the other major chords likely to sound nice in this region of Cmaj?*

The student could then rotate the chord circles ('a', 'b' and 'c') to discover major chords which have all three common pitch classes with the current diatonic region. Rotating clockwise from Cmaj, the student would first come to C#/(**C**) (see Figure 6). Clearly this chord does not share all notes with the region Cmaj, since neither C# nor G# as in Cmaj. Continuing to step the major chord shape (rotate the outer three circles) around the chromatic or diatonic circles the student would first come to chord Fmaj (see Figure 7) then Gmaj (see Figure 5) and find that these two chords, in addition to Cmaj, are the only three major chords that have all notes common with the Cmaj region.

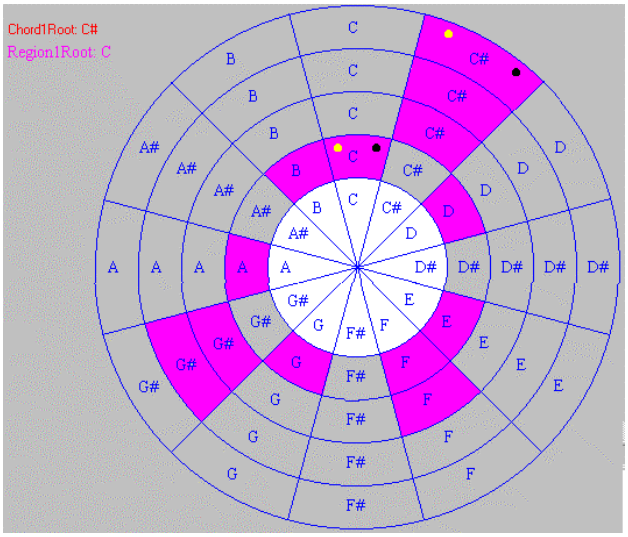


Figure 6: C#/(C)

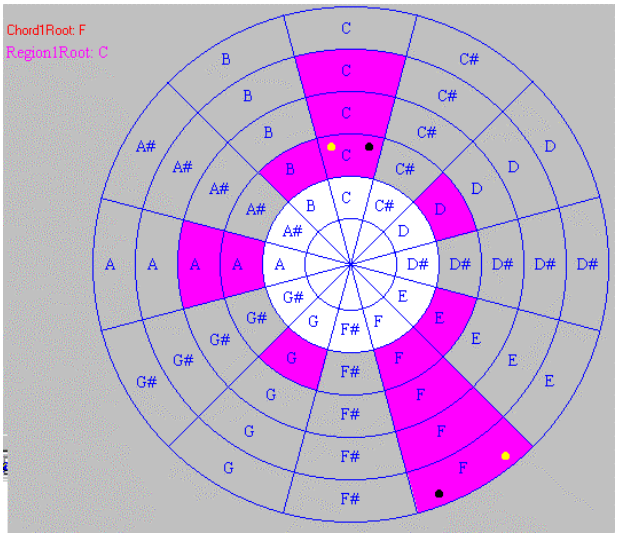


Figure 7: IV/(**I**) as F/(C)

Regions for which a particular chord will fit

Again, working within the current (chromatic) constraints of the current implementation, another question we might ask a student is to find in which other regions a particular chord would share all three notes. For example if, once again, the musical novice is presented with the system set up with Cmaj chord in the region Cmaj they could be asked the following question:

> *In addition to the region Cmaj what are the other regions in which the chord Cmaj is likely to sound nice (i.e. fit all three pitch classes)?*

Rotating the diatonic region circle clockwise from Cmaj, the student would first come to C#maj (see Figure 6). Clearly the chord Cmaj only shares one note with the region C#maj (note C, see Figure 8).

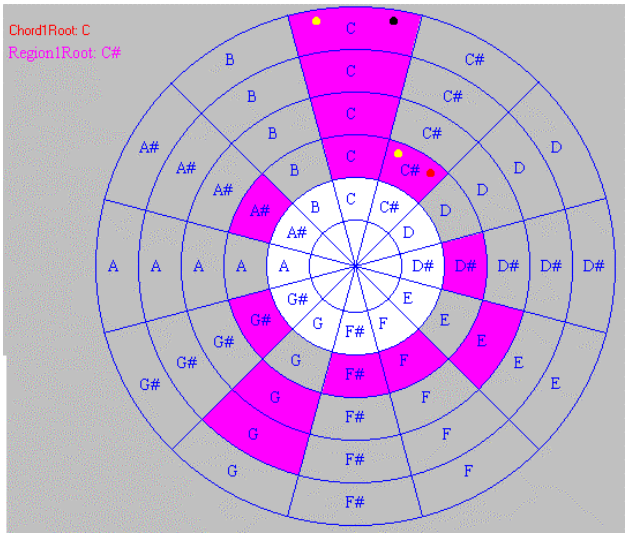


Figure 8: C/(C#)

Continuing to step the region around this way the student will come across the 2 diatonic regions in which chord Cmaj does share all notes — region Fmaj (see Figure 9) and region Gmaj (see Figure 10).

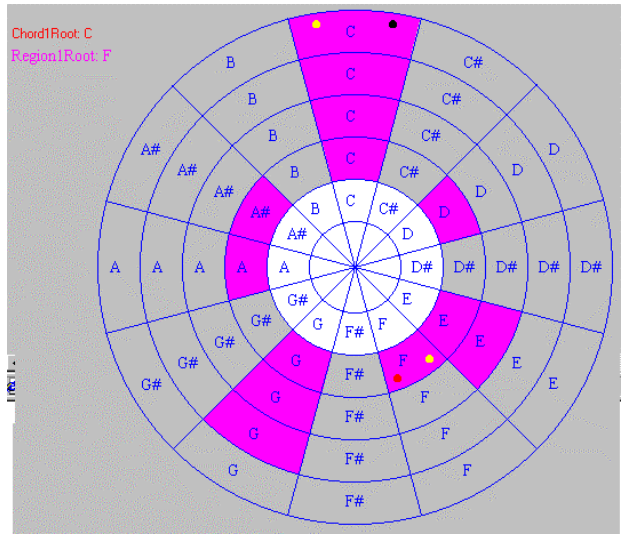


Figure 9: C/(F)

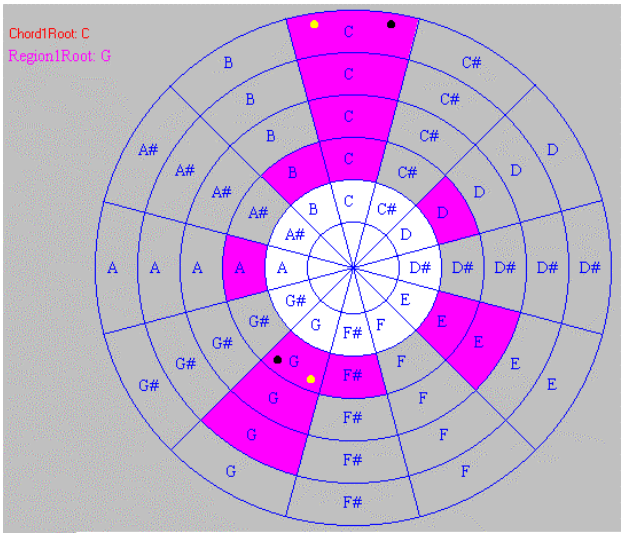


Figure 10: C/(G)

Tasks possible with extensions to the current implementation

If the system has added features to allow straightforward comparison of two regions (perhaps two overlapping pitch circles with different colours/shading), we could ask the student to derive the cycle of fifths as progressions to the most similar regions:

Which are the closest regions to Cmaj — i.e. which regions shall all but one pitch class.

If the system could properly apply the chord circle rule, so that the chord shapes will change between major, minor and diminished as spaces a, b and c are stepped along the diatonic space, then a student could be asked to derive which chords are major, minor and diminished for each region. Both aspects of chord proximity (circle of fifths distance and number of distinct pitch classes) could easily be understood by students using the tool with such facilities.

Conclusions & Further work

Initial, informal experiments with musical novices have been encouraging. Clearly the extended circular model maintains the features and strengths of Lerdahl's original, tabular pitch space model. Once the prototype implementation is complete stand alone and comparative experiments as suggested above will be conducted. The fact that the model and computer program represent the asymmetry of the diatonic system may be important to help students move more easily from theory to practice on physical instruments where such asymmetry is unavoidable. We hope we have illustrated the advantages of building direct manipulation interactive computer-based learning tools on sound and rich theoretical foundations – it becomes straightforward to set simple tasks that enable a student to engage

with the domain theory and relationships directly. With appropriate actions and constraints, actions with the tool are always meaningful in the domain. The challenge now is to add direct manipulation interface components to allow users to perform a range of meaningful actions on the state of one or more pitch circles.

Oversize screen shot figures

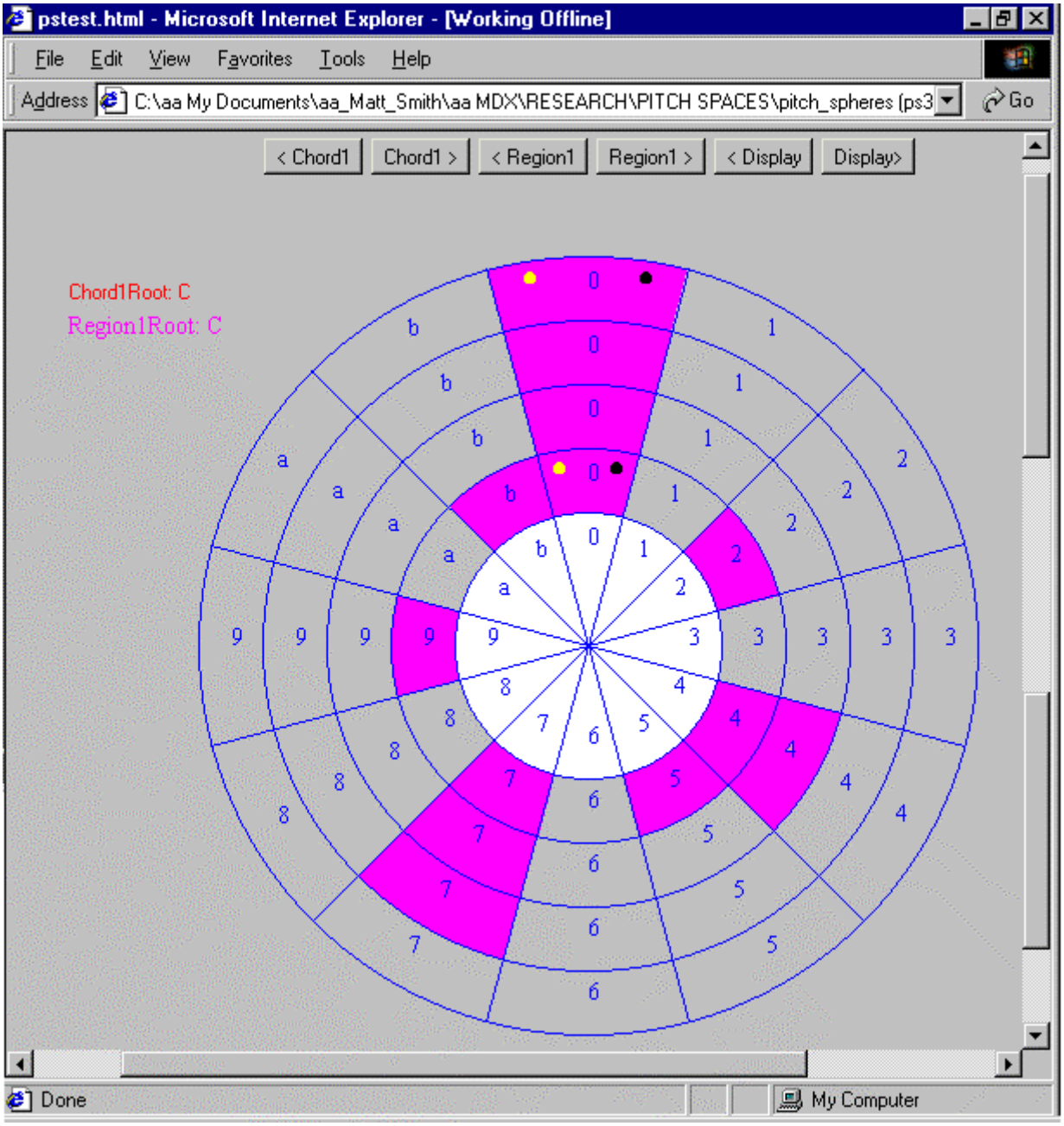


Figure 11: Pitch Circles tool in web browser window, numeric notation mode I/(**I**)

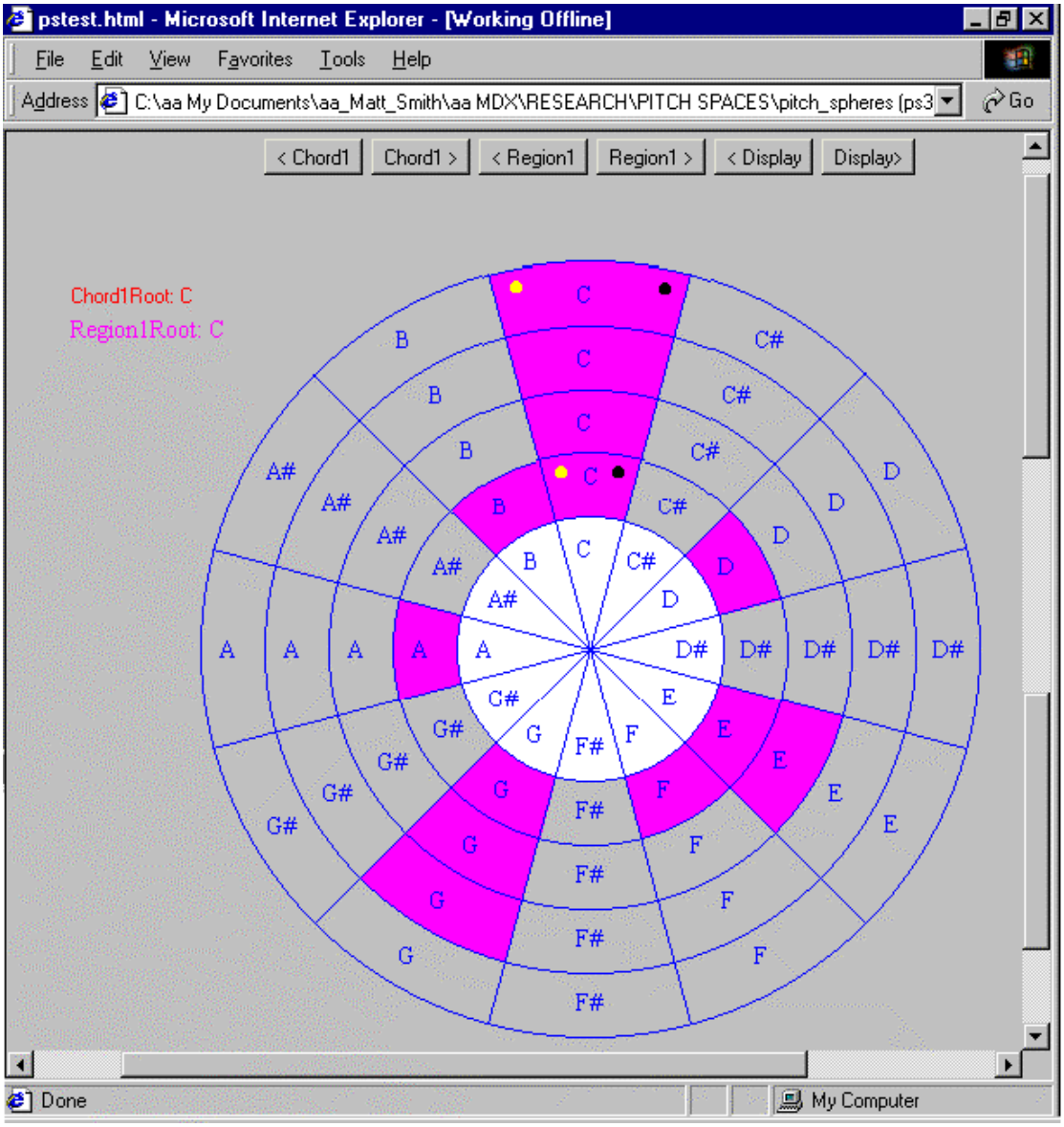


Figure 12: Pitch Circles tool in note letter mode I/(**I**) as C/(**C**)

Related publications

This paper is an updated and extended version of a publication at the Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science, part of AISB-2000, Birmingham, UK, April 2000.

References

Balzano (1982) G. J. Balzano, The pitch set as a level of description for studying musical pitch perception, In *Music, Mind and Brain: the neuropsychology of music*, M. Clynes (Ed.), Plenum, New York, USA.

Deutsch (1982) D. Deutsch, The processing of pitch combinations, in D. Deutsch (Ed.): *The Psychology of Music*, Academic Press, NY, USA.

Deutsch & Feroe (1984) D. Deutsch & J. Feroe, The internal representation of pitch sequences in tonal music, *Psychological Review*, 88:503-522.

Holland (1989) S. Holland, *Artificial Intelligence, Education and Music*, Unpublished PhD thesis, IET, Open University, UK, 1989. , Perceptual structures for tonal music, In *Perception*, 1(1):28-62.

Krumhansl et al. (1982) C. Krumhansl, J. J. Bharucha & E. Kessler, Perceived harmonic structure of chords in three related musical keys, In *Journal of Experimental Psychology: Human Perception and Performance*, 8:24-36.

Krumhansl & Kessler (1982) C. Krumhansl & E. Kessler, Tracing the dynamic changes in perceived tonal organisation in a spatial representation of musical keys, In *Psychological Review*, 89:334-368.

Krumhansl & Shepard (1979) C. Krumhansl & R. Shepard, *Quantification of the hierarchy of tonal functions within a diatonic context*, Presented at the Conference on Music and the Cognitive Sciences, 17-21 September, Cambridge, UK.

Lerdahl (1988) Fred Lerdahl, Tonal Pitch Space, *Music Perception*, 5 (3):351-350.

Lerdahl and Jackendoff (1983) Fred Lerdahl & Ray Jackendoff, *A generative theory of tonal music*. Cambridge, MA: The MIT Press, 1983.

Longuet-Higgins (1962) H. Christopher Longuet-Higgins, Two letters to a musical friend, In *The Music Review*, November 1962, 23: 244-228 & 271-280.

Schoenberg (1911/1978) Arnold Schoenberg, *Theory of Harmony*, originally published 1911. Translated by R. Carter, University of California Press, Berkley, CA, USA.

Shepard (1982) R. N. Shepard, *Mental images and their transformations*, The MIT Press, Cambridge, MA, USA.

Shneiderman (1982) Ben Shneiderman, The future of interact systems and the emergence of direct manipulation, In *Behaviour and Information Technology*, 1:237-256.

Smith & Cuddy (1997) Nicolas A. Smith & Lola L. Cuddy. Patterns of tension/relaxation in music: A consideration of psychoacoustic and cognitive influences. *Canadian Acoustics*, 25(3):38.

Weber (1830/1851) G. Weber, *The theory of musical composition*, Mainz: B. Schotts Sohne. (originally published as *Versuch einer Geordneten Theorie der Tonsetzkunst*, 1830).

Cosmic Radiation

By Dr. Anthony J. Keane*

Introduction to Cosmic Radiation

The planet Earth orbits the Sun in what is often considered to be empty space but is in fact full of very small charged particles speeding in all directions. The situation can be compared to the Earth having to constantly travel in a light shower of rain and with the atmosphere acting like an umbrella. The 'rain' is made up of charged particles called 'Cosmic Rays'. The name *cosmic ray* was given long ago to invisible ionising radiation that could mysteriously discharge an electroscope even when the electroscope was heavy insulated. Early scientists quickly established a relationship between the rate of charge loss and altitude, i.e. a gold-leaf electroscope would loose its charge much faster at the top of a mountain than at sea level, [1]. Many clever experiments led scientists to suspect that invisible radiation was coming from the sky and penetrating the electroscope thus neutralising the charge, [2]. The nature of the ionising radiation and its origin has remained one of the primary areas of research in astrophysics over the last hundred years. This article gives an overview of elemental cosmic radiation and also presents the results from an experiment to measure the upper charge regions of the cosmic radiation spectrum. This work was done at the Dublin Institute of Advanced Studies in a project called the Ultra Heavy Cosmic Ray Experiment, (UHCRE). This experiment involved an extensive study conducted during the 1980s and 1990s with the co-operation of the European Space Agency, which built the satellite (LDEF) and NASA, which deployed and retrieve LDEF from low Earth orbit, [3].

What are Cosmic Rays

In explaining what constitutes elemental cosmic rays we need to recall the periodic table. This is an ordered list of the elements that make up all matter in the Universe. The first element is Hydrogen, then Helium, Lithium, and so on. Each element in the list has a number according to the number of protons it contains. Hydrogen has one proton, Helium has two protons, Lithium has three protons and so on, see figure 1.

The proton is a positively charged particle and to make an atom (which is neutrally charged) the protons need an equivalent number of negatively charged particles called electrons. Each electron has an equivalent charge magnitude to the proton but electrons are very much smaller in size than protons, so much so, that it takes almost 2000 electrons to be equal in mass to one proton.

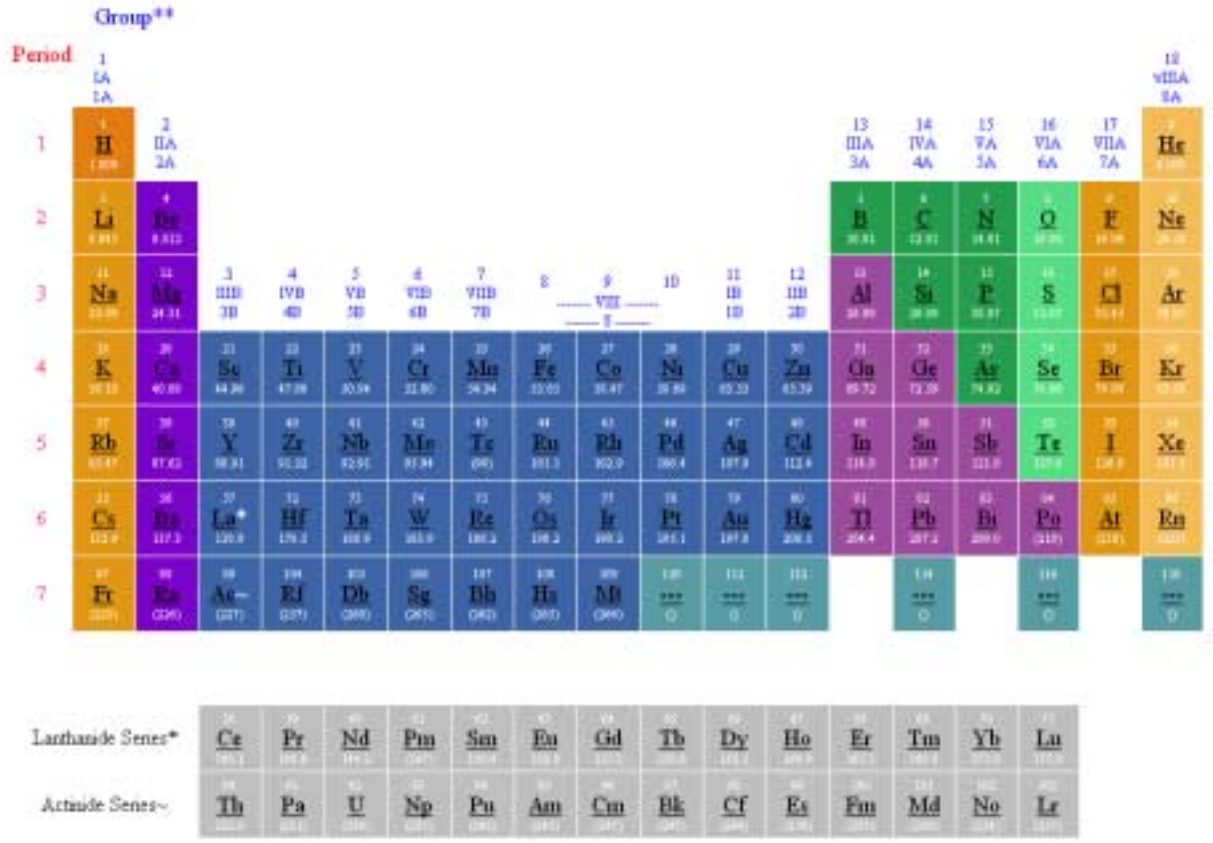


Figure 1: Periodic Table of the Elements

This means that the mass of electrons do little to change the mass of a neutral atom. Electrons orbit the central positive core (called the nucleus) of an atom in discrete shells. The nucleus contains protons and neutrons. Neutrons are particles similar in size to the proton but are neutrally charged. The neutrons are used to prevent the nucleus of an atom from breaking up due to the repulsive positive charges of the protons. Elemental cosmic rays consist of all the elements from the periodic table and these elements, (or particles) can exist in varying degrees of ionisation. An ionised particle is where one or more electrons have been removed from the neutral atom giving the atom a net positive charge.

Origin of Cosmic Radiation

To talk about the origin of cosmic rays we need to look at how matter is created. In the beginning of the Universe, according to the Big Bang theory, there initially existed just energy at very high temperatures, too high for matter to exist. As cooling occurred, basic constituents of matter were formed, like electrons and protons. Mutual attraction between the particles led to the formation of simple atoms like hydrogen, helium, and their isotopes. To create atoms bigger than Lithium requires a mechanism that can add protons to the nucleus of an atom but the problem is that a nucleus has a net positive charge and easily repels any

approaching proton. The proton needs sufficient energy to overcome the repulsion force of the nucleus to gain entry. Neutrons on the other hand do not experience any repulsion from positively charged nuclei. Also neutrons can change into protons while inside the nucleus thus allowing one element to change into another element (the dream of the alchemists).

In the beginning of the Universe there was just Hydrogen, Helium and their isotopes that formed into clouds. When a cloud had sufficient mass it would collapse into a denser cloud and so on until the temperature at the centre of the cloud was high enough to start a fusion reaction and thus a star is born. A star has a life cycle; it is born, lives by burning its stellar material and dies when the forces of radiation that hold-up a star can no longer resist the gravitational forces that are compressing the star. This creates a cycle in which stellar material is reused again and again in a true recycling process. Many stars are good at producing elements from Lithium up to the iron group, [4]. To make bigger elements than the iron group requires a more energetic process than found in the centre of a star. One way sufficient energy is supplied is when a star explodes. Stars come to the end of their life in many different ways but the most impressive way is when the star explodes with the brightness of a thousand galaxies called a supernova, [5]. Supernova can be so bright that we can easily see them with the unaided eye, even during daytime hours. Supernova cause the individual parts (the atoms and molecules) to be split apart by the tremendous energy and flung into space with a great velocity. During this process there is the opportunity for new elements to be created with higher proton numbers than was originally in the star itself through a process called nucleosynthesis. Stars are factories for matter production and primary sources of cosmic rays. Other sources that contribute to cosmic radiation are partially ionised particles that gain energy by passing close to our Sun and particles and dust grains accelerated at the edge of the Solar System (called anomalous cosmic rays), [6].

Cosmic Ray Energies

Each cosmic ray source is distinguished by the energy (how fast the cosmic ray is travelling) of the cosmic rays. Energies are measured using a unit called the electron volt, eV. An electron volt is the energy that an electron gains when it travels through a potential of one volt. You can imagine that the electron starts at the negative plate of a parallel plate capacitor and accelerates to the positive plate, which is at one volt higher potential. Numerically one eV equals 1.6×10^{-19} joules or a joule is 6.2×10^{18} eV. For example, it would take 6.2×10^{20} eV/sec to light a 100 watt light bulb.

Space is essentially empty so there is little material that can get in the way and slow down the particles from supernova as they are scattered throughout the Universe. Some of these

particles will become seed material for other stars while some remain travelling through the Universe moving with a high velocity approaching the speed of light. Interactions with magnetic fields add further energy to the particles but the exact mechanism that allows particles to achieve such high velocities, such as 10^{21}eV, is not fully understood even today.

Solar cosmic rays generally have low charge and low energy (~1MeV) and are about half ionised. Anomalous cosmic rays have energies around 10GeV and are singly ionised. Galactic cosmic rays have energies in the approximate range of 1GeV to 10^{15}GeV and are fully ionised. Figure 2 shows a log-log plot of the flux of cosmic rays bombarding Earth as a function of their energy per particle. Researchers believe cosmic rays with energies less than ~3x10^{15} eV come from supernova explosions while the origin of cosmic rays with energetic more than 10^{18} eV (beyond the "knee" in the figure 2) remains a mystery.

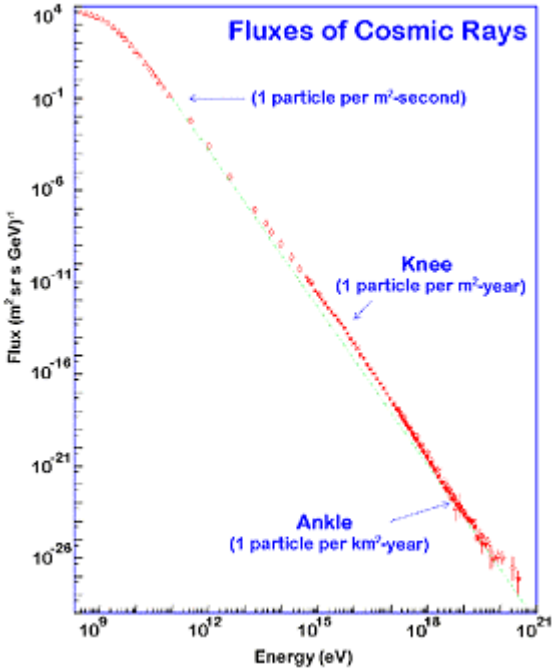


Figure 2: Energy Spectrum of Cosmic Rays

Detection and Measurement of Cosmic Radiation

The primary interests for cosmic ray scientists are the charge, energy and isotropic spectra for both galactic and solar cosmic radiation. The measurement of these spectra has proven to be a long and difficult effort and much still remains to be done. The difficulty in measurement lies in the wide range of charges and energies of cosmic rays. Also the numbers of the particles for each elemental type fall dramatically after hydrogen and helium and this makes detection even more difficult. In figure 3 we see that solar particles are rich in Hydrogen and Helium relative to galactic cosmic rays, probably due to selective acceleration mechanisms.

However, lithium, beryllium, and boron are more plentiful in galactic cosmic rays than in the Solar System.

There is no one instrument that can measure everything. In fact sometimes there is no instrument at all that can measure some of the data while the use of other instruments is prohibited due to the need to place the instrument outside the Earth's atmosphere in order to make the measurements. The Earth's atmosphere acts like a shield against the cosmic rays, it absorbs some cosmic rays, causes others to have collisions with the molecules making up the atmosphere and is partially invisible to other very small cosmic rays, some of which can travel right through the Earth. Some experiments conducted on the Earth's surface operate by detecting the sub-particles created in a collision between a cosmic ray and a molecule high up in the atmosphere (called a cosmic ray shower) and trying to decipher the original cosmic ray from the reconstruction of the collision. The highest energy ever attributed to a cosmic ray was detected using this method, [7]. The best place to measure the cosmic rays is outside the interference of the Earth's atmosphere. To put detectors in Space requires the help of space agencies like ESA and NASA. Satellites often carry instruments designed to measure features of the energy, charge, or isotropic spectra. Sometimes a complete satellite is dedicated to the measurement of cosmic rays and one such satellite was the Long Duration Exposure Facility, LDEF.

Long Duration Exposure Facility

LDEF was a specially designed satellite that was launched into low Earth orbit in 1984 and retrieved in 1990 having orbited the Earth some 3000 times, (see figure 4). LDEF contained fifty-seven individual experiments that made a wide array of measurements of galactic and solar cosmic radiation. A team of scientists headed by O'Sullivan and Thompson from the Dublin Institute for Advanced Studies designed the single largest experiment (UHCRE) on the LDEF. It covered one-fifth of the surface area of the LDEF and it was designed to measure the charge abundance of galactic cosmic rays with a charge greater than sixty, [8]. Previous attempts by prestigious research groups like Price et al. (Skylab) [9] and Binns et al. (HEAO satellite) [10] suffered from short duration exposures with small area collectors in their experiments which resulted in low statistics of cosmic rays for charges greater-than sixty. LDEF allowed for a long duration (six years) large area detector (10 meters squared) to collect primary cosmic rays without the interference of the atmosphere.

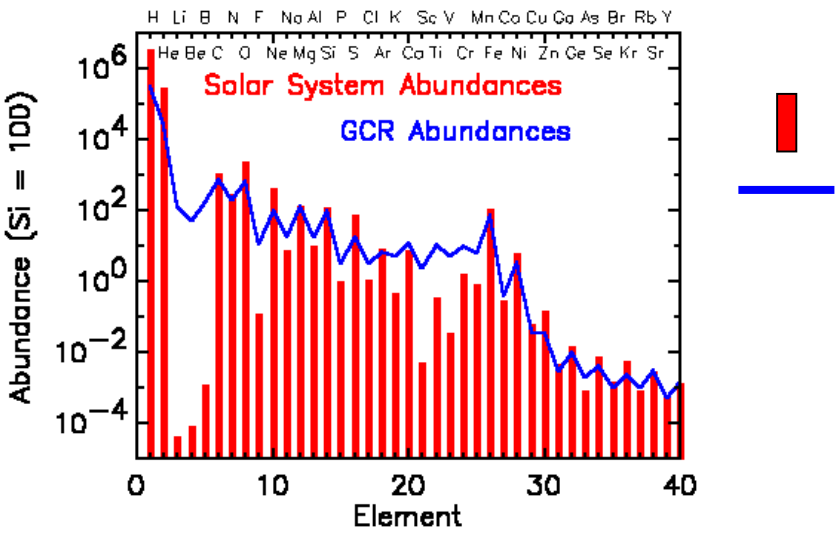


Figure 3: Abundance Spectrum of Cosmic Rays

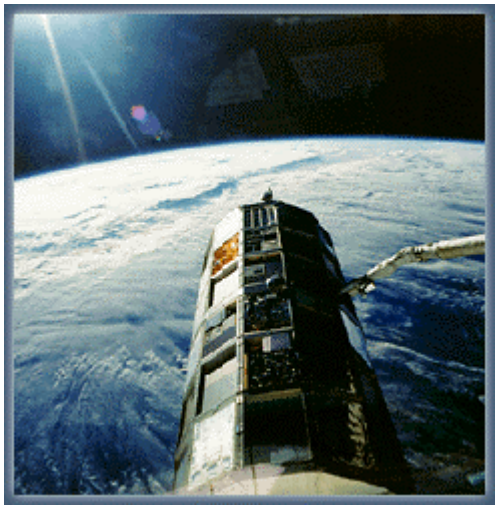


Figure 4: LDEF being deployed in Earth orbit by Shuttle *Challenger*

Data from over 3000 galactic cosmic rays with charge greater than sixty were collected by UHCRE. Among the UHCRE data was evidence of approximately thirty very rare cosmic rays called actinides (charges greater than 89). The total combined world sample of actinides previously detected was three so the UHCRE delivered ten times the world sample or a 1000% increase in data. UHCRE was the first experiment to measure sufficient quantities of ultra heavy cosmic rays that allowed for good statistical analysis to be carried out, (see figure 5), [11].

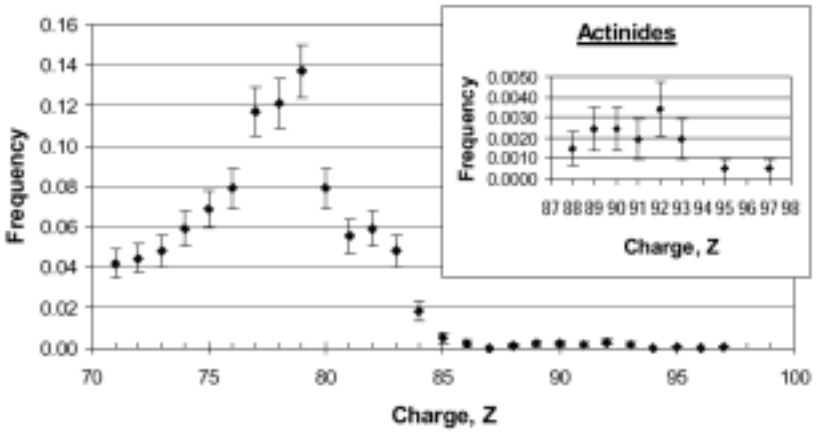


Figure 5: Charge frequency distribution of charges over 70 with an inset showing an enlargement of the histogram in the actinide region.

UHCRE detectors were made of a material called Lexan. This is a plastic type material that has some interesting properties. It was discovered in the 1960s that Lexan and other plastics recorded the passage of ionised particles and the trajectory of the particle could be enlarged using chemicals and viewed under a microscope, [12]. By careful controlled treatment of the plastics in the track enlargement process, (called track etching), parameters of the trajectory could be measured to give information leading to the identification of the energy and charge of an unknown particle. The sensitivity of Lexan was tested in various environmental conditions as a function the charge and energy of known particles, [13]. This information was used to predict the response of the detector to any particle within the range of charge calibration, (see figure 6).

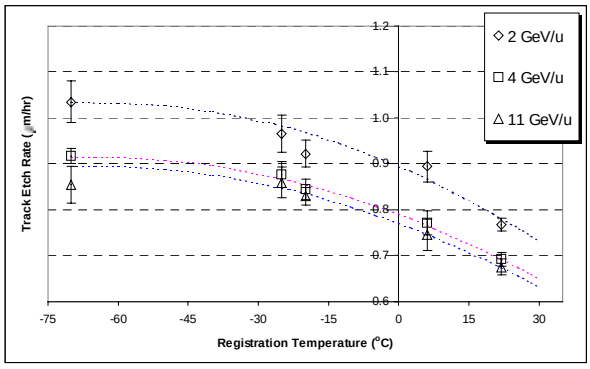


Figure 6: Registration temperature effect using relativistic gold at three different energies

Information from Cosmic Rays

Unfortunately cosmic ray trajectories do not point back to their sources so we must use indirect methods to determine their sources and the way they have propagated through the Galaxy. The chemical composition of the cosmic rays can provide a source for propagation

studies. The chemical composition of the solar system has been determined from a combination of spectroscopy on the Sun, studies of the solar wind, and by chemical analysis of meteorites, which are presumed to have a purer sample of the early solar system than terrestrial rocks, [14]. The composition of cosmic rays is important because cosmic rays are a direct sample of matter from outside the solar system and contain elements that are not seen in spectroscopic lines from other stars. Cosmic rays also provide important information on the chemical evolution of the Universe. If we look at the elemental composition measured for cosmic rays and compare it to our best understanding of the composition of the solar system, we quickly see some large differences. Information regarding the creation and existence of the elements can be found from the relative abundance of the cosmic rays. Some isotopes are radioactive so their existence or lack of it in cosmic ray samples can be used to estimate the age of the cosmic rays.

Some Final Comments

Just over one hundred years of investigation into cosmic rays has provided answers to many questions of phenomena observed on Earth, some examples being the discharging of electroscopes in laboratories and on a bigger scale, the cause of the Aura Borealis or otherwise known as the Northern lights. Many big questions still remain to be answered on the origin, composition, and propagation of cosmic rays. For reasons of simplicity and space in this article, I have completely ignored a whole area of cosmic ray research involving subatomic particles.These particles are much smaller then elemental nuclei and are often the result of collisions between other particles but are important for testing hypothesis of atomic models. This, as they say is another story. The investigation of elemental cosmic radiation has recently taken on a more important and somewhat urgent role to that of simple academic curiosity. It involves the conquest of Space. As we have noted already, outside the safety of the Earth's atmosphere is a constant 'rain' of charged particles that can easily penetrate many grams of thickness of lead. Astronauts and their spacecraft, as well as electronic equipment on satellites are vulnerable to long term exposure of cosmic rays and are particularly sensitive to high fluxes of charged particles like those that occur in solar flare events. The Earth's magnetic field can provide some protection from low energy particles for satellites in near-Earth orbit as long as the solar flare has the same magnetic polarity as the Earth's field. Periodically, a solar flare is ejected from the Sun with ionised gas in a magnetic cloud of opposite polarity to the Earth's magnetic field and this has resulted in wide spread blackouts in terrestrial power and communication systems. Research into early detection and measurement of the polarity of ionised gas clouds is taking place and this may result in satellites and power stations taking some form of action to minimise damage. Another

important question being asked today is whether astronauts would survive the high doses of radiation from a manned mission to Mars given the amount of cosmic radiation they would be exposed to during their extended journey time. Other research involves measuring the amount of cosmic radiation that is penetrating aircraft that fly at high altitudes, [15]. Our understanding of cosmic rays is proving to be vital for future expansion in communications and exploration of our Solar System.

References

[1] Y.Sehido and H.Elliot, "*Early history of cosmic rays*", Astrophysica and Space Science Library, D.Reidel Publ. Co. Vol. 118 (1985).

[2] R.A.Millikan and G.H.Cameron, Phys. Rev.,2nd Ser. 28, p.851 (1926)

[3] A.J.Keane, A.Thompson, D.O'Sullivan, L.O'C. Drury and K-P Wenzel,"*A charge spectrum of ultra heavy cosmic ray nuclei, including actinides, detected on the LDEF*", Publ. Proc. 25th ICRC vol. 3 pp361-364 (1997).

[4] E.M.Burbidge, G.R.Burbidge, W.A.Fowler and F.Hoyle, "*Synthesis of the elements in stars*" Rev.Mod.Physics, 29 p547 (1957)

[5] J.P.Meyer, L.O'C.drury and D.C.Ellison,"*Galactic cosmic rays from supernova remnants*" Ap.J., 487 (1997)

[6] L.O'C.Drury and A.J.Keane,"Ultra heavy nuclei in the galactic cosmic rays", Nucl.Phys.B (Procc Suppl.), 39A pp165-170 (1995)

[7] J.W.Cronin, T.K.Gaisser and S.P.Swordy,"*Cosmic rays at the energy frontier*", Scientific American pp32-37 (1997)

[8] D.O'Sullivan, A.Thompson and K-P Wenzel,"*The LDEF ultra heavy cosmic ray experiment*" First LDEF Post Retrieval Symposium - part 1, p367-375, NASA CP 3134 (1991)

[9] E.K.Shirk and B.P.Price,"*Charge and energy spectra of cosmic rays with Z>60: The Skylab experiment*" Ap.J., 220, pp719-733 (1978)

[10] W.R.Binns, T.L.Garrard, P.S.Gibner, M.H.Israel, M.P.Kertzman and C.J.Waddington, "*Abundances of ultra heavy cosmic elements in the cosmic radiation: Results from HEAO 3",* Ap.J., 346, pp997-1009 (1989)

[11] J. Donnelly, A. Thompson, D. O'Sullivan, A.J. Keane, L. O'C. Drury and K.-P Wenzel, *"New Results on the Relative Abundance of Actinides in the Cosmic Radiation"*, Proc. 26th Int. Cosmic Ray Conf. (Salt Lake City), Vol 3, pp 109-112 (1999)

[12] A.J. Keane, D. O'Sullivan, A. Thompson, L. O'C. Drury and K.-P. Wenzel *"Application and Analysis of SSNTD in the investigation of Ultra Heavy Cosmic Rays in the Dublin-ESTEC LDEF experiment"*, Radiation Measurements, Vol 28, pp 329-332 (1997). [13] A.J. Keane, A. Thompson and D. O'Sullivan,*"Investigation of the Response of Lexan Polycarbonate to Relativistic Ultra Heavy Nuclear Particles"*, Radiation Measurements, Vol 31, pp 601-604 (1999).

[14] E.Anders and N.Grevesse "*Abundances of the elements: Meteorite and Solar*" Geochim. Cosmochim Acta., 53, pp197-214 (1989).

[15] D O'Sullivan, D Zhou, W Heinrich, S Roesler, J Donnelly, R Keegan, E Flood and L Tommasino,*"Cosmic Rays and Dosimetry at Aviation Altitudes*, Radiation Measurements, Vol 31, pp 579-584 (1999).

*Dr A.J.Keane was a researcher at the Dublin Institute for Advanced Studies, School of Cosmic Physics, from 1993 to 1998 where he worked on the extraction and analysis of the UHCRE data. He currently works as a lecturer in the School of Infomatics and Engineering, Institute of Technology Blanchardstown.