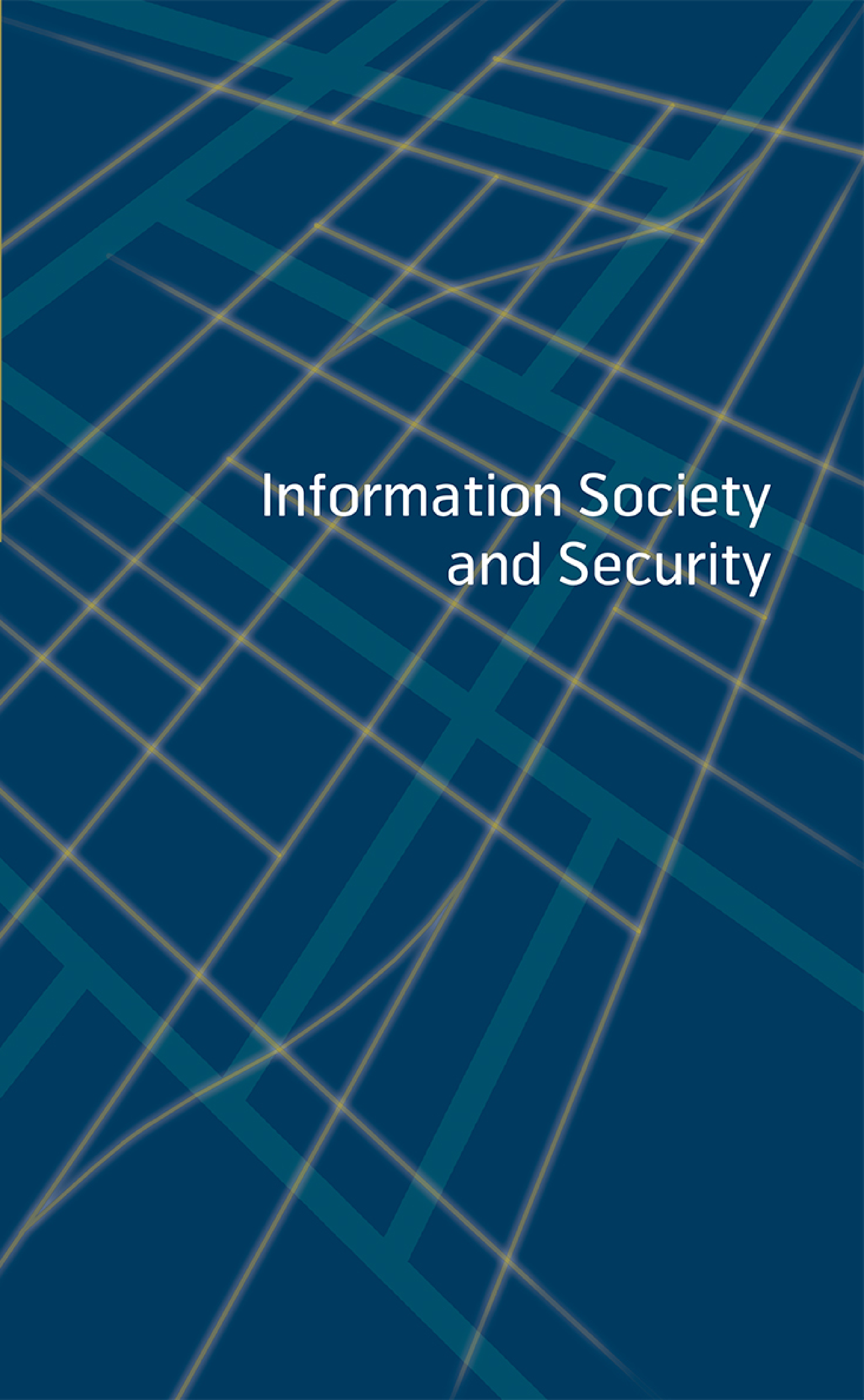
telektronikk

1.05

Information Society and Security



Telektronikk
Volume 101 No. 1 – 2005
ISSN 0085-7130

Editor:
Per Hjalmar Lehne
(+47) 916 94 909
per-hjalmar.lehne@telenor.com

Editorial assistant:
Gunhild Luke
(+47) 415 14 125
gunhild.luke@telenor.com

Editorial office:
Telenor R&D
NO-1331 Fornebu
Norway
(+47) 810 77 000
telektronikk@telenor.com
www.telektronikk.com

Graphic design:
Design Consult AS (Odd Andersen), Oslo

Layout and illustrations:
Gunhild Luke and Åse Aardal,
Telenor R&D

Prepress and printing:
Rolf Ottesen Grafisk Produksjon, Oslo

Circulation:
3,500

Networks on networks

Connecting entities through networks – in technological, societal and personal terms – enables telecommunication. Networks occur on different levels, form parts of larger networks, and exist in numerous varieties. The artist Odd Andersen visualises the networks on networks by drawing interconnected lines with different widths. Curved connections disturb the order and show that networks are not regular but are adapted to the communication needs.

Per H. Lehne, Editor in Chief

Contents

Information Society and Security

Section 1 – Risk and information security

Section 2 – Structural vulnerability

Section 3 – Three cases

Special

Status

Guest editorial

JAN A. AUDESTAD AND SVEIN E. PETTERSEN

Jan A. Audestad is Senior Adviser in Telenor

Svein E. Pettersen is Managing Director of Gjøvik Science Park

During the last ten years, ICT has changed the way in which we manage society. The change that has taken place is not only huge but also irreversible. During this period the average computational power per computer has become larger by a factor of 1000 (Moore's law), and the number of equipment which contains a Central Processing Unit (CPU), and thus can be regarded as a computer, has increased by a factor somewhere between 1000 and 10,000. This means that the available computational capacity on Earth has increased by a factor of more than one million during these ten years – corresponding to a doubling more often than twice a year. During the next five years, it is expected that the available computer power will increase by another factor of 1000.

A conservative estimate of the total number of active CPUs on Earth is 10,000 billion devices. Of these, only one billion CPUs (or less than 0.01 %) are personal computers; that is, devices directly controlled by people. Most of the CPUs are autonomous devices contained in sensors, control systems, infrastructure, production facilities, vehicles, smart cards, cameras, and so on and so forth.

All this computational capacity is interconnected by the internet. Every pair of computers may then, in principle, exchange information and commands making up a formidable computational resource. Only a very small part of this capacity is actually being used but work is going on under headings such as grid computing, agent networks and versatile computing by which this formidable resource may be utilised more efficiently. CERN[1] alone is extending its current supercomputer consisting of a network of 20,000 PCs to a global grid of more than 100,000 PCs in order to analyse particle collisions in the Large Hadron Collider that will commence operation in 2007. It is expected that a large number of grids and agent networks is established before 2010 in order to make huge computations in biology, ecology, astronomy, meteorology and social sciences possible. These networks will have massive impact on information security issues such as nondisclosure of data and protection of computational algorithms, preservation of data integrity, protection of authentication keys and passwords, anonymity and accountability, autonomous micro-payment systems supporting economic compensations for the use of other people's computers, protection against denial of service attacks, and development of countermeasures against parasitic computation and illegal storage of data.

A secondary effect of the evolution of ICT is the computerisation of society. All production and delivery of goods and every public and private service depend critically on ICT. As a consequence, if the ICT systems go down, most of society will run into severe problems and probably come to a halt until we have restored the ICT system or gone back to the old way of doing things. The latter may seem simple but certainly is not because we no longer have the skills and manpower nor the technology to run society in the way we did ten years ago.

Another worry is that we are no longer at grips with all this complexity. We may understand, on the one hand, how the overall infrastructure works and, on the other hand, how all individual software pieces, hardware devices and subsystems making up this complex infrastructure are designed, but we have very little (if any) knowledge of how these viewpoints of the same system are interrelated. We are facing a complexity barrier that is difficult to cross; not because our knowledge of details and processes is inferior but because the relationship between the pieces and the whole is so overwhelmingly complex. This is a feature common to all complex systems: relationship between the motion of single particles and turbulent fluid flow, relationship between properties of atoms and behaviour of macroscopic bodies, and the relationship between neurons and conscience. Complexity gaps exist in every science, and the study of complexity has become a science in its own right.

From the viewpoint of information security, it is not enough to protect single computers but also to protect the network itself. This became evident in 1999 when Albert and Barabási discovered that certain networks, among them the World Wide Web and the internet, had a property that became known as *scale-freeness*. It was well known that the number of links terminating at a node (called the degree of the node) in a normal random network (called Erdös-Rényi (E-R) graphs after the two mathematicians who were the first to study random graphs in 1959) followed a

[1] *Conceil Européen pour la Recherche Nucléaire – elementary particle physics laboratory in Geneva.*

ISSN 0085-7130 © Telenor ASA 2005

Poisson distribution. One property of the Poisson distribution is that the probability of finding a node with degree significantly larger than the average degree falls off faster than the exponential rate (g^{-g} for large g where g is the degree of the node). The probability of finding nodes with large degrees is then very small and the graph is uniform in the sense that any sufficiently large segment of an E-R graph is statistically similar to the whole graph.

What Albert and Barabási found was that the probability of finding nodes with large degrees in certain networks such as the World Wide Web fell off at a much slower rate, namely (approximately) as a power law $g^{-\gamma}$ where g is the degree and γ is a constant usually between 2 and 3. This gives rise to non-uniform graphs where different segments are certainly not statistically similar. These graphs contain a significant number of nodes with very large degrees. Such nodes are referred to as hubs.

One surprising property of scale-free networks is that they are at the same time very robust against attacks randomly destroying nodes but exceptionally vulnerable to attacks targeted at the hubs: if the search engines of the World Wide Web are removed, the Web will dissolve into small islands of web pages impossible to find. The attack against the hubs is in fact an attack against the structure of the network since the attacker must determine the structure of the network before an effective attack can be launched. Scale-free networks are thus *structurally vulnerable*. Furthermore, since nature attacks its own creations in a random manner, it is not surprising that many networks found in nature such as food webs, neural networks and metabolic systems are scale-free. This is just why biological systems have survived more than a billion years of the random whims of Nature. Nor is it surprising that apparently small interventions by humans sometimes have devastating impacts on ecology. We are clever enough to identify hubs and take them out. But we do not understand the further consequences of the intervention.

Scale-free networks are formed in certain natural growth processes as described in one of the papers below (Svendsen). Preferential selection is one such process where a new node added to the network will attach to existing nodes with a probability proportional to the existing degrees of these nodes. No one in particular has designed the architectures of the Web or the internet. There is then good reason to believe that these networks have grown in accordance with the rules of preferential natural growth. Therefore, it is not surprising that several investigations have confirmed that these networks are, to a very good approximation, scale-free networks.

Another surprising discovery is that epidemiologic thresholds do not exist in scale-free networks. In most diseases, the number of infected subjects must exceed a given threshold before the disease begins to spread. There is no such epidemiologic threshold in scale-free networks: the disease might spread throughout the entire network from just a single node, as shown in one of the papers below (Canright et al.).

In 1997, the Norwegian Defence Research Establishment conducted the first in-depth study of the ICT vulnerability of the Norwegian society. This was one of the inputs that led to the establishment of a political committee chaired by Kåre Willoch in 1999 with the mandate to analyse potential threats to society and propose how preparedness should be organised in order to meet these threats. The committee considered all kinds of threats with information security as just one such threat. None of the committee members and only less than 10 % of those who contributed to the work of the committee had any professional knowledge of information security and ICT. However, during the work, these people were able to bring information security and ICT into the focus of the committee, and several of the final proposals of the committee were directed towards how to establish preparedness against information security threats. Several of these proposals were implemented but there is still a long way to go, in particular after the discoveries of scale-free networks.

The discoveries of Albert and Barabási were not known at the time the report of the committee was finished in early 2000. Five years later, some of us are still struggling to make these discoveries recognised at the political level so that the findings can be taken into account both in making society less vulnerable and to develop preparedness against the new threat. Several circumstances make the ICT threat different from all other threats. First, ICT is a component of every activity of society and is a precondition for proper operation of society. Second, an attack on ICT can be launched from any corner of the world. The attacker need not deploy troops on foreign soil but carry out the attack from behind their desk. Third, the attack may be blind – that is, not directed at anyone or any computer in particular – or be directed toward a particular part of the system, for example the hubs of one of the many network structures of society. Fourth, it may not be possible to trace the attacker or, even worse, the attacker may masquerade as an innocent party in order to escalate conflicts. And finally, it may not be possible to uncover the motive for the attack – in some cases the attack may be children's play – there is enough malicious software available on the Web to make this happen; in other cases, the attack may be an act of terrorism; the 'attack' may even be caused by someone hitting the wrong key.

These are the scenarios in which information security risks must be analysed and defined. Building faith on historical events is nonsense when history goes back only ten years and the ICT evolution is taking place at an exponential rate.

This issue of *Telektronikk* is divided into three parts: The first part is concerned with the important problems of risk awareness, risk reduction and risk management in large systems. This is an area of growing importance and complexity as the size, the interconnectivity and the openness of the computer networks become larger.

The second part is about structural vulnerability, in particular the type of vulnerability caused by scale-free networks.

The third part contains three examples from emerging areas of information security:

- The impact of wireless access and mobility on intrusion detection;
- Digital forensics;
- Security models for electronic medical records.

But first a review of the Bluelight effort is given. Bluelight is a network of users of information securities, manufacturers of security equipment and software, independent consultants and education with the aim to build and maintain information-security awareness.

Jan A Audestad is Senior Adviser in Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology (NTNU) and of informatiton security at Gjøvik University College, where his main task has been to build up a new master degree in information security, sponsored, amongst others, by Telenor. He has a Master degree in theoretical physics from NTNU in 1965. He joined Telenor in 1971 after four years in the electronics industry. 1971 – 1995 he did research primarily in satellite systems, mobile systems, intelligent networks and information security. Since 1995 he has worked in the area of business strategy. He has chaired a number of international working groups and research projects standardising and developing maritime satellite systems, GSM and intelligent networks.

email: jan-arild.audestad@telenor.com

Svein E. Pettersen obtained his Master of Science in Computer Science from the University of Oslo. He is Managing Director of Gjøvik Science Park and Associate Professor II at Gjøvik University College. He is also Project Manager for the innovation network Bluelight. Svein E. Pettersen previously held various positions in the Corporate Security Division of Telenor ASA in the areas of system security, telecommunications fraud and police response.

email: svein@gkp.no

The Bluelight Network: Security awareness and business development – Knowledge sharing and coopetition in practice

JAN A AUDESTAD AND SVEIN E PETTERSEN

Jan A. Audestad is Senior Adviser in Telenor

Svein E. Pettersen is Managing Director of Gjøvik Science Park

Five years ago, a small group of experts came together with the aim to improve education in information security in Norway. The effort resulted in one bachelor study and one master study. The cooperation developed rapidly into a network that has since expanded to include about 40 members from industry, consultancy, academia and public administration. The field of interest incorporates creation of new businesses, establishing joint ventures, building and sharing knowledge, development of tools and best practices, promotion of education and research, and assisting the evolution of information security in public administration. The paper describes how this evolution took place and what it has led to.

One of the most successful network-building initiatives in Norway took place five years ago.

The background is as follows. A government committee on vulnerability led by Kåre Willoch had shown that the need for expertise in the field of information security was increasing fast. This brought information security into the political arena in 2001. At the same time, a few leading industries in Norway, among them Telenor, was worried that education in information security was lagging behind the development of information crime and other security violation. A bachelor degree in information security was not offered in Norway at all and the education at master level was given as part of other studies (usually one or two courses in general computer science and telecommunications studies). These were classical, technological courses on items such as cryptography, cryptographic protocols, authentication, access control, database security, protection of operating systems, and intrusion detection and prevention.

During the 1990s, information security had rapidly moved into new areas that include important subject-matters not taught at university level such as how to build and manage information security in large corporations, how to protect the infrastructure of society against information terrorism and warfare, identifying data crime and securing digital evidence, the wide range of legal aspects related to information security, protection of personal integrity and anonymity, and understanding the psychology and sociology of data terrorism, data crime, cracking and hacking.

The first private initiatives taken were the development of a bachelor degree in information security for the Defence Engineering School in 2001 and a master degree for Gjøvik University College in 2002.

The founders and sponsors of this education decided to create a network with a wider outlook:

- enhance security awareness among its members and in society in general;
- create a cradle for innovation; and
- establish a forum for developing new businesses and joint ventures.

This network developed into *Bluelight*. The network is managed by Gjøvik Science Park (Gjøvik kunnskapspark).

Bluelight has developed into an innovation network and cluster of key participants in the field of information security. The members include commercial interests (suppliers, user groups and consultants), public sector and academic institutions. The network focuses on competence development, applied science, innovation and commercialisation (entrepreneurship and business development) with international focus and dimension.

The network contains about forty member companies and organisations ranging in size from one person to international corporations. Among the members are the following:

- Users of information security: Telenor (telecommunications), Norsk tipping (lottery), DnB NOR (banking) and Helse Øst (health)
- Suppliers of hardware and software: Thales and Ergo Group
- Service providers: Ibas (data recovery, erazure and forensics)
- Consultants: PricewaterhouseCoopers and Accenture
- Academia: The Royal Institute of Technology (Sweden), Gjøvik University College and the University of Oslo

ISSN 0085-7130 © Telenor ASA 2005

- Government: Innovation Norway, Norwegian Research Council and Norwegian Defence Security Agency

The network is actively supported by several of the national ministries (Commerce, Research and Education, Health, Modernisation), government bodies (The Data Inspectorate, NSM (Nasjonal sikkerhetsmyndighet = National Security Authority)) and interest groups (ICT Norway).

The network consists not only of industries with complementary interests. Some of them are also competitors in the same market. Nevertheless, these companies find it advantageous to cooperate creating what is referred to as *coopetition* [S-V]. Bluelight makes the total market bigger and richer giving room to such business models.

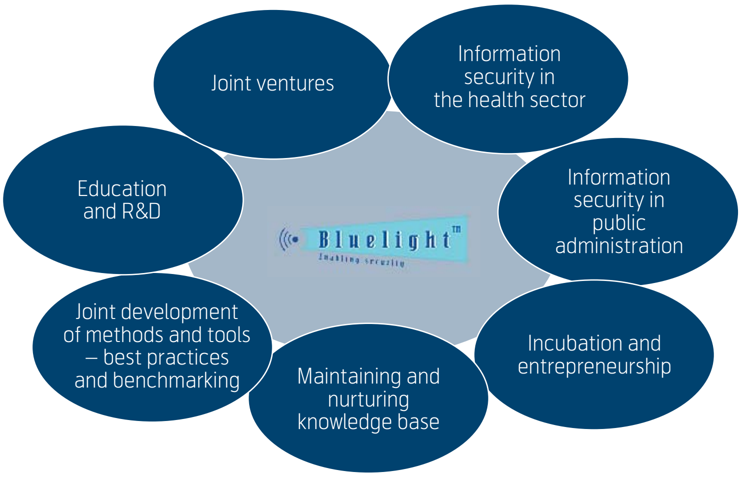

Joint ventures
Information security in the health sector
Education and R&D
Bluelight™
Enabling security
Information security in public administration
Joint development of methods and tools – best practices and benchmarking
Maintaining and nurturing knowledge base
Incubation and entrepreneurship

Figure 1 Areas of interest and cooperation

Strategies of Bluelight include (Figure 1):

- setting up joint ventures and projects – several such projects already exist;

- supporting education and research & development in information security – Norwegian Information Security Laboratory (NISlab) is one of the first results of this kind; new projects are planned;

- initiating joint development of methods and tools and establishing best practices and benchmarking – this is an activity particularly in demand by small consultants and small clients (the SMB[1)] market);

- building up and maintaining a common knowledge base;

- acting as incubator for new businesses – Security Partner is one such result. In addition, the network has been directly involved in the establishment of two other security related companies and several other businesses through indirect means;

- taking particular initiatives concerning public administration – one particularly active arena has been the municipalities. A direct spin-off is KInS (Kommunal Informasjonssikkerhet = Municipal Information Security) which is an active network of county administrations, municipal governments, consultants and industries. KInS arranges conferences and creates joint projects;

- information security in the health sector is an area that is under establishment – this is one of the most complex areas in information security at present and can best be dealt with if several of the partners of Bluelight cooperate.

New subjects that are just being included in the portfolio are energy and biotechnology. More partners on the user side (energy producers and distributors, and biotechnology research and production) will be included in the network.

The Government has decided that the Bluelight network shall operate the Centre of Information Security (SIS), a public institution responsible for coordinating activities related to ICT security in Norway. SIS will be located at Gjøvik Science Park.

Bluelight is currently working on the establishment of Sector – the only security incubator in the national incubator program created by SIVA.

Bluelight is a unique type of initiative in that it fosters coopetition and exchange of ideas based on the principle that knowledge behaves in a different way from physical assets: "if I give you a penny and you give me a penny, we still have one penny each; if you give me an idea and I give you an idea, we have two ideas each [Q-B-Z]." Coopetition and knowledge sharing are two of the basic characteristics of the knowledge industry as explained in [S-V] and [S-F].

Bluelight is both a business laboratory and an arena for public and private cooperation. Most of the authors of the following papers are associated with Bluelight in one way or another.

[1)] *Small and medium-sized businesses*

ISSN 0085-7130 © Telenor ASA 2005

References

[Q-B-Z] Quinn, J B, Baruch, J J, Zien, A K. *Innovation Explosion: Using Intellect and Software to Revolutionize Growth Stategies*. The Free Press, 1997.

[S-F] Stabell, C, Fjeldstad, Ø. Configuring Value for Competitive Advantage: On Chains, Shops, and Networks. *Strategic Management Journal*, 19, 413–437, 1998.

[S-V] Shapiro, D, Varian, H R. *Information Rules: A Strategic Guide to the Network Economy*. Boston, Harvard Business School Press, 1999.

For a presentation of the authors please turn to page 3.

Protection of vital information assets

– Balanced information security

ERLEND SKUTERUD

Erlend Skuterud is Senior Manager in PricewaterhouseCoopers AS in Norway

Information is a corporate asset and, like other valuable business assets, needs to be properly protected. Securing information from various threats is essential in order to increase business opportunities, ensure business continuity, minimize business damage and maximize return on investments.

Information can exist in numerous forms: on paper, film or orally in a conversation. And when it comes to digital or electronic form, the ways information can be stored are endless: computer hard drives, e-mail accounts, corporate applications, databases, operating systems and more. Regardless of a piece of information's form or how it is stored or transmitted, it can be devastating if it gets into the wrong hands.

Information security – a definition

If you ask corporate executives about their information security initiatives, you are most often asked to talk to the technical guys. – Information security is predominantly seen as a technical issue. Are there any other ways of looking at it?

Logical security is about handling the access to systems and information – vital business assets. The superior goal for logical security is to arrange measures to reduce risk exposure to what *corporate executives*, not the technical guys, perceive as an acceptable level. – What has to be protected? From whom? How?

Information security can be broken down into three key elements:

Confidentiality: Ensuring that information is accessible only to authorised users.

Integrity: Safeguarding the accuracy and completeness of information and its processing methods.

Availability: Ensuring that authorised users have access to information and associated assets when required.

Security of information is achieved by implementing a set of controls, which could be policies, practices, procedures, standards, organizational structures or software/technology functions. These controls need to be established in an enterprise in a coherent manner, in order to ensure that specific security objectives of information protection are achieved.

Who is compromising your information assets?

There has been an exponential growth in numbers of reported security breaches or "hacker attacks" as it is often referred to in the newspapers. The term "hacker" was originally used to describe a person who is a skilful programmer, but now the term is more often used to describe persons who illicitly attempt to access IT-systems and data, by others called a "cracker".

Roughly there are two types of crackers (or hackers); those who let the method control the goal, and vice versa; those who let the goal control the method. The first type is often familiar with one, or a couple of methods to penetrate a system. This individual search, most often without a specific object, after systems available on the Internet that is vulnerable to his methods and tools. When this person finds such a system it will be attacked. These types of hackers, often called "script kiddies", have specialized in a few techniques to attack a specific system and are often relatively young. – The biggest problem with this type of hacker is that everybody, from private PCs to the largest corporations is a potential target. The cracker does rarely make any selection of who they attack; it is the number of successful penetrations that is their primary motivation.

The other main category crackers are the ones who let the goal control the method. They are experienced and have solid competence within security and different types of systems.

The cracker's goal might for instance be to gain access to business secrets, credit card numbers or other attractive material. The goal might also be manipulation of contents of homepages for political or religious reasons. This type of cracker often has

ISSN 0085-7130 © Telenor ASA 2005

enough knowledge to work out new methods of attacking systems and does this according to change of personal goals.

A determined cracker of this calibre is very unpleasant to confront. Fortunately the number of people with this form of competence and intention is relatively limited, and the majority of those who are involved in illegal cracking fall into the first category; the one who lets the method control the goal.

But the largest threat to your vital information assets is your employees and colleagues. Not only in the form of a disgruntled, discharged employee. More often confidentiality, integrity, and availability of your data are weakened by incidents that stem from plain mistakes, user errors, lack of knowledge or curiosity.

How can vital information assets be safeguarded?

Of first priority is to establish a sound security culture and good attitudes. Many companies communicate empty words and false promises about their security work, but for creative attackers, external as well as internal, it means very little. Any organization that wishes to achieve satisfactory preventative measures must have a clear and well communicated security policy which is adequately anchored and communicated down into the organization. In addition, specific routines and procedures for how this should be practiced in day-to-day operations must be established.

The lack of this type of security culture leads quickly to a fundamental, daily indifference for the company's vital information assets. In many companies employees can contact the IT department and receive a replacement password for the "one they've forgotten" without showing any form of identification or proof of organizational belonging. Even though this is normal procedure, this is only one of many options which can be used to exploit slack attitudes and routines related to information security.

Another aspect of information security is robust technical measures. But no matter how captivating the thought might be of pulling satisfactory security out of a box, these products cannot just be installed as they are and be left to do the job. It requires careful configuration to adapt it to the organization's technical infrastructure, protocols and business needs, and when installed it requires monitoring, updating and adjustments according to latest security exposures and threats.

The following list represents important areas that need to be considered in the process of safeguarding your vital information assets:

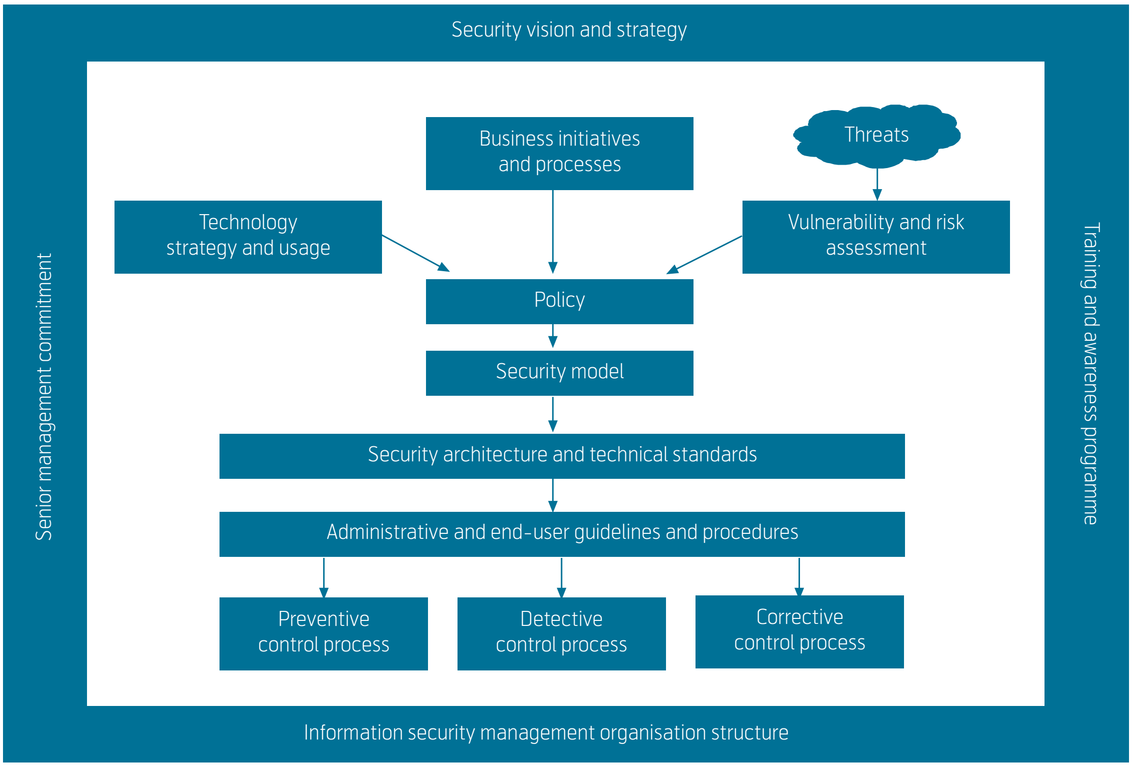
- Establish a security conscious culture through training and awareness programmes.
- Look at information security as a strategic element, and include this in your commercial initiatives.
- Know the value of the company's systems and data; perform thorough risk and vulnerability analysis.
- Establish a security policy balanced against the commercial needs, and based on recognized standards (for example ISO 17-799).
- Plan, build and implement security architecture in business critical systems. Information security investments should be based on cost/benefit and risk analysis.
- Have dedicated resources to coordinate and control the work with information security.
- Implement adequate technical security measures to manage the threats according to the results of the risk evaluation.
- Keep systems and applications adjusted and updated for known security exposures. Security measures in existing hard- and software should be used efficiently and according to established policies (operating system, firewalls, routers, ERP systems and other applications).
- Complete periodic tests to confirm real security level (i.e. security audits and penetration testing).

Information security measures are most often perceived as cost-intensive and as a restriction on productivity. This is partly due to the fact that it is difficult to quantify the costs by not having satisfactory information security in place before a problem really occurs. A fitting comparison could be to drive a car without proper insurance; you save money – so long as you are not involved in any accidents.

A framework for balanced Information Security

An important element in the work with information security is to find the balance between too much security (limits creativity and productivity, expensive solutions) and too little security (unlimited possibilities, great vulnerability, low direct costs). Figure 1

ISSN 0085-7130 © Telenor ASA 2005


Security vision and strategy
Senior management commitment
Training and awareness programme
Business initiatives and processes
Threats
Technology strategy and usage
Vulnerability and risk assessment
Policy
Security model
Security architecture and technical standards
Administrative and end-user guidelines and procedures
Preventive control process
Detective control process
Corrective control process
Information security management organisation structure

illustrates a framework that can be of value in the quest to achieve such a balance.

Successful implementation of an information security programme requires four fundamental "pillars":

- *Security vision and strategy*

Senior management must create an information security mission statement that clearly defines guiding principles and a philosophy on the importance of information in the organization, plus a broad strategy for implementation. The mission statement should be distributed throughout the organization, and should establish that information security is a priority for senior management.

- *Senior management commitment*

Achieving a consistent standard for information security requires clear vision and direction from senior management. Top executives must be specific about how the programme is packaged and communicated to all individuals (internally and externally) with access to the organization's information and systems. Communications should also include direction on how to implement the programme.

- *Information security management organization structure*

Information security is a management issue that cuts straight to the heart of an enterprise. Following a proper organizational assessment, a dedicated information security function should be established to "own" and direct security matters. Depending on industry, size and corporate culture, this function could be centralized, decentralized, or a mixture of both.

Since information security encompasses the breadth and depth of any organization, those charged with executing the programme must be senior decision-makers who can effectively make a difference – when and where needed. Authority and responsibility for enterprise information security should be vested with an executive who has accountability to top management. This could be the chief security officer (CSO), who reports to either the CEO or CIO.

- *Training and awareness programme*

This final component, stressing security awareness and education, must imbed the importance of security in an organization's culture. Training must be provided to all staff and should focus on security aspects relevant to each organization's business. It should be persuasive, provided continuously, and be an integral part of required training.

Information Security management lifecycle

But the implementation of information security management is not over. Just like any other management process in an enterprise, information security management requires continuous effort and commitment. The cycle of "assess, design, implement, maintain" is constantly executed for the organization and, as may be required, for business partners for contractual reasons.

ISSN 0085-7130 © Telenor ASA 2005

Experience shows that certain factors are critical to a successful implementation of an information security programme within an enterprise:

- A method of security implementation that is unified with organisational culture;
- Solid support and commitment from senior management;
- A corporate appetite for risk assessment and risk management;
- Security policies, standards and actions that are in line with business direction;
- A good knowledge and understanding of the technology and security requirements;
- An effective way of marketing security to everyone in the enterprise;
- Providing relevant and appropriate security awareness and training;
- A well defined system of performance measurement and suggestions for improvement.

Any conclusions?

There are no simple, magical tricks or products which can resolve all security challenges and problems. Information security is to a large extent a question of attitude and priorities.

In brief, a successful information security management programme should begin as an integral part of an organization's overall business strategy. Once security management is recognized as a core business operation, it requires the development and implementation of security practice guidelines and standards that are necessary to support the strategy.

These guidelines and standards are known as the enterprise security policy, which drives the development and implementation of an overall security-management architecture. Finally, this framework is monitored for vulnerability, attack, as well as misuse through preventive, detective and corrective controls. The end result for the enterprise is enhanced information confidentiality, integrity and availability, from both inside and outside the organization.

Good information security is about controlling and balancing risk – not eliminating risk entirely.

Erlend Skuterud has an MSc in Business and Economics. He is Certified Information Security Manager (CISM) and Certified Information Systems Auditor (CISA). He is Senior Manager in PricewaterhouseCoopers information security group in Norway.

email: erlend.skuterud@no.pwc.com

Security culture and risk management is a management responsibility

ARNE JOHAN HELLE

Arne Johan Helle is Partner in PriceWaterhouseCoopers AS

Daily your company is exposed to risk in business critical processes which rely on information technology and applications. Threats and events which may have great consequences for the organization need to be controlled and risk be mitigated where needed in accordance with the management decisions and business strategy.
Has the company established guidelines to create quality and security to avoid risks in information systems? Will the organisation be able to create a risk and security culture?

Introduction

For years now, security experts have been lamenting the use of point solutions to address security threats and vulnerabilities. Professionals have talked about the need to integrate security processes and technologies across an organization's infrastructure to provide a centralized and automated capability to monitor, protect and defend critical information assets. It is a good vision, but until recently, that is all it has been. In the new age of compliance, the vision is fast becoming a reality for companies that want to sustain compliance for the long haul.

The need for regulatory compliance provides today's management with this challenge to get the right business processes and integrated solutions in place to meet the compliance mandates of today … and tomorrow.

One of the challenges today is that several companies are about to "open their internal IT-systems" for outsiders in the form of customers, partners and contractors. This involves new risks for the company's work processes, and the customers have high demands to security and reliance to the information systems. Society has higher demands to openness, and we have an abundance of information which has to be managed in a good and effective manner. Privacy and personal information need to be secured. Privacy regulations give guidelines for personal information in Norway and other countries have similar legal framework.

Information security is a complicated area. Security measures are often considered bothersome and restrictive seen from the user's side, i.e. extra time with several log-ins and passwords which have to be remembered. This is a misunderstanding. There are techniques today which give you access on the first identification. Information security can be an integrated part of quality assurance in the work processes to be able to:

- Ensure that information is made accessible and available for only authorised users (confidentiality);
- Ensure high service degree to the user by securing complete and accurate information (integrity);
- Ensure the users have access to business critical information when needed (availability).

These are all elements which will secure efficient utilisation of available resources in a company which is dependent on their IT-systems in work processes.

Security culture

The Organisation for Economic Co-operation and Development's (OECD) guidelines for the Security of Information Systems and Networks were defined in the Council July 25, 2002. These guidelines aim to:

- Promote a culture of security among all participants as a means of protecting information systems and networks;
- Raise awareness about the risk to information systems and networks; the policies, practices, measures and procedures available to address those risks; and the need for their adoption and implementation;
- Foster greater confidence among all participants in information systems and networks and the way in which they are provided and used;
- Create a general frame of reference that will help participants understand security issues and respect ethical values in the development and implementation of coherent policies, practices, measures and procedures for the security of information systems and networks.

ISSN 0085-7130 © Telenor ASA 2005

The guidelines are also incorporated in the Norwegian National Information Security Strategy. The main principles consist of the following points:

1 Awareness – Participants should be aware of the need for security of information systems and networks and what they can do to enhance security.

2 Responsibility – All participants are responsible for the security of information systems and networks.

3 Response – Participants should act in a timely and co-operative manner to prevent, detect and respond to security incidents.

4 Ethics – Participants should respect the legitimate interests of others.

5 Democracy – The security of information systems and networks should be compatible with essential values of a democratic society.

6 Risk assessment – Participants should conduct risk assessment.

7 Security design and implementation – Participants should incorporate security as an essential element of information systems and networks.

8 Security management – Participants should adopt a comprehensive approach to security management.

9 Reassessment – Participants should review and reassess the security of information systems and networks, and make appropriate modifications to security policies, practices, measures and procedures.

The above are good guidelines and represents what we can call "common practice" to maintain good security of information and elements which have to be in place in order to create a security culture.

Management involvement in Information Security

Experiences from Information Security Forum (IFS)

The forum is an independent "non-for-profit" organisation which today has more than 250 members from all over the world. Everybody is engaged in improving the security according to the company's own needs. The forum has drawn up a common framework, "Standard of Good Practice for Information Security", which forms the foundation for what OECD points out as a coordinated policy and procedures.

The forum has over a period of 12 years completed a "status survey" and benchmarking within IT-security among members. The members will from the survey be able to learn how "the best" solutions in different organizations achieve success.

From the survey completed in 2002, I would like to summarize some conditions the management should take notice of; no revolutionary news, but common known security issues:

The thoroughfare of approximately 200 companies concluded that there is no easy way or "quick wins" within IT security. It is a complex area which is connected with multiple conditions, and one system or network is dependent on other applications and networks. The recipe is priorities from the management and neat and thorough work toward a moving target. This demands focus on risk assessments and focus on key controls which mitigate the defined risks and have documented effect.

Some of the most important findings:

- Increased management involvement, increased investment in security and number of staff with knowledge, will lead to considerable improvements in the security level due to reduced security occurrences/incidents and increased focus on security awareness within the company.

- Central focus from a corporate group function, with authorisation to instruct Business Units, will make better security awareness in the organisation also on management level. When the organisation gets common security policies and requires transparency and common reporting standards including monitoring, the security level improves.

- The companies which complete a holistic risk evaluation in all business critical systems manage to prioritize their safety measures in areas which give greater effect and value.

Control and risk management

One of the most important elements in the administration of the companies' values is risk evaluation and daily monitoring of the consecutive operational risk which the company is exposed to. The responsibility lies with the board and management according to Norwegian Legal Requirements and good internal control practise, to ensure the company's fund management.

It is important to prioritize where the company should use their resources in the best possible way. With a shortage of resources, competence, finances and time it will be an important task to complete a risk evaluation and look equally for risk in the new business opportunities.

Today's organizations need to ensure the continued availability of IT systems, as well as the confidentiality and integrity of the information carried over them. This is crucial to your organization's ability to harness the opportunities, and manage the hazards of modern business.

Regulatory bodies, shareholders, partners and customers alike expect diligent management of corporate systems and information. However, there are significant challenges in protecting corporate information against a background of ever increasing security threats, especially in the distributed environments common to global companies.

Tackling these challenges requires a comprehensive end-to-end information risk management approach that embraces the whole organisation, regardless of whether your IT is managed internally or outsourced to a third party provider.

The organisation needs an approach to information risk management designed to address key elements of the evolving risk environment, and a combination of business and technical skills to help the organisations implement a comprehensive trust infrastructure. One common approach is based on the following four elements:

Controls Environment: The essential policy and procedure foundation for managing information risk, including critical elements such as developing management commitment to security underpinned by policies, standards, guidelines and training.

Technology Platforms: Securing the core network infrastructure and operating systems is a vital part of the information risk management challenge. Specifically, businesses face an increasing need to protect their infrastructure from both internal and external threats including hacking, viruses and cyber crime.

Enterprise Application Controls and Security: Enterprise applications are increasingly complex and the success or failure of an implementation can be judged on the level of user uptake. A major influencer in this is achieving the right balance between functionality and security, improving the interaction with the users while still addressing the competing demands of convenience and the need to protect critical data and resources.

Identity Management: Whereas in the past the challenge was keeping people out of your IT systems – now it is bringing the right people in. Embracing business partners, suppliers, contractors and customers, often connecting remotely, forces organisations to find a new approach to handle their users' complex requirements while still controlling costs. Identity Management (IdM) addresses this new challenge by building a framework of solutions that enable organisations to authenticate, authorise, provision and store user access rights in a secure and scalable manner.

Sarbanes-Oxley Act of 2002

Information security and risk management are fundamental elements in establishing internal controls. These elements are also important in the SOX requirement regarding General Computer Control. The Sarbanes-Oxley Act of 2002 (SOX), which deals with "Corporate Governance rules, regulations and standards for specified public companies including SEC registrants" focuses on good internal controls related to financial reporting.

Most companies agree that the reliability of financial reporting heavily depends on a well-controlled IT-environment, including risk assessment. The guidelines for section 404 which addresses internal control over financial reporting refer to the frameworks COSO (The Committee of Sponsoring Organizations of the Treadway Commission) for internal control and CobiT (Control Objectives for Information and related technology) related to specific IT-controls. Both frameworks have focus on the management responsibility for establishing a control environment with focus on internal control including risk assessment.

Key elements in the COSO framework for internal control:

Control Environment:

- Sets tone of organization-influencing control consciousness of its people.

- Factors include integrity, ethical values, competence, authority, responsibility.

- Foundation for all other components of control.

Risk Assessment:

- Risk assessment is the identification and analysis of relevant risks to achieving the entity's objective – forming the basis for determining control activities.

ISSN 0085-7130 © Telenor ASA 2005

Control Activities:

- Policies/procedures that ensure management directives are carried out.

- Range of activities including approvals, authorizations, verifications, recommendations, performance reviews, asset security and segregation of duties.

Information and Communication:

- Pertinent information identified, captured and communicated in a timely manner.

- Access to internal and externally generated information.

- Flow of information that allows for successful control actions from instructions on responsibilities to summary of findings for management action.

Monitoring:

- Assessment of a control system's performance over time.

- Combination of ongoing and separate evaluation.

- Management and supervisory activities.

- Internal audit activities.

Organizations have been challenged to document their internal control with focus on risk, defining key controls to mitigate the risk and to make sure all controls are preformed effectively with good communication and monitoring.

Summary

Risk assessments with knowledge and understanding of important threats, and which consequences different occurrences will cause the organization, are important elements in a culture of risk, control and security. Security awareness activities and management focus and awareness are important. When this is linked to quality assurance, it will create a good understanding amongst the users.

Simple methodologies for risk analysis which are completed similarly across all Business Units and departments of an organisation are important for a common and holistic approach. This will empower management to prioritize measures and activities towards the most exposed threats in the organisation.

And finally; a successful information security management awareness programme should begin as an integral part of an organization's overall business strategy. Senior management involvement in the security area is important. Security management must be recognized as a core business operation and be tightly linked to quality assurance. This again requires development and implementation of security policy, practice guidelines and standards supporting the continuous changes and development of the organisation.

All these elements need to be in place to have good internal control in an organization.

Arne Johan Helle is a Partner in PriceWaterhouseCoopers AS with responsibility for the Information Technology and Security group in Norway.

email: arne.j.helle@no.pwc.com

Creating a security culture

HANS MARIUS TESSEM AND KJELL RUNE SKAARAAS

The vast majority of security incidents are caused by human error, not by flawed technology. Creating a secure business process cannot be achieved through technical means alone. While routines and procedures describing how employees should act are necessary, they are not sufficient. It is equally important to raise security awareness and motivate the employees to act responsibly and in accordance with these practises. To achieve this, the advantages and means of creating a security culture are explored.

Hans Marius Tessem is Managing Director of Security Partner

Kjell Rune Skaaraas is Manager of development and engineering at Security Partner

Introduction

Since the massive introduction of IT systems in all forms of business and governance, much of the focus has been directed at technical security. While this provides a foundation on which to build information security, time and time again studies show that the majority of security incidents are caused by people, not by flawed technology.

Creating a security culture and efficiently securing the business process is not as straightforward as technical solutions. Many companies seem to think that good security is something that can be bought through anti-virus software and firewalls, which exists independent of and unrelated to the actual business process.

But security is equally much about how the company manages user rights, documents and configuration changes. Employees must be aware what information they give out and to whom, and the dangers of ignoring or subverting the security measures that are in place. It is about how each employee handles their keys and logins, if they lock their terminals and how they behave when on the Internet.

Information security can be divided into three aspects; technology, process and people. Physical access controls and IT systems are typical examples of technology that have inherent properties in themselves. The business process represents how the company operates, both formally and informally. And finally, the human factor describes how people interact with these systems and processes. But regardless of the circumstances, the information is always the focal point, whether it is electronic, on paper or held by the employees.

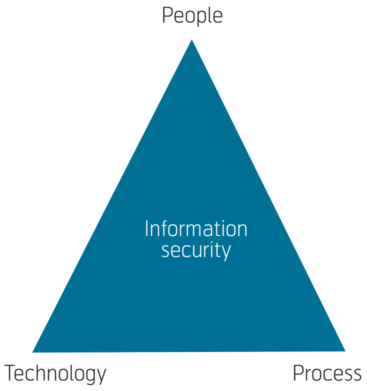


Figure 1 Aspects of information security

Technical security, which primarily consists of physical security and IT security, provides the underpinning for the other. These systems typically provide services as identification and authentication, traceability and non-deniability, which are vital to implementing access controls. Change management of technical security could be considered part of process security as it utilizes much of the same methodology and tools, but in practical application it is a relatively separate analysis. The difference is that process security focuses on direct risks to the business process, while this analysis looks at indirect risks to the business through compromises of the technical security.

Process security encompasses the actual workflow of the core business, including all policies, routines, procedures and guidelines. This also includes interaction with customers, suppliers and business partners, as well as contingency plans to recover from any crisis situation. These are far more business-specific than technical security measures and in an environment where rapid change of the business process may be in order, maintaining a good and current process security is challenging.

The human factor is often neglected in this process, perhaps because it is not so easily quantifiable as firewalls and manuals. It is important to consider that employees are the ones who will deal with the security measures on a daily basis. In many of the routines and procedures the employees are not only passively exposed to them, but are also the ones executing them. If they alone or collectively choose to perform

ISSN 0085-7130 © Telenor ASA 2005

their work differently, the security measure may no longer exist.

Despite all good intentions, security must at times come at odds with ease of use, and employees are always looking to make their work day easier. If they can neither see the logic nor the reasoning behind the routines, they are easily subverted or ignored. Good security is usually built in several layers, which means that shirking may be caught by another layer, or for a while nothing happens. Through this process good security is often hollowed out by neglect and indifference, until only a single point of failure remains.

Another reason security may deteriorate over time is that in order to prevent the routines from becoming too detailed and rigid, they are made more like general guidelines. Through a changing business process, interpreted by employees that are neither qualified nor inclined to evaluate the security risk, this may deviate gradually towards a process that requires less work, but also provides less security. There is a delicate balance between the routines becoming outdated and irrelevant and being too bland and obtuse to provide proper security.

To err is human. But many computer users are not even aware that they are exposing themselves or the company to risk, nor what security routines are in place. Many companies have extensive routines – if only to collect dust on a shelf. If they are not actively kept updated and regular reminders keep them persistent and fresh in memory, they are unlikely to be followed. This can be achieved through a good security culture.

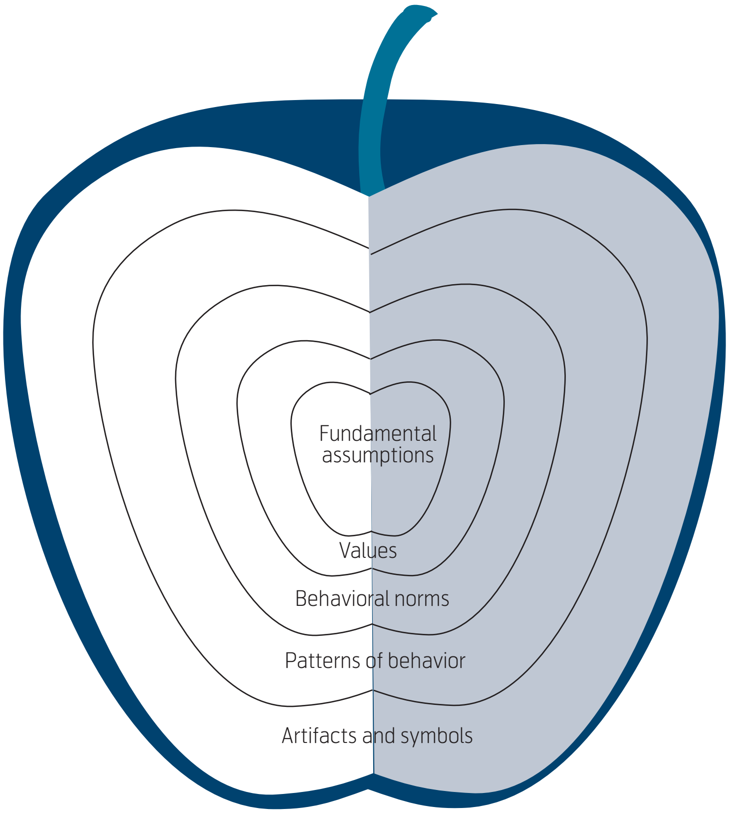


Figure 2 Levels of organizational culture (Schein 1985)

What is corporate culture?

There are several definitions of culture focusing on differing aspects such as values and beliefs to behaviour and symbols. These aspects can be visible as well as hidden. The following figure identifies five levels which centre on fundamental assumptions of what the organization represents.

It is easy to assume that this represents a strong centre with weaker layers surrounding it, but that is not always true. While the cultural influence typically radiates outwards, visible symbols and traditions may be preserved even though the deeper values may have disappeared or changed. When it comes to security, it is important to maintain this as a value, not only symbolic. An example of relegating security to a symbol is certifications that serve no other purpose than to display the company's apparent commitment to security.

Denison (1990) offers four basic views on corporate culture which can be summarized as follows:

- Consistency – the idea that a common culture will enhance coordination and promote a sense of identification among the members;
- Goal-seeking – the idea that a shared sense of purpose will focus and strengthen the members' effort toward a collective goal;
- Participation – the idea that involvement will contribute to a sense of responsibility and ownership, which brings organizational commitment and loyalty;
- Adaptability – the idea that norms and beliefs that enhance an organization's ability to transform external influences into internal changes provide a competitive advantage.

Which effects are desirable is highly dependent on the purpose of the culture, which must ultimately support the business model. If the culture is not well aligned with the business needs, it can also act as an obstacle. In terms of security, which exists to support the core business, the focus is primarily internal. A consistent security is clearly desirable to minimize the risk exposure, and participation is required to allow the business' security measures to change along with the business strategy.

A security culture is not at odds with existing corporate culture. Rather than being a mush of different cultures, it is built around critical organizational aspects. Examples of these can be conflict, diversity, innovation, organizational learning, change management, knowledge management, alliances and partnerships, relationship formation, and corporate as well as individual responsibility. Security will be another aspect that is brought into the culture as a whole.

Modern sociology theory usually points to three key factors in predicting employee behaviour. The first is the employee's own attitudes. The second is the subjective norm of what is expected of them, not only by their superiors but also by their co-workers. The third is control mechanisms, which has been almost exclusively singled out. But in order to perform their duties, they require access to locations, resources and information. Even in an environment that operates on the strictest access controls, user behaviour cannot be eliminated as a risk factor.

There have been countless examples of employees subverting security, everything from post-it notes with passwords to doorstops holding "secure" doors open, yet very few make any effort to change their attitude. The typical answer is more rigid security measures they cannot circumvent, sanctions or to cut back just enough so that it is tolerated. While no-one is suggesting that security should be relaxed to accommodate the complacency of employees, one can question if this is the most effective approach to improve security.

Forging a culture

When initiating a security program, a common mistake is not having a long term plan. Diving into it without a clear plan of what goals are to be met or how to achieve them can lead to some unpleasant surprises. It is important to remember that the desired goal is a long term effect and thereby a good cost/gain ratio, not a brief state of heightened security. There are examples of companies that after a one-year plan have experienced lasting effects for more than five years after the program ended.

When making changes in an organization it is very common to come across resistance to change, both as opposition and sheer inertia. Hence it is crucial that the right people are involved right from the start, and while beginning with management it is equally important to find good ambassadors among key people and unions. Through gathering their support, resistance is minimized.

	Stability	Change
Internal	Consistency	Participation
External	Goal-seeking	Adaptability

Figure 3 Effects of organizational culture

In order to create a culture throughout the organization, it is important to bring management on board. While they are not the ones who will form the culture, their approval and support will be vital to its success. Schein (1992) suggests that leadership includes the creation, the management, and at times the destruction and reconstruction of culture. The IT staff is often not in a position to address all the organization's information security needs, despite being in charge of IT security. In order to promote security throughout the organization, it must be brought to focus for the managers of all departments.

Particularly when it comes to attitudes and consciousness about security threats, the IT department is not part of the daily operation and the interaction too seldom to promote a culture. To that end, it is wise to identify resource persons who show an interest in security as well as key personnel to create local support for this activity.

As an IT department, the majority of the department's skills will be in the technical area, yet to form a culture requires skills in interpersonal communication and psychology. Should you seek external help to facilitate this process, it is important to maintain focus on this as an internal project. Security must come from the inside, communicated by permanently employed staff. While outside consultants can be a valuable help, they should not front the project towards the rest of the organization.

There are also many resources within the company, for example the marketing department. While their primary focus is on the customers, they have good skills in communication, and they typically have a clear perception of how the company should be perceived by the customers. Values like reliability and trustworthiness go hand in hand with strong information security, and are very common.

Security through participation

So how do we involve and convince these key people? The key is participation. A common trait is that by delegating responsibility, we also create a personal commitment. By giving these people the possibility to contribute to the process, we also create ownership

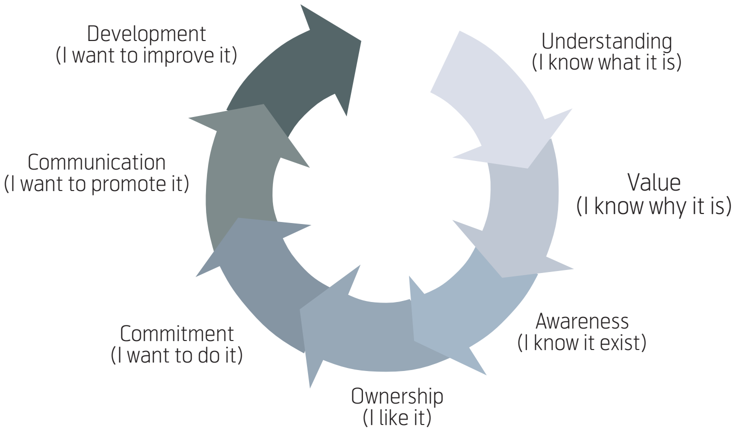


Figure 4 Stages of participation

to the process as a whole. People are far less likely to work against something they have been a part of forming, even if their influence on the process as a whole is minor.

The most central part of creating a culture is to seek participation from the employees. A company culture is far more than a top-down exercise in information and discipline. In order to create a culture, is it necessary for the employees to take part in this culture. Understanding, commitment and a sense of responsibility are key concepts in order to make employees sustain and evolve such a culture.

To many this may sound far fetched, as many users find computers incomprehensible and security dreary. The key is to provide security not as one huge manual to be read cover-to-cover, but to wrap it up in smaller tasks like group assignments, quizzes, branding security to make you quickly recognize anything related security and so on. While it is important not to let the message get lost in the process, society has come to require an element of entertainment to keep the attention long enough to teach them about security.

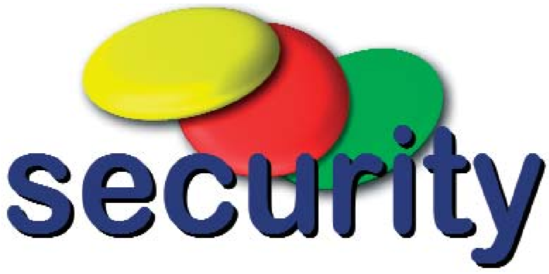


Figure 5 Example of branding – Non Stop Security™, based on Non Stop candy and developed by Norsk Tipping

While it has been claimed that we live in the information society, a more accurate claim might be that we live in the entertainment society. Studies have shown that about 90 % of all US citizens receive their news through infotainment like *The Late Night Show* and *The Tonight Show*. The use of entertainment is also on the rise in marketing in order to educate and convince a potential buyer about the product or service in question. The same tools of marketing can also be applied to educate your employees on security matters.

Does this mean that security has to be entertaining all the time? Not at all! But in order to effectively communicate your message, you need to do so on the receiver's own terms and premises. While what we find entertaining or interesting is highly individual, there are typically considerable similarities in company subcultures. For example, an R&D or sales department often consists of people who share a common background and interests, or who have developed it in the workplace.

By identifying these groups by their similarities and needs for security awareness they can be reached in the way that is most appropriate and effective. This is key both to achieving a good cost-benefit ratio and to avoid "carpet bombing" the employees with potentially irrelevant information. It is extremely hard to regain the attention of employees that have lost interest because the earlier topics have not really applied to them.

Branding information security

It is highly recommended that a form of brand or logo is used on all security material. This will quickly and effectively show that the information given is regarding information security, and should be given proper attention. By increasing recognition, the time window required to communicate the message is shortened. While making the brand or logo recognised throughout the organization requires effort, it is a lasting asset that can be reused in many security-related events.

One of the biggest advantages of creating a security brand is that security awareness can be tied to items and events not inherently related to security. As an example from Norway where winters are long and cold, branded ice scrapes with a brief security reminder not to leave the laptop in the car and so on have been very successful. In this way security has been tied to a useful product that is used daily during the winter season, constantly advocating security without resorting to what might be considered nagging warnings. Over time, such warnings are increasingly tuned out and ignored.

This branding also helps identify security as a separate and important task, which is not inherently covered by the IT department alone. Both when it comes to identifying resources spent and activities carried out relating to security, a clear distinction between security and other daily operations will lead to more accurate figures. In companies that do not make such a distinction, there is often a discrepancy where management believes a lot of resources are indirectly used on security as part of the daily operation, while the reality is quite different. Good estimates of the available resources compared to the security challenges the company is facing are vital in order to receive the funding required to implement good security.

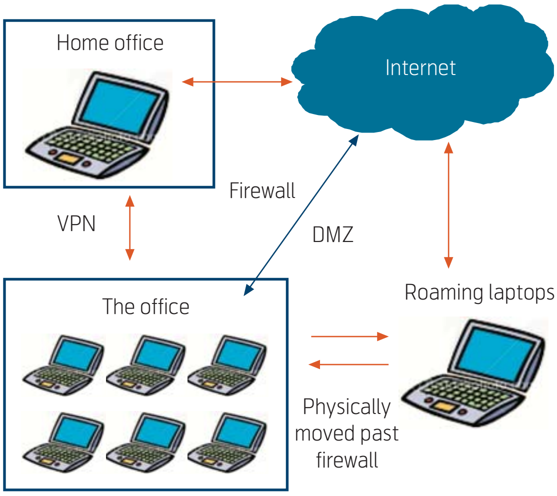


Figure 6 Home office and mobile access

The mobile user

A growing trend in nearly all businesses is to allow more and more external access and mobile access. Many companies provide laptops to their workers, allowing them to work from outside the office. This has put a lot more responsibility on the users, who may be tempted to install additional software or modify the system settings, outside the control of the IT department. Infected laptops or compromised home offices circumvent the traditional corporate firewall or VPN, making them much harder to contain.

A consequence of this is that each user must protect their machines, software, documents and passwords where and when they are working, far away from the relative security of the office. While the IT department may provide guidelines and restrictions, the actual use is normally outside their control. This makes a security culture even more important to guide employees to act in a responsible way.

IT department and security

The primary focus of creating a security culture is to include all employees, and it is important to not forget the IT staff. While they have a far heightened consciousness when it comes to security, the requirements are equally higher. Their internal routines to create and maintain a secure system are crucial for the entire business, and they are not infallible either. To this end they require an even higher degree of knowledge and awareness.

A key element for IT staff is change management. With IT systems constantly undergoing change and pressure on IT staff to shorten response times, it is very important that security does not get lost in the rush. While the phrase "make it work, then make it secure" may provide a short-time gain, it will soon tie up as much or more resources to troubleshoot problems and avert crises.

Soft security barriers like a front desk naturally lead to a multilayered security, while technical barriers are often absolute. That makes it very tempting to rely on one barrier which may be sufficient when operational, but that is prone to misconfigurations and changing circumstances. Combined with rapidly changing systems, this can quickly lead to exploitable vulnerabilities.

It is also important to remember that while IT staff often has great skills in technical security, they are not necessarily equally skilled when it comes to information security in a broad sense. It should not be forgotten that even low-level IT staff, through their administrative rights may have access to more information than many far higher ranked positions in the company.

Sustaining and reinforcing culture

Because major security incidents are rare, it is a challenge to maintain the required security awareness over time. It is part of human nature to lend greater weight to personal or anecdotal experience than their actual value, particularly when it comes to rare events. As it is unlikely that everyone will follow all the rules all the time there are constantly small tests to see what will happen. Since security is often built in layers and is designed so that even the improbable should not occur, it is probable that nothing of importance will happen. However, many small probabilities add up to what will be a major incident, sooner or later.

Unfortunately, these constant small "proofs" to the contrary have a very clear impact on employees, making them believe that the threat is being exaggerated by the security department. It is also true that the

ISSN 0085-7130 © Telenor ASA 2005

tabloid media are exaggerating at times, by turning what is really an obscure and difficult to exploit attack vector into a major news item. While these make good sensationalist press, they are actually a disservice for those seeking to heighten actual security levels. Ultimately the result is that everything is proclaimed dangerous, and the warnings are no longer effective at pointing out what is really dangerous and not.

In order to counteract this, it is important to provide real and plausible examples, either from your own organization or incidents reported in the news. While it is unwise for security reasons to provide details that may be sensitive, it is important that employees feel this is related to their company and their workday. Repetitions of generic warnings are unlikely to trigger the same response. While few companies would like to publish their losses due to security incidents, making employees aware that these losses are being tracked and have a very real business impact can also help underline the seriousness.

In addition, it is important for each company to inform their own employees which threats they see as likely, and what security measures are vital to uphold. These may be quite traditional and not newsworthy in the same way as recent threats, but they may still be just as vital to security. The media have a very shifting attention moving from one new threat to the next, and should as such not be a major source of security training for your employees.

Cost of security breaches

The total costs of a breach of security are spread over many departments, and are often absorbed in with noone knowing the total cost. A large survey from Norway suggests that for a high-technological country of 4.5 million people, the direct losses are estimated to be over 200 million US dollars, with indirect losses of over 650 million US dollars. (Mørketallsundersøkelsen 2003).

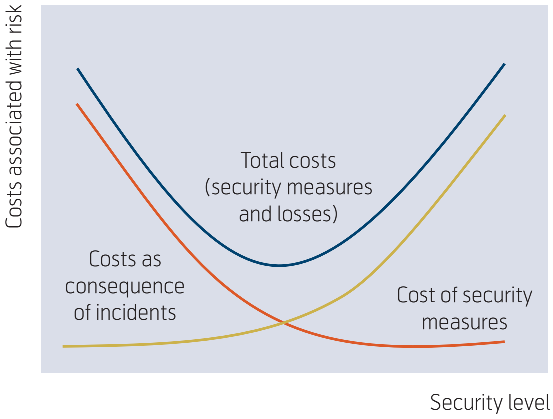


Figure 7 Total cost of security (SIS 2005)

Because the total loss of security breaches is hard to predict, these numbers are inherently inaccurate. However, it is still easier to quantify technical solutions through metrics such as viruses blocked than it is to measure the value of a security culture. Those that have made a conscious effort to create such a culture, feel it has been a valuable investment even though they cannot exactly quantify the gains.

Investing in a security culture is a long-term investment, which will persist even as technologies and processes change. While users would have to be informed about the specific risks of the new system, the underlying awareness and care with which they treat information and the systems containing it is carried on unchanged. In this respect user awareness fundamentally differs from user training, which is focused on performing one specific task in one specific way.

When evaluating the cost of security, it is clear that optimal security is a balance between costs of security measures and costs of incidents. However, there is also an implicit assumption that is more subtle – that the resources are optimally spent on the right systems, and of the right form. Prevention, deterrence, detection, limitation, correction, recovery, monitoring, awareness and transfer are all means to managing risk.

Through embracing a security culture many of these factors can be improved across all systems simultaneously, by making all employees aware of the risks. Not only does this make their own information handling safer, preventing and limiting damage, but it also includes reporting possible security issues and thus helping monitoring and detection. The latter is in fact quite important; an often heard quote is that this would not have happened, had the right routines been in place. To that end, it is fundamental for a secure business process to know where routines are needed, or when they are no longer appropriate for the current work process.

Measuring the effects of corporate culture

Due to the uniqueness of each culture, the majority of the material is anecdotal and based on business cases. Peters and Waterman (1982) as well as Cameron and Quinn (1999) show that many successful businesses scored low on traditional success factors, but instead had strong leadership that promoted strong strategies and a strong culture to support their strategy. There is also strong anecdotal support that failure to change the organizational culture has been the primary cause of many failed change efforts. (Kotter and Heskett 1992; Goss et al. 1993; Caldwell 1994)

Kotter and Heskett have tried to make this material more systematic and empirical, but were eventually inconclusive. Denison (1990) did find empirical support for the participation view of culture – the level of employee participation was positively correlated with good organizational performance.

In regard to a security culture it is even more difficult to provide empirical data, as security does not serve a purpose in itself, but rather exists to protect the core business of the company. In a quantitative analysis the effect would almost certainly be masked by the ups and downs and ultimately success or failure of the core business model. While the same is also a concern in a qualitative analysis, each effect can be isolated and studied more carefully.

Another factor which makes it difficult to accurately analyse the advantages is that we are analysing human behaviour, which has many facets. It is very difficult to create any form of absolute measure of security awareness. Most studies done have been made using a comparative analysis designed to measure change, not its eventual impact on business performance. Mathisen (2004) states that the following are common security awareness metrics in Norway:

1 Percentage of employees having finished the necessary security training;
2 Number of reported security incidents;
3 Percentage of employees leaving their desk clean at the end of the day;
4 Percentage of paper waste being shredded;
5 Percentage of illegal traffic on the internal computer network;
6 Percentage of weak user passwords;
7 Number of hits to security web pages;
8 Number of requests to security department;
9 Customer satisfaction.

It is clear that while these metrics can serve as indicators to measure changes in security awareness, they are obviously insufficient to measure the total effect of such awareness. Such metrics can also be easily manipulated by focusing exclusively on what is being measured, leading to excellent security ratings despite little improvement in total security.

Listing all the possible effects of a security culture would be an impossible task, since it applies to almost all aspects of security. Suffice to say that among the advantages presented are fewer security incidents due to neglect or carelessness, less dissemination of information leading to incidents, higher probability that incidents are discovered and reported, more active contributions to improving security, better incident handling and employees taking more responsibility for shared information and ownerless information.

Conclusions

Even though handling the human factor may seem more abstract and less measurable than technical security, it should be handled in a conscious way. Through clear planning of means and desired effects, a good cost-benefit ratio can be achieved. It is important that these means come together to form a coherent long-term plan. While technology is a rapidly changing factor, security awareness is a lasting characteristic which can be transferred across technologies. The probability of human error as a result of changing systems or work process is greatly reduced.

It is important to remember that no culture is built in a day. While it requires a good long-term plan, it does not need to be intensely time consuming. Compared to a technical solution that does exactly as instructed, the work may at times seem demanding, but the rewards are also great. The most important factor is to reach the users on their own terms – in practice, this means to entertain them while at the same time teaching them about security.

References

Caldwell, B. 1994. Missteps, Miscues: Business Reengineering Failures Have Cost Corporations Billions, and Spending is Still on the Rise. *Information Week*, June 20, 50–60.

Cameron, K S, Quinn, R E. 1999. *Diagnosing and Changing Organizational Culture: Based on the Competing Values Framework*. Reading, MA, Addison-Wesley.

Centre for Information Security (SIS). 2005. *Risk Management – A guide for local government* (Norwegian: Risikostyring – En veiledning for kommuner). June 15, 2005. [online] – URL: http://www.norsis.no

Denison, D R. 1990. *Corporate Culture and Organizational Effectiveness*. NY, John Wiley.

Goss, T, Pascale, R, Athos, A. 1993. The Reinvention Roller Coaster: Risking the Present for a Powerful Future. *Harvard Business Review*, 71 (6), 97–108.

ISSN 0085-7130 © Telenor ASA 2005

Kotter, J, Heskett, J L. 1992. *Corporate Culture and Performance*. NY, The Free Press.

Mathisen, J. 2004. *Measuring Information Security Awareness*. Høgskolen i Gjøvik.

Peters, T, Waterman, R. 1982. *In Search of Excellence: Lessons from America's Best Run Companies*. NY, Harper and Row.

Schein, E H. 1985. *Organizational Culture and Leadership: A Dynamic View*. San Francisco, Jossey-Bass.

Hans Marius Tessem (28) is the co-founder and managing director of Security Partner, a company specialising in information security. He received his MSc in industrial economics and technology management from the Norwegian University of Science and Technology in 2003. He has a background from Telenor R&D and EDB 4tel (Telesciences) where he worked with different software development projects.

email: hans@securitypartner.no

Kjell Rune Skaaraas (26) is the co-founder and manager of development and engineering at Security Partner, a company specialising in information security. He received his MSc in industrial economics and technology management from the Norwegian University of Science and Technology in 2003. His research interests involve applying models from economics, engineering, information technology and psychology towards security.

email: kjell.rune@securitypartner.no

Assessing social engineering robustness

OLA HOLM

Ola Holm is a Major in the Norwegian Defence Security Agency

User oriented security measures are assumed to affect individual resilience against social engineering attacks. By measuring the individual perception of existence and acceptance for each measure, a value is proposed to indicate a level of social engineering robustness for the organisation. This article describes how a metric for social engineering robustness can be established.

Introduction

The effect of user oriented security measures must not be taken for granted. A security policy is of no help if employees do not know of its existence, or if the content is so extensive that one cannot expect people without a special interest for information assurance to remember the essence. Who has not been introduced to extensive security documents where the essentials have been drowned in too much information? Because of the inevitable uncertainty that comes along with security measures in general, a common and well-meant strategy has been to add up too much measures and information. This is unfortunately not the right way to make security effective. Often the most likely result of too extensive security documents and procedures are frustrated employees and a low degree of compliance. If we could measure the effect of our security measures, we would know how to get the most benefit from our investments. Normally management does not have a clue about how effective their information security measures are.

Critical information is an asset that needs protection against threats such as competitors or other agents with malicious intentions. Social engineering is a non-technical threat that targets people, the human aspect of information systems. People can be viewed as the weak link, or the strong link in the security chain, depending on their skills and ability to act in given situations.

When we want to assess social engineering robustness, one obvious question is what indicators are suitable for this purpose? Social engineering robustness is a characteristic that is difficult to measure directly. Therefore an indirect approach is suggested, where factors that are assumed to affect employees' basis for decision making when facing an attempted social engineering attack, are chosen. A metric will be described in order to answer whether user oriented security measures are effective or not against a social engineering threat.

Defining social engineering robustness

By social engineering we mean manipulation of individuals to make them do or say something they normally would not have done or said. This definition is broad and needs to be limited into an information security perspective, which means manipulation of people who normally has an interest in upholding information assurance to carry out actions that threaten the information security properties; confidentiality, integrity or availability of an information system.

By robustness we mean resilience over a time period. A robust person will be more difficult to manipulate compared to a less robust person. The time aspect is important to show that robustness covers resilience over time, not just a single incident that could be flavoured with coincidence or luck. Robustness also covers the ability to reveal attempted attacks and perform damage containment after a successful attack. An attack that is revealed and reported will definitely put the organisation in a better position to act. This calls for the following definition:

> *Social engineering robustness is "Ability to reveal and report malicious requests over time that intentionally manipulate the receiver to perform actions in order to jeopardise the confidentiality, integrity or availability of information".*

To what extent can social engineering be measured?

Measurements are single observations of specific factors, while a metric is a construct of several measurements over time. Shirley Payne describes in [5] good metrics as specific, measurable, attainable, repeatable and time dependent (SMART). Some people, like McHuge in [1] dismiss the idea of measuring security due to lack of scientific foundation to support it. If someone comes up with a totally new way to perform an attack, we cannot use numeric values to classify security of an information system. Others, like McCallam in [2] oppose numeric values as a measure

ISSN 0085-7130 © Telenor ASA 2005

of information assurance. His main argument is that information assurance must be viewed from an extended PTO (People Technology Organisation) perspective, where these factors are integrated. Despite explicit resistance against numeric measures, he suggests a resilience index with a scale from zero to ten. At level zero the attacker has full control, while the defender has complete control at level ten. This opens up an interesting comparison related to social engineering. If it is easy for an attacker to succeed, it could relate to a low value on an index on social engineering robustness. On the other end of the scale we could imagine that social engineering attacks would be impossible to perform unless the defender allows it. In other words, qualitative levels of social engineering robustness are possible to derive.

Absolute security is neither desirable nor obtainable according to Odlyzko in [3]. The flexibility that comes with bending rules is necessary in order to be effective. One example is a trusted secretary that forges the manager's signature in situations where he or she obviously would have accepted. This releases more time to the manager, but is undeniably a violation of rules. Odlyzko claims that people like to have some slack in life, and that most of us have problems relating to rules and formal methods. This is exactly the weakness social engineering exploits. Odlyzko suggests creating speed bumps in order to slow down execution and possible effects of attacks. If we increase the attacker's risk of being revealed, an organisation's social engineering robustness will increase accordingly.

Constructing a metric

It is necessary to know how a typical attack is carried out before one makes a program to measure social engineering robustness. Knowledge about how to resist social engineering attempts is also needed. This insight can provide ideas about factors to choose from, and then select the SMART [5] ones to be part of this metric. Møllerhaug [6], Allen [7], Gulati [8] and Mitnick [4] all describe typical social engineering attacks. From these descriptions, an attack can be divided into three phases:

1 Contact
2 Empathy
3 Abuse

Contact is a necessary precondition for the attack to take place, and can be done through personal meetings, telephone or e-mail. The second phase is to achieve empathy, where the attacker turns his or her problem over to the victim. Often it is about making an exception from regulations and standing instructions in order to exploit a window of opportunity or to fix a problem that has occurred. A skilled attacker is conscious on sending out signals that appeal to the victim, in order to make it easier to accomplish the underlying objective of the meeting, manipulation and abuse. Empathy can be achieved from various angles such as common interests, flatter, flirt, helplessness, or other factors that will trigger the victim's empathy. The third phase is the abuse, where the attacker takes advantage of the situation created in order to carry out the original intention behind the attack.

The decision how to respond to a request for help can be right or wrong. The request will always be part of a context. By context we mean in what connection is the attempted contact made. This connection will vary, and be surrounded by noise. By noise we mean factors that can divert attention and affect the outcome of the decision. The attacker will consciously create noise in order to call for the victim's empathy. An interpretation of the request's legitimacy will finally form the decision on how to respond. The decision is right if attacks are revealed and legitimate requests are met and carried out. The decision will definitely be wrong if an attacker gets help or a legitimate request is wrongly turned down.

Decision-making foundation as a parameter

Interpretation of the legitimacy of a request is difficult to measure in a reliable and valid manner. Factors like noise, interpretation and empathy are not SMART [5], and for that reason less suitable for our purpose. Interpretations of various situations are by nature subjective, will differ from one situation to another and will not give us reliable measurements.

If we assume that preconditions for decision making affect the outcome, it will be easier to get reliable results by measuring the foundation for decisions instead of the decision itself. In other words, we measure indirect indicators that have an influence on security awareness and factors that we believe help make the right decisions. The existence of a written security policy can be observed, confirmed or dismissed. Whether the policy contains clear and simple rules is more difficult to measure, but is still possible through a qualitative approach according to McCallam [2].

The same argument can be used about a training programme, where a question whether a user-training session has taken place can be answered accurately. Our assumption is that a security training program can contribute to the overall security awareness, and thus affect social engineering robustness. Training could improve the ability to reveal and report malicious requests. Another question is how relevant the training programme is in a social engineering context. Nevertheless, a completed training programme will only give an indirect indication on social engineering robustness.

A security or authorisation conversation is in some organisations part of new employment procedures and periodically repeated, say annually. One purpose of an authorisation conversation, besides granting access to company assets, should be a listing of relevant threats and other issues that influence on the overall security awareness. The existence of a formal authorisation conversation programme could give another indirect indication on social engineering robustness. If the content of the conversations covers relevant threats such as social engineering, it would be another indirect indicator of robustness. Again, the security conversation is not sufficient by itself; it should be based on an existing security policy and refer to a programmed training session.

In our definition of social engineering there are two central aspects described, the "ability to reveal and report malicious requests". While the first is difficult to measure, the latter meets Payne's recommendation for metrics to be SMART. Reports can be measured reliably. Møllerhaug also highlights "reporting system" as decisive in order to put up a defence against social engineering attacks. This factor is therefore given a stronger emphasis in this metric, and for the same reason gives a higher score than security policy, user training, and security conversation. If results from reports do not find their way back to all employees, the reports will eventually dry out and the reporting system will lose its purpose.

Factor selection

We chose the factors security policy, user training, security conversation and reporting system because we believe they are all relevant to social engineering robustness. Further, they are concrete and observable. The content can be adjusted by management influence, which means the metric is relevant for a company management. Effects of changes can be measured validly and reliably. Finally, one measurement will give a level of robustness than can be compared to a threshold or a standard.

The factors are assumed to affect the individual precondition of not being manipulated when confronted with a social engineering attacker. In other words, this metric is about how well management has prepared the preconditions for employees and information system users. We believe that a good preparation of the chosen factors has a connection to individual and collective robustness.

Measurement method

Each employee in a department is given a questionnaire that is to be answered anonymously. Anonymity is chosen to achieve a higher degree of validity since we assume respondents will answer more truthfully in confidence that their name cannot be traced. It is no objective for this metric to link respondents to identities, only departments. The responses will draw a picture of how good the respondent assesses existence and quality of the listed factors. Measurements can be done for an organisation as a whole, a department or a sub-department.

It is the perception of the individual employee that this metric captures, not what management claim is implemented. An example would be if an employee answers that there is no security policy, while one in reality exists. If this person does not know of any policy, it means that the policy does not exist for him or her. The same goes for the other factors as well. This means that an individual who in the questionnaire confirms the presence of a security policy, training, security conversation and reporting system will achieve a certain score. But if the policy is considered to be bad, which means failure to satisfy criteria such as clear and simple rules that are not remembered by the respondents, a full score cannot be achieved. The respondent makes assessments of the given criteria by grading statements and claims about the quality of the security policy. The answers will give a good indication on how well the security policy fulfils its purpose as a tool to create the necessary precondition for the employee to make the right decisions.

The results of each questionnaire will provide a score for each employee. What will be an acceptable number can be assessed based on the company overall security ambition. Management can set a target on social engineering robustness. This target can in some cases be differentiated where some departments are expected to have a higher degree of robustness and thus given more resources to obtain and maintain this expectation.

Organisation description

We tested the metric on an organisation to see if we got results to support our assumptions that a robustness level could be assessed. Here are some facts about the unit:

- 150 employees. Of these are
  - 15 employees that have information assurance as their primary function.
  - 8 trainees in the information assurance field.

- A security policy for the information system exists
- An authorisation conversation programme exists.
- An information system-training programme exists.

Calculating score

The score is based on the respondent's perception of existence and individual assessment of each of the four factors security policy, training, security conversation and reporting system. Each factor has a potential of 20 or 40 points depending on the answers. Each questionnaire gave a picture of how respondents viewed the factors related to given criteria. These criteria were coded into statements the respondent should consider whether he or she agreed to or not. A total of 100 points could be scored, and the respondent could grade answers from 1 (disagree) to 6 (agree), referred to as answer value in the tables below.

This method to assess social engineering robustness is based on an assumption that the selected factors affect employees' ability to reveal and report possible attacks. It is difficult to predict exactly how people will react in certain situations and this is why the validity of our results must be addressed. Each decision has its unique setting, and we do not know the details of the context at a given situation. This metric is based on general factors that we believe affect people's ability to respond to a possible attack.

The result is an assessment of social engineering robustness and can be compared to a specified target. It is possible to define an exact number as a target on how robust an organisation should be. The ambition should be derived from the overall security target for information assurance, which should be based on the operational need to uphold a certain information

Nr	Attributes	Max	Calculation	Remark
1.1	Does a written security policy exist?	5	Yes = 5, No or don't know = 0	Yes is a precondition for score on other questions
4.2	Consist of short and simple rules?	3	Answer value / 2	Correlation with 4.1
4.3 4.4	Is it accepted?	3	Answer value / 4 Answer value / 4	
4.1	Is it understood?	3	Answer value / 2	Correlation with 4.2
4.5	Do you comply?	3	Answer value / 2	Correlation with 4.6
4.7	Is it available?	3	Answer value / 2	
4.6	Do others comply?	0		Correlation with 4.5

Security policy (20)

Nr	Attributes	Max	Calculation	Remark
1.3	Have you been through a user-training programme for the information system?	5	Yes = 5, No, or don't know = 0	Yes is a precondition for score on 5.1-5.6
5.1 5.6	Security relevant?	3	Answer value / 4 Answer value / 4	Correlation with 5.8
5.2 5.3 5.4	Instructions on reporting system?	9	Answer value / 2 Answer value / 2 Answer value / 2	
5.8	Knowledge test?	3	Answer 1 = 3, 2, 3, 4, 5, 6 = 0	Should be a clear stand
5.5	Correlation to security policy?	0		Should be correlation
5.7	Attitude towards user training programme	0		

User-training (20)

Nr	Attributes	Max	Calculation	Remark
1.4	Security conversations are practised formally	5	Yes = 5, No, or don't know = 0	
2.1 2.2	Have been through a security conversation	3	Answer value / 4 Answer value / 4	
2.3	Relevant content	3	Answer value / 2	
2.4 2.5	Employer engagement	3	Answer value/ 4 Answer value / 4	
2.6	Active focus on threats	3	Answer value / 2	
2.7	Knowledge test	3	Answer 6 = 3, 1, 2, 3, 4, 5 = 0	Should be a clear stand

Security conversation (20)

Nr	Attributes	Max	Calculation	Remark
1.5	Existence of formal reporting system?	7	Yes = 5, No, or don't know = 0	Precondition for countermeasures
3.2 3.8	Sanction free to report?	12	7 - Answer value Answer value	Trust is considered decisive for employees to report
3.7	Is reporting encouraged?	3	Answer value / 2	
3.3	Where should reports be sent?	3	Answer value / 2	
3.4	What should be reported?	3	Answer value / 2	
3.6	Feedback to the sender of the report?	3	Answer value / 2	
3.9	Feedback to all employees?	3	Answer value / 2	
3.1	Precondition to reveal an attack?	3	Answer value / 2	
3.5	Precondition to report a possible attack?	3	Answer value / 2	

Reporting system (40)

Category	Average	Std deviation	Min score	Max score
Information assurance personnel	74.66	10.07	52.00	87.00
Information assurance trainees	64.13	10.61	53.50	81.25
Other employees	61.09	14.23	27.25	81.25
All	64.81	14.01	27.25	87.00

Results

security level. This is important, because security has no meaning unless it is viewed as an integrated part of the ongoing production.

This metric will enable management to measure how employees assess effectiveness of the respective user-oriented security measures. It is also possible to reveal deviation between actual levels to an explicit target for social engineering robustness.

Contradicting ideals

It is difficult to measure social engineering robustness for several reasons. Robustness is a security aspect that in this context is about the ability to reveal and report malicious requests. To treat security issues without an operational context would be too narrow. Since employees have to face other demands and expectations than just social engineering robustness, it is necessary to find a balance between security and

ISSN 0085-7130 © Telenor ASA 2005

effectiveness. Expectations about being service minded and effective are difficult, but not impossible, to merge with social engineering robustness expectations. It is naïve to expect such contradicting ideals to be fully reached by co-workers that do not have clear and simple rules. Efficiency is a western society ideal, where the will to solve problems and do things easier and cheaper is great. New technology and continuing development is brought to the market by people who improve existing solutions and challenge established routines.

We also see a conflict between the ideal to comply rules and the ideal to be effective. Who has not violated the speed limit on deserted wide roads or stepped a little extra on the accelerator while passing the slow car in front? We break rules on a daily basis in order to achieve something we consider more important or desirable than following rules. In the speeding example the interest of saving time and the fact that the risk of something going wrong seems acceptable to us, make us violate rules we know apply. We do not see negative consequences as a likely result of our actions, or we mistakenly assess the consequences too low. This goes for a social engineering victim as well; where the attacker's malicious intentions are well hidden.

Very few of us follow every rule all the time. The previous example illustrates that many people are willing to break rules if the risk tied to the violation is assessed as acceptable and the payoff exceeds the estimated possibility for bad consequences. People like the flexibility in bending rules and are dependent on it in order to be effective according to Odlyzko [3]. He states that absolute compliance of every rule is neither desirable nor realistic. A natural conclusion is to put more emphasis on detection and reaction after a social engineering attack has taken place. This is the explanation why existence of an effective reporting system is viewed as the most important factor in this metric to assess social engineering robustness.

Bibliography

McHugh, J. *Quantitative Measures of Assurance: Prophecy, Process, or Pipedream?* CERT/CC, Software Engineering Institute, Carnegie Mellon University, 2001.

McCallam, D. *The Case Against Numerical Measures for Information Assurance*. Logicon Northrop Grumman Company. San Diego, University of California, 2001.

Odlyzko, A. *Economics, Psychology, and Sociology of Security*. Economics of Security, Financial Cryptography 2003 Conference, 2003.

Mitnick, K. *The Art of Deception – Controlling the Human Element of Security*. Wiley, 2002.

Payne, S C. *A Guide to Security Metrics*. SANS Security Essentials GSEC Practical Assignment, 2001.

Møllerhaug, S. *Social Engineering*. Næringslivets Sikkerhetsorganisasjon Sikkerhetskonferansen, 2003.

Allen, M. *The Use of 'Social Engineering' as a means of Violating Computer Systems*. SANS Institute, 2001. http://www.sans.org/rr/papers/51/529.pdf

Gulati, R. *The Threat of Social Engineering and Your Defense Against It*. SANS Institute, 2003. http://www.sans.org/rr/papers/51/1232.pdf

Ola Holm is a Major in the Norwegian Defence Security Agency. His current position is senior instructor at the Norwegian Defence Engineering School. He has an MSc degree in information and communication system security from the Royal Institute of Technology, Sweden.

email: ola-holm@online.no

Helping prevent information security risks in the transition to integrated operations

JOSE J. GONZALEZ, YING QIAN, FINN OLAV SVEEN AND ELIOT RICH

Jose J. Gonzalez is Professor at Agder University College, Norway

Ying Qian is PhD fellow at Agder University College, Norway

Finn Olav Sveen has just finished his Master study at Agder University College, Norway

Eliot Rich is Assistant Professor at University at Albany, USA

The transition to Integrated Operations in the Norwegian oil and gas industry is expected to yield up to 30 % reduction in costs and a 10 % increase in production. The success of the transition hinges on mastering the information security problems introduced by the ***e***Operation – the gradual substitution of traditional offshore operations – drilling, production, delivery, etc, mostly locally operated at the offshore platforms – with increasing remote onshore operation via computer networks. An ongoing research project uses various kinds of system dynamics models to help prevent information security risks in the transition to Integrated Operations.

1 Using models to define a problem

Consider a Chief Information Officer (CIO), a Chief Information Security Officer (CISO) or a manager concerned with planning and decision-making in matters of information security. When designing policies and strategies, the problem landscape might span disciplines as varied as computer networks, wireless technology, information technology, cryptography, software, hardware, organization science, psychology, law, etc. The issues would have a typical time horizon of several years at hand and they would involve technology, organization, processes and human behavior. Hence one deals with a system characterized by feedback, temporal change, nonlinear dynamics, time delays, soft factors, and interdisciplinary aspects. A discipline explicitly designed to manage systems characterized by the above factors (feedback, dynamics, time delays, soft factors, interdisciplinary aspects) is system dynamics [1–3].[1)] System dynamics models differ in conception, nature, scope, purpose and use from models that are commonly used in most disciplines. In particular, the ability of system dynamics to help articulate complex dynamic problems, using various kinds of models, is still little known in most enterprises. System dynamics recognizes that problems exist originally as highly incomplete mental models and it sees as one of its main aims to improve the mental models upon which decisions are based [4]. To achieve this, system dynamics uses external problem representations ("models") of various kinds, e.g. causal-loop, influence and sector diagrams; system archetypes; feedback analysis; and stock-and-flow models. Some of these system dynamics models are qualitative, others are quantitative [2, 3]. Together they can be amazingly effective in improving initially poor mental models to adequate mental models of the complex dynamic problem. It is difficult to grasp the unfamiliar features of system dynamics models and to appreciate their advantages other than by becoming a system dynamics modeler or – and that is the theme of this paper – by participating in a group model building workshop [5, 6] as "client" (a.k.a. expert in some domain that contributes with relevant information for the problem to be modeled).

Figures 1 and 2 illustrate the transition from poor, fragmented mental models to de-fragmented, common mental representations of a given complex problem. To deal perfectly with the problem, the sum total of problem-related knowledge in the enterprise would have to cover the whole problem domain. However, much of the existing knowledge is fragmented in individual minds, with little overlap and few networks utilizing the fragmented knowledge. In addition, people can "know" a lot that isn't so [7].

The transition to a "common" mental model of the complex problem would typically involve increased individual knowledge, greater overlap of individual knowledge, de-fragmentation of knowledge through networks that utilize knowledge that is dispersed in individuals, and correction of erroneous knowledge.

Figure 2 should be considered as an intermediate step in the transition to adequate mental representations of a given complex problem. Notice, however, that it will never be possible to achieve a perfect situation where individuals possess all the problem-related knowledge. It is widely known that modern society rests on division of labor; less known, but equally incontestable, is the fact that individuals can only have a fraction of the knowledge possessed by the totality of those who by discipline or profession deal

[1)] *Originally, the discipline was called "industrial dynamics" and such is the title of ref. [1]. For a comprehensive exposition of modern system dynamics – "business dynamics" – see ref. [3]. Ref [2] is still an excellent introduction to the discipline, but the reader needs to reinterpret the models (expressed in the first-generation Dynamo language) in terms of modern system dynamics languages such as ithink, Powersim or Vensim.*

ISSN 0085-7130 © Telenor ASA 2005

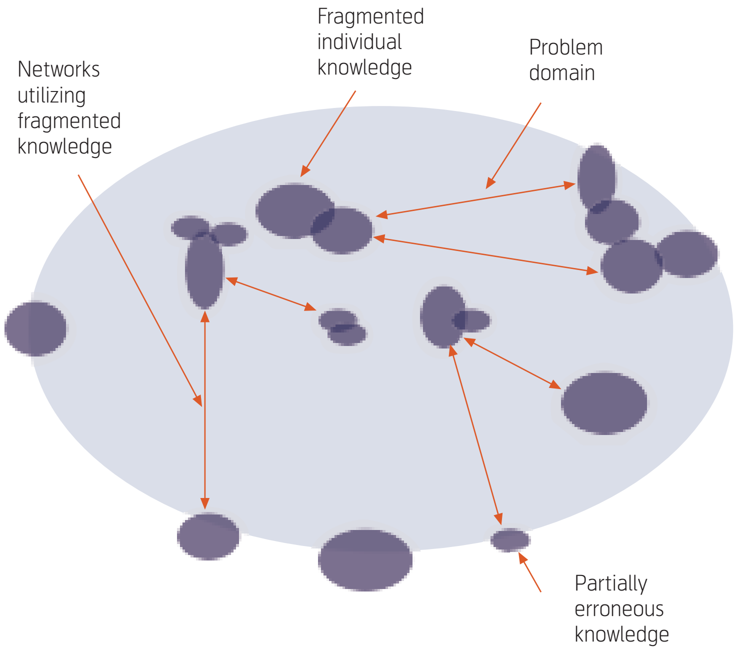

Fragmented individual knowledge
Problem domain
Networks utilizing fragmented knowledge
Partially erroneous knowledge

Figure 1 The problem-related knowledge in an enterprise is initially vastly fragmented and the total knowledge is quite insufficient with respect to the challenge (represented as the oval "problem domain"). There are many partial and unrelated views, and some of the "knowledge" might be wrong

with aspects of any complex enterprise problem. In fact, beyond such fragmentation of knowledge one should humbly recognize that complete knowledge is impossible. As Hayek [8, p. 12] expressed it: "Complete rationality of action … demands complete knowledge of all the relevant facts. A designer or engineer needs all the data and full power to control or manipulate them if he is to organize the material objects to produce the intended result. But the success of any action in society depends on more particular facts than anyone can possibly know. And our whole civilization in consequence rests, and must rest, on our believing much that we cannot know to be true." Accordingly, we can and should improve the mental models upon which decisions are based, but we will forever have to live with imperfect knowledge and not fully adequate mental models. In Figure 2 this fact is indicated in the sum of available knowledge as a sub-domain of the oval representing the ideal total problem-related knowledge.

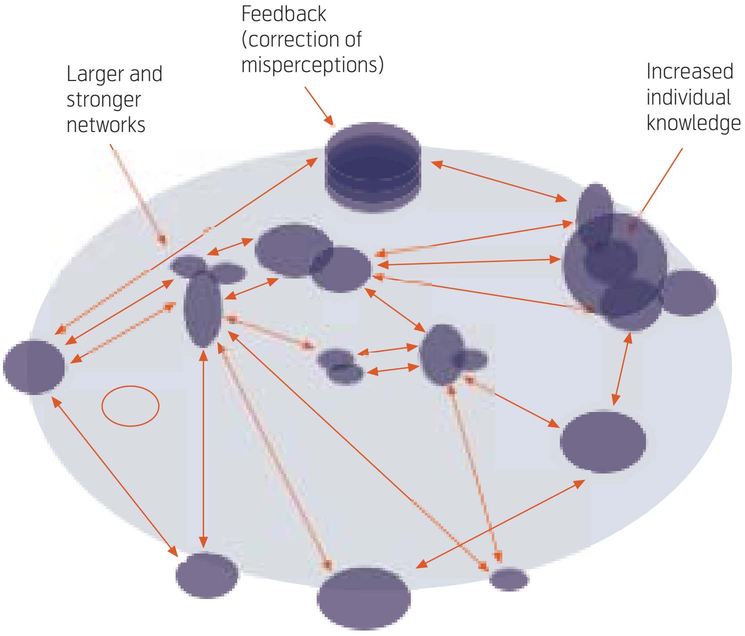

Feedback (correction of misperceptions)
Larger and stronger networks
Increased individual knowledge

Figure 2 Transition to common mental model of the complex problem. Note the increased individual knowledge, more overlap of individual knowledge, de-fragmentation of knowledge through networks utilizing resources that are dispersed in individuals and correction of errors and misperceptions through appropriate feedback.

To be specific, the remainder of this paper describes preliminary results from a particular "enterprise challenge", namely an ongoing project about information security in the Norwegian oil and gas industry.

2 Vulnerabilities in the transition to integrated operations

Consider Integrated Operations (formerly called *e*Operations) in the Norwegian oil and gas industry. The aim is to increase production by 10 %, reduce costs by 30 % and extend the lifetime of mature fields in the Norwegian offshore sector through better utilization of drilling and production data, and closer collaboration between offshore and land-based personnel. Integrated Operations is a process that began with a long-term scenario for the Norwegian continental shelf as described in Report no. 38 (http://www.dep.no/oed/norsk/publ/stmeld/028001-040009/index-dok000-b-n-a.html) to the Norwegian Parliament; the project evolves through two generations (integration of on- and offshore operations, ca. 2003 –2010; integration of companies, ca. 2007–2015).

From the point of view of information security, the transition over a long time span (10–12 years) from traditional offshore operations – drilling, production, delivery, etc, mostly locally operated at the offshore platforms – to Integrated Operations, with increasing remote onshore operation, is an "engine" that generates vulnerabilities. In this context, vulnerabilities are weaknesses of the Integrated Operations environment that facilitate unintended or intended incidents. An unintended incident could occur if an onshore operator – believing that the system is in test mode – inadvertently closes valves, thus causing an organizational accident and down-time. A mixture of an intended and unintended incident could be caused by a contractor who – under maintenance operations – connects to the Integrated Operations' intranet. The contractor might inadvertently introduce malware from his PC to the intranet; i.e. he might act as (super) Trojan Horse for all kinds of malicious agents. An intended incident could be a planned cyber attack, exploiting that an onshore operator is wireless connected to the Integrated Operations intranet (assuming that the wireless connection is a weak point of the system).

In addition to those risks, the countless software vulnerabilities (a.k.a. system vulnerabilities) in commercial-off-the-shelf (COTS) software are weak points for potentially disastrous attacks [9, p. 9]. In their paper, "Thirty Years Later: Lessons from the Multics Security Evaluation" [10], Karger and Schell argue that the decades-old Multics operating system (which was used in a relatively benign closed environment) is more secure than most operating systems of today: "Given the understanding of system vulnerabilities that existed nearly thirty years ago, today's 'security enhanced' or 'trusted' systems would not be considered suitable for processing even in the benign closed environment. Also, considering the extent of network interconnectivity and the amount of commercial-off-the-shelf (COTS) and other software without pedigree (e.g. libraries mined from the Web), today's military and commercial environments would be considered very much 'open'. … Thus, systems that are weaker than Multics are considered for use in environments in excess of what even Multics could deliver without restructuring around a security kernel." Integrated Operations is a (nearly) closed environment, but according to Karger and Schell, its reliance on COTS makes it insecure at the outset, even as detached environment. But then, the Integrated Operations network is not a fully detached environment, since contractors and suppliers are allowed to connect; further weak points are its remote accessibility to authorized personnel via (still quite vulnerable) wireless connections.

Summarizing, the Integrated Operations intranet is exposed to unintended human failures and it is vulnerable to malicious agents (insiders, combinations of insider/outsider, outsiders with different motives – hackers, criminals, terrorists). Incidents can lead to down-time and they might even jeopardize safety.

3 A group model-building workshop

A group model-building workshop was held at Agder University College, Grimstad, Norway, in May 2005 to consider the information security risks in the transition to Integrated Operations. This two-day event brought together experts in the offshore oil industry, computer security, psychology, and system dynamics. A facilitation team from the University of Albany led the sessions. This team has over the last decade developed several protocols for organizing and running such sessions [5, 11, 12].[2)]

The workshop had three goals: The first was the development of a common mental model, anchored on situational knowledge, roles and responsibilities, procedural knowledge, and cultural knowledge. Developing a collective understanding of each stakeholder's concerns and perspectives would advance the goals for secure and safe operations under the proposed Integrated Operations approach.

The second goal of the workshop was the development of a set of causal models that would explicate the mental model. This in turn would enable the development of formal simulations to elicit structural, parametric, and other information. Once constructed, these models can be extended to interactive learning environments that can be used to promote common mental models and enhanced organizational learning.

The third goal was to identify the substantive dynamic hypotheses that were driving the Integrated Operations. A dynamic hypothesis is a statement of the anticipated change in a system's state over time [2, 3].

In September 2005, a second group model-building workshop will be held. One of the authors (YQ) will develop preliminary, quantitative stock-and-flow models to be used as input for the discussions with domain experts in the workshop. It is expected that the second group model-building workshop should lead to calibrated stock-and-flow models for policy analysis, as basis for Interactive Learning Environments and to assist in auditing of information security systems in the Integrated Operations environment.

In a parallel paper [13], two of the authors have described some outcomes of the workshop, namely the feedback structures that appear to influence the outcomes of the transition to Integrated Operations and the use of CSIRTs[3)] to detect, analyze and mitigate the effects of computer security events and threats. In this paper we use system archetypes to express insights derived from the group model-building workshop during a follow-up secondary analysis in the four weeks after the workshop. Before introducing the concept of system archetypes we explain

[2)] *The group model-building workshop is part of the AMBASEC (A Model-based Approach to Security Culture) and the IRMA (Incident Response Management) projects that are funded by the Research Council of Norway (grants Nos. 164384/30 and 164372/30, respectively). AMBASEC is anchored at Agder University College, while IRMA is a SINTEF project.*

[3)] *CSIRT stands for Computer Security Incident Response Team.*

briefly the development of a common mental model in the first group model-building workshop.

4 Developing a common mental model

We refer again to Figures 1 and 2 as metaphors for the transition from fragmented mental models to a common mental model of the enterprise challenge.

The group model-building workshop creates a shared mental model in a piecemeal process of adding problem-related knowledge bits through a series of group exercises: stakeholder and policy diagrams (Figures 3 and 4), behavior over time (Figure 5), and dynamic stories (one example given in Figure 6). Figures 3–6 are intended for a general illustration of the group model-building process only. The reader does not need to dwell on the details of these figures, since the

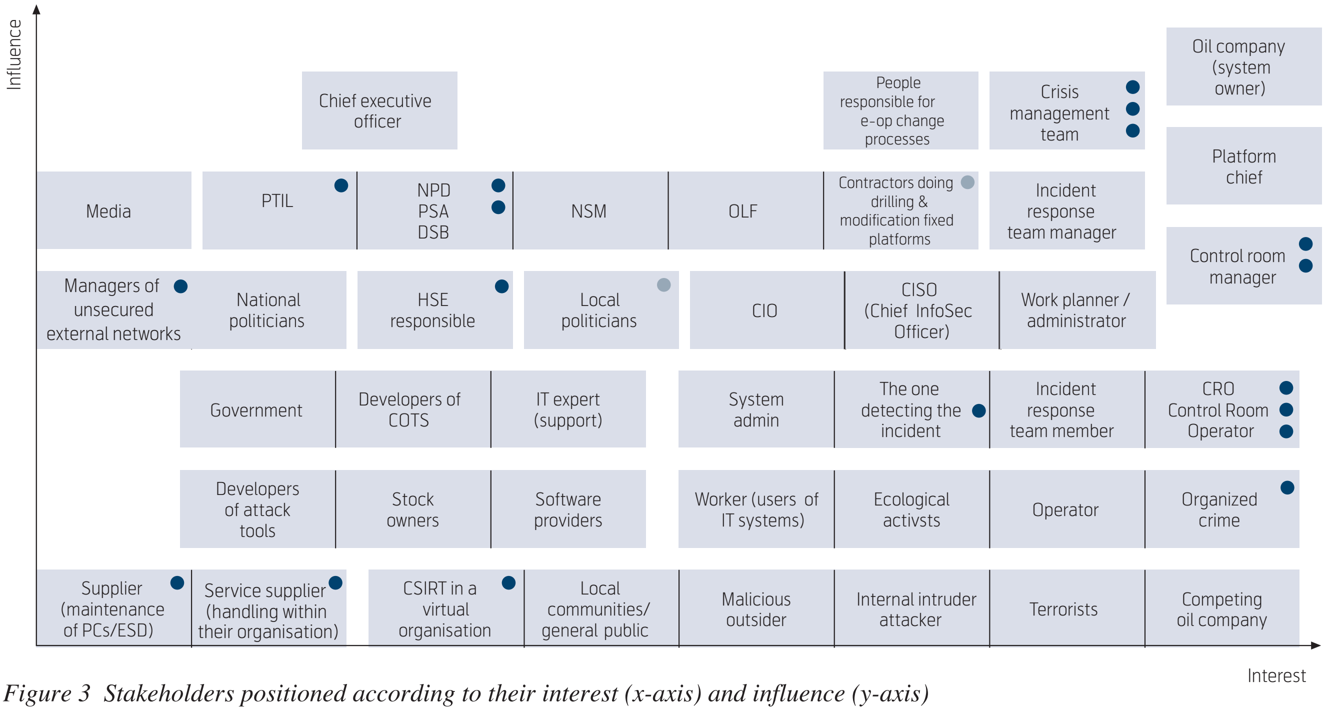
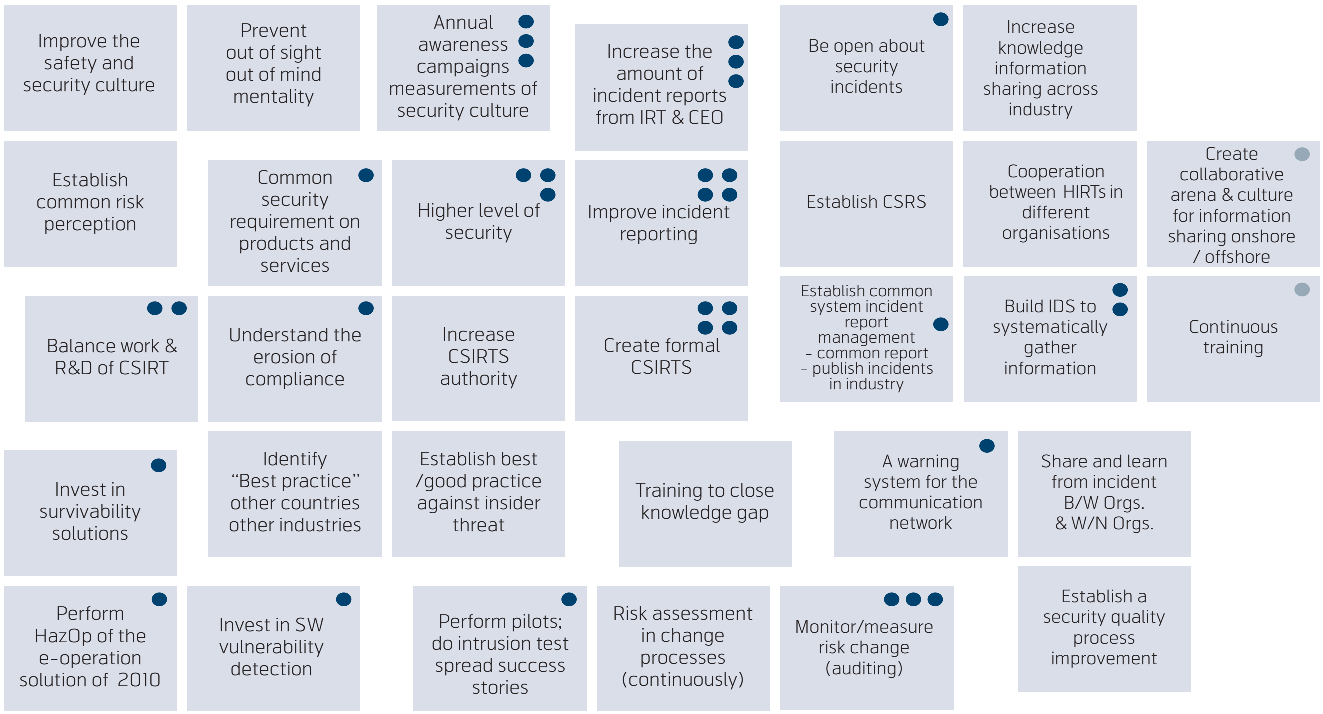

Influence
Chief executive officer
People responsible for e-op change processes
Crisis management team ●●●
Oil company (system owner)
Platform chief
Media
PTIL ●
NPD PSA DSB ●●
NSM
OLF
Contractors doing drilling & modification fixed platforms ○
Incident response team manager
Control room manager ●●
Managers of unsecured external networks ●
National politicians
HSE responsible ●
Local politicians ○
CIO
CISO (Chief InfoSec Officer)
Work planner / administrator
Government
Developers of COTS
IT expert (support)
System admin
The one detecting the incident ●
Incident response team member
CRO Control Room Operator ●●●
Developers of attack tools
Stock owners
Software providers
Worker (users of IT systems)
Ecological activsts
Operator
Organized crime ●
Supplier (maintenance of PCs/ESD) ●
Service supplier (handling within their organisation) ●
CSIRT in a virtual organisation ●
Local communities/ general public
Malicious outsider
Internal intruder attacker
Terrorists
Competing oil company
Interest

Figure 3 Stakeholders positioned according to their interest (x-axis) and influence (y-axis) on the enterprise challenge. Note that some stakeholder positions might change according to circumstances (e.g. media interest)

● Prioritization by group members

○ Added day two


Improve the safety and security culture
Prevent out of sight out of mind mentality
Annual awareness campaigns measurements of security culture ●●●
Increase the amount of incident reports from IRT & CEO ●●●
Be open about security incidents ●
Increase knowledge information sharing across industry
Establish common risk perception
Common security requirement on products and services ●
Higher level of security ●●●
Improve incident reporting ●●●●
Establish CSRS
Cooperation between HIRTs in different organisations
Create collaborative arena & culture for information sharing onshore / offshore ○
Balance work & R&D of CSIRT ●●
Understand the erosion of compliance ●
Increase CSIRTS authority
Create formal CSIRTS ●●●●
Establish common system incident report management - common report - publish incidents in industry ●
Build IDS to systematically gather information ●●
Continuous training ○
Invest in survivability solutions ●
Identify "Best practice" other countries other industries
Establish best /good practice against insider threat
Training to close knowledge gap
A warning system for the communication network ●
Share and learn from incident B/W Orgs. & W/N Orgs.
Perform HazOp of the e-operation solution of 2010 ●
Invest in SW vulnerability detection ●
Perform pilots; do intrusion test spread success stories ●
Risk assessment in change processes (continuously)
Monitor/measure risk change (auditing) ●●●
Establish a security quality process improvement

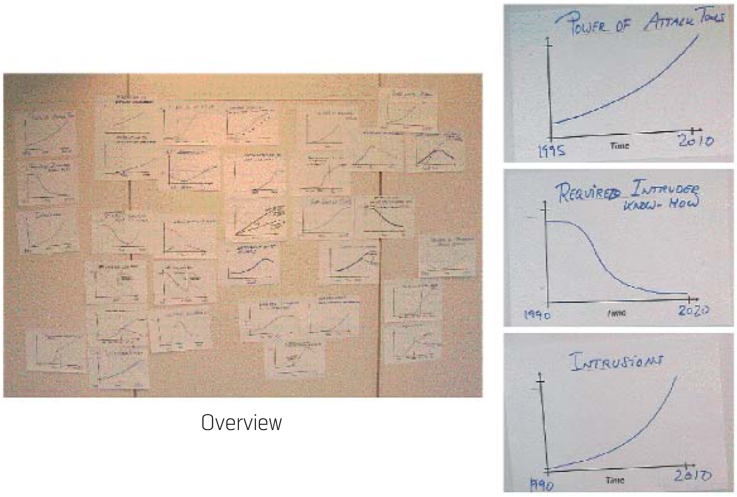
● Prioritization by group members

○ Added day two

Figure 4 Policy diagram (not yet structured according to categories)

specific information needed for the modelling in Sections 5–6 will be fully contained in the text below.

The products developed at the group model-building workshop constitute pieces of shared mental representations of the enterprise challenge. It remains to assemble these pieces to a coherent whole. Normally, after the first group model-building workshop one or more system dynamics modelers proceed to develop iteratively various system dynamics models (causal-loop diagrams, system archetypes, stock-and-flow models). In follow-up group model-building workshops the client group interacts with the group model-building experts to correct, extend and improve the model outcomes until a shared model (or several shared models) is (are) unanimously accepted. In this particular case the status corresponds to the follow-up stage of the first of a series of group-modeling workshops. Nevertheless, one is already able to characterize important aspects of the enterprise challenge in terms of "problem" archetypes with indicated "solution" archetypes.



Overview

Figure 5 The team exercise to suggest the behaviour over time of problem-related variables led to numerous diagrams ("Overview"). Three specific examples are shown on the right-hand side

5 System Archetypes

Experience shows that most actions result in unintended consequences. Typically: a remedy for a problem causes another, unintended problem; e.g., low tar and nicotine cigarettes actually lead to increased intake of carcinogens, CO, etc, as smokers compensate for the low nicotine content by smoking more frequently [3]. Road construction to relieve road congestion works only for a brief period; the new road increases the attractiveness of using cars, so that more people will buy and use private cars instead of public transportation. After some time, traffic congestion will at least as bad as it was before – but with higher traffic density, pollution and city sprawl [14]. Most people fail to recognize the feedback structure that causes the unintended consequences.

In the business world we often get repeating patterns; i.e. the same basic model structures occur in many different problems. Over the years a number of basic

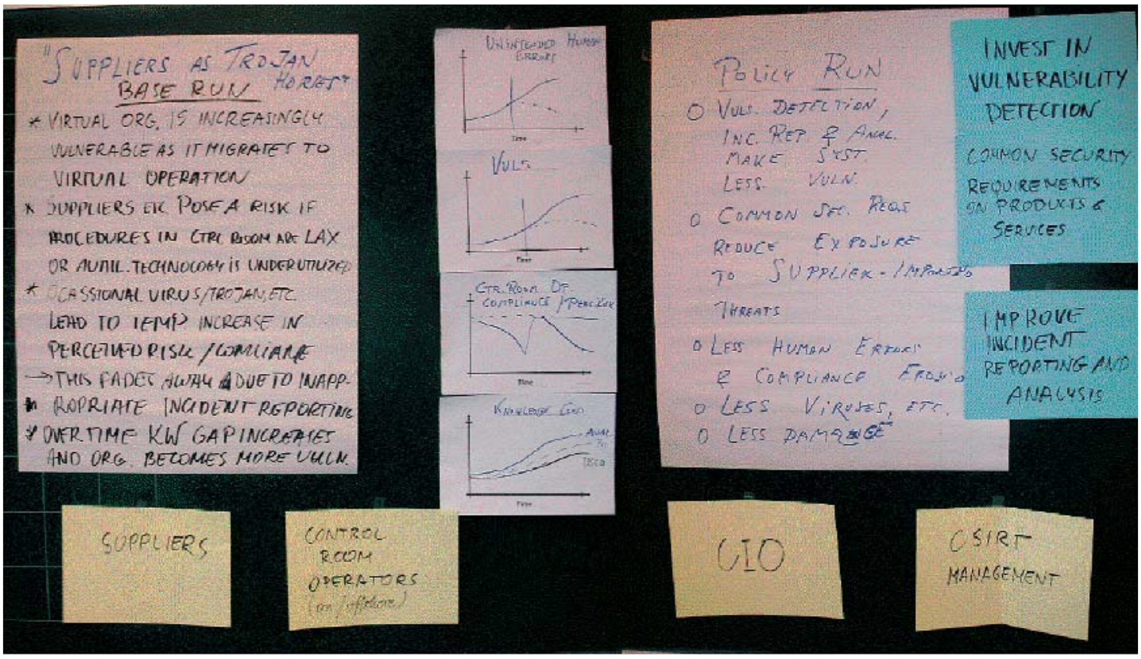

"SUPPLIERS AS TROJAN HORSE"
BASE RUN
* VIRTUAL ORG. IS INCREASINGLY VULNERABLE AS IT MIGRATES TO VIRTUAL OPERATION
* SUPPLIERS ETC. POSE A RISK IF PROCEDURES IN CTRL ROOM ARE LAX OR AUTH. TECHNOLOGY IS UNDERUTILIZED
* OCCASSIONAL VIRUS/TROJAN, ETC. LEAD TO TEMP INCREASE IN PERCEIVED RISK/COMLIANCE
→ THIS FADES AWAY DUE TO INAPPROPRIATE INCIDENT REPORTING
* OVER TIME KW GAP INCREASES AND ORG. BECOMES MORE VULN.
UNINTENDED HUMAN ERRORS
Time
VULS
Time
CTR. ROOM OP. COMPLIANCE / PERC. RISK
Time
KNOWLEDGE GAP
Time
POLICY RUN
o VULN. DETECTION, INC. REP & ANAL. MAKE SYST LESS VULN.
o COMMON SEC. REQS REDUCE EXPOSURE TO SUPPLIER-IMPORTED THREATS
o LESS HUMAN ERRORS & COMPLIANCE EROSION
o LESS VIRUSES, ETC.
o LESS DAMAGE
INVEST IN VULNERABILITY DETECTION
COMMON SECURITY REQUIREMENTS ON PRODUCTS & SERVICES
IMPROVE INCIDENT REPORTING AND ANALYSIS
SUPPLIERS
CONTROL ROOM OPERATORS (on/offshore)
CIO
CSIRT MANAGEMENT

Figure 6 Dynamic stories link pieces of information form other workshop outcomes to a scenario. The final system dynamic model should be able to reproduce the dynamic stories. One of the dynamic stories describes how suppliers might inadvertently introduce data viruses, worms, spyware, etc. when connecting to the Integrated Operations network and thus act as Trojan Horse for malicious insiders

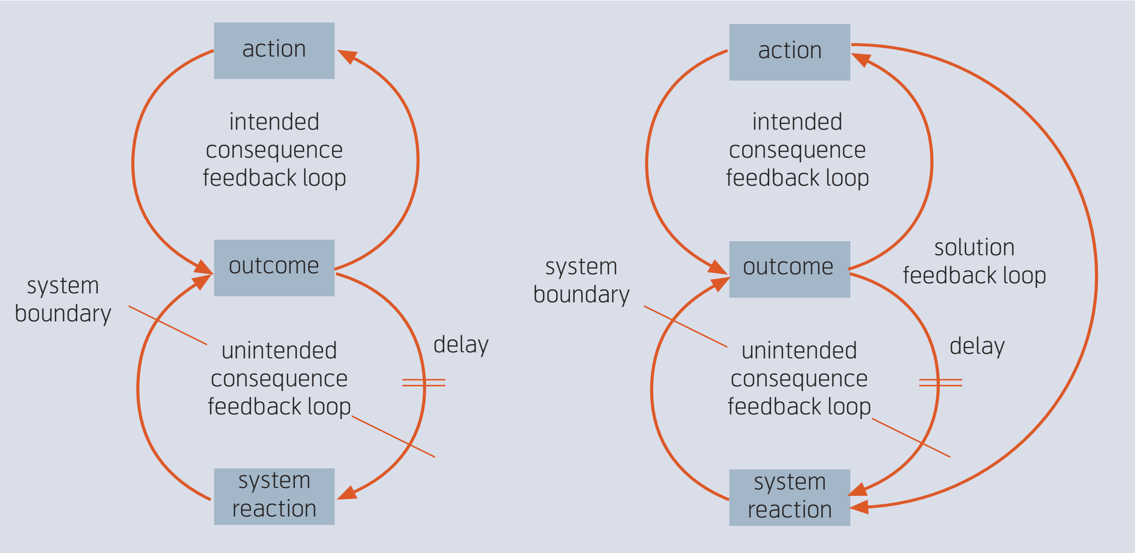

action
intended consequence feedback loop
outcome
system boundary
unintended consequence feedback loop
delay
system reaction
action
intended consequence feedback loop
outcome
solution feedback loop
system boundary
unintended consequence feedback loop
delay
system reaction

Figure 7 A generic problem archetype (left) and the corresponding solution archetype (right)

model structures – so-called system archetypes – consisting of multiple feedback loops have been identified [15, 16]. Wolstenholme [17, 18] has shown that there are four totally generic system archetypes. They consist of only two feedback loops, the first loop referring to the intended outcome and the second to the unintended outcome. See Figure 7 for an example of a system archetype ("problem" archetype on the left-hand side and "solution" archetype on the right-hand side).

The feedback loops are either reinforcing (R) or balancing (B). The four totally generic archetypes are called, 1) *Underachievement* (R for the intended, B for the unintended outcome); 2) *Relative Achievement* (R for the intended, R for the unintended outcome); 3) *Out of Control* (B for the intended, R for the unintended outcome); and 4) *Relative Control* (B for the intended, B for the unintended outcome). Typically, the unintended consequence occurs with a significant time delay that makes it difficult for the decision maker to link the action with the unexpected outcome. Wolstenholme emphasized also the existence of boundaries and the need to include them in system archetypes (or, for that sake, in system dynamics models). Boundaries exist in all organizations; they may be physical or mental. In a company, an action may have unintended consequences in another division, but they are not visible to those triggering the action because of the physical or mental boundary.[4)]

Wolstenholme also showed that a "solution archetype" exists for each "problem archetype". The solution archetype adds a further feedback loop (it can be reinforcing or balancing, depending on the particular problem archetype) that inhibits the unintended outcome triggered by the intended outcome.

Archetypes are short-hand versions of more complex models. Archetypes are almost never detailed enough to facilitate a formal simulation, but they are excellent for communicating insights and knowledge about the dynamics of a system. They can be easily understood by people who have little or no training in system dynamics.

In the following section we present some problem archetypes that appear to describe information security problems associated with the transition to Integrated Operations; these problems in turn feed back on the performance of Integrated Operations as unintended consequence. We also sketch potential solution archetypes.

6 System archetypes in the transition to integrated operations

New elements *(Operation Transition)* are added to *Integrated Operations*. The intended outcome is a rapid and successful implementation of Integrated Operations, in that success will facilitate faster and smoother subsequent transitions, reinforcing the move to Integrated Operations (a reinforcing, R, feedback loop). The unintended consequence of the transition to Integrated Operations is the introduction of numerous unknown vulnerabilities – unseen and unexpected due to the delay in their manifestation and the system boundary that hides the vulnerabili-

[4)] *Wolstenholme uses the term "system boundary", but a more appropriate term might be a "masking boundary," i.e. a boundary that masks consequences in one part of the system from an actor in another part of the system. The term "system boundary" suggests a boundary that encompasses the system as a whole.*

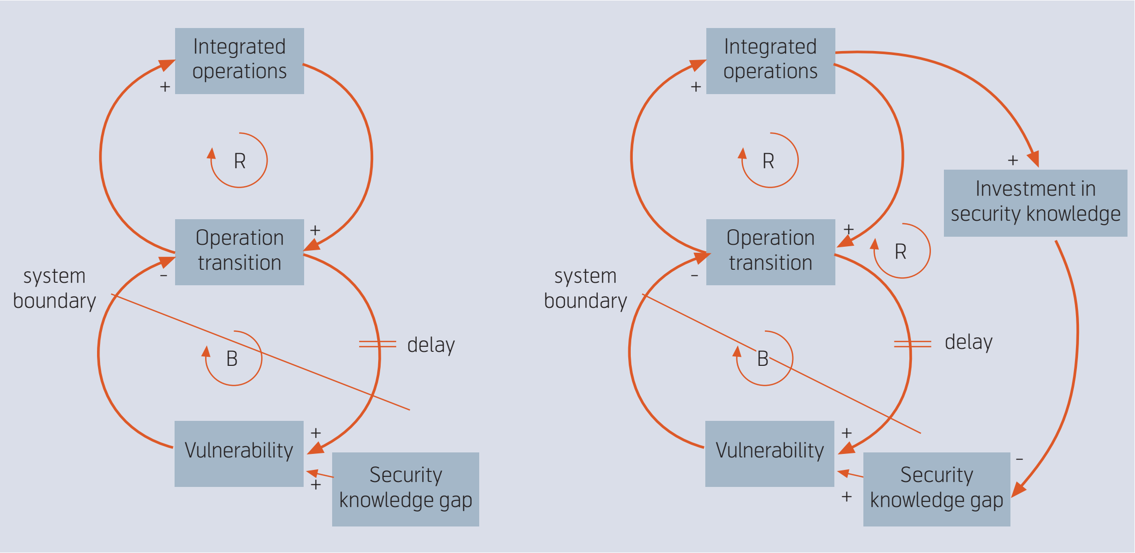


Figure 8 Underachievement due to the generation of security vulnerabilities

ties. Added elements *(Operation transition)* increase *Vulnerability*; increased vulnerability means an increase in the number of security incidents; the increase in security incidents will then compromise the intended transition to Integrated Operations. This is depicted as a balancing (B) feedback loop that threatens the success of Integrated Operations (implementation delays, unexpeced costs). The combined effect of the intended (R) and unintended (B) actions is an Underachievement archetype (shown in the left-hand side of Figure 8.)

The solution archetype (right-hand side of Figure 8) indicates that a successful transition to Integrated Operations requires a thoughtful effort to eliminate the security knowledge gap associated with new technologies and processes, and that is responsible to the introduction of the vulnerabilities. If the security knowledge gap is eliminated promptly and to the necessary extent while the transition to Integrated Operations proceeds, the generation of many vulnerabilities can be prevented and potential risks from extant vulnerabilities can be diminished.

Figure 9 shows a Relative Achievement archetype. Investment in CSIRT (Computer Security Incident Response Team) capacity improves detection of intrusions. As more intrusions are detected, risk perception increases, leading to more investment in CSIRT capacity. On the other hand, more detected intrusions will lead to less trust to system and partners, which in turn leads to more even investment in CSIRT capacity. The combined effect of the two reinforcing feedback loops could be over-investment in CSIRT capacity and internal trust problems. However, the more serious problem would be if the re-

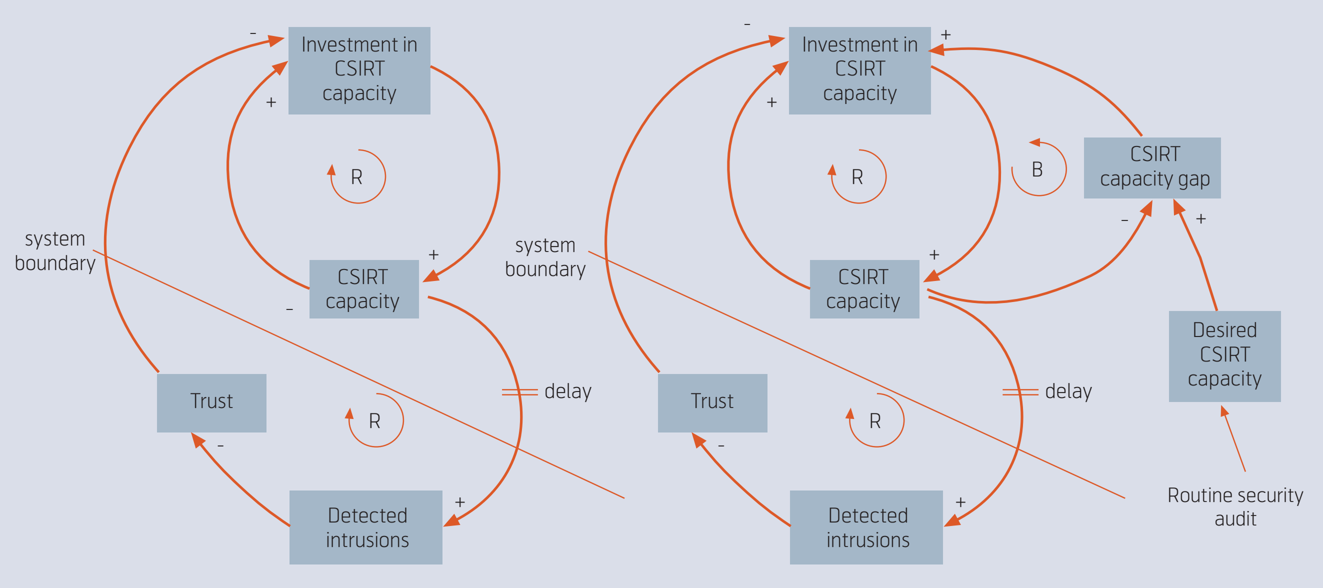


Figure 9 Relative Achievement – trust as double-edged sword

ISSN 0085-7130 © Telenor ASA 2005

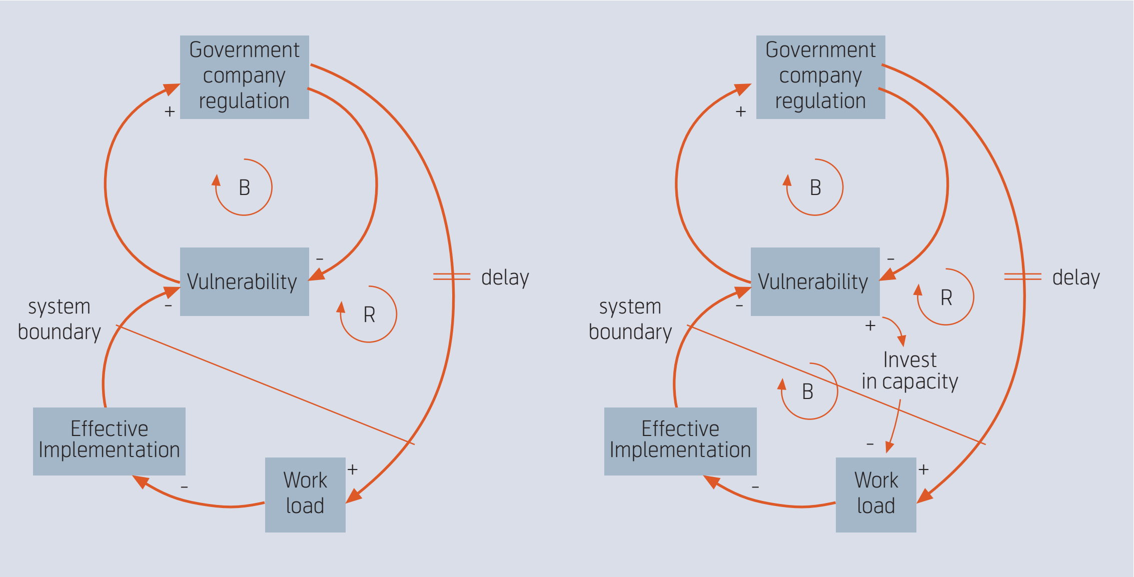


Figure 10 An Out of Control archetype – erosion of compliance due to excessive work load

inforcing loops operate in the opposite sense – the trust trap: low investment in CSIRT, low detection of intrusions and high, reckless trust to system and partners. Again, delays and system boundaries make it difficult to see that the low level of detected incidents might be caused by low capacity to detect them. The solution archtetype in the right hand side of Figure 9 adds routine security audits to assess the real situation and determine the desired CSIRT capacity. Investment decisions should be made according to the gap of desired CSIRT capacity and the actual CSIRT capacity. Typically, security audits correct the picture provided by intrusion detection systems (elimination of false positives and of low-priority attacks from further consideration); alternatively, security audits would tell if intrusion detection capacity is insufficient. In both cases, the resulting action would be a correction of investment in CSIRT capacity.

Finally, Figure 10 shows an Out of Control archetype. Facing higher vulnerability, the enterprise or even the government tends to propose various regulations, standards or supplementary rules for operation to control the vulnerabilities. However, due to the organizational boundary ("system boundary") and time delays, executives might fail to see that the newly introduced policies introduce extra workload to operators. Overworked operators will cut corners, ignoring some regulations to get their work done in time. This renders regulations less effective, with the unintended consequence of causing higher vulnerability to Integrated Operations (problem archetype in the left-hand side of Figure 10). The solution link for this archetype is to invest in necessary workforce capacity.

7 Concluding remarks

System dynamics models developed in the group model-building workshops occur in several shapes and serve different purposes (e.g. as causal-loop diagrams and system archetypes). They help articulate enterprise problems and they create shared understanding. Even at this early stage these simple models might suggest potential policies to prevent delayed unintended consequences of executive actions. In addition, causal-loop diagrams and system archetypes anchor the development of more ambitious, quantitative system dynamic models for refined study of scenarios, design of advanced policies, development of Interactive Learning Environments for improved security culture and audit systems to improve the detection of complex security risks.

8 References

1 Forrester, J W. *Industrial Dynamics*. Cambridge, MA, MIT Press, 1961.

2 Richardson, G P, Pugh, A L. *Introduction to System Dynamics Modeling*. Portland, OR, Productivity Press, 1981.

3 Sterman, J D. *Business dynamics: systems thinking and modeling for a complex world*. Boston, Irwin McGraw Hill, 2000.

4 Sterman, J D. A Skeptic's Guide to Computer Models. In: Richardson, G P (ed). *Modelling for Management*. Aldershot, UK, Dartmouth Publishing Company, 1996.

5 Richardson, G P, Andersen, D F. Teamwork in Group Model Building. *System Dynamics Review*, 11 (2), 113–137, 1995.

6 Vennix, J A M. *Group Model Building – Facilitating Team Learning Using System Dynamics*. Chichester, John Wiley, 1996.

7 Gilovich, T. *How we know what isn't so: the fallibility of human reasoning in everyday life*. New York, Free Press, 1991.

8 Hayek, F v. *Law, rules and order*. London, Routledge and Kegan, 1973.

9 Lipson, H F. *Tracking and Tracing Cyber-Attacks: Technical Challenges and Global Policy Issues*, 2002. March 27, 2004. [online] – URL: http://www.cert.org/archive/pdf/02sr009.pdf.

10 Karger, P, Schell, R. Thirty Years Later: Lessons from the Multics Security Evaluation. In *Eighteenth Annual Computer Security Applications Conference*, Las Vegas, Nevada, 2002.

11 Andersen, D F, Richardson, G P. Scripts for Group Model Building. *System Dynamics Review*, 13 (2), 107–129, 1997.

12 Luna-Reyes, L F, Andersen, D L. Collecting and Analyzing Qualitative Data for System Dynamics: Methods and Models. *System Dynamics Review*, 19 (4), 271–296, 2003.

13 Rich, E, Gonzalez, J J. Maintaining Security and Safety in High-threat E-operations Transitions. To appear in the *Proceedings of Thirty-ninth Annual Hawai'i International Conference on System Sciences (HICSS-39)*, Hawaii, January 4–7, 2006.

14 Downs, A. *Still Stuck in Traffic – Coping with Peak-Hour Traffic Congestion*, 2nd ed. James A. Johnson Metro Series. Washington, DC, Brookings Institution Press, 2004.

15 Senge, P M. *The Fifth Discipline: the Art and Practice of the Learning Organization*. New York, Doubleday/Currency, 1990.

16 Kim, D H. *System Archetypes*. Cambridge, MA, Pegasus Communications, 1992.

17 Wolstenholme, E F. Towards the definition and use of a core set of archetypal structures in system dynamics. *System Dynamics Review*, 19 (7), 7–26, 2002.

18 Wolstenholme, E F. Using generic system archetypes to support thinking and modelling. System Dynamics Review, 20 (4), 341–356, 2004.

Jose J. Gonzalez is Professor of system dynamics and information security at the Faculty of engineering and science, Agder University College, Norway. He leads the Security and Quality in Organizations group with two postdoctoral fellows and four PhD fellows. In addition to numerous publications in the fields of system dynamics and information security, Dr. Gonzalez was co-founder of Powersim, developer of one of the leading system dynamics tools.
email: Jose.J.Gonzalez@hia.no

Ying Qian is Ph.D. fellow at the Security and Quality in Organizations group, Faculty of engineering and science, Agder University College, Norway. She took a Master degree with the system dynamics group at the University of Bergen 2002-2004. Before that, Ying Qian took Bachelor studies at Fudan University in Shanghai.
email: Ying.Qian@hia.no

Finn Olav Sveen has recently finished his Master study in Industrial and information management at Agder University College, Norway. His Master thesis was a system dynamics study of expansion to adjacencies.
email: ooryl@spray.no

Eliot Rich is Assistant Professor in the Department of Information Technology Management, School of Business, University at Albany, USA. He uses qualitative and quantitative modeling techniques to study problems in information systems, computer security and knowledge management. Dr. Rich received his PhD in Information Science from the University at Albany, an MPP from Harvard University, and a BA from Brooklyn College.
email: e.rich@albany.edu

ISSN 0085-7130 © Telenor ASA 2005

Four reasons why 100 % security cannot be achieved

JAN A AUDESTAD

Jan A. Audestad is Senior Adviser in Telenor

The paper analyses various reasons why protection against malicious attacks on our computers, networks and infrastructures is not 100 % achievable despite how much resources we are willing to spend on countermeasures. The first reason is purely mathematical and fundamental in the meaning that there is no way in which it can be circumvented. It is simply not possible to develop a method that enables us to detect all types of malicious software.

The computer systems and society have developed into networks that are tremendously complex. Recently it was discovered that many of these networks are vulnerable in a new and surprising way. Furthermore, direct handling of security problems in these networks results in computationally intractable algorithms. This sets a practical but real limit to what can be protected. It is probably more important to make these network structures fault tolerant rather than just trying to remove the fault by insufficient protection means.

Finally, economic and managerial constraints limit the extent of protection we are willing to pay for. It is difficult to calculate how much is economically gained by investing in protection. Furthermore, the amount of historical data is insufficient in order to base decisions on experience as we do when we insure ourselves against fire and flood. One severe problem is that the structural problems of the societal and technological networks are still not recognised to their full extent by decision makers. Therefore little money is so far channelled into research in this important area.

If we say that the methods we have today protect us *x* %, *x* is probably a non-computable quantity. We simply do not know how efficient or inefficient current information security methods are.

1 Introduction

From time to time I hear that even professional computer scientists claim that it is possible, though extremely difficult, to construct firewalls and other gadgets that can protect the computer system against all types of malicious code – known or unknown – that can destroy the computer. The claim is also that this task, of course, cannot be done in practice because of insufficient knowledge, lack of time and limited resources. However, without these constraints, the ultimate defence system could have been constructed. Alas, this claim is not true as is demonstrated below: for fundamental mathematical reasons it is not possible to construct the ultimate defence system, and there is no way in which this problem can be circumvented: the conclusion is ultimate.

This does not mean that we should give up the design of computer defence systems. Despite the mathematical limitation, there is a lot we can do in order to protect ourselves against malicious attacks. Some of these protection mechanisms may even protect us against future attacks of a nature that is unknown to us at present. What is important is that we are aware of the fact that ultimate protection is not possible and that we cannot predict the severity of the information threat with absolute confidence. This may cause us to think about the vulnerability of the computerised society in a new way and concentrate our efforts on designing robust systems that can withstand attacks without being completely destroyed.

Around 1900 the great German mathematician David Hilbert put forward the thesis that, if a universal mathematical language not relying on the vagueness of human language is developed and from that an appropriate set of axioms is constructed, absolutely all mathematical theorems can be proved. The effort to develop a *formal axiomatic method* of mathematics was called the Hilbert Programme. Several mathematicians spent much energy on the programme, but the effort came to an end in the 1930s because of the discovery of two of the most fundamental theorems in mathematics. The first discovery was the incompleteness theorem (or the proof of unprovability!) of Kurt Gödel (1931) stating that an axiomatic mathematical theory may contain true theorems that cannot be proved within the theory. The second event, and the most important for us, was the proof of the undecidability theorem of Allan Turing (1936) stating that it is not generally possible to construct a general algorithm that can be used to decide whether an arbitrary problem has a solution or not.

This observation alone tells us that absolute security is unattainable. In addition, technical and operational complexity, limited resources and unawareness of possible threats are other reasons why it is even more difficult to achieve absolute security.

These aspects are discussed in what follows.

ISSN 0085-7130 © Telenor ASA 2005

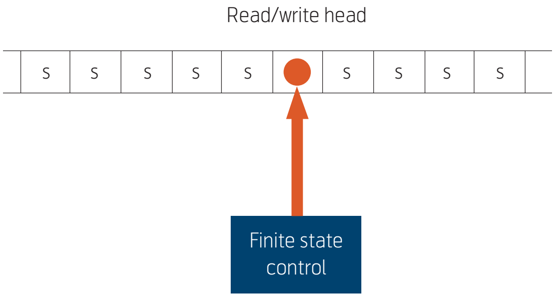


Figure 1 Basic Turing machine

2 Mathematics: Achieving the impossible

It is impossible to construct a computer program (or algorithm) that, given an arbitrary computer program as input, is able to decide whether the computer program contains a malicious code or is a harmless or useful program. This is a corollary of the undecidability theorem. Let us see how this comes about.

Turing used a simple computer model similar to that shown in Figure 1, now called the Turing machine, in order to study the formal, mathematical aspects of computation. Turing proposed his model long before the digital electronic computer was designed. The machine turned out to be so powerful that Alonso Church put forward the hypothesis that every problem that is computable on any digital computer can be computed on a Turing machine. This hypothesis cannot be rigorously proved because "computable" remains an informal notion but it has been shown that the Turing machine is equivalent to the most general mathematical notions of computing (recursion theory) and represents a formidable tool in the study of this area of mathematics. Furthermore, no alternative formalism has been found that cannot be mapped to a language that can be handled by the Turing machine.

The machine consists, in its simplest but general form, of a finite state control, a tape consisting of cells containing a symbol (that may be a blank), and a read/write head. The fundamental computing cycle of the machine is as follows. The machine reads the symbol in the cell and depending on the symbol, the current state, and the "next move function" defined for the finite state control:

- selects the new state (which may happen to be the same state as the current state);
- prints the same or a new symbol in the cell; and
- keeps the tape in the same position or moves the tape one cell to the right or to the left.

Then either a new computing cycle starts or the machine halts because the computation is finished.

In the initial state, the symbols written on the tape represent the problem to be solved. When the computation comes to completion, the machine halts and the tape contains the result.

We can also construct a Turing machine containing several tapes and several read/write heads; for example, one input tape with read-only head and one output tape with write-only head. It is easy to prove that this configuration is equivalent to another Turing machine containing just one tape and one read/write head. Interacting Turing machines may interact by sharing a set of common tapes and having several read/write heads. Interacting Turing machines are also equivalent to a single machine with a single tape and a single read/write head. These facts can be rigorously proved and are important for the rest of this section: the complex configuration is useful for formulating the problem while the simple configuration is easy for analysing it.

Turing proved the following theorem:

> *There does not exist a single algorithm (or Turing machine) which, given as input the description of another algorithm A (or Turing machine) and an input x to that algorithm, will answer* true *if and only if A terminates (returns an answer) on input x.*

An algorithm is the same as a computer program that can be executed on a digital computer and, in accordance with Church's hypothesis, is equivalent to a Turing machine. What this theorem states is that it is not always possible to find out before computation starts whether the computer will continue searching for a solution forever or halt and return an answer in finite time. The theorem is amazingly easy to prove and is found in introductory textbooks in the field of mathematical logic called recursion theory and in textbooks on theoretical computer science.

What is important to us is the following corollary to Turing's theorem:

> *There is no algorithm by which it can be decided for an arbitrary Turing machine M, a state q of M and an input sequence σ to M, whether the Turing machine M will ever enter state q.*

How does this relate to malicious software?

The question is now: is it possible to define an algorithm which, given as input the description of a computer and the program input to it, can answer whether the computer will enter a state called a *bad state* where the computer does not operate properly – for example, halt without producing the proper result? According to the corollary, this is impossible provided that bad states exist. We know from experience that bad states exist in computers (though it may be impossible to find all bad states – another case of Turing's theorem).

The algorithm we are thus seeking is one that takes any computer program (or input sequence) and answers the question whether a given computer will enter a bad state executing the program. The arguments above based on the undecidability of the halting problem of Turing machines show that this is not possible.

Several objections may be raised against these arguments.

The number of malicious software codes that can be constructed in finite time is finite so that we are asked to solve a finite problem, which obviously can be done in finite time. It is true that the malicious codes are a subset of the finite set containing all computer programs designed now and the next (say) 1000 years. However, this set is finite in the practical sense that only a finite number of people have designed software in finite time. In the mathematical sense it is infinite because an infinite number of computer programs can be constructed. Suppose that this was not the case – that is, the number of computer programs that can be constructed are actually finite – then take an arbitrary computer program in the finite list of possible computer programs and add a tautology, for example, $a ::= a$ (a is replaced by a) where a is a variable already contained in the program, and you have designed a new computer program that was not in the original list. If you are still not convinced because the new program does not provide a new result, choose a dummy variable a and compute, say, $a ::= a + 1$; *output a*. The new program is not in the original list and produces a new result. This is a trivial example of the diagonal method of Cantor – claim that a complete list (finite or infinite) of occurrences of a mathematical object exists and then show that there is at least one identifiable object that is not in the list. The Turing halting theorem is proved using a similar argument.

From these arguments we see that the problem is certainly not finite in the system (namely mathematics) where we try to solve it. What we are asked to do is to select a finite subset from an infinite set, and this can certainly be done in infinitely many ways.

All bad states can be removed, given enough time and resources. This is not true and is another application of the halting problem. A similar argument as the above shows that it is impossible to construct a computer program that detects all possible programming errors in an arbitrary computer program and all execution events that may lead to unexpected or unwanted behaviour of the computer. The error may be triggered by the internal program execution or by an external input (for example a virus). Note also that execution of a program solving a given problem involves simultaneous execution of many other programs of the runtime system of the computer. It is therefore not enough to remove programming bugs from each application program but from all runtime programs and compilers of the computer, and also make sure that no misalignment problem exists between the application programs and the runtime programs.

The computer may never be in a location where it will receive malicious software. If this is true, then there is no problem for that particular computer – and this is how certain systems should be designed, for example, the control system of a nuclear power station. The computer must then be effectively isolated form all other systems. If it is not totally isolated but connected to other systems even via firewalls executing severe access restrictions, there will always be a possibility that the computer will receive malicious software.

However, is an isolated system really isolated? What about maintenance and software and hardware upgrading? Do these activities require interconnection with computers that do not belong to the isolated system so long as maintenance takes place? What are the risks of contamination during such periods? Some of these computers may even belong to the manufacturer of the system. Are removable memory devices, diskettes, compact discs, personal digital assistants or other mobile devices ever allowed in the isolated system? When a computer is installed for the first time in a system it may contain Trojan horses, hidden logical bombs and backdoors. There is no way in which we can find this out and we may conclude that isolated systems do not exist.

The following example illustrates how difficult this question is.

One problem that was never resolved was whether it is possible to spread malicious software via signalling systems in the telecommunications network. The access control is extremely severe for such systems: only messages satisfying stringent conditions concerning format and content are accepted. However, there are exceptions. Signalling System No 7 and its cousin on the ISDN subscriber line (DSS.1) contain fields that allow the users or the network to insert information that are not screened by the access control. Is it then possible to insert malicious code in these fields in such a way that the computers in the signalling network and the terminals execute it? This problem was seriously discussed by telecom operators and manufacturers in ITU and ETSI during the early 1990s without coming to conclusions except that utmost care had to be taken when executing the content of such fields.

Because of all these restrictions, a completely isolated computer can hardly do any useful work!

Firewalls and daily updates of virus programs solve the problem. This is simply not true because these devices can only protect us against what we know and not necessarily against what is new and we have never experienced before. The deficiency of these defences is simply what this section has been all about!

3 Technology: Faced with complexity

The second item has to do with the complexity of computer systems. During the last decade, the complexity of computer systems has grown enormously. It is estimated that the number of active CPUs on Earth is in the order of 10,000 billion devices. CPUs are found everywhere: in personal computers and printers, in servers, routers and multiplexers, in mobile phones and personal digital assistants, in automobiles and airplanes, in production tools and machines, in traffic-lights and road-toll systems, and so on and so forth. Most of these CPUs are not controlled directly by people – they are performing autonomous functions. We do not even know that these devices exist until one of them fails and the car stops for inexplicable reasons.

The CPUs are interconnected via internet. This means that every CPU can, in principle, communicate with any other CPU. The CPUs then represents a network consisting of 10,000 billion nodes. The links in this network interconnect CPUs that are performing common tasks, for example, exchanging email, executing distributed programs, or retrieving or disseminating information or commands. A path in the network consists of a contiguous chain of linked CPUs. There may be several paths between two CPUs. The path may also consist of just one link. Though the CPUs at each end of such a chain may not exchange information directly, information dissemination or retrieval may take place via intermediate CPUs. This is in fact one way in which malicious software is spread between computers. The more interconnected this network is, the faster the software is spread.

The CPUs in this network are the kernel of computers containing from a few to millions of executable programs and data files. These programs and files may also, in principle, be viewed as nodes in a huge, distributed computing infrastructure. Two nodes are connected by a link if the two programs or files are used in, for example, the same computation. The structure of this network is not the same as that of the CPU network though the interconnection between programs and files is realised over the CPU links.

The whole structure is built on top of internet. Internet enables CPUs to communicate. This results in a layered configuration as shown in Figure 2. The lowest layer is internet where routers are nodes and links are communication channels. The middle layer represents the CPUs and their interconnection. The structure of this network is independent of the structure of internet although the CPUs are connected to routers as shown.

The upper layer consists of programs and files. The software runs on computers but the network structure at the software layer is independent of the CPU layer below.

The overall structure of the computing infrastructure is very complex. However, some of the complexity is avoided just by splitting the structure into structurally independent layers as done in the figure. The model allows us to study each layer independently searching

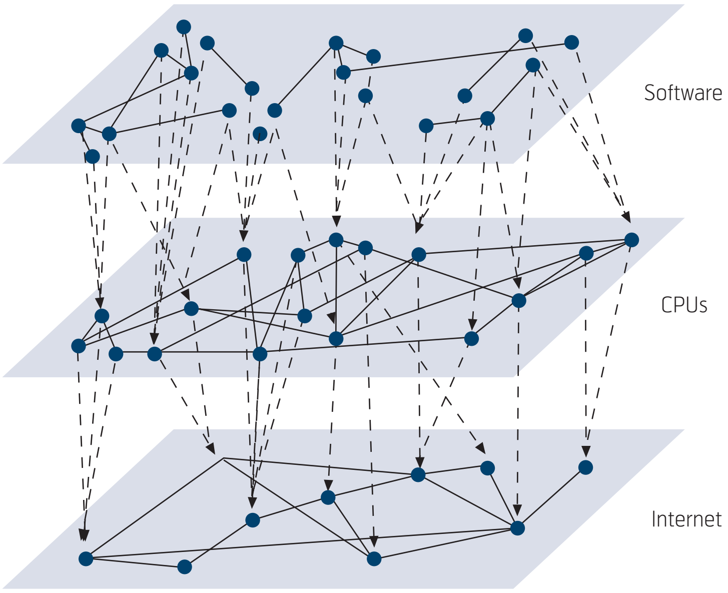


Figure 2 Layers of complex networks

for vulnerable spots and creating robustness. The model also shows that the computing infrastructure can be attacked independently at the three layers using different means. The mode of attack and the damage done will be different on each layer. Therefore, separate countermeasures against attacks of the computing infrastructure must also be designed for each level: it is not enough to protect one layer only, for example, the internet.

New investigations in graph theory has shown that the vulnerability – that is, how easy it is to destroy the network if an attack against it is orchestrated in the most efficient way – depends not only on the robustness of each node in the graph but more strongly on the topological structure of the graph. We may say that the network is *structurally vulnerable*. Independent security measures in each node may then not be the best strategy of protection. Some networks may be robust against any type of attack while other networks such as the World Wide Web, internet and email are vulnerable to attacks directed at certain nodes in the network. The structure of some networks, such as the World Wide Web, is particularly vulnerable to such attacks. Figure 2 shows that the computing infrastructure consists of at least three networks with independent structure. Structured attacks at any of these networks may cause damage to the computing infrastructure.

Figure 2 may give the impression that the computing infrastructure is static. But this is not the case: the system is very dynamic. The temporal development of each layer is independent of the other layers and involves several stochastic processes:

- addition or removal of links and nodes goes on all the time;

- dynamic reconfiguration of each of the three layers of networks takes place simultaneously at several timescales ranging from milliseconds to years, and there are reasons to believe that this is a fractal structure;

- the number of nodes and links within each layer increases with time resulting in three expanding network topologies;

- the connectivity between adjacent layers changes dynamically (CPUs roam between nodes in internet and software modules migrate between computers (CPUs)).

There is no method by which we can make such complex and dynamic systems secure against attacks by malicious software. Even if it is theoretically possible to design a secure system – that is, not violating the halting theorem of Turing machines – it will not be possible to find computationally feasible algorithms that will do the job. The reason has to do with computational complexity.

Some recursive problems (i.e. a problem solvable by a Turing machine) can be solved in polynomial time. This means that if the size of the problem is n, then the time required to solve the problem is less that n raised to some power; that is, $T_n < n^k$ where k is a constant exponent. If n is large, then the time it takes to solve a slightly larger problem $n + m$ ($m << n$) is $T_{n+m} < (n + m)^k \approx T_n (1 + km/n)$. If $k = 10$, $n = 1{,}000{,}000$ and $m = 10$, then the computation time has increased by a factor less than 1.0001, or, in other words, it is just a little harder to solve the new problem. If it took one second to solve the original problem, it takes only 0.1 milliseconds more to solve the new problem.

Problems that can be solved in polynomial time is said to belong to complexity class P.

However, there are problems that cannot be solved in polynomial time. Or more accurately; it has not been proved that such problems do not exist. This is the most important outstanding problem in computer science. That is, the time required to solve them increases at least exponentially with the size of the problem. The time to solve a problem of size n is then proportional to e^{an} where a is a constant; that is, $T_n \sim e^{an}$. The time to solve a problem of size $n + m$ is then $T_{n+m} \sim e^{a(n+m)} \sim T_n e^{am}$. If $n = 1{,}000{,}000$, $a = 1$ and $m = 10$, then the time to solve the new problem has increased by a factor of about 22,000. If it took one second to solve the original problem, it takes 60 days to solve the new problem. This is in sharp contrast to the polynomial case.

Problems that cannot be solved in polynomial time are said to be NP-complete.

Many apparently simple network problems are proved to be NP-complete, for example, to find the shortest closed path in a complete network containing all nodes only once (the travelling salesman problem), and to determine whether it is possible to colour all the nodes in a network with just three colours such that adjacent nodes (i.e. nodes connected directly by a link) do not have the same colour (the 3-colouring problem). (The size of these problems is the number of nodes in the network.)

There are also many problems that have not been proved to be NP-complete but for which an algorithm solving the problem in polynomial time has not been found. Such problems are referred to as computationally hard or intractable. Factoring huge numbers is such a case, and the security of the RSA algorithm relies on the non-existence of fast factoring algorithms – hopefully, factoring is an NP-complete problem but this has not been proved.

Most network problems are either NP-complete or hard (but not proved to be NP-complete or P). Therefore, we cannot expect to solve all the security problems of the network structures of Figure 2 in feasible time.

The complexity implies that it is often impossible to define a single security policy for an application because it may involve internet nodes, CPUs and program modules located in different administrative domains. In many applications, different domains may be involved in different invocations of the application. One simple example is e-banking: the type of software modules used for each interaction between the user and the bank may be the same but different CPUs, versions of the software and internet nodes are involved in different interactions. A security policy for electronic banking must then involve not only the computer system at the bank but also the computers of all potential users of the service. The bank has then two choices: abandon the service because it is deemed to be too insecure, or deeming the risk for fraud not to be too severe and offer the service. The latter is often referred to as "taking a calculated risk". The problem is that the risk has neither been calculated nor assessed.

Complex systems can best be handled by building robustness into them. One method is to make the systems fault tolerant, that is, the system does not fail in a catastrophic way if it is attacked by malicious software. Avoiding that faults occur is much more difficult and expensive. This brings us smoothly to the next theme.

4 Economy: Needing unlimited resources

Complexity and dynamics imply that even if we understand the attack against the computing infrastructure, we may not be able to protect the infrastructure against damage. Furthermore, the cost of implementing countermeasures against attacks we actually can stop may be formidably expensive. The problem should then be studied from a different angle.

Malicious software may cause the system to fail in much the same way as programming errors, wrong input data, faulty operational routines or physical events. Information security may then also be examined applying classic dependability theory.

There are principally two ways in which the dependability of a system can be improved. The most direct way is to remove the cause of the failure. This is called fault avoidance. In information security, fault avoidance consists in installing firewalls and antivirus programs and by applying strict access control. We have seen above that this can only protect the system partially since there are faults that can never be detected and that it may be computationally infeasible to detect other faults.

The more indirect method is to make the system fault tolerant. This means that even if the system is subject to attack and partially damaged, the system is still able to deliver correct services. There are many methods that can be used in order to achieve fault tolerance: automatic or manual restart (reboot), redundancy and standby systems, and degeneracy where other parts of the system, in addition to performing own tasks, can deliver partial or complete services of the damaged part.

The cost of dependability is shown in Figure 3.

From an economic viewpoint, it is cheap to implement fault avoidance if the dependability requirement is poor. An antivirus program may be sufficient in many cases. However, the cost of fault avoidance increases fast if more dependability is required. Moreover, for reasons explained in Sections 2 and 3, dependability cannot be extended beyond a feasibility barrier.

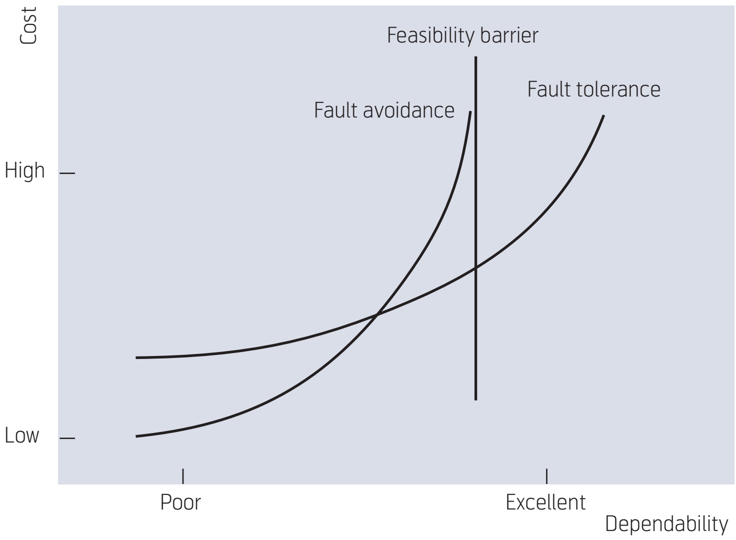


Figure 3 Cost of dependability

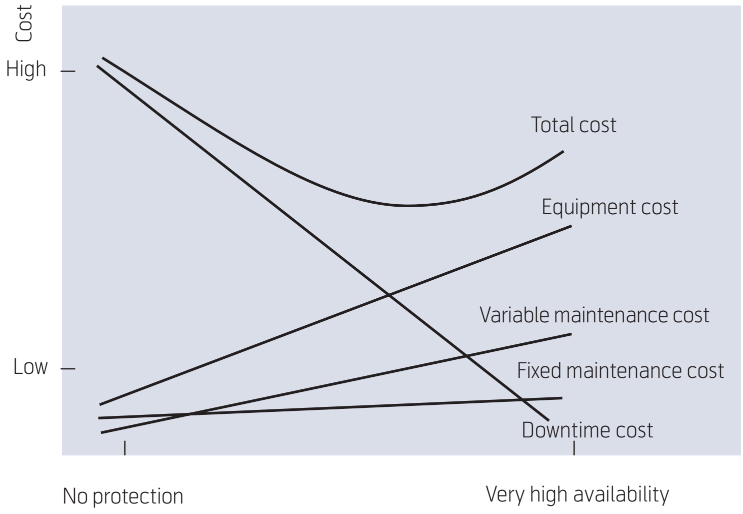


Figure 4 Trade-off between cost and service availability

It is expensive to apply fault tolerance to systems that do not require much dependability. On the other hand, it is the cheapest way of making systems very dependable. Fault tolerance is not subject to the same type of feasibility barrier as fault avoidance. In a fault tolerant system, energy and time is not spent on finding possible causes of faults but to develop means and methods by which the system can recover from the fault independently of the cause of the fault. In fault tolerant systems it is accepted that the system fails provided that the recovery procedures are fast, reliable and preferably autonomous.

Another way of looking at the economy of protection is shown in Figure 4. The total cost of protection is composed of several components[1]:

- Downtime cost associated with lost production, annoyance and reduced work moral;
- Cost of equipment for protection and recovery;
- Cost of fixed maintenance staff and tools;
- Variable maintenance costs if external resources are required to solve the problem or if equipment must be replaced.

The total cost and the cost components are functions of some dependency measure such as the availability (or uptime) of the system as shown in the figure. The total cost will then be a minimum for some value of the availability. The company may then simply determine the optimum protection strategy by determining this minimum. However, this works nicely only for simple theoretical examples where all parameters and dependencies can be determined accurately. In other words, the world must be completely deterministic.

The real world is not deterministic, and such an optimum solution cannot be found because of uncertainty and computational complexity. The reduction in production may vary vastly depending upon which category of employees is affected by the failure. The downtime cost cannot be determined by a single deterministic function. There is no easy way to relate availability and the cost of protection equipment, the number of maintenance staff required and the cost of other resources required for recovery. Intractable computation may even be needed to determine these functions. Furthermore, the functions are not static and may vary with time in a complicated manner.

Furthermore, the figure is concerned with the cost of dependability in a single node in the computational infrastructure. Since the vulnerability of individual nodes in this infrastructure depends on the overall structure of the network as pointed out in Section 3, the node may be damaged as a result of a structural attack. “Perfect” local countermeasures may be worthless in this case.

Therefore, it is practically impossible to determine how much economic resources should be put into protection against attacks by malicious software. The situation becomes worse when we take into account other managerial uncertainties.

5 Management: Unrecognised threats

The business managers and politicians are faced with other problems. This causes another set of threats against society and still another reason why protection against information warfare is lagging behind in the technological evolution.

5.1 Threats and historical evidence

The history of malicious software goes back about 30 years. That malicious software may represent a threat to society and business became evident less than ten years ago. This threat was taken seriously just five years ago. The concept of structural vulnerability is new and still only recognised in small communities of mathematicians, physicists and biologists analysing technical, social and biological networks known as scale-free graphs. The concept of structural vulnerability has not penetrated into the communities of industrial managers, authorities and politicians.

[1] *The cost components are drawn as straight lines for simplicity of drawing. In practice, each component may be a complicated curve.*

All natural and political threats such as earthquake, flooding, fire, pandemics, terrorism and war are well documented from historical evidence. We understand how we can protect us against such events.

Little evidence exists concerning information security threats. We have experienced a number of pandemic data viruses and internet worms. All these pandemics have hit us with surprise. Some of them have made considerable but bearable damage to data systems but the economic loss that each company has suffered has been small. Antivirus programs have been designed in a matter of hours after the outbreak and the pandemic has been over in a couple of days. From the viewpoint of the attacker, all these attacks have been successful in the sense that existing countermeasures could not stop them. They have been unsuccessful because it was too easy to develop countermeasures.

Historical evidence is then such that many are lulled into the belief that we can always efficiently combat attacks by malicious software. However, as we saw in Section 2, there is no way in which we can draw such a conclusion. We cannot predict from any theoretical or heuristic information when the next attack comes and how big the damage may be.

From a managerial viewpoint it may then be supposed that there is no reason to spend money on information security research and the authorities see no reason to implement preparedness against information security attacks. We have no experience concerning how fast a crisis caused by malicious data can develop and how much society and individuals will be affected.

5.2 Speed of the technological evolution

Modern data communications started to evolve in 1995. Since then the information and telecommunications technologies have penetrated into every corner of society with explosive speed. Nothing can be made without involving ICT in one way or another. Banks cannot function without ICT. Transportation comes to a complete stop and natural resources cannot be extracted and refined without the assistance of ICT. All communication between people over distance, including dissemination of news, will be impossible and local anarchy is likely to develop.

This evolution has taken just ten years. It takes months and years to agree on even minor political issues. This means that the management of society is lagging more and more behind technology.

Political decisions are singular events that easily engage people. The technological evolution is smooth and we hardly recognise the changes going on. This has resulted in a vulnerability of society that is in the early process of being understood. Development of malicious software is part of the technical evolution and we may expect that malicious software may become harder to detect and more damaging as time goes on. Furthermore, the more complex information technology becomes, the more security holes it may contain.

There are several evolutionary paths that technology has followed the last ten years. The first one is the development of larger and more complex software systems. The computational power of computers doubles almost once a year (Moore's law). We can then expect that the evolution in size and complexity of software is exponential with some fixed doubling time. The second evolution is that the computers are connected to the common internet so that they can interchange information for a variety of reasons: email, information retrieval, or participation in distributed computation sessions. Most of this information interchange is intentional but much of it is unintentional: spam, viruses, theft of memory and computer time, interference and so on. A third evolutionary path is concerned with the versatility of computing objects. CPUs are contained not only in traditional computers such as PCs and mainframes but in printers, mobile phones, measuring equipment, sensors, steering systems of automobiles, boats and aircraft, television sets and so on. This makes the number of units that may exchange information unintentionally extremely large. This leads to the network configurations described in Section 3. Finally, more and more of the interfaces between the computers and the internet have become mobile. This means that it is more convenient for us to communicate but it has become harder and harder to avoid accidental or malicious interference between computational devices.

Most of us welcome this evolution because it has made society more efficient and our lives richer and better. The reverse side of the medal is that society has become dangerously vulnerable.

5.3 Perceiving the big picture

The infrastructure described in Section 3 is reproduced in Figure 5 with the addition of ownership or administration of objects. On each layer there are owners or administrators of the objects. Some parties may own objects in several layers. However, what is important is that there are extremely many parties owning this infrastructure, ranging from individuals to large corporations. Furthermore, all three layers have grown independently in accordance with some stochastic process under no-one's control.

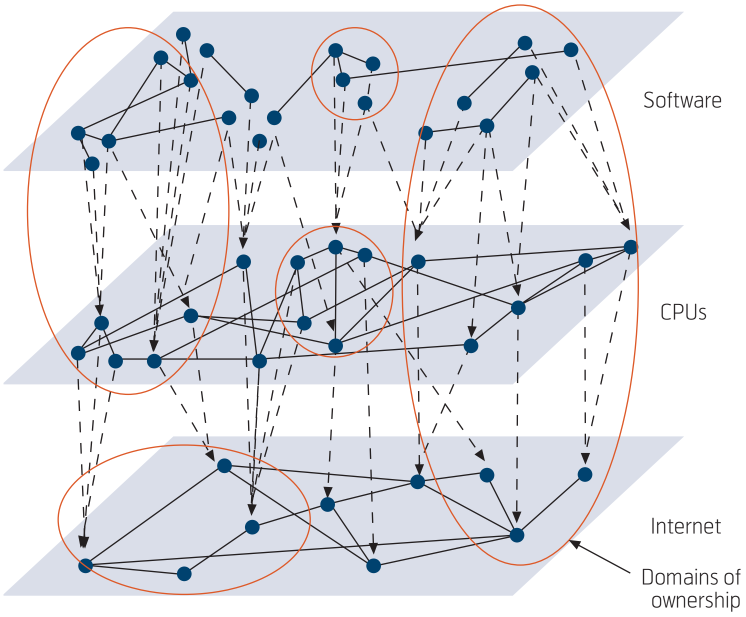


Figure 5 Domains of ownership or administration

This means that it is impossible to design a single security policy for the overall configuration. The impact this has on information security at large is obviously dramatic.

Society can be modelled in terms of several independent and interrelated networks: networks for physical infrastructure, social networks such as acquaintanceship networks, political and other networks of influence, networks of industrial ownership, competition and cooperation networks, networks of dependence between companies, financial networks and so on. It is believed, though not verified, that several of these networks are scale-free and as such contains nodes that are much more vulnerable than all other nodes in the network.

The technical infrastructure consists also of several networks: water supply and sewage, roads, airports and harbours, energy distribution networks, and telecommunications. Some of the networks are local while others are global. The networks are interdependent of one another. Telecommunications and the electricity grid are particularly important because if one of them is taken down completely, all the other networks go down.

The interdependency of the various networks of society allows us to draw another network where the nodes are all other networks in society and the links show the interdependence between the networks. This allows us to judge the overall vulnerability of society. It is urgent to construct this network so that it can be analysed with regard to both robustness and vulnerability.

This is the big picture that is now evolving. It is still not fully recognised by those who have the power to initiate the study of how this evolution is moulding society.

6 Conclusions

What we have found above is that information security face us with several problems of undecidability that reduces the possibilities we have to protect our computer systems. This results in a hierarchy as shown in Figure 6. Most of the information security threats that faces individual systems, networks, businesses and society cannot be eliminated for four broad reasons:

- Computational undecidability is an absolute statement that it is impossible to construct the ultimate firewall or access control system that can make us invulnerable to future malicious attacks. Firewalls and access control systems protect us against malicious influence that we understand and are capable of analysing completely or events that we already have experienced and examined; they may, by coincidence, also prevent new information threats.

- The structure of ICT systems consists of several independent but interrelated network layers. Several aspects of public management, common infrastructure, commerce and industry can be described in terms of similar independent but interrelated networks. Mathematical studies during the last few years have shown that many of these networks are *structurally* vulnerable but robust against random attacks: the network is rather safe against arbitrary attacks but there exist methods of attack that can be used to damage the network seriously. The complexity of these networks is such that it is hard to analyse them. Some of these problems are proved to belong to the NP-complete complexity class, defying direct analysis of them. Complexity reduces then the capabilities we have to design countermeasures against information attacks. However, complexity may hopefully also make life difficult for the attacker so that it is not easy to design attacks against the vulnerable spots of the network.

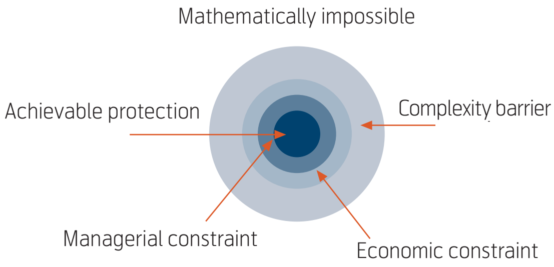


Figure 6 Unprotectability hierarchy

- Someone must pay for protection. The problem is that it is virtually impossible to determine how much resources should be spent on protection. One reason is that it is difficult to estimate the cost of an outage. Another argument is that it is impossible to calculate the effect of protection for reasons just explained.

- The decision makers are facing an even more uncertain situation. There is very little historical evidence concerning the severity and cost of information security violations. The technological evolution that has resulted in a vulnerable society has taken place in less than ten years and the growth of the new computational infrastructure has been exponential. This is a big change taking place over a very short time compared with political, economic and cultural changes. It took five years before information security was regarded as a possible problem; it has taken another five years before the first slight recognition of structural vulnerability has been shown in governmental bodies.

We are then left with a small kernel of achievable protection against information security threats. This protection is deemed to be insufficient in the future. What is required is research in areas such as

- describing and surveying actual technological and societal network structures and their interrelationships;

- studying how these structures can be modified to withstand both random and targeted attacks; and

- defining methods by which fault tolerance can be built into these structures.

Research into information security is not just the study of classical fields of confidentiality, integrity and availability, but also the new study of structural problems. This becomes more and more important as computation becomes more and more autonomous, ubiquitous and mobile. The networks are increasing fast and are becoming more complex and dynamic.

Bibliography

Barabási, A-L. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

Chaitin, G J. *Exploring Randomness*. Springer, 2001.

Dorogovtsev, S N, Mendes, J F F. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

Drummond, H. *The Art of Decision Making: Mirrors of Imagination, Masks of Fate*. John Wiley, 2001.

Ebbinghaus, H-D, Flum, J, Thomas, W. *Mathematical Logic*. Springer, 1990.

Gaarder, K, Audestad, J A. Feature Interaction Policies and the Undecidability of a General Feature Interaction Problem. *TINA93 Workshop*, L'Aquila, Italy, 27–30 September 1993.

Helvik, B E. Perspectives on the dependability of networks and services. *Telektronikk*, 100 (3), 27–44, 2004.

Helvik, B E. *Dependable Computing Systems and Communications Networks: Design and Evaluation*. Draft Lecture Notes, Department of Telematics, NTNU, 2001.

Hopcroft, J E, Ullman, J D. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979.

Kahneman, D, Slovic, P, Tversky, A (eds). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

Pham, H (ed). *Handbook of Reliability Engineering*. Springer, 2003.

Rogers, H, Jr. *Theory of Recursive Functions and Effective Computation*. MIT Press, 1987.

Savage, J E. *Models of Computation: Exploring the Power of Computing*. Addison-Wesley, 1998.

For a presentation of the author please turn to page 3.

ISSN 0085-7130 © Telenor ASA 2005

New challenges for emergency preparedness in the information society

JANNE HAGEN, HÅVARD FRIDHEIM AND KJELL OLAV NYSTUEN

Janne Hagen is principal scientist at Norwegian Defence Research Establishment

Håvard Fridheim is senior scientist at Norwegian Defence Research Establishment

Kjell Olav Nystuen is principal scientist at Norwegian Defence Research Establishment

The information and communication technology has irreversibly changed our society. We are living in a globally interconnected digital world, where our dependency on efficient ICT solutions increases every day. While the benefits of this development are indisputable, this also leads to serious challenges for our emergency preparedness work. The ICT systems are growing increasingly complex, and it will be progressively harder to understand and predict the effects of even simple component failures for the total system functionality. Vulnerabilities in critical infrastructures and societal services will emerge because of this, as it is hard to determine which emergency preparedness measures are the most efficient for complex systems. This paper discusses these challenges based on two cases: ICT dependency in the electric power supply, and our growing dependency on RFID technology.

Around the world in 8.0 seconds

In 1873, Jules Verne wrote his book *Around the World in 80 Days*, describing the attempts of Phileas Fogg to circumnavigate the globe in no more than the titular number of days. While this would have been an impressive feat at the time, today Fogg would have been able zip around the world in a couple of days.

An even more impressive feat is the ability to transmit information over long distances. In today's global electronic communication systems, digital data can be transmitted across continents in seconds. We've come a long way from the globetrotting adventures of the 19th century.

A small but dangerous digital world

The technological advancements in the communications sector make us perceive the world as "smaller". Our national borders are no longer the limits for our activities – the whole world has become our playground. We import goods from all over the world, our national businesses deliver services in markets in other continents, and at all times we have up-to-date information from faraway countries. The availability of information in real time also enables highly efficient solutions for businesses and public services, for instance electronic financial transactions or automated process control.

While the benefits of the technological advancements in recent years are indisputable, there is also some cause for worry. As our dependency on efficient ICT solutions grows, the impacts should they fail will be severe. The digital world also exposes us to manmade threats from every corner of the world. We depend on the Internet for our everyday communication services, financial transactions and delivery of public and private services, but this is also an arena for decidedly illegal activities and the battleground for cyber warriors worldwide.

Vulnerabilities in widespread software applications are detected continuously, and the world has become so small that it takes mere minutes from vulnerabilities are announced to our ICT systems are attacked. So far we are able to manage cyber attacks without experiencing catastrophic consequences. What we do not know is the extent to which our critical ICT systems will be subject to espionage or sabotage attempts from skilled agents, working for organizations or nations with high ICT attack capabilities. Several nations are developing computer network attack (CNA) capabilities in order to be able to shut down ICT systems or manipulate critical data. This constitutes the dark side of the technological opportunities from globally interconnected ICT systems.

Emergency preparedness challenges for comples systems

As our dependency on the Internet and efficient ICT solutions grows, the security related issues must be addressed correspondingly. There is a need for a focused approach to ICT emergency preparedness, to reduce the vulnerabilities in our critical ICT systems. Sadly, it is growing progressively harder to do this.

In his book *Normal accidents*, Charles Perrow discusses complex systems [1]. He claims that failures may be inevitable as systems grow increasingly complex. The main problem is that it will be impossible to predict the widespread impacts should one isolated system component fail.

Systems can be described by their complexity and by the coupling of their components and processes:

ISSN 0085-7130 © Telenor ASA 2005

- Simple systems have visible, often linear interactions between system components, and it is possible to follow logical cause and effect chains should one component fail. Traditional risk and vulnerability analysis will be able to determine which measures are ideal for reducing the system's total vulnerability. Complex systems, however, have component interaction (often non-linear) that may not be comprehensible for an analyst. Cause and effect will be uncertain.

- Loosely coupled systems have slack and are able to handle delays in system processes. Tight-coupled systems have time dependent processes and give no room for slack.

Most societal services and critical infrastructures will adhere to Perrow's description of complexity and tight couplings, and this is especially true for critical ICT systems. For instance, the myriad of information devices, PCs and sensors connected to the ever-growing Internet constitutes a highly complex system, and the impacts of system failure will be immediate. This is worrying, as it will be hard to measure the total vulnerability of ICT systems and to determine which emergency preparedness measures one should implement.

Even if we are able to establish a complete system description and predict the consequences of component failure, the dynamic evolution of ICT systems poses a major challenge. Any risk and vulnerability analysis of a system today will be of limited value tomorrow, as new security patches have been installed, configurations have changed, hardware has been removed or added and new users have been given access to the system. This adds to the bleakness of the situation, and presents us with a serious dilemma for present-day ICT emergency preparedness work. Is it at all possible to reduce the vulnerabilities in complex ICT systems? Or do we have to accept the possibility for serious failures, and try to handle the consequences instead?

To help clarify this issue, we will look at two examples of ICT use in our society.

Case 1: The Electric Power Supply

The electric power supply is crucial for modern societies. Even a short interruption in the delivery of electric power will stop the production of goods and services. For instance, the networks for electronic communication services are completely dependent on a stable power supply, especially since emergency power capacities are limited.

Norway's power system has grown into a complex and tightly coupled system. In the beginning, electric power infrastructure was developed locally, exploiting water resources in the vicinity of the users. The different local supplies were isolated from each other. Gradually, the systems were interconnected into today's electric power supply, a massively complex system where an international power grid connects geographically separate supplies and users [2].

Earlier, this infrastructure was supervised and controlled by manual means. Hundreds, even thousands of employees worked in this system, located at the different infrastructure nodes. They communicated with central system control personnel by robust telephone or telefax services, carrying out orders to operate circuit breakers, delivering data on the state of the different sites and so on.

Today, most of this work is automated or remote controlled through an integrated ICT system, linking all nodes of the physical power infrastructure to a few system control centres. Automation and remote control has enabled the electric power system to be run by a handful of people at the system control centres in a very efficient way. Both government policies and market demands are important driving forces for this development [3].

The deregulation of the power market has also led to an increased communication flow within and between power companies. Electric power is now a tradeable commodity at the Nordic Power Exchange (Nordpool), and the power market needs up-to-date information on the power flow, production levels, planned revisions and so on. This information is available in the process control systems. Earlier, these systems were completely isolated from the outside world – even from most employees within the company. But because of market demands, there are now logical links between the process control systems and other networks.

Also, reductions in staff and expertise within the power companies have led to an increased dependence on external competence. Work by external suppliers may often be carried out online. This increases the need for tying all the different participants of the electric power supply together in a massive ICT network. The Internet has grown to be an increasingly important part of this network. Figure 1 illustrates the connections between electricity production, transmission networks, distribution grids and consumers, and it shows how process control systems (e.g. SCADA systems) support the complete chain of electric power delivery. It also shows that the process control sys-

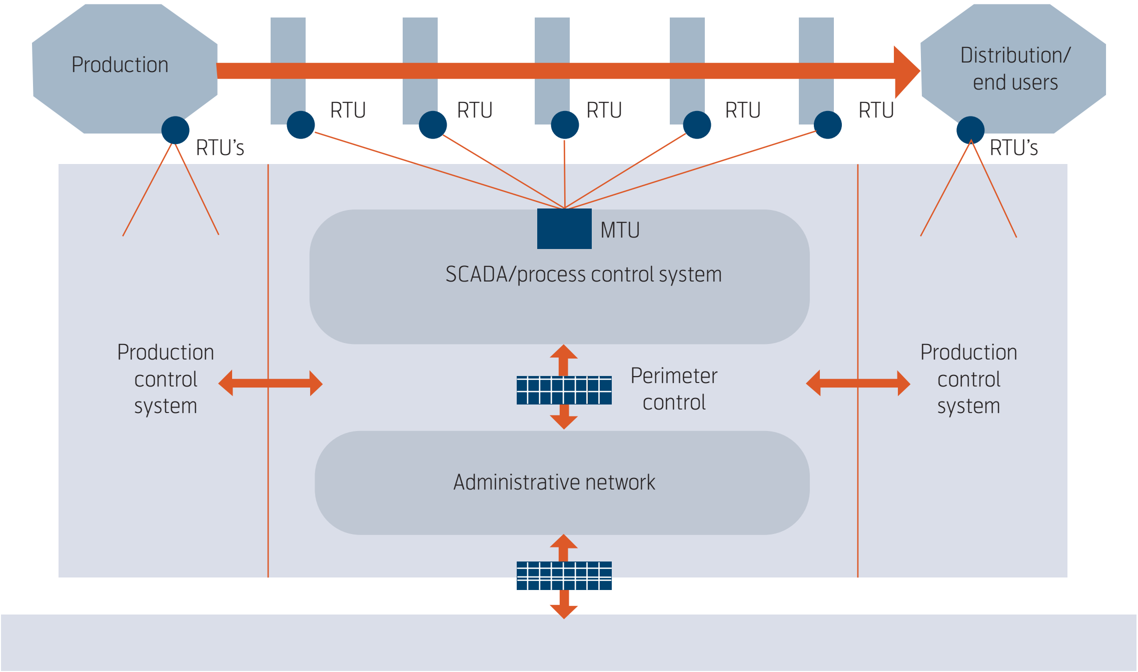


Figure 1 Connections between electricity production, transmission networks, distribution grids and end users in the electronic power supply, where SCADA systems, administrative systems and the Internet support the production and distribution chain

tems have been opened to the outside world by being connected to administrative systems and the Internet.

It is worrying that process control systems are opened to the outside world. Obviously, there are security measures *en masse* on the path from the Internet to any process control system, and it is unlikely that any outside player will be able to breach these defences and gain access to the critical ICT systems in the power supply. However, since there are logical connections between the different ICT systems, there is a risk that skilled hackers may be able to penetrate the defences. The complexities of the ICT systems make it difficult to develop broad and comprehensive defences for all relevant threats and to identify successful attacks and their related consequences.

The dependency on ICT is expected to increase further in the future power supply, for instance by the use of flexible AC/DC transmission (FACTS) technology to increase the capacities of power lines, and widespread use of external suppliers with online access to the companies' ICT systems. A result of this is that it will grow increasingly harder to identify relevant measures to reduce vulnerabilities in these ICT systems. The complexities may leave doors open for external attacks.

However, a more realistic source for power blackouts will be natural incidents or even technical failures. In recent years we have seen several examples of unexpected large-scale power outages. The year 2003 is of particular interest, since Italy, Sweden/Denmark, London and USA/Canada all experienced massive blackouts in a few months' time. Some of these were the results of relatively minor natural incidents or technical failures. In retrospect, some of these incidents may seem almost too simple to explain the dramatic impacts they initiated. However, this clearly illustrates the difficulties of predicting the impacts of single failures in complex systems. The tight coupling between the power supply and ICT systems was further demonstrated in January 2005, when heavy winds led to a prolonged power outage for tens of thousands of people in the southern parts of Sweden. A serious problem for the re-establishment work was the loss of communication services.

How should we deal with this development? Technical and organisational measures could and should be implemented in the different organisations according to best practices or risk and vulnerability analysis of ICT systems. Knowledge on information security and implementation of measures is important and necessary. Here the Norwegian Water Resources and Energy Directorate plays a major role, and the directorate is building knowledge on information security. However, ICT related security in the power supply might be a too complex issue for one government body to ensure. It is also necessary to consider emergency preparedness on other levels, both within the electric power sector itself, but also within other sectors with respect to emergency power supply for critical functions and systems. By doing so, the risk for the worst-case scenarios can be reduced, but not totally removed.

Case 2: The Radio Frequency Identification (RFID) society

In case 1 we discussed ICT vulnerabilities in one of the key infrastructures for modern society. One characteristic of this example is that the number of bodies with security responsibilities is relatively few and relatively clearly defined. The Norwegian Water Resources and Energy Directorate is the regulatory authority for the operation of the whole system.

In contrast to this, we will next focus on a technology that to a larger extent will be integrated into everybody’s beings and doings, individuals as well as large corporations and government institutions. This example shows the security challenges that arise when no specific actor is responsible for security on a societal level.

The technology in question is sensor networks based on radio identification tags (RFID tag). The smallest passive tags are less than 1 square mm and may be attached to almost every physical item. These RFIDs communicate with a reader that queries the tags for information. The information will typically contain data about the object that the tag is attached to. This will often include an identifier that can be related to a set of information representing semantics or attributes of the object. The distance between the reader and the tag will vary, from a few tens of centimetres for passive tags to significantly longer distances for battery powered active tags [4].

The RFIDs may also contain tiny microprocessors. This gives a significant scope for additional applications in the future, which may become parts of more complex smart sensor networks where other characteristics about the object and its surroundings are included. The RFID tags have over the last few years proved themselves as a possible replacement of the current bar codes in supermarkets logistics control, but they have also been applied in goods tracking and electronic toll collection from cars. It is known that the US military uses this technology in military inventory support on the battlefield in Iraq.

In a few years time it is likely that every commercial product and most physical items will have an RFID tag attached to them. Most people see this as a prosperous technology, which will benefit the society by making service production and the production of goods more efficient, and in the end less expensive. Today it seems the biggest gain will be in the logistics sector, but it is also possible to see the use of this technology in several other sectors, including applications for the single individual and their everyday actions. In the future animals and even humans may have RFID tags attached to their bodies.

In this text, we use the example of fitting mobile phones with RFID tags [5]. The RFID tag and the properties of the SIM card are combined in a GSM mobile phone. By combining these two technologies, the individual gets a complete solution for mobile communication, identification and electronic payment.

A typical RFID day might include the following activities, driven by RFID signals between the cell phone and RFID transceivers located around our society:

- Using the cell phone as an electronic wallet for online and on-site purchase and to pay for goods and services: transport tickets (bus, train, air travel etc), hotel rooms, concert and cinema tickets, groceries, etc.

- Using your cell phone as an identity card: access to the workplace, access to airplanes based on ticket purchase information stored in your cell phone, etc.

- Using your cell phone as a primary source for stored online information: travel schedules, business opening hours, telephone directories, etc.

This will be a highly efficient solution for most people, enabling them to travel freely through society without having to bring separate ID cards or cash. The dependency on the cell phone, however, will be dramatic – it will be the principal tool for survival. What would happen if a person lost his cell phone or if it should fail to work? Will there be backup solutions for those who are temporarily disconnected from the RFID world? The immediate impacts for the single individual will probably be that he is not allowed aboard buses, not allowed to buy groceries and so forth. For the society at large, these will be minor annoyances.

Regardless, RFID networks will imply that different types of information are interconnected in databases. This constitutes the first of at least two principal problems, namely the privacy issue. By interconnecting potentially huge masses of information about each individual and their doings, the individual’s privacy will be more vulnerable [4] [6]. The threat may come from the Government as well as from private companies and individuals, with or without “legal” access to all or pieces of the information.

The other problem, which is not so often addressed, is the vulnerability issue. This type of sensor network implies a high degree of centralised processing and

Figure 2 Keyless entry into a truck based on RFID technology. This is only one of numerous potential applications of RFID. Photo courtesy of Texas Instruments Inc.

information storage. When we are integrating vital services from the modern society, for instance banking services, travel services, food deliveries, energy supply and communication services based on this technology, we also become much more vulnerable. The cost of increased effectiveness is a huge dependency on available computer and communication services. Failures in the ICT systems, or the electrical power systems which the ICT systems are highly dependent on, may cause dramatic impacts for both individuals and society as a whole. Also, what will happen if somebody manages to penetrate and break the integrity of these systems?

A possible future RFID network will be a complex, multi-layered, ever-expanding network, which will also be very dynamic in its nature. Changes to the network will happen every day, probably every minute. New nodes and links will be added, and others will be removed. Very few, if any, will have the complete oversight of this network. This development, which is completely driven on the fast technology development and great commercial interest from large private companies, constitutes a number of problems for society. Will we accept the vulnerabilities that this development integrate into our society? On the other hand, will society have any means at all to control or adapt this complex development to its emergency preparedness policy, if such a policy exists? In contrast to the power supply example in case 1, this problem will be spread across all societal sectors and cannot be limited to one specific regulatory authority. Who will be responsible for this challenge in the societal planning?

To act or not to act?

After World War II, a key dogma for Norwegian emergency preparedness work has been that preventive strategies planned by the Government are cost-efficient for society as a whole. It may have been costly to reduce vulnerabilities in societal services and critical infrastructures, but this has been viewed as necessary to avoid even worse consequences. A government lead of this work has been widely accepted.

This dogma has been challenged by the recent technological and commercial developments, where the same processes that give us great benefits in our daily lives also lead to complex systems that are hard to analyse, and where preventive measures are not easily identified. We experience that the premises for preventive emergency preparedness strategies are challenged by decisions made in international markets, where our national authorities have little or no influence.

When faced with the challenges presented above, it is easy to adopt a fatalistic approach to ICT emergency preparedness: Whatever happens will happen. It will be too hard to reduce vulnerabilities in critical ICT systems. This would, however, constitute a revolutionary new approach to Norwegian emergency preparedness work, where the Government accepts no responsibility for failures in critical ICT systems. Instead, our emergency preparedness would have to try to mitigate the consequences, should failures occur.

Another approach is to accept the fact that there is no such thing as a risk-free society when considering ICT systems vulnerability, but still embrace the Government's lead in ICT emergency preparedness, simply because no other body will speak for the population in these matters. However, the Government will have to modify its approach to emergency preparedness work, and consider other means than those we see today.

We must accept that failures will occur, and look at ways to reduce their impacts

Obviously, there is no simple answer to the question of what to do. Due to the complexities and tight couplings of today's societies through electronic networks, we will probably have to face unforeseen situations. The question is: Are the Norwegian society and the Government prepared?

Today, the complete list of government bodies that have some sort of role in protecting our critical ICT systems is staggering. Responsibilities and roles are fragmented, even if there are variations among different sectors, as pointed out in the two cases presented earlier.

We claim that the emergency preparedness is not up to date when it comes to the threat from cyberspace. One question should be raised to illustrate this. Who is in charge if the following scenario happens, where the Norwegian energy and transport sector is hit by a cyber attack, causing an unstable power supply and air traffic accidents, dramatic disturbances in logistics etc? Who will have the capabilities and resources to handle this situation, and to whom should the bill be sent?

We do not intend to discuss this in detail, but merely point out the need for warning, response capabilities and emergency preparedness in a cyber attack scenario. This work must include a broad spectre of participants, including private enterprises and civil-military cooperation. The complex and tight coupled world forces us to think of the unthinkable, and to be prepared for the improbable scenarios. Even if we are not able to identify one easy way to solve this problem, a simple but helpful measure could be joint crisis exercises, involving players from different sectors, addressing the questions mentioned above, over and over again.

Another possibility is to apply emerging network theories to understand and simplify the complexities of society and the network of ICT systems, humans and organisations. The theory of scalefree graphs seems to be suitable for analysis of complex networks [7]. Scalefree graphs are characterized by the following properties: If network nodes are attacked randomly, the net will survive. If the busiest nodes are attacked, the net will rapidly collapse. Given that the theory of scalefree graphs is correct and relevant for the complex ICT systems we see today, we will come a long way in protecting complex networks by identifying the busiest nodes and protecting those.

Understanding the nature of risky enterprises

According to Charles Perrow, we will be able to reduce the dangers of complex systems by understanding the nature of risky enterprises better. This statement gives hope. Multi-disciplinary research teams at the FFI have done research on the protection of the society (BAS) with a comprehensive and multi-methodological approach since 1994. The FFI research has recommended changes in the emergency preparedness within the studied sectors. The present BAS5 project is bringing this research further with its focus on critical information infrastructure protection.

In the latest years it has become clear that there is a growing need to establish a broader cooperation between academia, governmental agencies and private enterprises. Thus, in the BAS5 project, FFI cooperates with the University of Stavanger, The Norwegian University of Science and Technology, and Gjøvik University College. In addition, BAS5 involves a long list of partners from both the public and private sector.

The focus of BAS5 is to develop and apply risk and vulnerability methods for ICT systems and to develop methods to rank measures for reducing vulnerabilities. In a crisis scenario with overloaded networks it is also important to give priority to the most urgent functions and supporting ICT systems. Thus, one of the project's tasks is to develop and apply a method to prioritise societal functions and related ICT systems.

We would stress that there is a need for a continuous approach to ICT emergency preparedness, including research and development. This is a necessity due to the rapid and ongoing technological and organisational development we experience today. This must be done from an international viewpoint. International cooperation will be a key issue to handle the challenges from cyberspace now and in the future.

References

1 Perrow, C. *Normal accidents. Living with High-Risk Technologies*. With a New Afterword and a Post script on the Y2K Problem. Princeton, NJ, Princeton University Press, 1999.

2 Fridheim, H, Hagen, J, Henriksen, S. *En sårbar kraftforsyning – sluttrapport etter BAS3*. Kjeller, Forsvarets forskningsinstitutt, 2001. (FFI/RAPPORT – 2001/02381) (in Norwegian). Link: http://www.mil.no/multimedia/archive/00001/Fridheim-R-2001-02381_1863a.pdf

3 Rodal, S K. *Sårbarhet i kraftforsyningens drifts- og styringssystemer*. Kjeller, Forsvarets forskningsinstitutt, 2001. (FFI/RAPPORT – 2001/04278) (in Norwegian). Link: http://www.mil.no/multimedia/archive/00003/Rodal-R-2001-04278_3649a.pdf

4 Sarma, S E, Weis, S A, Engels, D W. *RFID Systems, Security & Privacy Implications, White Paper*. MIT, USA, Auto-ID Centre, 2003.

5 Noll, J. *The mobile phone – access and authentication*. Fornebu, Telenor R&D, Norway, 2005. Guest lecture at FFI.

ISSN 0085-7130 © Telenor ASA 2005

6 Duce, H. *Public Policy. Understanding Public Opininion, Executive Briefing*. MIT, USA, Auto-ID Centre, 2003.

7 Audestad, J. Vulnerability exposed: Telecommunications as a hub of society. *Telektronikk*, 100 (3), 45–54, 2004.

Bibliography

Barabási, A L. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

Fridheim, H. *Survivability of the modern society – deregulation of services vs critical infrastructure vulnerability*. Paper presented at 1st European Survivability Workshop, 26–28 February 2002, Köln-Wahn, Germany.

Hagen, J. *Economic Analysis of Civil Emergency Efforts – A Methodological Approach*. Paper presented at 14 International Symposium on Military Operations Research (ISMOR), The Royal Military College of Science, UK, Sep 1–5, 1997.

Hagen, J, Fridheim, H. *Cost-Effectiveness Analysis of Measures to Reduce Vulnerabilities in the Public Telecom System*. Paper presented at 16 International Symposium on Military Operations Research (ISMOR), The Royal Military College of Science, UK, Sep 1–3, 1999.

Nystuen, K O, Hagen, J. *Critical Information Infrastructure Protection in Norway*. Paper presented at Critical infrastructure protection (CIP) – Status and perspectives, An international Two-Days Workshop within the Annual Meeting "Informatik 2003", German Informatics Society (GI) at JWG University, Frankfurt, Sep 29 – Oct 2, 2003.

Hagen, J. *Securing Energy Supply in Norway – Vulnerabilities and Measures*. Paper presented at The Conference NATO-membership and the Challenges from Vulnerabilities of Modern Societies, The Norwegian Atlantic Committee (DNAK) and the Lithuanian Atlantic Treaty Association, Vilnius, Lithuania, Dec 4–5, 2003.

Janne Hagen is a principal scientist at the Norwegian Defence Research Establishment (FFI). She worked at the Norwegian Institute for Transport Economics before joining the FFI in 1996. At FFI she has primarily been engaged in research projects on critical infrastructure protection, but she has also been attached to defence analysis projects.

email: Janne.Hagen@ffi.no

Håvard Fridheim is a senior scientist at the Norwegian Defence Research Establishment. He has been part of the research team on critical infrastructure vulnerabilities at the FFI since 1996, excluding a one-year stay at the Armed Forces National Joint Headquarters in Stavanger.

email: Havard.Fridheim@ffi.no

Kjell Olav Nystuen is a principal scientist at the Norwegian Defence Research Establishment. In recent years he has been engaged in project activities on security and vulnerability issues related to military and civilian utilisation of computer networks.

email: Kjell-Olav.Nystuen@ffi.no

Structural vulnerabilities in communication networks

NILS KALSTAD SVENDSEN

Nils Kalstad Svendsen is a PhD student at Gjøvik University College

This article emphasises the difference in the design of the telephone network and the Internet and gives a brief survey of recent results showing that self-organisation of Internet may be the source of structural vulnerabilities in this network. Finally the article points out that today we do not have the infrastructure needed to manage this vulnerability.

1 Introduction

The history of telecommunication started with Sir William Cooke's and Sir Charles Wheatstone's design of the first workable telegraph in 1837. Since this first invention, telecommunication engineers have given us inventions such as telephony, fax, mobile telephony, and finally Internet. These applications do not only come in handy for communication between individuals, but have become a key factor for just about all sectors in our society. A study carried out by FFI[1] resulted in a matrix stating the dependability between 14 Norwegian sectors. This matrix is presented in the governmental report [16] as the foundation for estimating vulnerabilities in the Norwegian society.

To visualise the key role of telecommunications the strong dependabilities in this matrix are extracted, and represented as a graph in Figure 1. The motivation for eliminating weak connections is that the full representation gives a very connected graph, from which it is difficult to read information. The graph in Figure 1 has two clear hubs, telecommunication and power supply. Both these sectors have nine other sectors which are heavily dependent on them and all sectors are available within two steps from these hubs. Further on it is only from the hubs that all other sectors can be reached[2]. This shows that all sectors included in the study are directly or indirectly dependent on telecommunication.

The FFI study was carried out in 1997. The last part of the 1990s is in [2] described by Audestad, as the moment when a paradigm change takes place in the occidental society. This was the moment when society changed from having a low information and communication technology (ICT) dependency to being highly ICT dependent. At the base of this dependency we find the telecommunication network making machine-machine, data and mobile communication possible. The number of mobile and network applications has been growing steadily since the late 1990s, and one can therefore assume that society's actual dependence on telecommunication is even stronger than what is shown in Figure 1.

The cost of this dependability is obvious; we all become extremely vulnerable to faults that might induce errors and failures in the telecommunication network. The components of telecommunication systems are computers, communication links, switches and routers, software, storages, user interfaces, etc. Wanting to protect this network the operator must be able to decide which of the components the functionality of the network depends on. If not, resources may be lost as non-critical infrastructure gets over protected without increasing the system reliability. Classical theory on reliability engineering, such as [20], focuses on parallel and serial multi state systems, hardware and software dependability, etc. This kind of networks are highly structured. It is rather unclear how the theory can be applied to a complex self-organising network of which we hardly know its interconnections. This article emphasises how some special network topologies introduce vulnerabilities in the network that cannot be found using the classical reliability approach.

2 Two different telecommunication networks

The two dominant telecommunication networks are actually the telephone network and the IP-network or Internet. When describing the difference between these networks most textbooks, such as [14], starts by explaining the difference between circuit and packet switching, and thereafter the technology and protocols needed to set up a connection from a point A to a point B. The parts of the network that are not used by this communication are represented as loose ends, or some kind of cloud. Few dwell on the availability, connectivity and capacity in these networks, but sim-

[1] *FFI = Forsvarets Forskningsinstitutt (Norwegian Defence Research Establishment)*

[2] *This can be seen by running a shortest path algorithm on the graph.*

ISSN 0085-7130 © Telenor ASA 2005

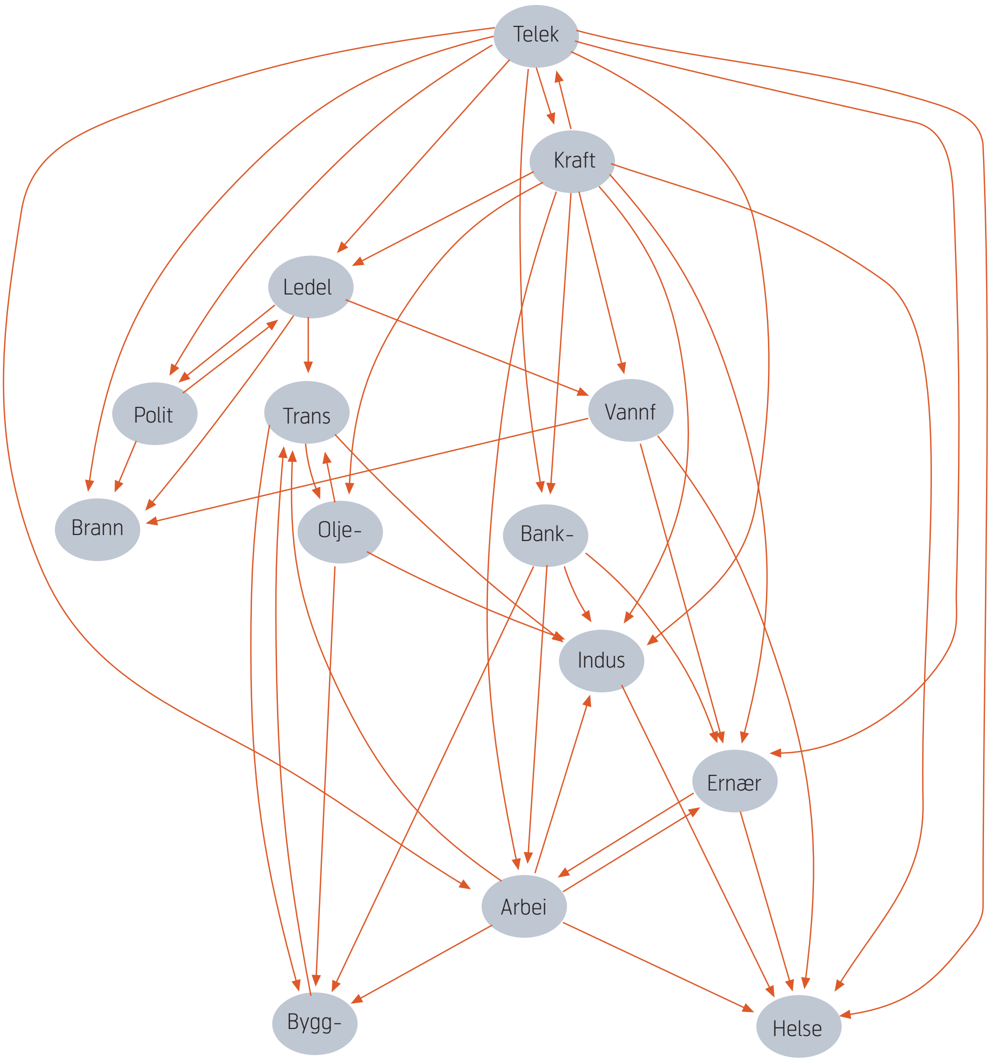


Figure 1 Overview of the dependability between different Norwegian sectors. A sector is judged to be dependent of all the sectors pointing to it. Legend: Ledel = Management/Information, Kraft = Power supply, Telek = Telecommunications, Olje- = Oil and fuel, Trans = Transport, Arbei = Workforce, Vannf = Water supplies, Bank- = Bank and finances, Bygg- = Construction, Indus = Industry and commerce, Helse = Health, Ernær = Nourishment, Brann = Fire and rescue, Polit = Police and order

ply assume that the network is so large that there is always an available path between A and B. This assumption should not be made without knowledge of the topology of the network. This section explores these main characteristics of the classical telephone network and the Internet, and tries to emphasise on what makes them different.

2.1 Basic network topologies

Networks are normally modelled as graphs. A graph consists of *N* nodes and *L* edges connecting the nodes together. Edges are distributed between nodes according to rules based on local or global properties of the graph, and these can be directed or undirected, with or without weights.

Examples of traditional network topologies are shown in Figure 2. Two intuitive criteria for evaluating if a network topology is appropriate to be used for a communication network are resilience against failure and scalability in terms of users. The networks in Figure 2 have nice and simple structures, but they are inadequate for practical purpose. The linear topology and the ring become disconnected if respectively one or two nodes are removed, and if the centre of the star is removed, all nodes will be disconnected. The mesh is more robust to failure, but it is infeasible to divide a large area into a fine grid in this way. The k-trees have good scaling properties, but no redundancy. Every time a node is removed, the graph is split into *k* disconnected subgraphs, unless the node is a leaf node. The fully connected graph has very high redundancy but is infeasible to implement for large communication systems. As these classical "engineering" topologies have turned out to be inappropriate for telecommunication, we ask ourselves which topologies are used?

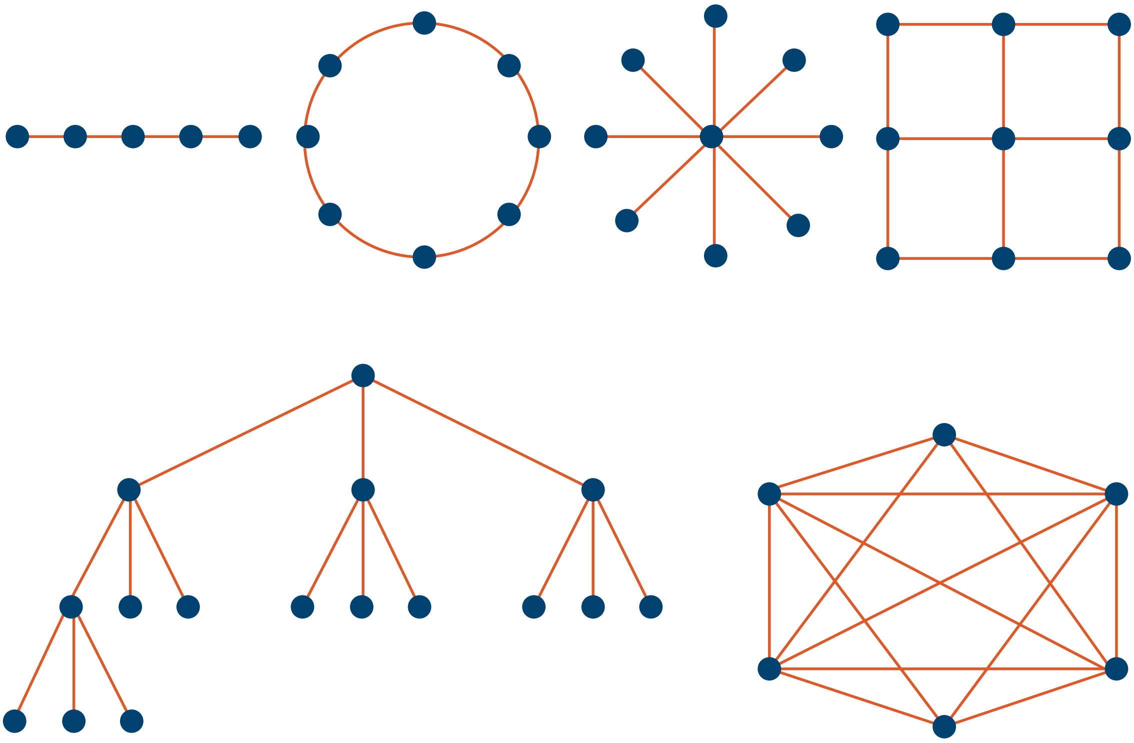


Figure 2 Examples of regular topologies: from left to right and top to bottom, linear, ring, star and mesh, 3-tree and fully connected (complete)

2.2 Classical telephone network

The topology of the classical telephone network is the consequence of an economic and a technical factor:

1 Providing telecommunication services has traditionally been a task given to a governmental or private monopoly in different countries.

2 For stability and noise reasons a connection should go through more than seven switches.

Originally telecommunication aimed to provide telephony (voice transmission) to all inhabitants of a region wanting to have access to this. A local optimisation to obtain full coverage in a region has been a natural consequence of this. The stability criterion has imposed interregional, and intercontinental co-operation to obtain global coverage.

At the centre of the telephone network, there is a strongly connected core, while the structure in the outskirts of the graph is more tree-like. The topology of the telephony transport network can be visualised as in Figure 3. This network has a clear structure and although large and complex, it is possible to manage. The majority of the local exchanges are connected to

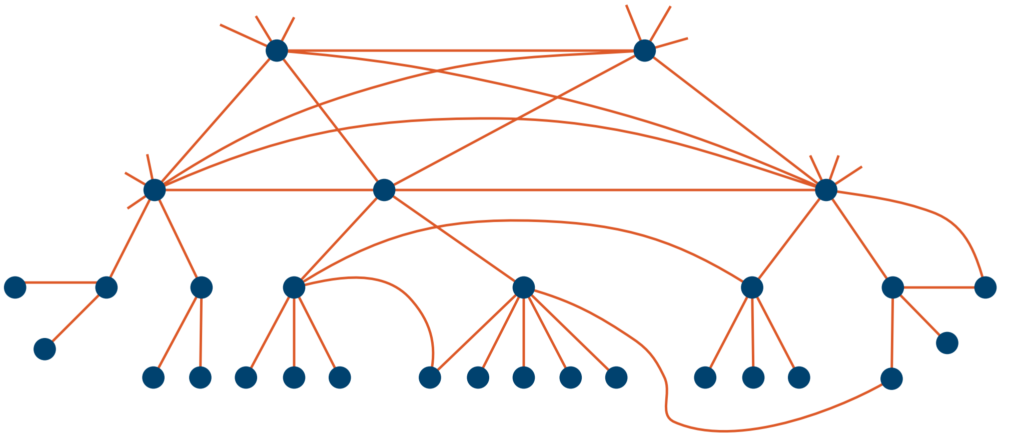


Figure 3 A representation of the topology of the telephony transport layer. The lower links represent local exchanges while upper nodes are core exchange terminals

their closest regional exchange point. We note that direct links can exist for especially busy routes and for critical connections. Routing in this part of the network is based on local optimisation. The regional exchange point is connected to an exchange point in the core. In this part of the network routing becomes more complex, but it is based on optimisation of the traffic in the core networks.

To sum up the topology we may say that the network has two components; a tightly connected core and low trees with many branches in the outskirt. In this way billions of telephones are interconnected with less then ten switches between each other. Even though the diameter[3] of the graph is small, it would be hasty to conclude that it is a small world network[4] without having done a thorough analysis of the network.

Throughout the 1990s liberalisation in occidental countries opened the market to allow the existence of several long distance national and international carriers, it allowed the existence of several GSM networks, and it gave operators access to local networks owned by other operators. The motivation for this process was to lower prices, improve customer service and introduce new telecommunication services. One of the most popular services introduced is data (binary) traffic on the network, but most of the income for the operators comes from voice traffic. In the majority of the cases, data traffic use the same access points as the telephone service, but the local exchanges are connected to a different transport network.

The liberalisation changed the business model from having two to three parties. Now there are subscribers, network providers and service providers. A subscriber in densely populated areas may have several networks and service providers to choose between, while people living in more remote areas may have no options. This multi-actor situation and the digitalisation of the transmissions have changed the well engineered topology we had up till the mid 1980s, but the basic structure of the network is as described in this section.

2.3 Internet

The history of the Internet started in 1958 [19] when ARPA[5] started to found a project concerned to create a computer network to access and share data and programs among computers located in different places. The motivation for this project was to secure control over information in the event of large-scale international conflicts.

Several network topology candidates, among these the ones previously illustrated in Figure 2, were proposed. As a part of this process Paul Baran outlined the principle of "redundancy of connectivity" and explored various models for designing communication systems and evaluating their vulnerabilities. In [6], from 1964, Baran proposes a distributed communication system with no obvious central unit where all nodes have the same routing capabilities. This differs largely from the network topology described in Figure 3, as the leaf nodes in a tree topology do not have sufficient redundancy.

Such design criteria would create a fully connected network, which is unrealistic for practical applications. The approach for the following development was therefore to design a sufficiently, but far from fully connected, distributed network. Together with the packet switching technology this was the basis for the development of ARPANET in the late 1960s and early 1970s.

ARPANET was soon followed by other governmental and commercial networks such as MFENET[6], SPAN[7] and Telnet. At the same time other network technologies such as CSNET[8], based on UUCP[9] requiring only ready-to-use technologies such as modems and telephone lines, emerged. These networks provided essentially basic store-and-forward facilities and e-mail but were major contributions in the process of building the Internet community.

The next turning point in the history of the Internet was the decision of the NSFNET[10] to use the TCP/IP protocol and to encourage commercial network traffic at the local/regional level along with the ban of the commercial use of their backbone. This

[3] *The longest shortest path.*

[4] *In [22] Watts defines a small world network as a network with degree distribution similar to a random graph but with a high cluster coefficient.*

[5] *Advanced Research Project Agency.*

[6] *MFENET was a research network for the US Department of Energy.*

[7] *Research network for NASA space physicists.*

[8] *Computer Sciences Research Network.*

[9] *Unix-To-Unix Copy Protocol.*

[10] *NSFNET was a network founded by National Science Foundation to establish a transcontinental network and five super computing centres.*

stimulated the establishment of long-haul private network offering alternative to the traditional telecommunication backbones for the commercial traffic. In 1989 NSFNET co-opted ARPANET and in 1995 it reverted back to a pure research network, leaving the national-scale connectivity to private backbone providers.

Up till the late 1980s the size of the Internet had been limited and well administrated. This is underlined by the neat map of the Internet shown in Spilling and Lundh's article [21]. When the central organisation disappeared, the Internet became a growing self-organising system. Over the last decade it has experienced explosive growth, as shown in Figure 4, and the number of hosts can be counted in tens of millions and providers are connecting to the Internet on a daily basis. The networks are connected together without any governing office determining how routers must be connected. Each network organisation (or autonomous system) decides autonomously its connections and makes arrangements to pass along other's network traffic. As a consequence of this the large-scale dynamics and structure is driven by many interacting units aiming at optimising local communication efficiency. An open question is what the global effect of this local optimisation is, what the actual structure or topology of the Internet is, and how this influence its performance.

Internet, as we know it today, consists of three different classes of networks: Large national scale networks or autonomous system (AS) networks, regional scale or router (or IP) networks and local area (LAN) networks. These networks and their connections are represented in Figure 5. The organisation of the LANs is done by the owner of the router. The structure of this network will depend on weather it is a large enterprise or a private modem accessing the network. The router connections, connections to public exchange points and private peering relations are organised by the ASs themselves. Private peering is mostly used between local ASs. For continental and intercontinental connections the public exchange points are used in the majority of the cases.

There are no general international agreements for how private peering should be done, or what services

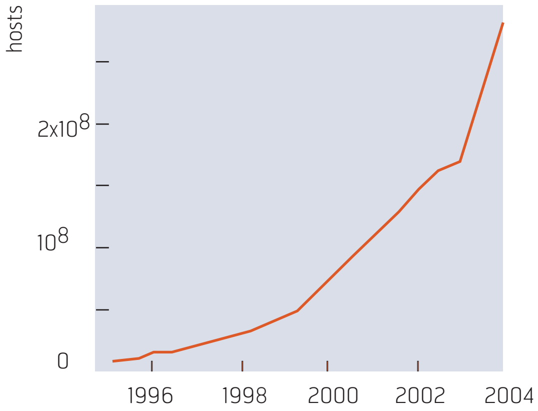


Figure 4 The growth in Internet hosts starting from 1995 to 2004. Data is taken from Hobbes' Internet Timeline http://www.zakon.org/robert/internet/timeline/

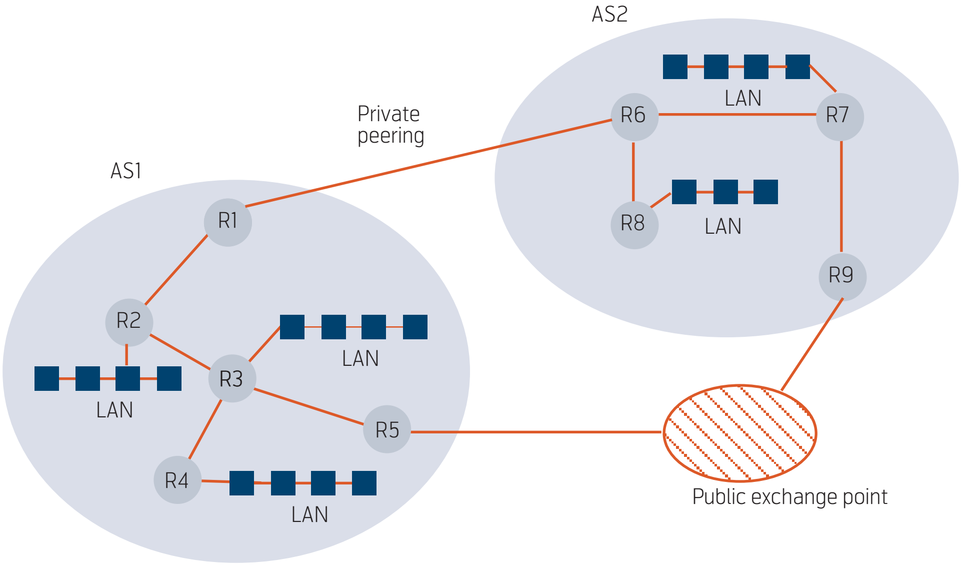


Figure 5 The Internet networks and their connections as described by [19]. Autonomous systems are interconnected by a private peering relation (routers R1 and R6) or public exchange points (Routers R5 and R9). Routers R1, R5, R6 and R9 use an exterior gateway protocol such as Border Gateway Protocol (BGP), while the remaining routers adopt and Interior Gateway Protocol (IGP) only aimed at internal routers

ISSN 0085-7130 © Telenor ASA 2005

the parties must provide to each other. The driving dynamic in the creation of these links is the economic motivation. Private peering can be done between ASs that expect to have a lot of traffic between each other. The reasons for this can be diverse; one of the ASs can host a public service, meaning it should easily be reachable from any host, the two ASs can be large national providers expecting a lot of traffic between each other, etc.

The role of the public exchange points is to connect traffic from different ASs. In its original form the exchange points did not discriminate between the size the ASs and provided the same services to all who connected to it. This became one of the clearest vulnerabilities of the Internet. Today the exchange points are becoming more and more discriminating and we also observe private exchange services [14]. The choice of which exchange point to connect to also depends on economic interests and local optimisation.

2.4 Comparison of telephone network and Internet

The principal difference between the telephone network and the Internet is that Internet has no organisation dedicated to the global optimisation of the network. There are standards and protocols that are recommended and the actors have to trust that all other actors apply the protocols. Otherwise all trust is local. An AS trusts the other ASs to which he connects, to forward his traffic and hope that all packages will reach their destinations. If a failure occurs, it must be detected by one of the parties and the packages must be retransmitted. In the telephone network international agreements assure the connectivity to the end users. In this way technical, economic, legal and error handling trust is established between all parties.

3 The topological properties of Internet

Internet is a growing network and new ASs are added every hour. From the above discussion we have seen that ASs connect to each other based on a set of local preferences. Growing networks with preferential attachments have been studied with great interest over the last years, and this section summarises some of the results.

3.1 Network statistics

Network statistics are used to describe essential properties of vertices and edges, regions and entire graphs. A network statistic should describe essential properties of the network and differentiate between certain classes of networks. Especially two statistics are essential for classification of the Internet networks:

- Degree: The probability $p(k, s, N)$ that the vertex s in the network of size N has k connections.

- Clustering coefficient: The likeliness $c(s)$ that two neighbours of the node s are connected. The clustering coefficient $C(G)$ of a graph G is the average of $c(s)$ taken over all nodes in the graph.

Empirical observations of the Internet networks, such as in [13], show that the degree distribution has a heavy tail that differs significantly from the normal distribution. When fitting the data, the best model is obtained by using power law distributions, i.e.

$$P(k) = ck^{-\gamma} \qquad \gamma > 0, c > 0. \tag{1}$$

We note that this expression is independent of the observed node and the size of the network. Faloutsos et al. [13] observed that the router level and the AS level respectively have $\gamma = 2.5$ and $\gamma = 2.2$, meaning that their variance diverges. Observations of the clustering coefficient of the AS network, as in [17], show that it is significantly different from 0.

3.2 Modelling self-organising, growing networks

The traditional approach for modelling complex networks was introduced by Erdös-Réyni in a series of papers in the early 1960s. These graphs, consisting of N nodes, can be constructed in two different ways:

1 The set of graphs $G_{N,p}$ is constructed by connecting any two vertices of the network by an edge with probability p.

2 The set of graphs $G_{N,E}$ is constructed by distributing edges between L randomly chosen pairs of vertices.

Using the second model to deduce the degree distribution of Erdös-Réyni graphs, we see that the degree distribution is given by the binomial distribution with parameters N and p. Further, it is a well known fact that the binomial distribution can be approximated with the Poisson distribution if N is large and $pN = \bar{k}$ is constant, where $\bar{k}$ is the average degree of the nodes in the network. Hence $P(k)$ is given by

$$P(k) = e^{-\bar{k}} \frac{\bar{k}^k}{k!}.$$

This result can be obtained by more rigorous arguments, as by Bollobás in [7]. Thus this is not an appropriate model to describe the Internet networks.

Inspired by the fact that many social networks are highly clustered while at the same time exhibiting a small average distance between vertices, Watts and

Strogatz [22] proposed a model interpolating between ordered lattices and purely random networks. The Watts-Strogatz model starts with a ring of N nodes in which each node is symmetrically connected to its $2m$ nearest neighbours (m on the clockwise and m on the counter clockwise sense). Then, for every node, each edge connected to a clockwise neighbour is rewired with probability p and preserved with probability $1 - p$. The rewiring connects the edge endpoint to a randomly chosen vertex, avoiding self-connections. The parameter p therefore tunes the level of randomness present in the graph, keeping the number of edges constant. With this construction we obtain a graph of average degree $\bar{k} = 2m$ and minimum degree m. Examples of this process are represented in Figure 6. The degree distribution of the Watts-Strogatz model is given by Pastor-Satorras and Vespignani in [19]:

$$P(k) = \sum_{n=0}^{min(k-m,m)} \binom{m}{n} (1-p)^n p^{m-n} \frac{(pm)^{k-m-n}}{(k-m-n)!} e^{-pm},$$

for $k \geq m$.

In the limit of $p \to 1$ it can be shown that the above expression reduces to

$$P(k) = \frac{m^{k-m}}{(k-m)!} e^{-m},$$

a Poisson distribution for the variable $k' = k - m$.

While the degree distribution has essentially the features of an homogeneous random graph, the effect of the parameter p is more acute on the clustering coefficient. Pastor-Satorras and Vespignani give the following expression for the average clustering coefficient:

$$\bar{c}_p \approx \frac{3m(m-1)}{2m(2m-1)}(1-p)^3.$$

Watts and Strogatz [22] noticed that in a wide range of $p << 1$, the average shortest path length after decreasing abruptly reaches almost the value corresponding to a random graph, while the clustering coefficient remains constant and equal to that of the original ordered lattice. Therefore there is a broad region of the parameter space in which it is possible to find graphs with a large $\bar{c}$ and a small $\bar{l}$ as observed in most natural networks. However, in the context of the Internet, the Watts-Strogatz model misses several important features from the empirical observations. For instance, it displays a Poisson degree distribution, signalling that some other shaping principles, different from the rewiring, are missing in the model.

Simultaneously with Watts and Strogatz proposing their model, there was a change of perspective in the theoretical study of complex networks that shifted the focus from reproduction of network structure to the modelling of its evolution. This is the outcome of the realization that most complex networks are the result of a growth process. In 1999 Barabási and Albert [4] introduced the first growing network model. The key ingredient in these new models consists in considering the network as the result of the subsequent addition of new vertices and edges following a prescribed set of dynamic rules. Barabási and Albert's network was based on the following dynamic rules:

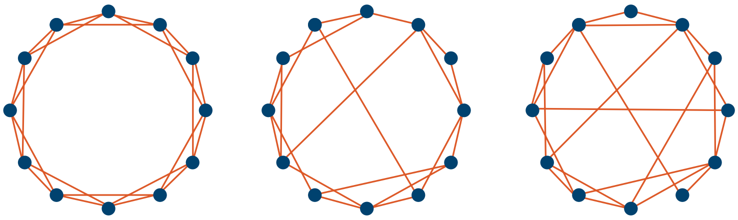


Figure 6 A daisy chain and rewired chains with probability 0.1 and 0.2. The daisy chain has clustering coefficient 47, while the clustering coefficient in the Watts-Strogatz graphs are a bit lower

- Growth: The network starts with a small core of m_0 connected nodes. Every time step we add a new node, with m edges ($m < m_0$) connected to an old node in the system.

- Preferential attachment: The new edges are connected to the old s-th node with a probability proportional to its degree k_s.

The graph in Figure 7 is grown according to these principles. From the construction the degree distribution at any time t can be derived (see for example [19]):

$$P(k,t) = 2m^2 \frac{t + \frac{m_0}{m}\bar{k}_0}{t + m_0} k^{-3}.$$

In the limit when $t \to \infty$ we can obtain the time-independent solution

$$P(k) = 2m^2 k^{-3},$$

which indicates that the preferential attachment spontaneously generates a network with a power-law degree behaviour. On the other hand, [19] also estimate the average clustering coefficient by

$$\bar{c}_N = \frac{m}{8N}(\ln N)^2.$$

We see that $\bar{c}_N$ decreases with the network size, and vanishes in the limit of the infinite network size.

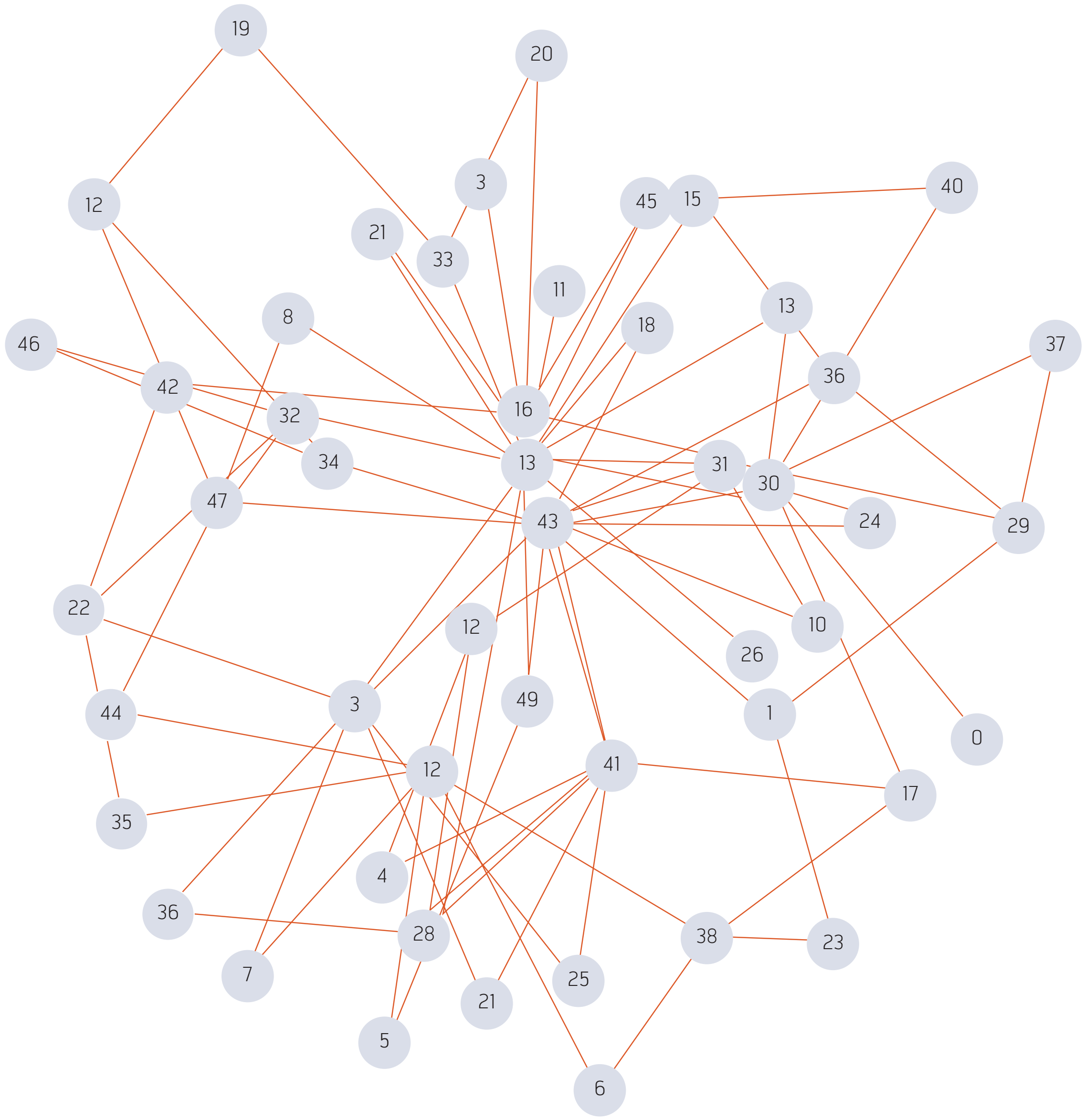


Figure 7 A graph grown according to the principles of Barabási and Albert with 50 nodes and m = 2

The interest raised by the Barabási-Albert construction resides in its capacity to generate graphs with power-law degree distribution and small world properties from very simple dynamic rules. Although the model does not capture all the features of real world networks, such as clustering coefficient and that older nodes are always the most connected ones, it will be an important part of real world network models.

3.3 Vulnerability of scale-free structures

The exponent γ in Equation 1 indicates how quickly the tail of the distribution approaches to 0. If $2 < \gamma < 3$ the variance of the distribution goes to infinity as N goes to infinity, indicating that for very large networks one can observe hubs with very high connectivity [12]. This is the origin of the name scale free, meaning that there is no upper bound on the most connected components. For a thorough description of the properties of scale free networks, we refer to survey articles such as [1, 11, 15] and books such as [5, 8, 12, 19]. Four important characteristics of scale free networks can be derived:

- The epidemic threshold is very low;
- The average distance between two randomly chosen nodes is very small;
- The network is very resilient to random failure;
- The network is very vulnerable to targeted attacks.

Pastor-Satorras et al. [18] elaborate on epidemics and immunisation in scale-free networks. They conclude that the topology of the network has a great influence on the overall behaviour of epidemic spreading. Scale free networks are very weak in the face of infections, presenting an effective epidemic threshold that is

ISSN 0085-7130 © Telenor ASA 2005

vanishing at the limit $N \to \infty$. In an infinite population, this corresponds to the absence of any epidemic threshold below which major epidemic outbreaks are impossible.

The last two points in the list is known as the Achilles' heel of complex networks [3]. The resilience follows from the fact that as long as $\gamma < 3$ the critical threshold for fragmentation is $f_c = 1$ [10]. This demonstrates that scale-free networks cannot be broken into pieces by random removal of nodes. This extreme robustness to failure is rooted in the inhomogeneous network topology. As there are far more small nodes than hubs, random removal will most likely hit these. The price to pay for this is a vulnerability to attacks. Removal of a tiny fraction of the most connected nodes will break the network into pieces.

In the case of Internet we know that viruses and worms can spread extremely quickly from a single source, and that once they are circulating in the network an unprotected host is likely to be infected within a short time. Further, Internet has proven itself rather resilient against failure. While about 0.3 % of the nodes of the routers are down at any moment, we rarely observe major disruptions. From the observations above, a possible explanation for this is the topology of the network. While taking advantage of this property it is important to be aware that well-informed attackers can design scenarios to handicap the network.

4 Conclusion

The observations made in the previous section create a new setting for fault tolerance as it turns out that full protection of all nodes in the network is not necessary to ensure the functionality of the network. Protecting the hubs would ensure that major parts of the network are operational. To do this the hubs must be identified. This sounds straightforward, but it turns out to be a tough problem. In [9] Claffy, principal investigator at CAIDA[11], claims that interdomain Internet science cached after the retirement of NSFNET in 1994, and that today we

- Cannot figure out where an IP address is;
- Cannot measure topology effectively in either direction at any layer;
- Cannot track propagation of a routing update across the Internet;
- Cannot make a router give you all available routes, just best routes;
- Cannot get precise one-way delay from two places on the Internet;
- Cannot get an hour of packets from the core;
- Cannot get accurate flow counts from the core;
- Cannot get anything from the core with real addresses in it;
- Cannot get the topology of the core;
- Cannot get accurate bandwidth or capacity information;
- Cannot trust who is registry data;
- Have no general existing tool to tell "what is causing my problem now?"
- See that privacy and legal issues deter research.

Many of these unsolved problems in Internet operations and engineering are rooted in economics, ownerships and trust. Solving these problems is a long-term project and it is not clear how it should be done. Beyond the domain an operator is administrating, the information that can currently be achieved is limited to trace route and BGP table lookup like information. This might not be satisfactory in a world that keeps getting smaller and more dependent on the Internet.

The telephone operators were obliged to establish technical, economic and legal trust, while the cooperation between autonomous systems is purely driven by economic interests. Some of the questions that need answering are:

1 What benefits could the Internet society get from some centralised authority managing the large-scale functionality of the Internet? What are the risks connected to this?

2 What is the chance of a coordinated attack against Internet hubs today?

3 The scale free structure is likely to be represented at a local part of the Internet. Can nations or regions administrate and coordinate their "part" of the Internet? This would imply the introductions of restrictions on the liberty of the autonomous systems. Will they tolerate this?

The purpose of this article has been to show that the structure of a large, complex network is driven by simple mechanisms. These mechanisms are based on popularity of some of the actors, attributes we also find in economics. This mechanism is the origin of an Achilles' heel of robustness and vulnerability, which nobody currently knows how to handle.

[11] *Cooperative Association for Internet Data Analysis*

ISSN 0085-7130 © Telenor ASA 2005

References

1 Albert, R, Barabási, A-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 2002.

2 Audestad, J A. *E-bomber og e-granater: Om ikt og sårbarhet*. Et innspill til BAS 5 prosjektet, 2004.

3 Barabási, A-L. Emergence of scaling in complex networks. In: Bornholdt, S and Schuster, H G (eds). *Handbook of Graphs and Networks – From the Genome to the Internet*, chapter 3, 69–84. Wiley-VCH, first edition, 2003.

4 Barabási, A-L, Albert, R. Emergence of scaling in random networks. *Science*, 286, 509–511, 1999.

5 Barabási, A-L. *Linked – The New Science of Networks*. Perseus Publishing, first edition, 2002.

6 Baran, P. On distributed communication networks. *IEEE Transactions of the Professional Technical Group on Communication Systems*, 13 (3), 1964.

7 Bolobas, B. *Random Graphs*. Cambridge studies in advanced mathematics. Cambridge University Press, second edition, 2001.

8 Bornholdt, S, Schuster, H G (eds). *Handbook of Graphs and Networks – From the Genome to the Internet*. Wiley-VHC, first edition, 2003.

9 Claffy, K C. *Top problems of the internet and how to help solve them*. June 21, 2005 [online] – URL: www.caida.org/outreach/presentations/2005/topproblemsnet

10 Cohen, R, Erez, K, ben Avrahim, D, Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett*, 85 (4626), 2000.

11 Dorogovtsev, S N, Mendes, J F F. Evolution of networks. *Advances in Physics*, 51, 1079–1187, 2002.

12 Dorogovtsev, S N, Mendes, J F F. *Evolution of Networks – From Biological Nets to the Internet and WWW*. Oxford University Press, first edition, 2003.

13 Faloutsos, M, Faloutsos, P, Faloutsos, C. On power-law relationships of the internet topology. In: *SIGCOMM*, 251–262, 1999.

14 Goleniewski, L. *Telecommunications Essentials – The Complete Global Source for Communications Fundamentals, Data Networking and the Internet, and Next-Generation Networks*. Addison-Wesley, first edition, 2002.

15 Newman, M E J. The structure and function of complex networks. *SIAM Review*, 45 (2), 167–256, 2003.

16 St.meld. nr 17 (2001-2002). *Samfunnssikkerhet, veien til et mindre sårbart samfunn*. Oslo, Det Kongelige Justis- og Politidepartement.

17 Pastor-Satorras, R, Vespignani, A. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63, 2001. [Article id 066117]

18 Pastor-Satorras, R, Vespignani, A. Epidemics and immunization in scale-free networks. In: Bornholdt, S and Schuster, H G (eds). *Handbook of Graphs and Networks – From the Genome to the Internet*, chapter 5, 111–130. Wiley-VCH, first edition, 2003.

19 Pastor-Satorras, R, Vespignani, A. *Evolution and Structure of the Internet – A Statistical Physics Approach*. Cambridge University Press, first edition, 2004.

20 Pham, H (ed). *Handbook of Reliability Engineering*. Springer, 2003.

21 Spilling, P, Lundh, Y. Features of the internet history, the Norwegian contribution to the development. *Telektronikk*, 100 (3), 113–133, 2004.

22 Watts, D J. *Small Worlds – The Dynamic of Networks between Order and Randomness*. Princeton University Press, first edition, 1999.

Nils Kalstad Svendsen holds an MSc in mathematics from NTNU in 2002. He is currently a PhD student at Gjøvik University College working on a project entitled "Connections Between Network Topology and Network Security".

email: nilss@hig.no

Epidemic spreading over networks – A view from neighbourhoods

GEOFFREY S. CANRIGHT AND KENTH ENGØ-MONSEN

Geoffrey S. Canright is senior researcher with Telenor R&D

Kenth Engø-Monsen is research scientist at Telenor R&D

We give an overview of a new approach to understanding epidemic spreading on networks. The ideas are basic, and yet we believe that they have ready application to a wide range of problems, including the control of the spread of data viruses and other harmful electronic information. We build a *'topographic'* picture of the network, in the sense that there are neighborhoods of high and low *spreading power*. From this picture we develop a detailed understanding, at the level of neighbourhoods, of epidemic spreading over the network. Our analysis of spreading thus gives a finer resolution than typical whole-graph studies. Our picture is strongly confirmed by a series of simulations on empirical social networks, and is also supported by some limited mathematical results. Finally, we offer a set of design suggestions for both helping and hindering spreading, and report on some tests of these suggestions. The results of these tests are encouraging. Thus we find clear support for our belief that our approach can have significant practical value for the problem of network security.

1 Introduction

It is a well known fact of this century that electronic information can spread to many people in a very short time. This fact is good news for some people (spammers, bloggers), but can be rather bad news for those responsible for security. The battle against viruses, spam, and other forms of harmful or undesirable, self-propagating information is never-ending. We have in our own work approached the problem of security from the direction of *network analysis*. In particular, we have sought to answer two questions: (i) Given a network topology (data network, social network, etc), how can one usefully describe its *structure?* – and more specifically, how can one identify (based on the topological input) the *well connected clusters* in the network? (ii) Given such an analysis, how can it help us to understand how harmful information (viruses, diseases) are spread over the network? We note that the latter question is very closely related to a third question: (iii) Knowing the network topology, what measures can we take, in a cost-effective way, to *hinder* the spreading of viruses and other harmful information?

In this paper, we will report on and summarize some answers that we have obtained to question (i) (Ref. [1]) and to question (ii) (Ref. [2]). In addition, we will present some partial answers to question (iii); this work has not been published before. Since questions (i) and (ii) are confined purely to *understanding* the phenomenon of spreading, any answers we find may be useful either towards the goal of *helping* information propagation, or towards the goal of *hindering* it. Hence much of what we report here has implications for both goals. When we come to the question of *designing* or *modifying* networks however, we will focus primarily (although not exclusively) on the aim of making the network more secure. Furthermore, we note that our analysis is done at such a level of abstraction that it may be applied equally well to social networks (gossip spreading, disease spreading, spreading of innovations [3]) as to data networks. In fact, most of our analysis has application to any network for which it makes sense to model the links as being symmetric with respect to spreading. The simulations that we describe below are in fact carried out over empirically measured social networks, as these were readily available to us. We note in this regard that many modern networks are in fact a nontrivial mixture of social and technological. An outstanding example is a peer-to-peer file-sharing network, which is simply a network of computers, but one where the choice of active links is made (at least partially) by the human users. Such networks are self-organized in a very real sense; and this gives them rather interesting topological properties. We will have more to say on this when we report on our simulations of spreading over snapshots [4] of the Gnutella file-sharing network.

There are many kinds of models for epidemic spreading. In perhaps the simplest class of such models, one assigns to each node only one of two possible states: 'uninfected' or 'infected'. If you are uninfected ('susceptible'), you are deemed liable to be infected by any infected neighbors. Correspondingly, if you are infected, you remain so for the duration of the experiment – and you remain capable of infecting any or all of your neighbors. Of course, on some appropriate time scale, nodes become 'immune' to the infection: a human develops antibodies, a machine gets anti-virus software, the gossip becomes boring, or the innovation becomes outmoded. We focus on a shorter time scale here, so that we can ignore the state of acquired immunity. The technical name for our model of spreading is 'SI', since the nodes have only two states: Susceptible or Infected.

ISSN 0085-7130 © Telenor ASA 2005

Since spreading takes place over the links of a network, it is clear that the topology of the network can have a profound influence on the spreading process. In particular, we believe that the best understanding of spreading will come from a perspective which is based on a view of the whole network, and on an understanding of that network's structure. In earlier work [1], we have presented an approach to the analysis of network structure which is applicable to any network with symmetric (undirected) links. We also suggested that the analysis should be useful for the understanding of spreading over such a network. Recently [2], we have developed a detailed semi-quantitative theory for how spreading takes place on such networks. The theory is based entirely on our structural analysis. In this paper, we will offer a summary of the results of [1] and [2], along with some new results; most of the latter address the question of active *design* or *management* of networks for the purpose of controlling (helping or hindering) spreading. Our analysis offers clear suggestions for how to control spreading. We report here some of these suggestions, including some preliminary tests via further simulation.

Our approach departs from previous work in that we focus on both the *time* and *spatial progression* of the epidemic spreading. We take a spatial resolution which is not microscopic, but rather at the level of 'neighborhoods' – connected subgraphs with roughly the same spreading power. More traditional approaches (reviewed in [5]) start from the 'well-mixed' approximation that every node can infect every other with some probability, at all times. This approach may be said to have no network perspective; or, it may be said to postulate a graph with extremely good mixing – such as a random graph of high degree, or a complete graph. The review of Newman [5] also discusses more recent work, involving a network perspective. All such work is based on whole-graph properties, such as the node degree distribution; also, these approaches have focused on obtaining whole-graph results, either over time [6,7], or focusing especially on the infected fraction at very long times [8]. This latter question is of course only interesting for models more complex than the SI model; and indeed most work is directed towards the behavior of the SIS model (where nodes lose their infection after some time, and so become Susceptible again), or the SIR model (where nodes, after losing their infection, go through a refractory period).

Brauer [9] has examined the SI model for the case that the nodes (organisms, especially humans) are born and die. Because of the addition of these dynamic features, the steady infection rate is not necessarily 100 %. This work uses the well-mixed approximation, which gives rise to coupled ordinary differential equations.

A work which is perhaps closest to the present work is that of Wang et al. [10]. Their model is SIS, in that nodes can be "cured"; but it is based on a fully microscopic view of the network. In fact, their time evolution operator is the same as the one we develop in Ref. [2], with two differences. One is their addition of the "curing" term. This term is simply a multiple of the unit matrix, and so does not change the dominant eigenvector – which remains that of the adjacency matrix *A*. Because their model is SIS, the long-time infection fraction is not obvious, and must be solved for. The second difference in the time evolution operator of Wang et al. is that they neglect the cross terms – i.e. those arising from multiple transmissions to an infected node. This approximation is valid for low infection fraction – while (as we discuss below) it *may* also be good even as the infection fraction becomes large. Wang et al. report simulations which offer some support for this statement.

We emphasize that our theoretical work, like that of Wang et al. [10], uses the full adjacency matrix *A* in modelling the time evolution of the infection. Thus we start from a microscopic foundation. However, we will quickly appeal to a *'mesoscopic'* picture, in which it is meaningful and useful to speak of neighborhoods and their properties. As far as we know, our work is unique in this regard.

2 Topography from topology

In this section we give a review of the network analysis presented in [1]. An essential aspect of our approach to analysing the structure of a network is to define a measure of *centrality* for each node in the network. There are in fact many different measures of centrality, most of them coming from social science [11]. Our aim has been to find a measure of centrality which implies well-connectedness. Furthermore, we want a notion of well-connectedness which is not purely local. That is, we want a definition of well-connectedness (centrality) for node *i* which tells us something about the *neighbourhood* of node *i*. We reason that this kind of centrality can be useful for defining *well connected clusters* in the network, and, based on that, for understanding spreading on the same network.

Our strategy is to choose eigenvector centrality [12] as a useful measure of well-connectedness. Eigenvector centrality (EVC) has the desirable property that – since it depends on the properties of the *neighborhood* of a node, and not just of the node itself – it is rather 'smooth' over the graph (or network; we use these terms interchangeably). This is in contrast to the related quantity *degree centrality*, which simply counts the links leaving a node and so is completely local.

Let us elaborate on this difference. We start with degree centrality. It measures the 'importance' or connectedness of a node simply by counting the node's neighbors. Hence the degree centrality of node i is its node degree k_i. Clearly this quantity is completely local: a given node may have a very high degree centrality, and yet all of its neighbors may have a very low degree centrality – there is no correlation between this quantity from one node to its neighbors. Eigenvector centrality is seemingly (at least, in words) only a slight modification. To find a node's EVC, one (again) counts the node's neighbors. but *weighting the count by the centrality (EVC) of the neighbors*. That is: it's not just how many people you know, but who you know that matters. Mathematically we express this by

$$e_i = (const) \times \sum_{j=nn(i)} e_j. \qquad (1)$$

Here e_i is the EVC of node i, and $j = nn(i)$ means only sum over the nearest neighbors of i. This definition is clearly circular – my centrality depends on that of my neighbors, but theirs depends also on mine. However Equation (1) is readily solved to find the EVC, as long as one includes the constant (*const*) in the weighted sum. Furthermore, assuming only that the graph is connected and the links are symmetric, we know that the EVC values will all be positive (although they can be 'practically zero' for very peripheral nodes).

Thus we see that the EVC depends not only on how many neighbors a node has, but also on longer-ranged questions such as how many neighbors a node's neighbors have, etc. In fact, in principle, the EVC of a node depends on the *whole graph*. More relevant for our purposes, however, are two things: (i) the EVC clearly does measure well-connectedness in some kind of non-local fashion, and (ii) because of (i), the EVC values of nodes on any given path through the network cannot vary randomly and arbitrarily. That is, Eq. (1) forces the EVC of any node to be positively coupled to the EVC of that node's neighbors. We like to rephrase this as follows: the EVC is 'smooth' as one moves over the graph. (More mathematical arguments for this 'smoothness' are given in [1].)

The smoothness of the EVC allows one to think in terms of the 'topography' of the graph. That is, if a node has high EVC, its neighborhood (from smoothness) will also have a somewhat high EVC – so that one can imagine EVC as a smoothly varying 'height', with mountains, valleys, mountaintops, etc. We caution the reader that all standard notions of topography assume that the rippling 'surface' which the topography describes is continuous (and typically two-dimensional, such as the Earth's surface). A graph, on the other hand, is not continuous; nor does it (in general) have a clean correspondence with a discrete version of a d-dimensional space for *any* d. Hence, one must use topographic ideas with care. Nevertheless, we will appeal often to topographic ideas as aids to the intuition. Our definitions will be inspired by this intuition, but still mathematically precise, and appropriate to the realities of a discrete network.

First we define a 'mountaintop'. This is a point that is higher than all its neighboring points – a definition which can be applied unchanged to the case of a discrete network. That is, if a node's EVC is higher than that of any of its neighbors (so that it is a local maximum of the EVC), we call that node a Center. Next, we know that there must be a mountain for each mountaintop. We will call these mountains *regions*; and they are important entities in our analysis. That is, each node which is not a Center must either belong to some Center's mountain (region), or lie on a 'border' between regions. In fact, our preferred definition of region membership has essentially no nodes on borders between regions. Thus our definition of regions promises to give us just what we wanted: a way to break up the network into well connected clusters (the regions).

Here is our preferred definition for region membership: all those nodes for which a steepest-ascent path terminates at the same local maximum of the EVC belong to the same region. That is, a given node can find which region it belongs to by finding its highest neighbour, and asking that highest neighbour to find *its* highest neighbour, and so on, until the steepest-ascent path terminates at a local maximum of the EVC (i.e. at a Center). All nodes on that path belong to the region of that Center. Also, every node will belong to only one Center, barring the unlikely event that a node has two or more highest neighbors having exactly the same EVC, but belonging to differing regions.

Finally we discuss the idea of 'valleys' between regions. Roughly speaking, a valley is defined topographically by belonging to neither mountainside that it runs between. Hence, with our definition of region membership, essentially no nodes lie in the valleys. Nevertheless, it is useful to think about the 'space' between mountains – it is after all this 'space' that connects the regions, and thus plays an important role in spreading. This 'valley space' is however typically

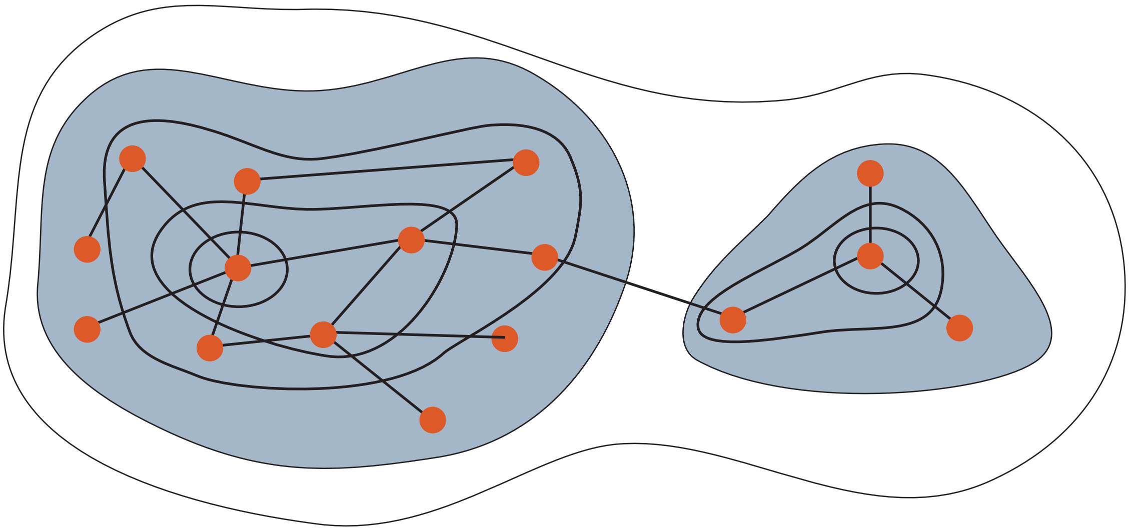


Figure 1 Our topographic view of a network, with two 'mountains' (regions)

composed only of inter-region links. We call these inter-region links *bridging links*. (And any node which lies precisely on the border may be termed a bridging node.)

Figure 1 offers a pictorial view of these ideas. We show a simple graph with 16 nodes. We draw topographic contours of equal height (EVC). The two Centers, and the mountains (regions) associated with each are clearly visible in the figure. The figure suggests strongly that the two regions, as defined by our analysis, are better connected internally than they are to one another. Furthermore, from the figure it is intuitively plausible that spreading (e.g. of a virus) will occur more readily *within* a region than *between* regions. Hence, Figure 1 expresses pictorially the two aims we seek to achieve by using EVC: (i) find the well connected clusters, and (ii) understand spreading.

In the next section we will elaborate on (ii), giving our qualitative arguments for the significance of EVC and regions (as we have defined them) for understanding spreading. We emphasize here that the term 'region' has, everywhere in this paper, a precise mathematical meaning; hence, to discuss subgraphs which lie together in some looser sense, we use terms such as 'neighborhood' or 'area'.

3 Topography and epidemic spreading

In order to understand spreading from a network perspective, we would like somehow to evaluate the nodes in a network in terms of their "spreading power". That is, we know that some nodes play an important role in spreading, while others play a less important role. One need only imagine the extreme case of a star: the center of the star is absolutely crucial for spreading of infection over the star; while the leaf nodes are entirely unimportant, having only the one aspect (common to every node in any network) that they can be infected.

Clearly, the case of the star topology has an obvious answer to the question of which nodes have an important role in spreading (have high spreading power). The question is then, how can one generate equally meaningful answers for general and complex topologies, for which the answer is not at all obvious? In this section we will propose and develop a qualitative answer to this question.

Our basic assumption (A) is simple, and may be expressed in a single sentence:

> ***Eigenvector centrality (EVC) is a good measure of spreading power.*** ***(A)***

We have tested this idea, via both simulations and theory [2]. In this section we will give qualitative arguments which support assumption (A); we will then go on to explore the implications of this assumption. We will see that we can develop a fairly detailed picture of how epidemic spreading occurs over a network, based on (A) and our structural analysis – in short, based on the ideas embodied in Figure 1.

First we recall that, because a node's EVC depends on that of its neighbors, the EVC values over a network may be thought of as 'smoothly varying' over the network. That is, a node with very high EVC cannot be surrounded by nodes with very low EVC. Of course, it is true that EVC tends to be positively correlated with a simpler measure of centrality, namely

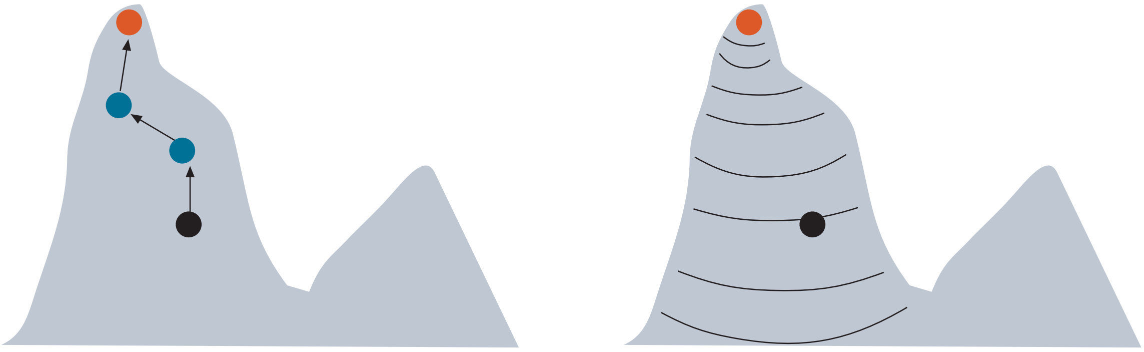


Figure 2 Schematic picture of epidemic spreading in one region. The two regions of Figure 1 are now viewed from the 'side', as if they were really mountains. (Left) Infection spreads from an initially infected node (black) toward more central ('higher') regions. (Right) Spreading of the infection reaches a high rate when the most central neighborhood (red) is reached; thereafter the rest of the region is infected efficiently

the node degree. In fact, one might say that the principal difference between the two measures is that EVC is constrained by its definition to be smooth, while node degree centrality is not[1]. This difference can however be nontrivial. For instance, a node with high degree, surrounded by many leaf nodes, and linked only tenuously to the bulk of a large and well-connected network, will have a low EVC, in spite of its high degree. The point is that EVC is sensitive to properties of neighborhoods, while node degree is not.

Thus, in short, there are no isolated nodes with high EVC. That is, a node with high EVC is embedded in a neighborhood with high EVC. (There can however be relatively isolated nodes with *low* EVC, as this situation is self-consistent. Low-EVC nodes can be isolated in the sense of having very few neighbors; but it is still the case that their neighbors will not have very much higher EVC.) Now if we take our basic assumption (A) to be true, then there are no isolated nodes with high spreading power. Instead, there are neighborhoods with high spreading power.

We then suppose that an infection has reached a node with modest spreading power. Suppose further that this node is not a local maximum of EVC; instead, it will have a neighbor or neighbors of even higher spreading power. The same comment applies to these neighbors, until one reaches the local maximum of EVC/spreading power.

Now, given that there are neighborhoods, we can discuss spreading in terms of neighborhoods rather than in terms of single nodes. It follows from the meaning of spreading power that a neighborhood characterized by high spreading power will have more rapid spreading than one characterized by low spreading power. Furthermore, we note that these different types of neighborhoods (high and low) are smoothly joined by areas of intermediate spreading power (and speed).

It follows from all this that, if an infection starts in a neighbourhood of low spreading power, it will tend to spread to a neighbourhood of higher spreading power. That is: spreading is faster *towards* neighborhoods of high spreading power, because spreading is faster *in* such neighborhoods. Then, upon reaching the neighbourhood of the nearest local maximum of spreading power, the infection rate will also reach a maximum (with respect to time). Finally, as the high neighborhood saturates, the infection moves back 'downhill', spreading out in all 'directions' from the nearly saturated high neighborhood, and saturating low neighborhoods.

We note that this discussion fits naturally with our topographic picture of network topology. Putting the previous paragraph in this language, then, we get the following: infection of a hillside will tend to move uphill, while the infection rate grows with height. The top of the mountain, once reached, is rapidly infected; and the infected top then efficiently infects all of the remaining adjoining hillsides. Finally, and at a lower rate, the foot of the mountain is saturated.

[1] *The star illustrates this difference to some extent. Suppose the graph is a star with n 'leaves' – that is, a graph with one node in the center, linked to each of n other nodes, each of which have no neighbour other than the center node. The degree centrality of the center is of course n, and that of the leaves is 1. The EVC of the center is however only $\sqrt{n}$ larger than the EVC of the leaves. Hence using EVC – which makes the centrality of the center dependent on that of its neighbors – gives a reduction (by a factor $1/\sqrt{n}$) in the (potentially large) difference in degree centrality between leaves and center.*

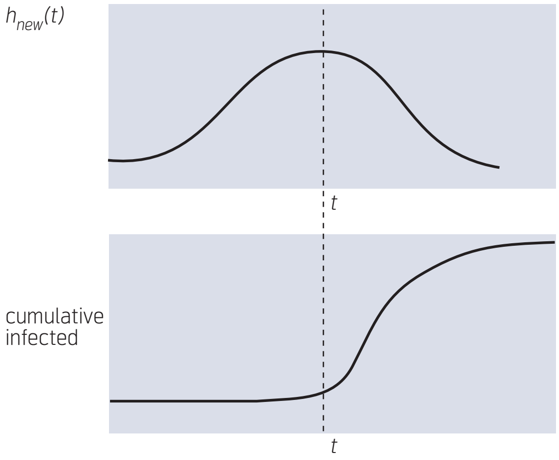


Figure 3 Schematic picture of the progress of an infection. (lower) The cumulative number of infected nodes over time. The infection "takes off" when it reaches highly central nodes. (upper) Our prediction of the "height" (centrality) of newly infected nodes over time. The vertical dashed line indicates roughly where in time the most central node's neighbourhood is reached

Figure 2 expresses these ideas pictorially. The figure shows our two-region example of Figure 1, but viewed from the 'side' – as if each node truly has a height. The initial infection occurs at the black node in the left region. It then spreads primarily uphill, with the rate of spreading increasing with increasing 'height' (= EVC, which tells us, by our assumption, the spreading power). The spreading of the infection reaches a maximum rate when the most central nodes in the region are reached; it then 'takes off', and infects the rest of the region.

We see that this qualitative picture addresses nicely the various stages of the classic S curve of innovation diffusion [13]. The early, flat part of the S is the early infection of a low area; during this period, the infection moves uphill, but slowly. The S curve begins to take off as the infection reaches the higher part of the mountain. Then there is a period of rapid growth while the top of the mountain is saturated, along with the neighboring hillsides. Finally, the infection rate slows down again, as the remaining uninfected low-lying areas become infected.

We again summarize these ideas with a figure. Figure 3 shows a typical S curve for infection, in the case (as we study in this paper) that immunity is not possible. Above this S curve, we plot the expected centrality of the newly infected nodes over time. According to our arguments above, relatively few nodes are infected before the most central node is reached – even as the centrality of the infection front is steadily rising. The takeoff of the infection then roughly coincides with the infection of the most central neighbourhood. Hence, the part of Figure 3 to the left of the dashed line corresponds to the left half of Figure 2; similarly, the right-hand parts of the two figures correspond.

One might object that this picture is too simple, in the following sense. Our picture gives an S curve *for a single mountain*. Yet we know that a network is often composed of several regions (mountains). The question is then, why should such multi-region networks exhibit a single S curve?

Our answer here is that such networks need not necessarily exhibit a single S curve. That is, our arguments predict that each region – defined around a local maximum of the EVC – will have a single S curve. Then – assuming that each node belongs to a single region, as occurs with our preferred rule for region membership – the cumulative infection curve for the whole network is simply the sum of the infection curves for each region. These latter single-region curves will be S curves. Thus, depending on the relative timing of these various single-region curves, the network as a whole may, or may not, exhibit a single S curve. For example, if the initial infection is from a peripheral node which is close to only one region, then that region may take off well before neighboring regions. On the other hand, if the initial infection is in a valley which adjoins several mountains, then they may all exhibit takeoff roughly simultaneously – with the result being a sum of roughly synchronized S curves, hence a single S curve.

Let us now summarize and enumerate the predictions we take from this qualitative picture.

a Each region has an S curve.

b The number of takeoff/plateau occurrences in the cumulative curve for the whole network may be more than one; but it will not be more than the number of regions in the network.

c For each region – assuming (which will be typical) that the initial infection is not a very central node – growth will at first be slow.

d For each region (same assumption) initial growth will be towards higher EVC.

e For each region, when the infection reaches the neighborhood of high centrality, growth "takes off".

f An observable consequence of (e) is then that, for each region, the *most* central node will be infected at, or after, the takeoff – but not before.

g For each region, the final stage of growth (saturation) will be characterized by low centrality.

4 Simulations

We have run simple simulations to test our qualitative picture. As noted above, we have implemented an SI model. That is, each node is Susceptible or Infected. Once infected, it remains so, and retains the ability to infect its neighbors indefinitely. We have used a variety of sample networks, extracted from data obtained in previous studies. Results reported here will be taken from simulations performed on: (i) seven distinct snapshots [4] of the Gnutella peer-to-peer file-sharing network, taken in late 2001; (ii) a snapshot of the collaboration graph of our own R&D department; and (iii) a snapshot of the collaboration graph for the Santa Fe Institute [14]. We will denote these graphs as (i) g1 – g7; (ii) fou (Norwegian for R&D); and (iii) sfi. We use social networks, rather than data networks, as the former were both empirically measured and readily available.

Our procedure is as follows. We initially infect a single node. Each link *ij* in the graph is assumed to be symmetric (consistent with our use of EVC), and to have a constant probability *p*, per unit time, of transmitting the infection from node *i* to node *j* (or from *j* to *i*), whenever the former is infected and the latter is not. All simulations were run to the point of 100 % saturation. Thus, the ultimate *x* and *y* coordinates of each cumulative infection curve give, respectively, the time needed to 100 % infection, and the number of nodes in the graph.

4.1 Gnutella graphs

The Gnutella graphs, like many self-organized graphs, have a power-law node degree distribution, and are thus well connected. Consistent with this, our analysis returned either one or two regions for each of the seven snapshots. We discuss these two cases (one or two regions) in turn. Each snapshot is taken from [4], and has a number of nodes on the order of 900 – 1000.

4.1.1 Single region

The snapshot termed ‘g3’ was found by our analysis to consist of a single region.

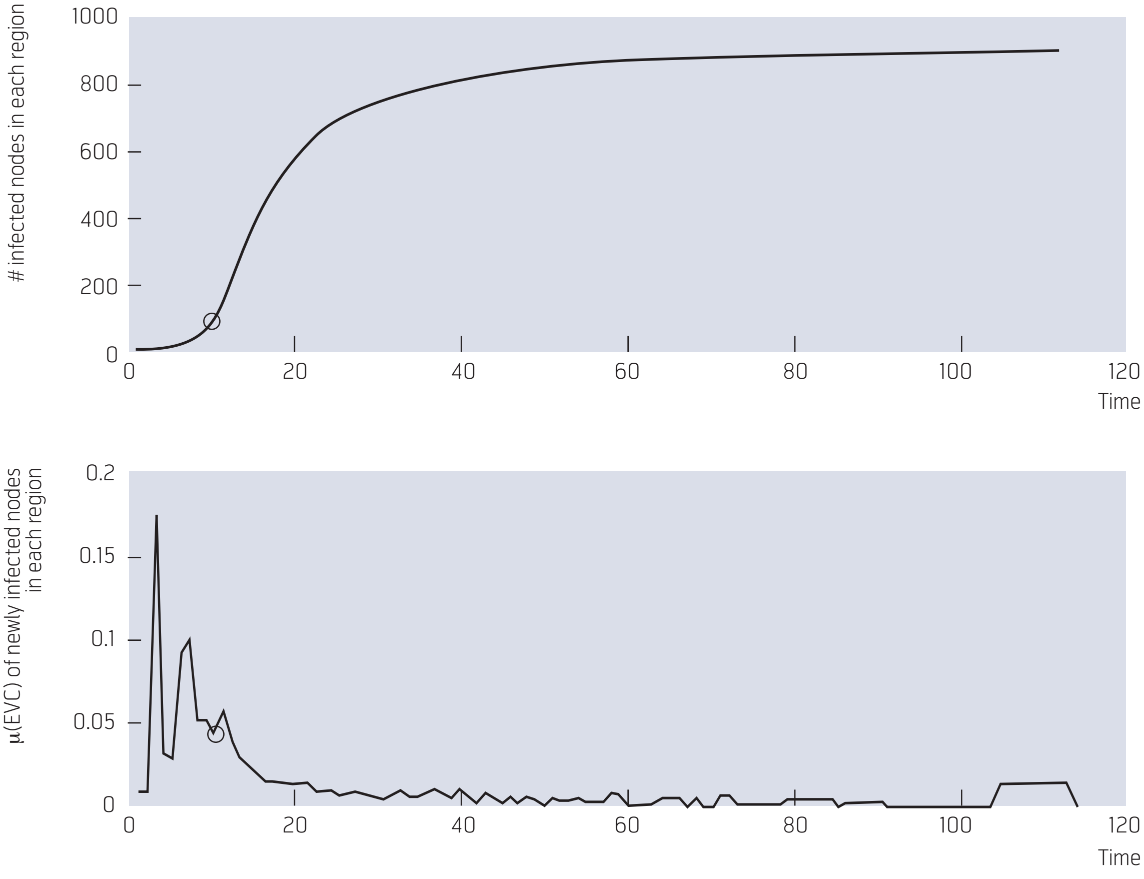


Figure 4 (top) Cumulative infection for the Gnutella graph g3. The circle marks the time at which the most central node is infected. (bottom) Average EVC (‘height’) of newly infected nodes at each time step

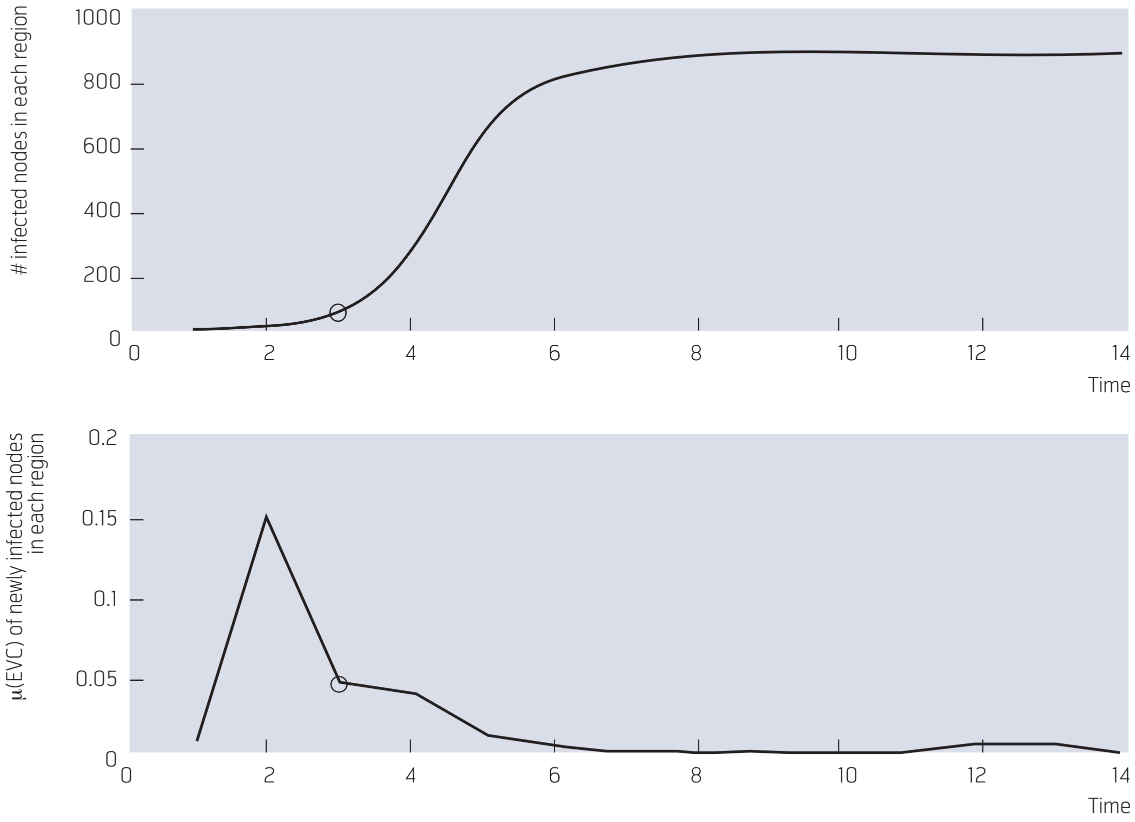


Figure 5 Same simulation as Figure 1, except for: (i) p = 0.6, and (ii) new random numbers determining the infection events

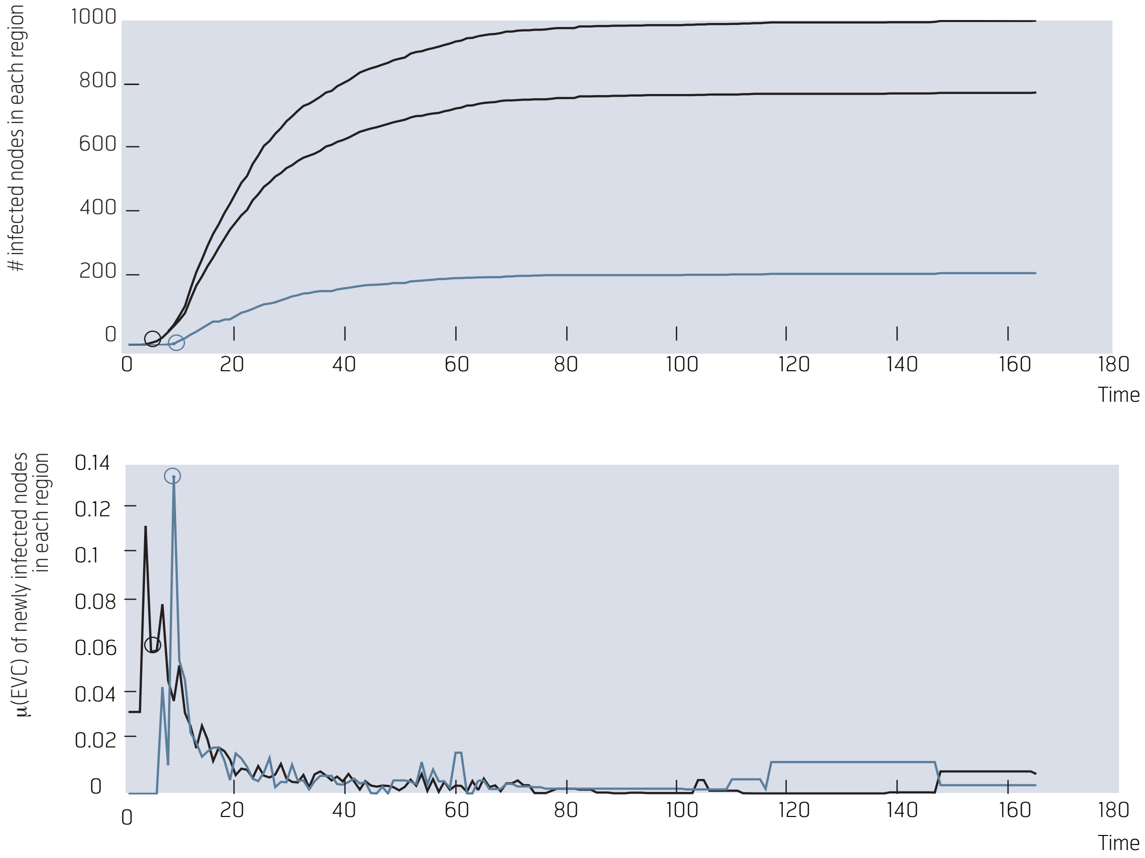


Figure 6 (top) Cumulative infection curves, and (bottom) µ curves, for a two-region graph g1, with p = 0.04. The upper plot shows results for each region (black and blue), plus their sum (also black). The lower plot gives µ curves for each region. Note that here, and in all subsequent multiregion figures, we mark the time of infection of the highest-EVC node for each region with a small colored circle

Figure 4 shows a typical result for the graph g3, with link infection probability $p = 0.05$. The upper part of the figure shows the cumulative S curve of infection, with the most central node becoming infected near the 'knee' of the S curve. The lower part of Figure 4 shows a quantity termed μ – that is, the average EVC value for newly infected nodes at each time. Our qualitative arguments (Figure 3) say that this quantity should first grow, and subsequently fall off, and that the main peak of μ should coincide with the takeoff in the S curve. We see, from comparing the two parts of Figure 4, that the most central nodes are infected roughly between time 2 and time 20 – coinciding with the period of maximum growth in the S curve. Thus, this figure supports all of our predictions a – g above, with the minor exception that there are some fluctuations superimposed on the growth and subsequent fall of μ over time.

It is interesting to note that this picture is rather insensitive to the probability parameter p. For example, Figure 5 shows results for the same graph and same initial node, with the probability now $p = 0.6$. We see that the main effect of this much higher p is simple, and offers few surprises: the time scale is of course much compressed, with the expected result that the cumulative curve is less smooth. We note that even the extreme case of setting $p = 1$ gives a picture very much like that of Figure 5.

4.1.2 Two regions

Figure 6 shows typical behavior for the graph g1, which consists of two regions. We see that the two regions go through takeoff roughly simultaneously. This is not surprising, as the two regions have many bridging links between them. The result is that the sum of the regional infection curves is a single S curve. We also see that each region behaves essentially the same as did the single-region graph g3. For instance, the new centrality (μ) curves for each region first rise, and then fall, with their main peak (before time 20) coinciding with the period of most rapid growth for the respective regions (and for the whole graph). These results thus add further support to our predictions a – g.

4.2 Telenor R&D

We have formed a collaboration graph for the researchers working at Telenor R&D. (The rules for forming a collaboration graph are simple: the nodes are researchers, and if two researchers have authored one or more papers together, they get a link between

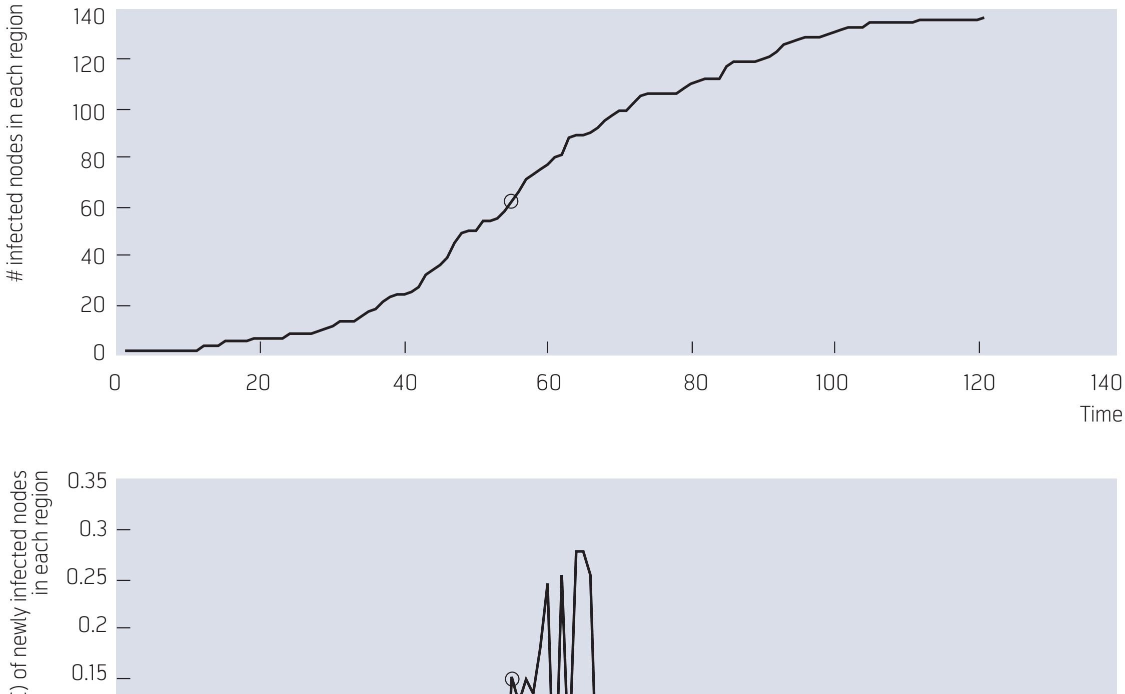


Figure 7 Spreading for the single-region fou graph, with p = 0.04. Note that the S curve is somewhat 'flat' on a time scale of 100 % saturation

ISSN 0085-7130 © Telenor ASA 2005

them.) For this graph, we analyzed the largest connected component, consisting of 137 nodes. Our analysis gave a single region for this graph. Spreading behavior was much like that seen in Figure 4. One difference is that the S curve is less smooth – an effect of the small size of the fou graph. Another difference is that, for many simulations (but not all), the time period of the rise of the S is disproportionately large compared to the time needed for 100 % of saturation. Also, in such cases, the time of infection of the most central node tends to fall rather late after the onset of the rise (the 'knee' of the S). We show an example of this behavior in Figure 7.

The differences between Figures 4 and 7 are interesting; however both figures are fully consistent with our predictions a – g. We note also an interesting correlation here, which is not surprising: steeper S curves tend to be associated with earlier infection times (relative to the knee) for the most central node.

One possible explanation for the slow rise in the S curve of Figure 7 is that the single mountain for the fou graph is relatively flat. That is, we expect the rate of growth of the infection to be high when the infection front is at a neighbourhood of high centrality – but for a single, flat region, the latter is never very large, and so the takeoff will not be very pronounced.

4.3 Santa Fe Institute

The sfi graph gives three regions under our analysis. The spreading behavior varied considerably on this graph, depending both on the initially infected node, and on the stochastic outcomes for repeated trials with the same starting node.

Figure 8 shows an untypical case for this graph. The aspect that is untypical here is that the whole-graph cumulative infection curve resembles (somewhat) a single, smooth, S curve. This result was obtained however for the rather artificial starting condition that the first infected node was the most central node in the largest region – hence in the entire graph. The result is that this region takes off immediately. (Our prediction c thus does not hold; but its *assumption* has been violated by our infecting the most central node first.) The next largest region (blue curve) is however infected fairly soon thereafter, so that its takeoff is not clearly seen in the total curve. Finally, the third region (red curve) takes off considerably later. But, because it is small, and its takeoff occurs before the blue region is fully saturated, the takeoff

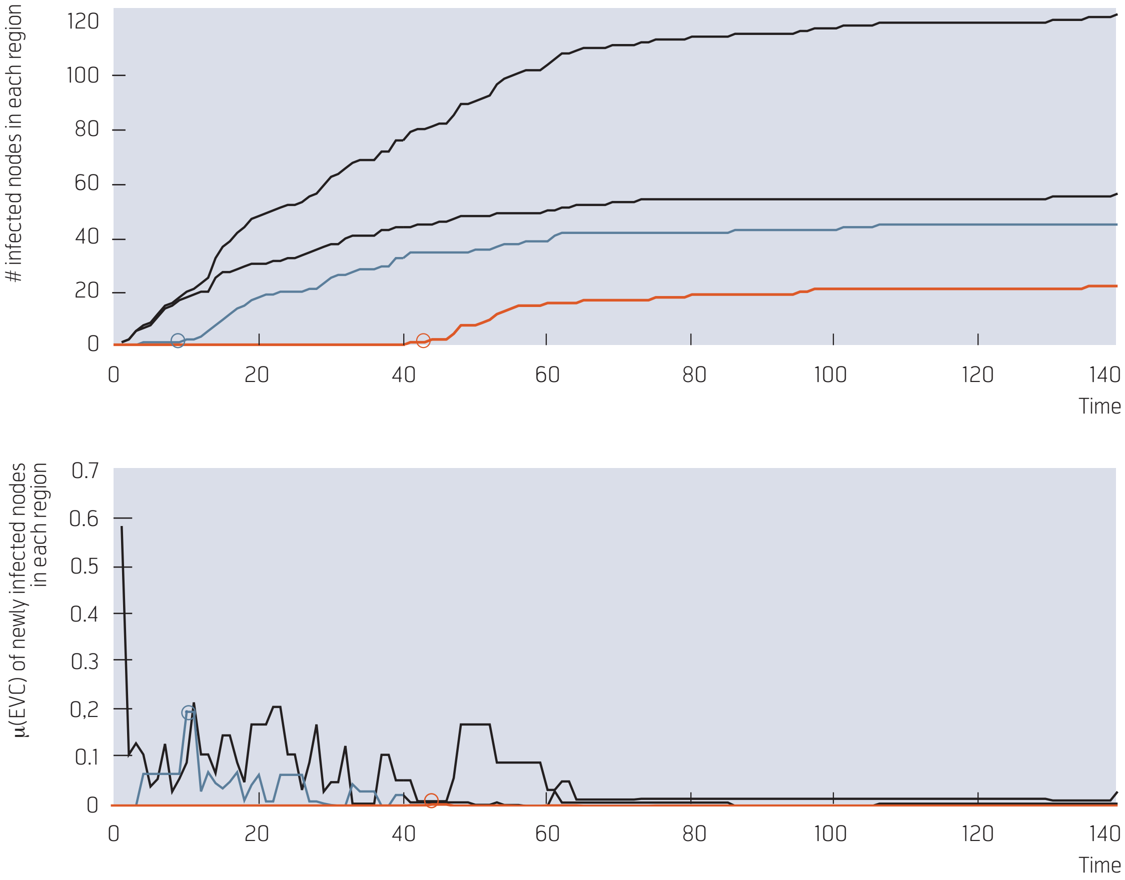


Figure 8 Spreading on the sfi graph, p = 0.04. The start node is the most central node in the largest of the three regions (black); hence its infection time is not marked

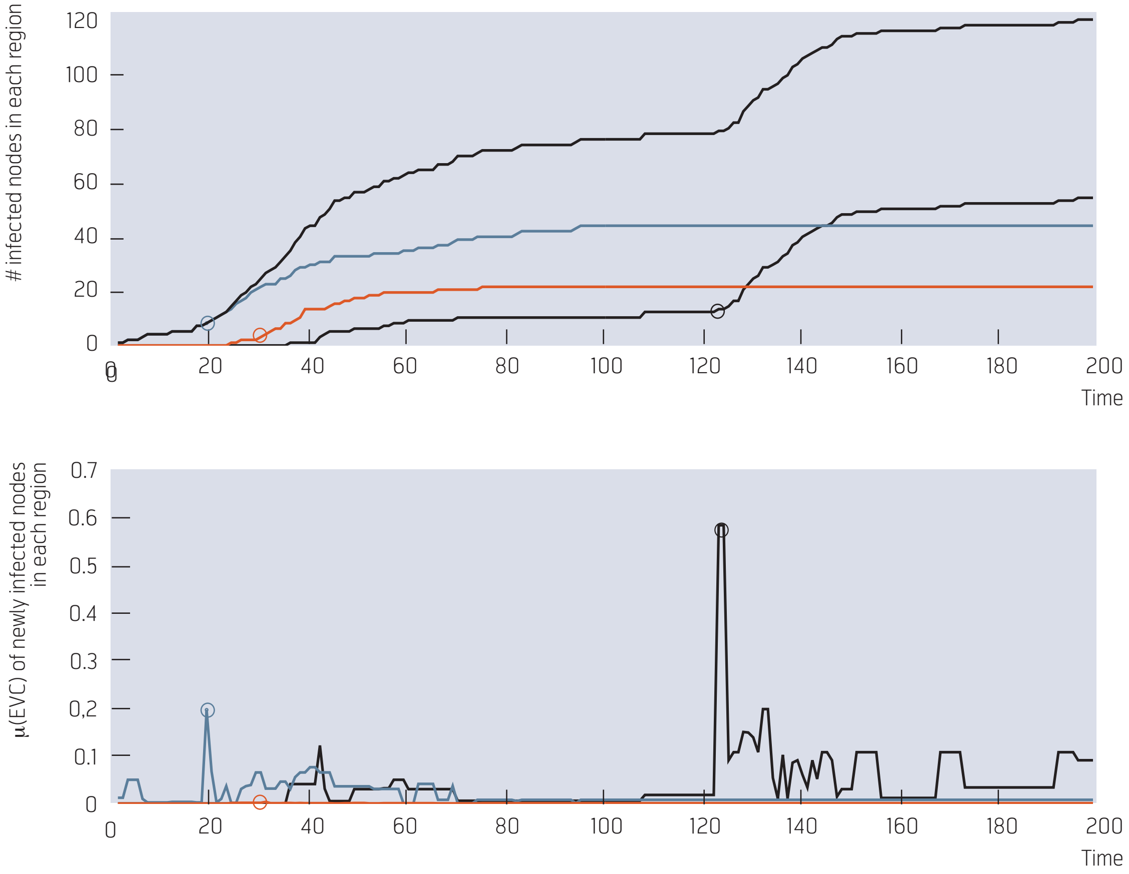


Figure 9 Same as Figure 8, except a randomly chosen start node

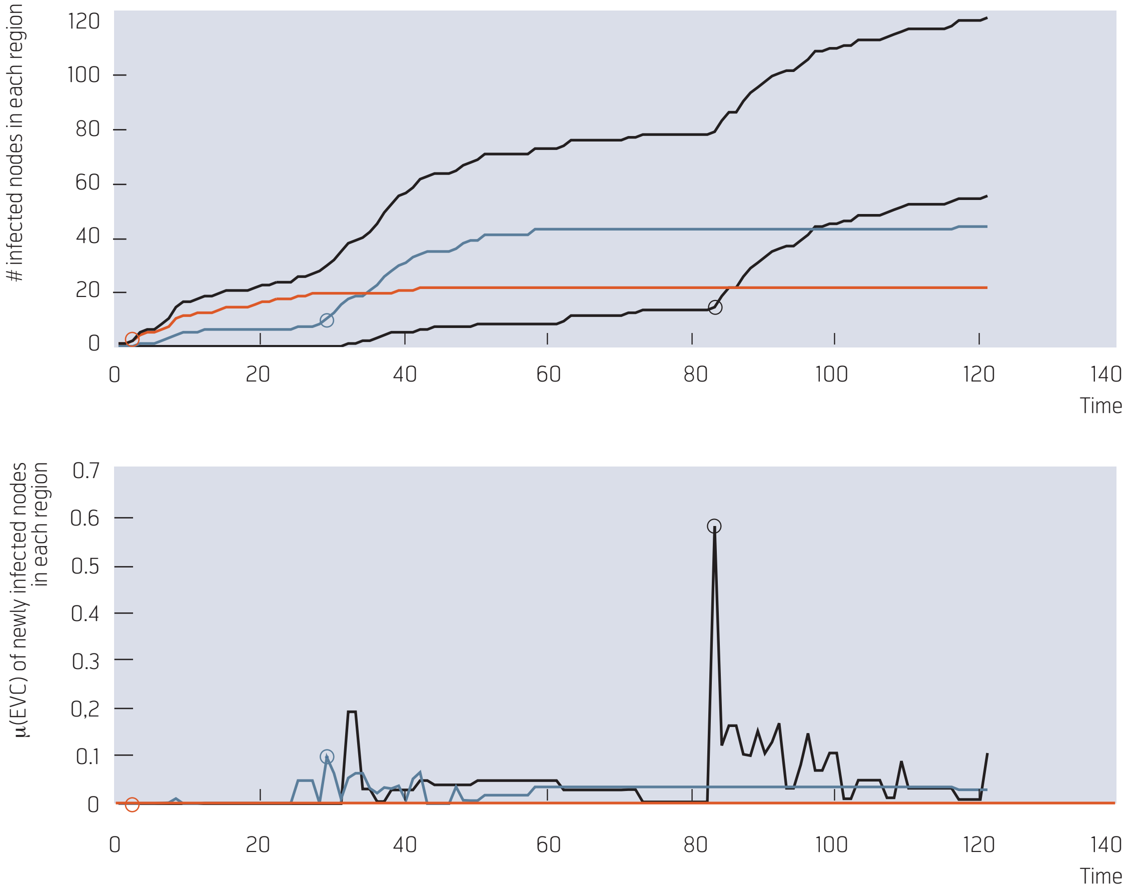


Figure 10 Spreading on the sfi graph, p = 0.05. In this case, the existence of three regions in the graph is clearly seen in the total cumulative infection curve

of this third region is also not clear from inspection of the total infection curve.

From observing the μ curves, we see that that of the blue region is as predicted. The red region is similar but not visible on the scale of the figure. The largest (black) region's μ curve lacks the initial rise in centrality – but this is to be expected, as the infection began at the top.

Now we move to a more typical case for the sfi graph. Figure 9 shows the behavior for the same infection probability, but with a randomly chosen start node.

The interesting feature of this simulation is that the cumulative curve shows very clearly two takeoffs, and two plateaux. That is, it resembles strongly the sum of two S curves. And yet it is easy to see how this comes about, from our region decomposition: the blue and red curves take off roughly simultaneously, while the largest (black) region takes off only after the other regions are saturated.

The μ curves show roughly the expected behavior – the qualification being that they are rather noisy. Nevertheless, the main peak of each μ curve corresponds to the main rise of the corresponding region's S curve.

We reiterate that the behavior seen in Figure 9 is much more typical for this graph than that seen in Figure 8.

We examine yet one more example from the sfi graph. Figure 10 shows a simulation with a different start node from Figures 9 or 8, and with $p = 5$ %. The message from Figure 10 is clear: the cumulative infection curve shows clearly three distinct S curves – takeoff followed by plateau – one after the other. It is also clear from our regional infection curves that each region is responsible for one of the S curves in the total infection curve: the smallest (red) region takes off first, followed by the blue region, and finally the largest (black) region. In each case, the time of infection for the most central node of the region undergoing takeoff lies very close to the knee of the takeoff. And, in each case, the peak of the μ curve coincides roughly with the knee of the takeoff.

We note that the behavior seen in Figure 10 is neither very rare nor very common. The most common behavior, from examination of over 50 simulations with this graph, is most like that seen in Figure 9; but we have seen behavior which is intermediate to that in Figures 8 and 9 (i.e. neither clearly one nor clearly two takeoffs), and also behavior which is intermediate to that of Figures 9 and 10. In particular, multiple simulations with the same start node and probability as that for Figure 10 have yielded two clear takeoffs (as in Figure 9), three clear takeoffs (as in Figure 10), and intermediate behavior. It is interesting to note that the behavior for this start node, with $p = 1$ (hence with deterministic behavior), gives a cumulative infection curve which is best described as showing between two and three takeoffs.

Thus the behavior of spreading on the sfi graph gives the strongest confirmation yet of our prediction b: that, for a graph with r regions, one may see up to r, but not more than r, distinct S curves in the total cumulative infection curve. All of our observations, on the various graphs, are consistent with our predictions; but it is only in the sfi graph that we clearly see all of the (multiple) regions found by our analysis to be present in the graph.

One very coarse measure of the well-connectedness of a network is the number of regions in it, with a high degree of connectivity corresponding to a small number of regions, and with a large number of regions implying poor connectivity. By this very coarse criterion, the sfi graph is as well connected as another three-region graph studied by us (the 'hio graph', reported in [2] and in [3]). Based on our spreading simulations on this hio graph (reported in [2]), we would say that in fact the sfi graph is less well connected than the hio graph. We base this statement on the observation that, for the hio graph, we never found cases where the different regions took off at widely different times – the three regions are better connected to one another than is the case for the sfi graph. (In detail: the three regions define three pairs of regions; and these three *pairs* are bridged by 57, 50, and 7 bridge links.)

Examination of the sfi graph itself (Figure 11) renders this conclusion rather obvious: the three regions are in fact connected in a linear chain. (The three pairs of regions have 6, 10, and 0 bridging links.) Thus, it is not surprising that we can, in some simulation runs, clearly see the takeoff of each region, well separated in time from that of the other regions. The hio graph, in contrast, has many links between each pair of regions of the three. We note finally in this context that the Gnutella graphs are very well connected; those that resolve to two regions always have numerous links between the two.

Finally, we note that the sfi graph, while clearly conforming to our assertion b (we see up to three regions), does not wholly agree with assertion a (that each region will have a clear S curve). For instance, in Figure 9, the black region has a small 'premature'

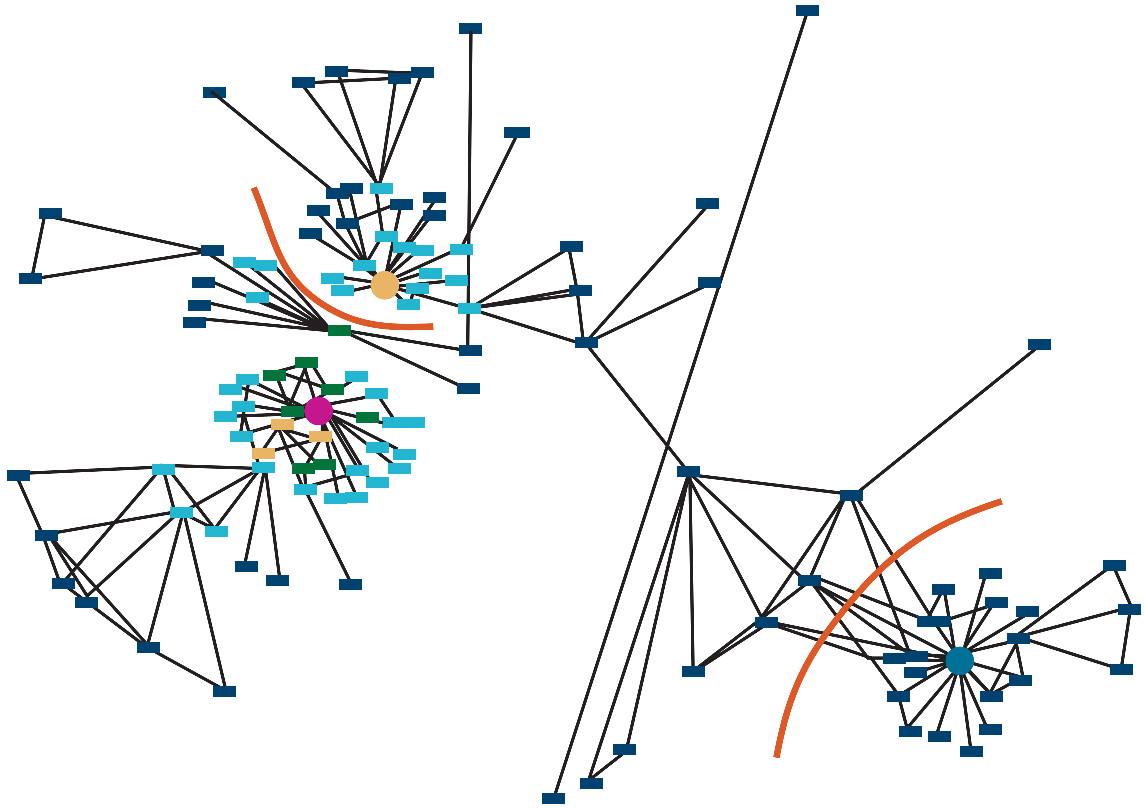


Figure 11 A visualization of the collaboration graph for the Santa Fe Institute. Colors of nodes code for eigenvector centrality, with warm colors implying higher EVC. Also, the most central node in each region is denoted with a round circle. Thick red lines show boundaries between regions. The leftmost region (with a magenta center) is the largest, and has black curves in Figures 8 – 10; the rightmost region is the smallest (red curves in Figures 8 – 10); and the middle region has blue curves in Figures 8 – 10. Hence, in the text, we refer to these three regions, respectively, as the 'black region', the 'blue region', and the 'red region'.

takeoff around time 40, giving a visible plateau before the 'main' takeoff after time 120. Similarly, both the blue and the black regions show such 'premature' takeoffs in Figure 10. These observations are not exceptional: the blue region in particular is prone to such behavior, showing two stages of growth in over half of the simulations we have run on the sfi graph. This weakens the support for assertion a; we would then say that a single S curve for a single region is a rule that is followed in most cases, but not all. However these two-stage takeoffs do not contradict the other predictions on our list. For instance, there is always a corresponding 'premature hump' in the μ curve for the region in question.

It is clear from Figure 11 where the 'premature takeoffs' arise. The black region includes a subcluster which is easily identified visually in Figure 11 (close to the next region). This subcluster is connected to the remainder of the black region by a single link. These same statements hold true for the blue region – it has a subcluster, joined to it by a single link, but close to the rightmost region. Hence, in Figure 10, for example, the infection starts on the 'far side' of the red region. Then it spreads to the adjacent subcluster of the blue region, but does not reach the most central nodes of that region until after about 20 more units of time (about $1/p$). Soon after the center of the blue region is infected, the subcluster of the black region begins to be infected; but again there is a delay (about twice $1/p$) before reaching the main center of the black region.

We have explored other rules for assigning nodes to regions [1]. For instance, if we simply use shortest distance rather than steepest ascent as our criterion, then the two above mentioned subclusters move: the blue subcluster moves to the red region, and the black subcluster becomes a set of 'border nodes' equidistant from both centers. We have not used the shortest distance rule in this paper, because it tends to place too many nodes in general in border regions. Possibly, one could obtain cleaner S curves for the three sfi regions using the shortest distance rule. However, this rule for defining regions ignores the view that the EVC should be viewed as a height function; hence it is less topographically motivated, and so may also give poorer results in many cases.

5 Mathematical theory

In [2] we have developed a mathematical theory for the qualitative ideas expressed here. We have focused on two aspects there, which we will simply summarize here.

5.1 Definition of spreading power

The first problem is to try to quantify and make precise our assumption (A). Since (A) relates to two quantities – spreading power and EVC – and the latter is precisely defined, the task is then to define the former, and then to seek a relation between the two.

Such a relation is intuitively reasonable. A node which is connected to many well-connected nodes should have higher spreading power, and higher EVC, than a node which is connected to equally many, but poorly connected, nodes. We have offered a precise definition of spreading power in [2]. Our reasoning has two steps: first we define an 'infection coefficient' $C(i,j)$ between any pair of nodes i and j. This is simply a weighted sum of all non-self-retracing paths between i and j, with lower weight given to longer paths. Thus many short paths between two nodes gives them a high infection coefficient. Our definition is symmetric, so that $C(i,j) = C(j,i)$.

Next we define the spreading power of node i to be simply the sum over all other nodes j of its infection coefficient $C(i,j)$ with respect to j. As long as the graph is connected, every node will have a nonzero $C(i,j)$ with every other, thus contributing to the sum. Hence each node has the same number of terms in the sum; but the nodes with many large infection coefficients will of course get a higher spreading power.

We then show in [2] that one can make a strong connection between this definition of spreading power and the EVC, if one can ignore the restriction to non-self-retracing paths in the definition. We restrict the sum to non-self-retracing paths because self-retracing paths do not contribute to infection in the SI case. This restriction makes the obtaining of analytical results harder.

5.2 Mathematical theory of SI spreading

We have in [2] given exact equations for the propagation of an infection, for arbitrary starting node, in the SI case. These equations are stochastic – expressed in terms of probabilities – due to the probabilistic model for spreading over links. They are not generally solvable, even in the deterministic case when $p = 1$. The problem in the latter case is again the need to exclude non-self-retracing paths. However, we have performed an expansion in powers of p for the time evolution of the infection probability vector. This expansion shows that the dominant terms are those obtained by naïvely applying the adjacency matrix (i.e. ignoring self-retracing paths because they are longer, hence higher order in p). The connection to EVC is then made: naïvely applying the adjacency matrix gives weights (infection probabilities) which approach a distribution proportional to the EVC. Hence we get some confirmation for our claim that, in the initial stages of an infection, the front moves towards higher EVC.

An interesting observation that remains to be explained is the insensitivity of the time evolution (see Figures 4 and 5) to the probability p. We note that short paths receive the most weight in two limits: $p \rightarrow 0$ and $p \rightarrow 1$. The observed insensitivity to p suggests that such short paths dominate for all p between these limits; but we know of no proof for this suggestion.

6 Design and improvement of networks

In this section we go beyond the problem of analysis, and address the problem of design of networks [15]. Our ideas have some obvious implications for design. Since we focus in this paper on security – i.e. on preventing the spreading of harmful information such as viruses – we will focus primarily on design for prevention here. However, we begin with a discussion of the opposite problem: how to *help* spreading, by modifying the topology of a given network.

6.1 Measures to improve spreading

We frame our ideas in terms of our topographic picture. Now we suppose that we wish to design, or modify the design of, a network, so as to improve its efficiency with respect to spreading. It is reasonable, based on our picture, to assume that a *single region* is the optimal topology for efficient spreading. Hence we offer here two ideas which might push a given (multi-region) network topology in this direction:

1 We can add more bridge links between the regions;
2 As an extreme case of 1, we can connect the Centers of the regions.

Idea 2 is a "greedy" version of idea 1. In fact, the greediest version of idea 2 is to connect *all* Centers to *all*, thus forming a complete subgraph among the Centers. However, such greedy approaches may in practice be difficult or impossible. There remains then the general idea 1 of building more bridges between the regions. Here we see however no reason for not taking the greediest practical version of this idea. That is: build the bridges between nodes of high centrality on both sides – preferably as high as possible. Our picture strongly suggests that this is the best strategy for modifying topology so as to help spreading.

We note that the greediest strategy is almost guaranteed to give a single-region topology as a result. Our reasoning is simple. First, the existing Centers cannot all be Centers after they are all connected one to another – because two adjacent nodes cannot both be local maxima of the EVC (or of anything else). Therefore, either new Centers turn up among the remaining nodes as a result of the topology modification, or only one Center survives the modification. In the latter case we have one region. The former case, we argue, is unlikely: we note that the EVC of the *existing* Centers is (plausibly) strengthened (raised) by the modification *more* than the EVC of other nodes. That is, we believe that connecting existing centers in a complete subgraph will 'lift them up' with respect to the other nodes, as well as bringing them closer together. If this 'lifting' idea is correct, then we end up with a single Center and a single region.

We have tested the 'greediest' strategy, using the sfi graph as our starting point. We connected the three Centers of this graph (Figure 11) pairwise. The result was a single region, with the highest Center of the unmodified graph at its Center. We then ran 1000 SI spreading simulations on this modified graph, with p = 10 %, and collected the results in the form of time to saturation for each run.

In our first test, we used a random start node (first infected node) in each of the 1000 runs. For the unmodified sfi graph, we found a mean saturation time of 83.8 time units. For the modified, single-region graph, the corresponding result was 64.0 time units – a reduction by 24 %. Note that this large reduction was achieved by adding only three links to a graph which originally had about 200 links among 118 nodes.

Figure 12 gives a view of the distribution of saturation times for these two sets of runs. We show both the binned data and a fit to a gamma probability distribution function. Here we see that the variance as well as the mean of the distribution is reduced by the topology improvement.

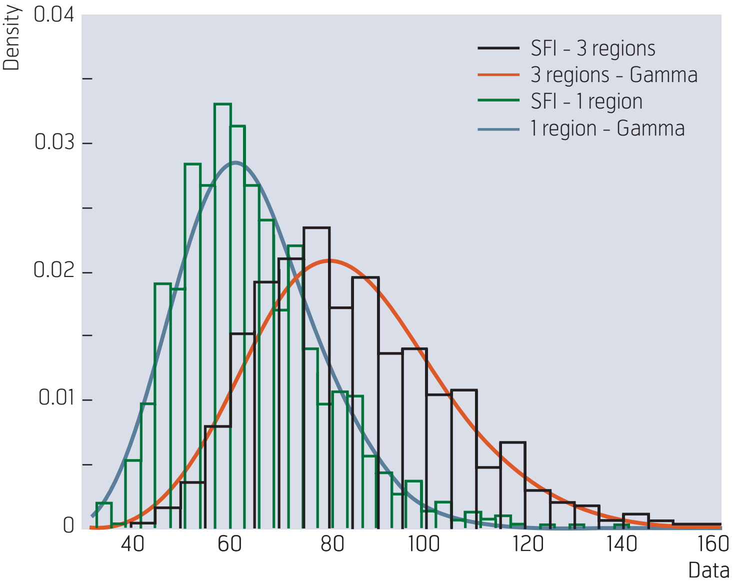


Figure 12 Distribution of saturation times for 1000 runs on the sfi graph, with a random start node. Left (blue): with the addition of three links connecting the three Centers. Right (red): original graph, with three regions.

	Random start node	High-EVC start node
Original graph	83.8	68.9
Connect Centers	64.0	56.0

Table 1 Mean time to saturation for the sfi graph, with and without two kinds of assistance: connecting the Centers of the three regions, and choosing a well placed start node

Our next set of tests used a start node located close to the highest Center (the only Center in the modified graph). These tests thus involve another form of 'help', namely, choosing a strategically located start node. We wish then to see if the new topology gives a significant difference even in the face of this strategic start. Our results, for mean saturation time over 1000 runs, were 68.9 time units for the unmodified sfi graph, and 56.0 units for the modified graph. Thus, connecting the three Centers of the sfi graph reduced the mean saturation time by almost 20 %, even though a highly advantageous node was used to start the infection. Also, as in Figure 12, the variance was reduced by the topology improvement. We summarize these results in Table 1.

From Table 1 we get some idea of the relative efficacy of the two kinds of help. That is, our base case is the unmodified graph with random start node (83.8). Choosing a strategic start node gave about an 18 % reduction in mean saturation time. Improving the topology, without controlling the start node, gave almost 24 % reduction. Thus we see that, for these limited tests, both kinds of help give roughly the same improvement in spreading rate, with a slight advantage coming from the greedy topology change. Finally, note that using *both* kinds of help in combination gave a reduction of mean saturation time of about 1/3. We note that both of these kinds of help are based on our topology analysis, and cannot be implemented without it.

6.2 Measures to prevent spreading

Now we address the problem of designing, or re-designing, a network topology so as to hinder spreading. Here life is more complicated than in the helping case. The reason for this is that we build networks in order to support and facilitate communication. Hence we cannot simply seek the extreme, 'perfect' solution – because the ideal solution for hindering spreading is one region per node, i.e. disconnect all nodes from all others! Instead we must consider *incremental* changes to a given network. We consider two types of 'inoculation' strategies – inoculating nodes (which is equivalent to removing them, as far as spreading of the virus is concerned), or inoculating links (which is also equivalent to removing them). Again we offer a list of ideas:

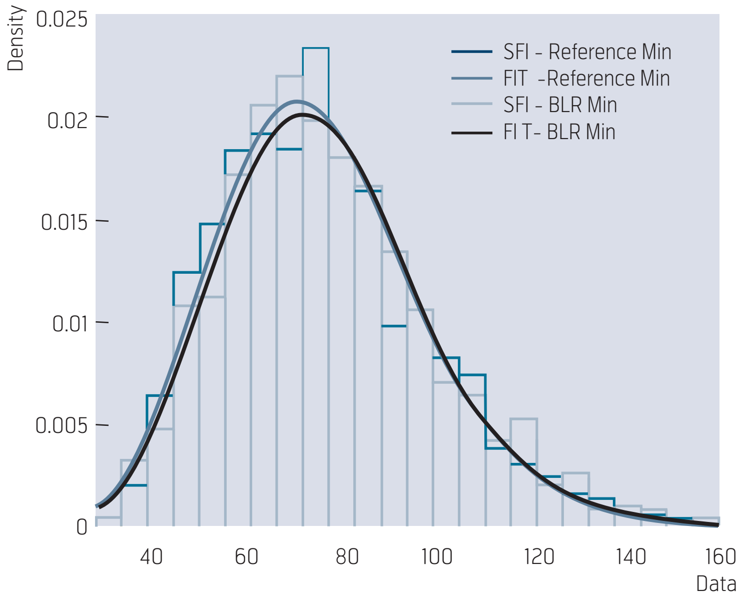


Figure 13 Effects on saturation time of removing the three bridge links (for each region pair) with lowest EVC

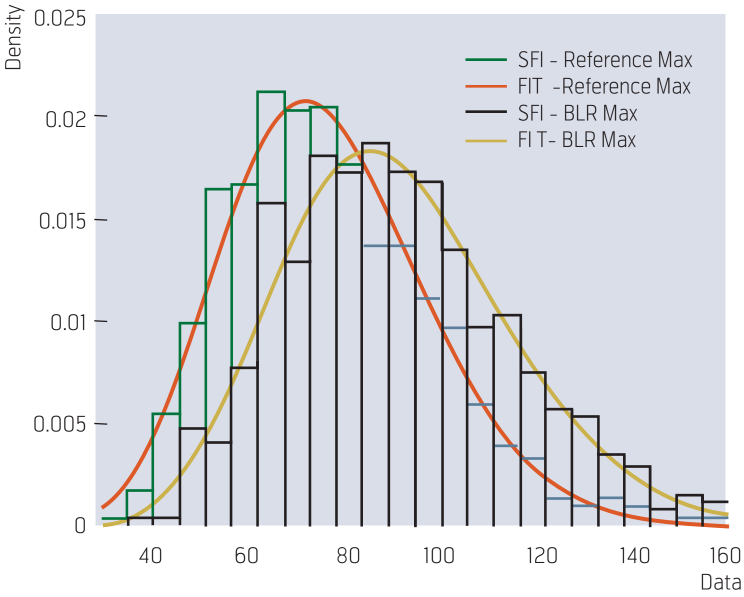


Figure 14 Effects on saturation time of removing the three bridge links (for each region pair) with highest EVC

1 We can inoculate the Centers – along with, perhaps, a small neighbourhood around them.

2 We could instead find a *ring* of nodes surrounding each Center (at a radius of perhaps two or three hops) and inoculate the ring.

3 We can inoculate bridge links.

4 We can inoculate nodes at the ends of bridge links.

We note that ideas 1 and 2 are applicable even in the case that only a single region is present. Ideas 3 and 4 may be used when multiple regions are found. Note that inoculating a bridge link (idea 3) is not the same as inoculating the two nodes which the link joins (idea 4): inoculating a node effectively removes that node and *all* links connected to it, while inoculating a link removes only that link.

Also, with link inoculation, one has the same considerations as with link addition – namely, the *height* of the link matters. We define the "link EVC" to be the arithmetic mean of the EVC values of the nodes on the ends of the link. Ideas 3 and 4 are then almost certainly most effective if the bridging links chosen for inoculation have a relatively high link EVC.

In order to get some idea of the merits of these four ideas, we have tested two of them: ideas 1 and 3. (Ideas 1 and 2 are similar in spirit, as are ideas 3 and 4.) First we give results for idea 3. We again start with the sfi graph, and measure effects on saturation time, averaged over 1000 runs, with a random start node for each run. We have tested two strategies for bridge link removal: (i) remove the *k* bridge links between each region pair that have the *lowest* link EVC; and (ii) remove the *k* bridge links between each region pair that have the *highest* link EVC. As may be seen by inspection of Figure 11, the highest and middle regions have 6 bridge links joining them, while the middle and lowest regions have ten (and the highest and lowest regions have no bridging links between them). Hence we have tested idea 3, in these two versions, for *k* = 1 and for *k* = 3 (which is possible without disconnecting the graph). Our confidence that strategy (ii) – remove the *k highest* bridge links – will have much more effect is borne out by our results. For example, in Figures 13 and 14, we compare results for *k* = 3. We see from Figure 13 that the effect of removing the lowest three bridge links is

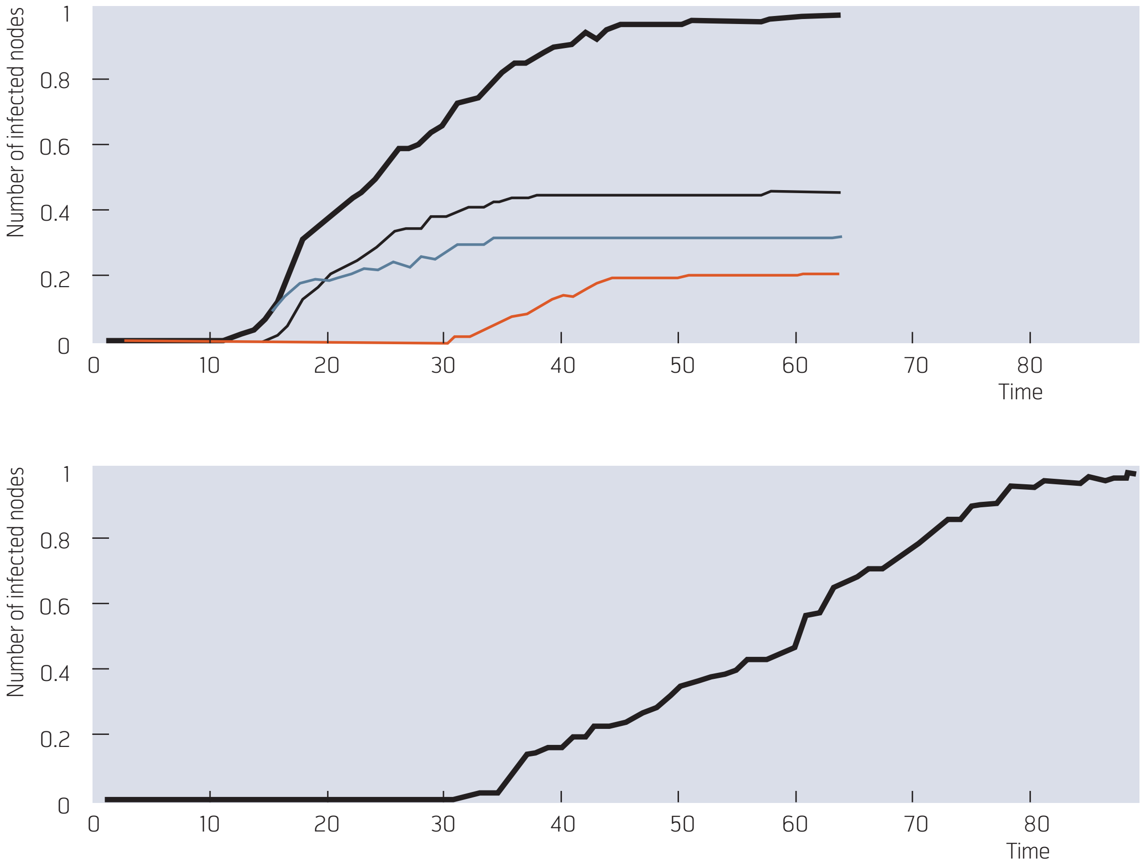


Figure 15 Infection curves for the sfi graph. (top) Six randomly chosen links have been removed. (bottom) The three highest links between the black and blue regions, and also those between the blue and red regions, have been removed. The result is a graph with a single region

negligible. Figure 14, in contrast, shows a significant retardation of saturation time as a result of removing the top three bridge links from between each pair of regions.

We have also, for comparison purposes, removed the same number of links, but at random. For example, for the case $k = 1$, we remove one bridge link between each connected pair of regions, hence two total; so for comparison we do the same, but choosing the two links randomly (with a new choice of removed links for each of the 1000 runs).

	$k = 1$	$k = 3$
Reference	82.9	83.3
Remove random	84.3	87.1
Remove lowest	84.4	85.8
Remove highest	87.7	96.5

Table 2 Mean saturation times for two sets of bridge-removal experiments on the sfi graph

Table 2 summarizes the results of our bridge-link removal experiments, both for $k = 1$ and for $k = 3$. In each case we see that removing the highest bridges has a significantly larger retarding effect than removing the lowest – and that the latter effect is both extremely small, and not significantly different from removing random links.

We have examined growth curves for some of the bridge-link removal experiments. These curves reveal a surprising result: the number of regions actually *decreases* (for $k = 3$) when we deliberately remove bridge links. Specifically, removing the lowest three links between region pairs changes three regions to two; and removing the highest three links changes three regions to one.

Figure 15 shows fairly representative infection growth curves for the $k = 3$ case, contrasting random link removal (top) with removal of bridge links with highest EVC (bottom). The upper part of Figure 15 offers few surprises. In the bottom part, we see a single region, which nevertheless reaches saturation more slowly than the three-region graph of the upper plot. It is also clear, from the rather irregular form of

ISSN 0085-7130 © Telenor ASA 2005

the growth curve in the bottom part of Figure 15, that the single region is not internally well connected.

Thus, our attempt to isolate the three regions even more thoroughly, by removing the “best” bridge links between them, has seemingly backfired! Some reflection offers an explanation. It is clear from examining Figure 11 that removing the three highest links between the highest and middle regions will strongly affect the centrality (EVC) of the nodes near those links. In particular, the Center (round yellow node in Figure 11) of the middle region (giving blue growth curves in Figures 8 – 10) will have its EVC significantly lowered by removing the strongest links from its region to the highest region. If its EVC is significantly weakened, it (plausibly) ceases to be a local maximum of the EVC, so that its region is “swallowed” by the larger one. A similar effect can make the lowest Center also cease to be a Center. This picture is supported by the fact that the Center of the modified, one-region graph (bottom curve in Figure 15) is the same Center as that of the highest region in the original graph (round, magenta node in Figure 11).

Next we report on our tests to date of idea 1. We did not use the sfi graph for these tests, because the graph breaks into large pieces if one inoculates (removes) either of the two high-EVC Centers (see Figure 11). Hence we used the hio graph from [3]. This graph also has three regions, and hence three Centers which can be inoculated. We find that, as with the sfi graph, removing these Centers also isolates some nodes – but (unlike with the sfi graph) only a few. Hence we simply adjusted our spreading simulation so as to ignore these few isolated nodes (13 in number, in the case of inoculating all three Centers).

As noted earlier, we regard the hio graph as being better connected than the sfi graph (even though both resolve to three regions). Hence we might expect the hio graph to be more resilient in the face of attacks on its connectedness – in the form of inoculations. In fact, we only found a slight increase in average saturation time, even when we inoculated all three Centers. For this test on the hio graph, we found the mean saturation time to increase from 71.5 to 74.8 – an increase of about 5 %. We know of no valid way to compare this result quantitatively with the corresponding case for the sfi graph – because, in the latter case, so many nodes are disconnected by removing the Centers that the mean saturation time must be regarded as infinite (or meaningless).

It is also of interest to compare our two approaches to hindering spreading (inoculate Centers, or inoculate bridges) on the same graph. Our results so far show a big effect from inoculating bridges for the sfi graph, and a (relatively) small effect for inoculating Centers for the hio graph. Can we conclude from this that inoculating bridges is more effective than inoculating Centers? Or is the difference simply due to the fact that the sfi graph is more poorly connected, and hence more easily inoculated? To answer these questions, we have run a test of idea 3 on the hio graph. Specifically, we ran the $k = 3$ case, removing nine links total.

	$k = 3$
Remove random	71.2
Remove lowest	72.1
Remove highest	71.9

Table 3 Mean saturation times for two sets of bridge-removal experiments on the hio graph

The results are given in Table 3. Here we see that the $k = 3$ strategy, which worked well on the sfi graph when targeted to the highest-EVC links, has almost no effect (around 1 %) for the hio graph. This is true even when we remove the highest links, and remove nine of them (as opposed to six for the sfi case). Hence we find yet another concrete interpretation of the notion of well-connectedness: the hio graph is better connected than the sfi graph, in the sense that is it much less strongly affected by either bridge link removal or Center inoculation (which causes the sfi graph to break down).

We also see that removing the three Centers from the hio graph has a much larger effect (5 % compared to 1 %) than removing the three highest bridge links between each pair of regions. (Note however that removing a Center involves removing many more than three links.) This result is reminiscent of a well-known result for scale-free graphs [16]. Scale-free graphs have a power-law degree distribution and are known, by various criteria, to be very well connected. In [16] it was shown however that removing a small percentage of nodes – but those with the highest node degree – had a large effect on the connectivity of the graph (as measured by average path length); while removing the same fraction of nodes, but randomly chosen, had essentially no effect. That is, in such well connected networks, the best strategy is to vaccinate the most connected nodes – a conclusion that is also supported, at this general level of wording, by our results here. In our center-inoculation experiments, we remove again a very small fraction of well connected nodes, and observe a large effect. However, our node removal criterion is neither the node degree, nor even simply the EVC: instead we focus on the

EVC of the node *relative to the region it lies in*. Also, our measure of effect is not path length, but rather the saturation time – a direct measure of spreading efficiency. Hence our approach is significantly different from that in [16] (and related work such as [17]); but our general conclusion is similar.

7 Discussion and future work

In this paper we have summarized two closely related research efforts: one involving a new method for structural analysis for undirected graphs [1], and another [2] applying this analysis to the problem of epidemic spreading on a network. We have also presented some new results in this paper. That is, the results of [1] and [2] strongly suggest strategies for either helping or hindering spreading (of information, innovation, diseases, viruses, etc). We have tested some of these strategies, inspired by our topographic picture, and report the results here. These results suggest clearly that our approach can be useful for *managing or controlling spreading*, via design or modification of network topology [15]. We look forward to developing further tests, and to implementing practical applications of the ideas presented here for modifying (plus or minus!) the spreading properties of networks.

We believe that our topographic picture of the structure of a network, based on using eigenvector centrality (EVC) as a height function over the network, with mountains, peaks, slopes, and valleys, is an excellent starting point for an understanding of spreading. From this starting point, we have developed a set of qualitative arguments which yield seven specific predictions (Section 3). Our picture, in short, is that an initial infection on the side of a mountain will run 'up' the mountain, while the rate of infection of new nodes grows with height. This is a self-reinforcing process, so that infection rate takes off at some point high up on the mountain, and the whole top is saturated quickly; finally the remaining hillsides are saturated at a decreasing rate. These predictions have been tested, and convincingly confirmed, in a series of simulations that we have run on various social networks for which we have data on the topology.

To supplement these qualitative arguments and simulations, we have developed a mathematical theory, which is presented in detail in [2] and summarized in Section 5 here. In the theoretical work we address two things: the definition of the spreading power of a node, and the dynamics of simple SI spreading. In each case, exact solutions are not possible, due to the problem that double infections must not be counted. However, in each case, we have shown that ignoring the double-counting problem gives an approximation which supports our basic claim (spreading power may be approximated by EVC). In particular, we present arguments why the correction terms due to double counting are likely to be small compared to those which ignore double counting; and these latter terms support the claim that infection probability is positively correlated with EVC. These results need to be deepened and extended in future work. Also, some obvious extensions of the present studies, which should be studied both theoretically and via simulations, include the case in which nodes are infected from 'outside' the graph at some steady rate, and the case where nodes lose their infection status after some time (SIS), perhaps after a refractory period (SIR).

We note that the sfi graph offered the most extreme behavior, stemming from the weak coupling both between and within its three regions. Thus we find the sfi graph to be the least well connected, in terms of criteria derived from our structural analysis and from our observations of spreading. We have argued in [1] that the property of being poorly connected is related to poor mixing, and thus to a small eigenvalue gap (difference between the dominant and second eigenvalues of the adjacency matrix). It would be of interest to test these ideas with the set of graphs studied here. If the gap is small in poorly connected (by our definition) graphs, then many things will be relatively sensitive to small changes in topology: their EVC values (from the dominant eigenvector); their topography, as obtained by our analysis, and their spreading behavior. This too could be tested.

More generally, we wish to deepen and render more quantitative our notion of the *well-connectedness of a graph*. Clearly, our coarse starting point (number of regions) give useful information in itself; but much more remains to be done. Besides making a connection to the eigenvalue gap, one could seek to quantify the degree of inter-region connectedness. Furthermore, one should seek connections and correlations between these different measures.

We note that our method of network structure analysis suggests a method of *graph visualization*. Figure 11 is an example of this: nodes in each region are placed close together on the page, so that the regions (and the height function defining them as 'mountains') are visually clear. Figure 11 represents work in progress. We hope to be able to present improvements on, and refinements of, this visualization approach in future work. One very attractive goal is to display epidemic spreading simulations (or empirical data!) on a network – in the form of snapshots, or a movie – with the network visualized topographically (as in Figure 11). Then one can hope to see our predictions a – g, not in the form of static plots as in this paper, but in the form of time development of the infection over the topography (as displayed in 2D) of the graph. We believe such a visualization technique could be a highly practical tool in those cases where data on the network topology (and, if possible, also the infection status) are available.

Our analysis suggests strategies for improving the spreading of information. These ideas may be useful when such improvement is desirable – for example, for the spreading of useful information within an organization. We have discussed two such ideas, and tested one of them – connecting Centers when there are multiple regions. We find that adding just three new links (but in the right place) can give around a 25 % reduction in the time needed to saturate the network. Hence we are encouraged to explore these ideas further.

Our picture also suggests a number of simple ways of *hindering* spreading. We have listed four in Section 6, and tested two of them in a limited fashion. The results of these tests are promising: we have shown that we can significantly retard spreading, by inoculating a small but strategically chosen number of links against spreading, or by inoculating a few strategically central nodes which are picked out by our analysis. Hence we believe that further testing is warranted, towards the goal of applying our ideas to real-world security problems. We are confident that such real-world application will be worthwhile; in fact, we believe that the sum collection of results reported here establishes that we have understood spreading in a highly useful way, at the neighbourhood level.

Further tests of "design for the control of spreading" should look at other performance metrics than those used in the preliminary studies reported here. For example, the saturation time is prone to large fluctuations due to the slow stochastic nature of infecting the last small fraction of nodes. Hence, one should look at (e.g.) the time to 80 % saturation, or other criteria that involve less than 100 % saturation. Also, the set of design ideas that we list here should be more thoroughly tested, and augmented with further ideas. We expect the result of such work to be a compact but highly useful set of design/improvement tools for choosing a network topology that has desired spreading properties.

Acknowledgments

We thank Michelle Girvan for sharing data on the SFI collaboration graph with us, Mark Burgess and Lada Adamic for helpful discussions, and Åsmund Weltzien for extremely fruitful collaboration in applying these ideas (along with help with Figures 1 and 2). This work was partially supported by the Future & Emerging Technologies unit of the European Commission through Project BISON (IST-2001-38923), and partially supported by the EU within the 6th Framework Programme under contract 001907 (DELIS).

References

1 Canright, G, Engø-Monsen, K. Roles in Networks. *Science of Computer Programming*, 53, 195–214, 2004.

2 Canright, G, Engø-Monsen, K. Spreading on networks: a topographic view. Submitted to *European Conference on Complex Systems (ECCS05).*

3 Canright, G, Engø-Monsen, K, Weltzien, Å, Pourbayat, F. Diffusion in social networks and disruptive innovations. *IADIS e-Commerce 2004 proceedings*. Lisbon 2004.

4 Jovanovic, M A, Annexstein, F S, Berman, K A. *Scalability issues in large peer-to-peer networks – a case study of Gnutella*. University of Cincinnati, 2001. Technical Report.

5 Newman, M E J. The structure and function of complex networks. *SIAM Review*, 45, 167–256, 2003.

6 Pastor-Satorras, R, Vespignani, A. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett*, 86, 3200–3203, 2001.

7 Pastor-Satorras, R, Vespignani, A. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63, 066117, 2001.

8 Newman, M E J. Spread of epidemic disease on networks. *Phys. Rev. E*, 66, 016128, 2002.

9 Brauer, F. A model for an SI disease in an age-structured population. *Discrete and Continuous Dynamical Systems,* B2, 257–264, 2002.

10 Wang, Y, Chakrabarti, D, Wang, C, Faloutsos, C. Epidemic spreading in real networks: an eigenvalue viewpoint. P*roceedings 22nd Symposium on Reliable Distributed Systems (SRDS 2003)*, 25–34, 2003.

11 A good introduction to many of these definitions may be found in: http://www.analytictech.com/networks/centrali.htm

12 Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113–120, 1972.

13 Rogers, E M. *Diffusion of Innovations*, 3rd ed. Free Press, New York, 1983.

14 Girvan, M, Newman, M E J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99, 8271–8276, 2002.

15 For a discussion of closely related ideas, see: Burgess, M, Canright, G and Engø-Monsen, K. A graph theoretical model of computer security: from file sharing to social engineering. *International Journal of Information Security*, 3 (2), 70–85, 2004.

17 Albert, R, Jeong, H, Barabasi, A-L. Attack and error tolerance of complex networks. *Nature*, 406, 378–382, 2000.

18 Holme, P, Kim, B J, Yoon, C N, Han, S K. Attack vulnerability of complex networks. *Phys. Rev. E*, 65, 056109, 2002.

Geoffrey S. Canright is a senior researcher with Telenor R&D. His background is in statistical physics. His current interests include network analysis and graph theory, social networks, self-organizing systems, and bio-inspired algorithms.

email: geoffrey.canright@telenor.com

Kenth Engø-Monsen is a research scientist at Telenor R&D. He holds a PhD (2000) in computer science from the University of Bergen, Norway, and a master in industrial mathematics (1995) and a master in technology management (2001) from NTNU, Norway. Since joining Telenor R&D in 2000, his interests have been in network analysis and graph theory, searching, and mathematical finance and risk.

email: kenth.engo-monsen@telenor.com

ISSN 0085-7130 © Telenor ASA 2005

Vulnerabilities in wireless networks and intrusion detection

SLOBODAN PETROVIĆ

Slobodan Petrović is professor of information security at Gjøvik University College

In this paper, the most frequently exploited vulnerabilities of wireless networks that use the IEEE 802.11 standard are enumerated. Intrusion detection and prevention systems are proposed as an important line of defence in a multifaceted wireless network protection system. However, some concepts from the classical IDS theory must be redefined and there are many changes that these systems must be exposed to in order to operate correctly and efficiently in a wireless network environment. The importance of anomaly detection systems is especially stressed because of very specific attacks that are difficult to detect and respond to by misuse (signature based) detection systems.

1 Introduction

According to a recently conducted opinion poll among the top 40 US research policy makers [1], wireless technologies will be among the most important ones in the next decade. This will dramatically change computer networks, considering not only mobility, availability and quality of service but also security. It is much more difficult to ensure security in this type of network than in ordinary (wired) networks. Since most of the communications in the future will use both wired and wireless networks on the communication paths, new hardware and software technologies will be needed to ensure security. Among these, the intrusion detection and prevention (IDS/IPS) technologies will have to be exposed to the most dramatic changes.

In 1999, IEEE completed and approved the 802.11 standard [2] and this is considered the starting point of wireless LANs (WLANs) in the form that we know them today. There are several members of this standard family, of which 802.11b and 802.11g are the most frequently used. WLANs exist in either *infrastructure* mode or in *ad hoc* mode. An infrastructure WLAN consists of several clients communicating with a central device, so-called *Access Point* (AP). Ad hoc networks have multiple wireless clients communicating with each other as peers in order to share data among them without using an access point. In this paper, by wireless network we mean an IEEE 802.11 (b or g) infrastructure network.

In wireless networks, an attacker does not need physical access to communication lines. He/she can be located anywhere within the range of the wireless communication equipment, and that range cannot be precisely defined and guaranteed because of inherent properties of radio communication that are influenced by many factors. As the consequence of this imprecision, the traditional concepts of insider and outsider attacks must be redefined in wireless networks. Another traditional concept, that of host based and network based intrusion detection, must also be exposed to changes in the wireless network environment. Wireless networks are faced with specific types of attacks that are not possible in wired networks, such as creation of unauthorized ("rogue") access points (AP), so-called war driving (probe requests that have not set the values of specific fields), flooding APs with associations, MAC address spoofing, etc.

In order to defend not only the wireless network but also the wired network related to it, a combination of physical, technical and organizational measures has to be implemented. These measures usually include firewalls, vulnerability scanners, virus detection software/hardware, and intrusion detection/prevention systems (IDS/IPS). Bearing in mind the alterations of traditional concepts related to IDS as well as specific attacks against wireless networks, wireless IDS must implement new solutions capable of detecting and responding to the new threats. In this paper, the most frequently exploited vulnerabilities of wireless networks are first enumerated and then some wireless IDS solutions intended to defend such networks are presented. Special focus is given to anomaly detection, since this type of IDS, although more difficult to implement, may become a more complete solution to the problems of wireless networks protection.

2 Vulnerabilities in wireless networks

Wireless connectivity is related to specific vulnerabilities and backdoors for potential attackers that are not available in wired networks. The very physical access to a wireless network is easier because of the nature of radio communication. Unlike that, a wired network can only be accessed and attacked through a physical connection, usually via Internet (unless the attacker has obtained the credentials through social engineering or dumpster diving). The hardware and software tools for penetrating into wireless networks are known and publicly offered ("Net-Stumbler" soft-

ISSN 0085-7130 © Telenor ASA 2005

ware, for example [3] and various hardware tools for so-called "war driving" – antennas, amplifiers, etc). In [4], seven major security problems related to the 802.11 wireless network standard have been identified and some solutions have been proposed from the corporate point of view:

2.1 Easy access to 802.11 networks

The access point, the key hardware device that serves as an interface between the wired and the wireless part of a network, uses so-called Service Set Identifier (SSID) to differentiate networks from one another. This SSID is broadcast every few seconds in so-called "beacon frames" in order for authorized users to find the correct network. By default, the SSID is set to a fixed value known by everybody and this often enables easy unauthorized access to such networks. On the other hand, the authentication process in the 802.11 standard is known to contain a flaw [5]; namely, before any communication can take place between the access point and a wireless client, they must first begin a dialogue and this process is called *associating*. There is a feature in the 802.11 standard that allows networks to require authentication immediately after a device associates, before it attempts any communication through the access point. There are two possibilities with this setting: *shared key authentication* and *open authentication* (default). The shared key authentication mechanism uses a challenge string that is sent unencrypted to the prospective client, upon its request to the access point. The client then encrypts the challenge string and sends it back to the access point. The fact that the challenge string is available in the radio channel in both clear and encrypted form obviously enables the reconstruction of the running key (the output sequence from the enciphering algorithm) used for encryption. Since the encryption algorithm (RC4 [6] in most cases) is known to be possible to cryptanalyse if certain circumstances are met [7-11] (and interestingly enough they have been met in the early versions of the standard [7,8]) it is also possible (although not so easy) to reconstruct the very secret key. Thus, it is sometimes better to use the open authentication than the shared key authentication, since in the open authentication everybody can access the network, but without the possibility of reconstructing the secret encryption key.

2.2 Unauthorized ("rogue") access points

This is a problem in large organizations, in which access points may be installed without prior notification to the system administration. If no security measures are activated on such access points, the whole organization's network may be put at risk, including the wired part of the network. Even if the organization does not officially use wireless networks at all, it may happen that such a network is installed somewhere within it without knowledge of the responsible person/department. An additional problem is that discovering unauthorized access points often requires special equipment (e.g. intelligent sensors deployed throughout the area) and goniometric procedures (triangulation).

2.3 Unauthorized use of service

If an unauthorized access point is installed or an authorized access point is misconfigured, this may result in enabling an attack against the whole organization's network. The problems that may result from such a security breach are not only of a technical nature (bandwidth misuse). They can be of legal nature too. The unauthorized use of service opens the door to spamming and similar activities in the name of the attacked organization, which may result in severe legal consequences for it.

2.4 Denial-of-service vulnerability

Wireless LANs have the transmission capacity limited to 11 Mb/s for 802.11b and 54 Mb/s for 802.11g (these variants of the 802.11 standard are widely used in Europe). This capacity is shared by all the clients associated to a single access point. Obviously, an unauthorized client could start a massive data transfer and occupy the whole available bandwidth, which would result in a denial-of-service for the rest of the (authorized) clients. On the other hand, radio capacity of the access point can be overwhelmed by traffic coming from the wired network too, at a rate greater than the capacity permits. A *ping flood* launched from the wired segment of a network is an example of such an attack. Several directly connected access points can be attacked at the same time by using broadcast addresses. Traffic injection into the radio network without being attached to a wireless access point is also possible. The classical radio jamming is also an option for performing a denial-of-service attack.

2.5 MAC spoofing and session hijacking

No frame authentication was originally defined by the 802.11 standard. Every frame contains a source address, but it can be forged since no frame authentication is performed. Thus, attackers can use spoofed frames to redirect traffic and corrupt Address Resolution Protocol (ARP) tables. They can easily get the MAC addresses of the stations currently in use on the network and adopt those addresses for malicious transmissions. In addition to hijacking sessions, attacks can exploit the lack of authentication of access points. This fact is exploited by an attacker by presenting him/herself as an access point, since no mechanism in the 802.11 standard prevents it. Then, the attacker gets the credentials of the clients and uses them to gain access to the network through a man-in-the-middle attack.

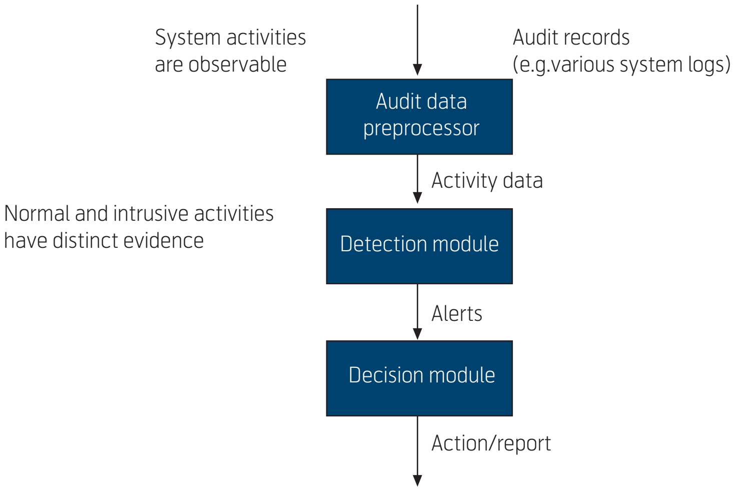

System activities are observable
Audit records (e.g.various system logs)
Audit data preprocessor
Activity data
Normal and intrusive activities have distinct evidence
Detection module
Alerts
Decision module
Action/report

Figure 1 Components of a general IDS

2.6 Relatively easy traffic analysis and eavesdropping

The 802.11 standard does not provide any protection against attacks which passively observe traffic. Frame headers are not encrypted and the security against eavesdropping was initially supposed to be provided by a weak encryption algorithm [7] (WEP – Wired Equivalent Privacy, which is an implementation of the well known RC4 algorithm [6]). Although in the latest implementations the key management protocol changes the secret key every 15 minutes [4], the flaws in the RC4 algorithm persist, such as the existence of weak keys, possibility of related keys attack [8] or the existence of distinguishers and statistical anomalies [9-11]. The 802.11 standard evolved in order to cope with the weaknesses of the WEP encryption. As a transitional solution, the WPA (Wi-Fi Protected Access) standard extension has been introduced. This encryption framework uses the same algorithm, RC4, but the length of the initialization vector has been increased from 24 to 48 bits. However, the flaws of the algorithm that do not deal with the cipher key repetition probability still persist. As a final solution to the encryption related problems within the 802.11 standard, the WPA-2 standard extension has been recently introduced [12]. The encryption algorithm has been changed to AES [13]. The problem, however, is in the requirement that hardware must be changed in order to implement the WPA-2 framework. The reason for this incompatibility is in the need for more resources in order to implement the AES cipher.

2.7 Possibility of higher level attacks

A successful attack against a wireless network can serve as a launch point for attacks on other systems. Placing a wireless LAN inside the security perimeter is therefore considered weakening the security within the perimeter.

3 Intrusion detection – the classical concept

Formally, we can define computer intrusion as a set of actions aimed at compromising the security goals (confidentiality, integrity, availability of a computing/networking resource). Then intrusion detection can be defined as the process of identifying and responding to intrusion activities. A system that performs automatically the process of intrusion detection is called Intrusion Detection System (IDS). An intrusion prevention system (IPS) combines an IDS with a firewall, a virus detection algorithm, a vulnerability assessment algorithm, etc. The ambition of such a system is to manage both preventive and responsive actions against attacks on a computer network.

Two basic assumptions related to the successful operation of an IDS/IPS are the following:

1 System activities are observable.
2 Normal and intrusive activities have distinct evidence – the goal of an IDS/IPS is to detect the difference.

The components of a general IDS are presented in Figure 1.

Intrusion detection systems are classified:

- By scope of protection (or by deployment)
  1 Host-based IDS
  2 Network-based IDS

- By detection model
  1 Misuse detection systems (pattern matching)
  2 Anomaly detection systems (pattern recognition).

Misuse detection IDS and anomaly detection IDS structures are presented in Figures 2 and 3, respectively.

Host-based intrusion detection systems use operating system's auditing mechanisms (various system logs) to monitor users' activities and execution of system programs. Network-based IDS operate by monitoring signals from sensors deployed at strategic locations in the network. They inspect network traffic and monitor users' activities.

Misuse intrusion detection systems use signature or rule based detection. Because of that, they are inca-

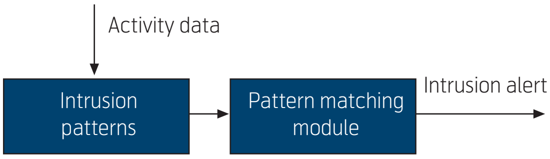


Figure 2 Misuse detection IDS structure

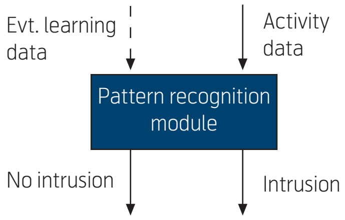


Figure 3 Anomaly detection IDS structure. It may contain a learning phase that uses labelled data

pable of recognising new attacks, which is a very serious drawback. However, all the commercial and publicly available ready-made intrusion detection systems are by now misuse detection based. There are technical and economic reasons for that. The technical reasons are easier design and good behaviour of such systems considering the probability of false alarms. The main economic reason for such a wide offer of misuse detection systems is in the dependence of clients on regular updating of attack signatures database, which includes an extra charge and increases the manufacturer's revenue.

Anomaly detection systems use pattern recognition techniques as their operation mechanisms. They look for deviations from "normal behaviour". Although they are capable of recognising novel attacks, the principal problem of their operation is the fact that these systems cannot reduce the number of false alarms easily. There are many other problems that prevent commercialization and a wider use of these systems. For example, it is very difficult to determine a sharp limit between "normal" and "abnormal" behaviour in a computer network. It is also difficult to ensure a real time operation of these systems since sufficiently precise pattern recognition requires application of complex algorithms. Anomaly detection is carried out by application of the results of various scientific methods, of which the most important ones are statistical methods (cluster analysis), artificial intelligence methods, cognitive science methods, data mining and mathematical abstractions of biological systems (neural nets, immunological system simulation, process homeostasis, etc.).

4 Intrusion detection in a wireless network environment

Unlike the ordinary wired networks, in the wireless network environment defined by the 802.11 standard it is possible to detect some intrusions even at the physical level. For example, unauthorized access points can be detected by carefully deploying radio sensors throughout the protected area and by using goniometric algorithms in order to locate unknown sources of radio transmission. This physical defence line of the wireless network has to be combined with the lines of defence at the network level in order to be able to detect other kinds of attacks.

Because of the nature of radio propagation, the exact border between the internal and the external network is not known. As a consequence, exact classification of attackers into insiders and outsiders is impossible. Classification of attacks into insider and outsider attacks is not possible either. Thus the security policies that use host based IDS to protect against the insider attacks and network based IDS to protect against the outsider attacks make no sense in the wireless environment. Intrusions are to be detected not only within the wireless network protected area, but also outside of it, bearing in mind the possibility of attacks against other parts of the network from a wireless network as well as pure interference with other wireless networks.

Wireless intrusion detection systems can be divided into misuse based and anomaly based systems in the same way as the IDS for wired networks. Beside classical misuse and anomalies detectable in any network, wireless IDS must also detect wireless specific misuse and anomalies. In wireless misuse detection systems, the main problem is the problem of distribution of the elements of the IDS. Three approaches are possible:

1 *Wireless IDS sensors and processors are integrated into the access points*
The advantage of this approach over other approaches is in total network coverage at relatively low cost and in easier network management. No separate hardware is needed for the IDS, which lowers the total cost of ownership. The disadvantage of this approach is obviously in the impossibility of integration into existing networks for hardware incompatibility.

2 *Overlay IDS with centralized processing*
This approach uses dedicated radio frequency sensors deployed throughout the wireless network to be protected. The sensors transfer the data to the dedicated IDS server that performs all the processing and manages eventual responsive actions. The advantage of this approach is in the possibility of integration into the existing network framework. The IDS of this type can achieve the total network coverage but at a higher cost than the integrated solutions. Centralized processing of all the data from the sensors requires an expensive hardware for the IDS server in order for the overall IDS to be efficient enough.

3 *Overlay IDS with decentralized processing*
The IDS structure used in this approach delegates

ISSN 0085-7130 © Telenor ASA 2005

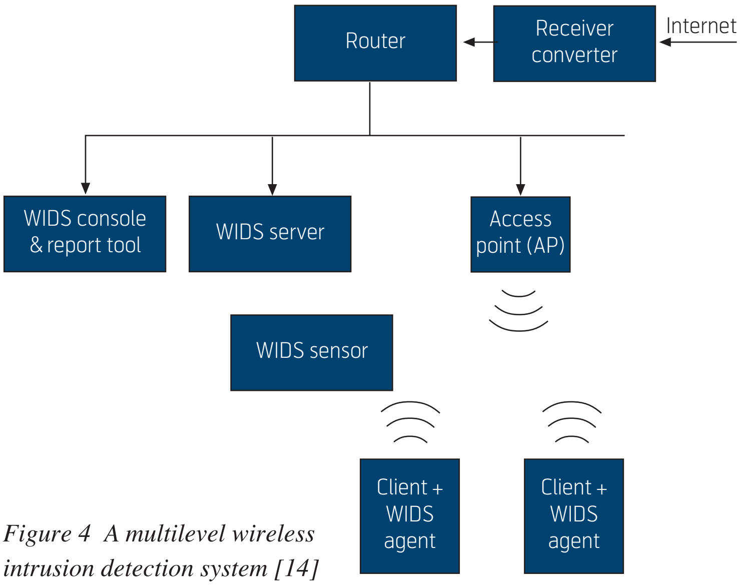


Figure 4 A multilevel wireless intrusion detection system [14]

some data processing to the very sensors reducing in such a way the processing power needed at the IDS server. The advantage of this multilevel approach over the centralized processing approach is in the cost of hardware needed for the IDS server.

Misuse based wireless IDS are incapable of detecting new attacks. Besides, the signatures of the attacks have to be updated very often, since new attacks are detected every day and many of the attacks remain undetected for quite a long time. So anomaly based systems could also be a solution to this problem. However, the problems of false alarms and real time operation persist in this case too.

In [14], a multilevel wireless IDS/IPS is proposed that uses agents on hosts, sensors, an IDS server and a reporting tool in order to combine host based and network based detection in a wireless network environment. The IDS cooperates with the firewall, the antivirus program, and other security tools in order to coordinate activities with them. The goals of the overall system are the following:

1 To make an efficient system to defend the wireless network;

2 To define attack and intrusion "axioms scope" (misuse detection);

3 To define conclusions mechanisms ("theorems");

4 To learn in order to anticipate (anomaly detection) – there is a trade-off between the level of intelligence of the system and its efficiency;

5 To recognise the wireless specific attacks;

6 To launch responsive actions in order to defend the system and/or the network.

The simplified scheme of the system is presented in Figure 4.

Neural networks and fuzzy logic have been combined in this system as the means to achieve self learning and recognition of previously unknown attacks (anomalies), especially the wireless specific ones. The responsive actions are launched at both the local and the global level (multilevel detection and response), which improves the efficiency of the system as the whole.

In [15], another multilevel approach to intrusion detection is combined with *situation assessment*, a classification procedure that maps a label to the current state of a system based on data received from multiple sources. This is a general approach, applicable to many processes, such as prognosis and handling of emergencies, monitoring, securing and recovering of critical systems like nuclear power plants and electrical power grids, prediction of terrorist intents, command and control, etc. Classification of input data in an IDS of this type is performed at two levels: source-based classifiers label security status of users' activity either as Normal or Alert. These decisions are forwarded to the upper level of decision as asynchronous data streams. The upper level classifier combines the decisions of the lower level ones and produces the final decision on the situation status.

The procedure used in [15] may be especially applicable in wireless IDS, because these systems need detection of events at two completely different levels – physical level and network level. Thus completely different sensors and processing are needed for detection of these two types of events. At the same time, it is quite probable that the events of different types are asynchronous. For example, an unauthorized access point may be installed by an insider without any malicious intent and this is detectable at the physical level from the very moment of installation. But the attack may arrive later, after the vulnerability has been detected by an attacker and only then the attack may be detected at the network level.

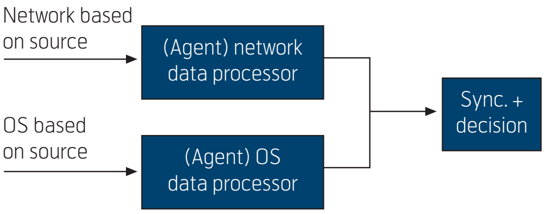


Figure 5 A multilevel intrusion detection system that applies situation assessment [15]

5 Conclusion

In this paper, the most frequently exploited vulnerabilities of wireless networks that use the IEEE 802.11 standard have been enumerated. In spite of improving the standard by strengthening security measures, there are still many problems that require careful deployment and continuous security monitoring in these networks. Wireless intrusion detection/prevention systems are still a necessary tool for this monitoring and management of countermeasures against the attacks. However, in order to cope with the wireless network specific attacks, they must be capable of detecting intrusions at the physical level too, not only at the network level. The problem of integration of host-based and network-based intrusion detection also has to be solved for this type of systems, as well as the unified treatment of insider and outsider attacks. Since it is very difficult to design a misuse-only based intrusion detection system for wireless networks, anomaly detection may play a much more significant role in this environment than in the ordinary wired network environment. This article presented two typical intrusion detection systems that use anomaly detection methods extensively.

References

1 Applewhite, A. The View from the Top. *IEEE Spectrum*, 41 (11), 2004, 16–31.

2 IEEE Standards Association. *IEEE 802.11 Standard*. http://standards.ieee.org/getieee802/download/802.11a-1999.pdf

3 *NetStumbler*. http://www.netstumbler.com

4 AirMagnet. *The Top Seven Security Problems of 802.11 Wireless*. AirMagnet Technical White Paper. http://www.airmagnet.com/products/wp-index.htm

5 ArsTechnica. *Wireless Security Blackpaper*. http://arstechnica.com/articles/ paedia/security.ars

6 Rivest, R L. *The RC4 Encryption Algorithm*. RSA Data Security, Inc., 1992 (proprietary).

7 Borisov, N, Goldberg, I, Wagner, D. Intercepting Mobile Communications: The Insecurity of 802.11. *Proceedings of the 7th Annual International Conference on Mobile Computing and Networking*, July 2001, 180–189.

8 Fluhrer, S, Mantin, I, Shamir, A. Weaknesses in the Key Scheduling Algorithm of RC4. *Proceedings of SAC 2001*, LNCS 2259, Springer Verlag, 2001, 1–24.

9 Golić, J. Linear Statistical Weakness of the Alleged RC4 Keystream Generator. *Proceedings of EUROCRYPT 97*, LNCS 1233, Springer Verlag, 1997, 226–238.

10 Knudsen, L, Meier, W, Preneel, B, Rijmen, V, Verdoolaege, S. Analysis Methods for (Alleged) RC4. *Proceedings of ASIACRYPT 98*, LNCS 1514, Springer Verlag, 1998, 327–341.

11 Fluhrer, S, McGrew, D. Statistical Analysis of the Alleged RC4 Keystream Generator. *Proceedings of the FSE 2000*, LNCS 1978, Springer Verlag, 2000, 19–30.

12 IEEE Standards Association. *IEEE 802.11i Standard*. http://standards.ieee.org/getieee802/download/802.11i-2004.pdf

13 *The Advanced Encryption Standard (AES)*. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

14 Pleskonjić, D. Wireless Intrusion Detection Systems. *Proceedings of the 19th Annual Computer Security Applications Conference*, Las Vegas, USA, December 8–12, 2003.

15 Gorodetsky, V, Karsaev, O, Samoilov, V. On-Line Update of Situation Assessment Based on Asynchronous Data Streams. *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2004)*, LNAI vol. 3213, Springer Verlag, 2004, 1136–1142.

Slobodan Petrović is professor of Information Security at Gjøvik University College. He received his PhD degree in 1994 from the University of Belgrade, Serbia and Montenegro. His research interests include cryptography, intrusion detection systems, coding theory, pattern recognition, and combinatorial optimisation. From 1986 to 2000 he participated in various projects at the Institute of Mathematics in Belgrade concerning fundamentals of computer science and pattern recognition. From 2000 to 2004 he was at the Institute of Applied Physics (CSIC), Madrid, Spain, working on the projects 'Cryptographic Protection of Copyright in Digital Networks' and 'Application of Intelligent Mobile Agents in Intrusion Detection Systems'.

email: slobodanp@hig.no

Digital forensics research

SVEIN YNGVAR WILLASSEN AND STIG FRODE MJØLSNES

Svein Yngvar Willassen is PhD student at the Norwegian University of Technology and Science, Trondheim, Norway

Stig Frode Mjølsnes is Professor at the Norwegian University of Technology and Science, Trondheim, Norway

Digital Forensics is the field of analysing and evaluating digital data as evidence. Time stamps stored on digital media play a crucial role in evidence analysis, but digital time stamps may not be correct for various reasons. A more scientific understanding of digital time stamps in digital forensics is therefore needed.
In this paper we present the emerging field of digital forensics, and take the first steps toward a methodology of digital time stamps in an evidential context.

1 Digital forensics

Digital forensics can be defined as the practice of scientifically derived and proven technical methods and tools toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of *after-the-fact* digital information derived from digital sources for the purpose of facilitating or furthering the reconstruction of events as forensic evidence.

Examples of digital sources include VLSI chips, hard disks, mobile phones, digital cameras, computers, printers, copiers, backup tape, CDs, DVDs and network routers, as well as software and communication protocols. Digital forensics must be based on the science of ICT within the requirements and interpretation of law [1]. Data can be recovered even if deleted from a user's point of view. Techniques for recovery of deleted information are therefore central to digital forensics [9]. Digitally stored information can easily be manipulated, so great care has to be taken when handling digital evidence, in order to be able to prove the origin of the information.

The practice of digital forensics is new. When computers became common in homes and businesses, the police more and more often came across computers which contained forensic evidence. Thus, police organizations saw the need of establishing special police units to handle electronic evidence. USA was first when the FBI established the Computer Analysis and Response Team (CART) in 1984. Later, a similar unit was established at Scotland Yard in UK.

The need for standardization and dissemination of knowledge in this area was first recognized during the 1990s. Meetings between police units in different countries were arranged in 1991, 1993 and 1995. As a result, several international organisations for standardisation of computer forensic techniques were established [5]. Interpol European Working Party on Information Technology Crime [3], a workgroup within Interpol, was established in 1993. The International Organisation on Computer Evidence was established in 1995. This organisation is a meeting place for computer forensic units in law enforcement all over the world, and work with standardization of digital evidence analysis.

In Norway, a process was started in 1993 that led to the establishment of a "computer crime team" at ØKOKRIM in 1995 [4]. This team had from the start a responsibility for computer forensics and investigation of computer crime cases. The computer crime team became the National Computer Crime Center (PDS) in 2002, taking on the national responsibility for digital forensics at the Norwegian Police. The national Police Academy has since 1996 given educational courses in digital forensics in cooperation with ØKOKRIM. Therefore, several local police districts have a certain capability for performing digital forensics. Still, many cases are solved with assistance from the National Computer Crime Center.

At the commercial side, the company Ibas AS has performed computer forensic analysis on behalf of businesses and governmental institutions since the late 1990s. From 2001, digital forensics was organized as a business area within Ibas. Ibas now exports this service from Norway to most countries in Europe.

The scientific ICT community has taken an interest in the field of digital forensics only recently. As a result, digital forensics has been and still is ad-hoc practice-driven procedures which may be lacking scientific foundation. In response to very apparent public needs, direct requests from the government authorities, and realizing that incident investigation is a vital part of efficient protection against computer crime, the ICT security community has taken interest in providing a scientific approach to digital forensics. In 1998 the Scientific Work Group on Digital Evidence (SWGDE) was established in USA. The research communities established the *International Journal on Digital Evidence* in 2002.

ISSN 0085-7130 © Telenor ASA 2005

Digital forensics practice is currently performed in three stages:

- *Securing of evidence*. Securing involves the process of producing exact copies of the seized digital medium, so-called imaging. The copies must be exact, and cryptographic hashing techniques are used to be able to prove that the copy contains the exact same information as the original. This phase is very important to ensure that the evidence is admissible in court [2].

- *Analysis of evidence*. Analysis involves the enumeration of evidence items in the data set. This process may be difficult as the data set is usually very large and it is unknown which pieces of information may have value as evidence.

- *Evaluation*. Evaluation is the process of assessing what implications the enumerated evidence items have in the investigation. What does the evidence tell us about the use of the computer and the actions of the user? As the goal of any investigation is to provide evidence of a chain of events, the evaluation phase is very important.

Up till now, most scientific effort has been put into the problems of the first phase. As a result, sound methods exist today for image copying and other securing of digital evidence, including scientifically evaluated software packages for forensic imaging [7, 8]. However, little effort has been spent on the analysis and evaluation phases yet. These are going to be the important areas for digital forensics research in the near future.

2 The challenge of time stamps

The Department of Telematics at NTNU has established the research project "TID – Time stamps in Digital Forensics", with funding from the Norwegian Research Council. The project is an important part of ongoing research within Norwegian digital forensics, and within NTNU Research Programme for Information Security it will be carried out in cooperation with internationally renowned research institutions within digital forensics.

2.1 Problem definition

A digital time stamp is a date/time stored or communicated by an electronic medium. The digital time stamp can take many different formats and resolutions. Computer systems store time stamps in many different ways and according to various rules. In most file systems, time stamps are stored whenever a file is created, written or accessed. Most computer systems also have logging functions which log activities on the computer with time stamps. There may also be different file formats which include time stamps. These may be operating system specific formats such as executables and configuration files. It may also be application specific formats, such as documents and spreadsheets. Time stamps are also stored on other electronic media, such as mobile phones, PDAs and flash memories. Protocol messages and packets include time stamp fields that are important to the functionality and security of networks and networked systems.

Digital time stamps are very important within digital forensics because establishing the correct sequence of events and time spans are a fundamental method to activity reconstruction during case investigation and in court. Knowledge of when an action was committed is often of vital importance as evidence. With credible digital time stamps, an investigator can determine when executables were run, or when documents were written, or when emails were sent.

Based on time stamps carried by a digital medium, the investigator can construct a *timeline of activities*. The ability to prove not only what has happened, but also when it happened is crucial in the evaluation of digital evidence. A timeline is necessary to tie the chain of events that can be found on a digital medium to the chain of events that has taken place in the real world. In most cases, it is also necessary to tie the chain of events in a digital source to a specific user [10].

Time stamps exist in many formats. Although international standards for date and time representations exist [15, 17], the implementers of common file systems, applications and devices have mostly chosen not to use the standardized formats. We all remember the Y2K problem of format, representation and use.

The information security community has traditionally met the challenge of time stamp analysis by asserting the need for network base synchronisation of time sources. [15, 16] This is a valid solution if the computer at any time is in a controlled environment. In digital forensics however, it is usually the case that the digital source was not in a controlled environment. An investigator must question what previous actions have been done to the time source of a computer or digital device.

Forensic examiners who wish to use time stamps as evidence are experiencing the following deficits:

- There is little systematic documentation on time stamp formats available.

- There is a lack of documentation of the use of time stamps and time zones in systems.
- Reference systems including different calendars, leap years and leap seconds vary.
- It is unclear how time stamps in file systems are handled at different operations on different operating systems.

In addition the investigator will encounter more fundamental problems regarding the use of digital time stamps as evidence. Computer programs can manipulate most time stamps. Time stamps are related to a time source, often a clock or a network service. Clocks can be unreliable, not working, or not synchronized.

These obstacles present great challenges to computer forensic investigators. The reason is the nature of investigations digital forensics is a part of. If the timeline cannot be established beyond reasonable doubt, the evidence may be worthless since it cannot be established who was using the computer at a particular time. In several recent court cases, it has been alleged that the timeline cannot be established *beyond reasonable doubt*, since the time source may be manipulated or time stamps may be manipulated by computer programs. Such allegations may create reasonable doubt on computer usage and may lead to incorrect acquittals or convictions.

2.2 Properties of time and clocks

Starting our examination of time stamps in an evidential context, it is important to understand some properties of time. We define here *Real Time* as an abstract concept of an ideal clock always representing the real and correct time. Such a clock does not exist, but close approximations do, and these serve as the orginal time source for most other clocks.

Real Time can be viewed as a mathematical relation on the set of points of time with two important properties. The first, *anti-symmetric*, implies that it is impossible from any given point in time to "go back" to a previous point in time. This is a fundamental property that is consistent with how most people percieve time. Although clocks can be adjusted back and forth, the underlying concept of Real Time is a constant directed current, where time travel backwards is impossible. The other important property is the property of *transitivity*. Transitivity implies that any given point B in time, that comes after another point A in time, must also come after all other points that A comes after. With the definition of Real Time as a transitive relation, we may have all time points placed somewhere on a linear time axis, and therefore have a relation to all other time points as either *happened-before* or *happened-after*. If a point in time is concurrent with another point in time, it is the same point. Events however, may be concurrent, since they may stretch over a period of time.

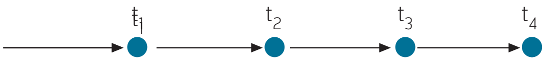


Figure 1 Real time considered as a mathematical relation with the properties of reflexivity, anti-symmetry and transitivity

A Clock is a device that can be defined to be a device that gives an approximation of Real Time. A clock can be adjusted at any time, by actions of the user or by failure. For convenience, we do not consider the case where a clock stops working. For our purposes, we define this as a situation with continuous adjustments of a clock.

In practice, most clocks are periodically synchronized with other clocks. This adjustment can be manual or automatic, such as when using time adjustment protocols. In most cases, clocks are adjusted from a time source that represents a closer approximation to Real Time than the clock that is being adjusted. In addition, the adjustment in itself also introduces an error in most cases. This is in turn true when considering the time source, and so on all the way to the root. Taking this into account, a way to view a clock is therefore to see it as the sum of approximations to Real Time, where each has been made in the path from the clock back to the root time source. [18] Provided that a clock is adjusted often enough to a valid source, and the error introduced at adjustment is

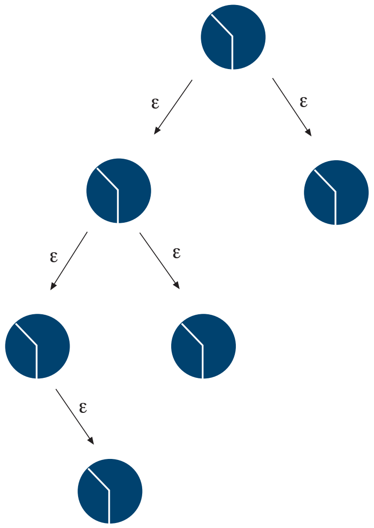


Figure 2 Clocks are adjusted from a source – Error introduced at each step

fairly small, most clocks will approximate Real Time to a level that is sufficient for everyday life.

A clock that is a few minutes off to either side will normally not be a problem in everyday life. In the Digital Forensics context, the same would be true. A few minutes off will likely not be crucial to the facts of the investigation. The problem arises where a clock shows a larger error due to lack of adjustment, failure or malicious adjustment.

2.3 Time stamps, events and causality

A time stamp is a representation of a state of a clock and that is somehow related to an event. Informally one may say that the generation of a time stamp was caused by an event. An event is something that occurred that changed the state of the world. Since changes cannot happen instantly, the definition of an event must be such that events are allowed to span over time. We define an event as a change that occurred over a time span. The start and end of an event are points in Real Time. With our definition of Real Time, an event can be said to have happened before another if the end point of the first event happen before the start point of the next event. If the time span of two events overlap, the events can be said to be concurrent.

Now, a time stamp can be said to be the representation of the local clock at some point in Real Time between the starting point and ending point of the event that caused it.

Having defined the time span of an event, we now turn to look at the change in the state of the world that an event produces. All events have necessary prerequisites. In order for an event to make a particular change in the state of the world it is necessary for the state of the world to be such that the particular event is possible. Say for instance that I pour coffee in a mug. In order for this event to occur, the state of the world must be such as to allow this event. For instance, I must have a cup and I must already have brewed coffee. This introduces the concept of causality. When an event starts, the state of the world must be such that it is allowed to happen. But the state of the world at a particular time is just the sum of all previous events plus (possibly depending on one's religious beliefs) a starting condition. Thus, an event is only allowed to happen if previous events that led the world to the necessary starting state have already happened.

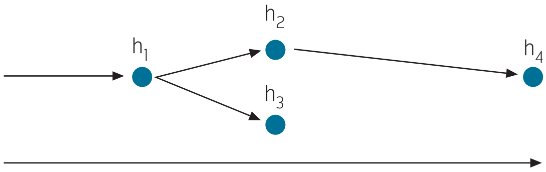


Figure 3 Causality – Some events are necessary for others to occur

It is common in a computer system that all time stamps are recorded according to the same time source, a local clock according to the definitions above that may deviate from Real Time. If such a system records time stamps for events that are causally connected, the system can be analyzed for changes in the relationship between the local clock and Real Time.

2.4 The time stamp context

The key to the solution of the time stamp challenge lies in the understanding of the context where time stamps live in computer systems, and the causal relationship between the events that caused them. Fortunately, common digital systems record a large amount of time stamps with clear patterns of causal connections. A file system is a very good example of this.

2.5 Time stamps in file systems

Most file systems contain time stamps for each file. On a normal workstation, these time stamps amount to hundreds of thousands of causally interconnected events. The most common file system time stamps are:

- *Last Written:* An event that caused the file content to change;

- *Last Read:* An event that caused the file to be read;

- *File Created:* An event that created the file at its current location;

- *Entry Modified:* An event that caused file metadata to change.

Already from these definitions, several causal connections can be postulated:

- If a file is created at a new location, its metadata must change.

- If a file is read, it must first have been created.

Taking into account the organisations of file systems, and even operating systems, more causal connections spring to mind:

- A file is created in a directory; thus the directory must already have been created. This must hold for all files and directories in a file system.

- Files cannot be created in a file system unless the file system has already been created.

- Files cannot partially overwrite other files unless those files were there already.

- If a file is created or deleted in a directory, the system must change the directory.

- If a file is read, the system must read the directory first.

- If a program is run, files that it depends on must be installed first.

If all time stamps are related to the local clock, a change in the relationship between the local clock and Real Time is bound to be reflected in the observed time stamps. This will lead to inconsistencies.

Analysing time stamps in file systems is only one example of where important time stamps can be found to determine the time stamp consistency. Many other sources exist, both in terms of actual time stamp existence, and as sources of interrelations.

2.6 The development of a framework for time stamp analysis

By analysing causal relationships between events in a computer system that produce time stamps, one can therefore detect changes in the local clock caused by failure or malicious adjustment. With the modelling of more causal interconnections, it becomes less likely that changes will be undetected. Due to the high number of time stamps and interconnections, analysis of causal relationships as defined here should be automated. It is therefore proposed to define a logical framework for the analysis of causal relationships between events and the time stamps resulting from them. The framework will be a foundation of implementations of causal analysis in software, a system that will be of great importance in digital forensic investigations.

3 Conclusion

Time stamps are vital elements in Digital Forensics, since they are the only entities that can relate evidence found on digital media to events that have taken place in the real world. Time stamps may be wrong for various reasons. We have shown that it is possible to improve the evidential quality of time stamps by correlating time stamps that occur on an evidence medium. Before such correlation can be done on a large scale, it is necessary to define a logical framework for time stamps analysis. The development of such a framework is one important goal in the TID research project funded by the NFR IKT Sos research programme.

4 References

1 Vacca, J. *Digital forensics – Computer Crime Scene Investigation*. Charles River Media, 2002.

2 Prosise, C, Mandia, K, Pepe, M. *Incident Response and Digital forensics*. McGraw-Hill, 2003.

3 Interpol. I*nterpol Computer Crime Manual, 1993–2004*.

4 Lilleng, S. *Datakriminalitet (Computer Crime)*. ØKOKRIM skriftserie, 1995.

5 Assosiation of Chief Police Officers. *Good Practice Guide for Computer Based Evidence*. June 1999.

6 *EnCase, software package*. June 16, 2005. [online] – URL: http://www.encase.com/

7 National Institute of Justice. *Test Results for Disk Imaging Tools: EnCase 3.20*. June 2003.

8 National Institute of Justice. *Test Results for Disk Imaging Tools: dd Provided with FreeBSD 4.4*. Jan 2004

9 Willassen, S. Forensics and the GSM Mobile Telephone System. *International Journal on Digital Evidence*, 2 (2), 2003.

10 Vatis, M. *Law Enforcement Tools and Technologies for Investigating Cyber Attacks*. Dartmouth College, June 2002.

11 Casey, E.. Error, Uncertainty and Loss in Digital Evidence. *International Journal on Digital Evidence*, 1 (2), 2002.

12 Hosmer, C. Proving the Integrity of Digital Evidence with Time. *International Journal on Digital Evidence*, 1 (1), 2002.

13 Weil, M. Dynamic Time & Date Stamp Analysis. *International Journal on Digital Evidence*, 1 (2), 2002.

14 Boyd, C, Forster, P. Time and Date issues in forensic computing – a case study. *Digital Investigation*, 1, Jan 2004.

15 Klyne, G, Newman, C. *Date and Timestamps of the Internet*. IETF, July 2002. RFC 3339.

16 Rousseau, L. Secure Time in a Portable Device. *Proc. of GemPlus Developer Conference*, Paris, France, June 20–21, 2001. (http://www.gemplus.fr/smart/rd/publications/pdf/Rou01heu.pdf)

17 International Standardization Organisation. *Data elements and interchange formats – Information interchange – Representation of dates and times*. 2004. ISO 8601:2004.

18 Stevens, M. Unification of relative time frames for digital forensics. *Digital Investigation*, 1, 2004.

Svein Y. Willassen has a Siv.ing. degree in Telematics from the Norwegian University of Science and Technology (NTNU). He has worked as a special investigator at the Norwegian National Computer Crime Center and as Computer Forensic Investigation Manager at Ibas AS. Willassen is currently working on a PhD within Digital Forensics in the research project "Time Stamps in Digital Forensics" at NTNU.

email: svein@willassen.no

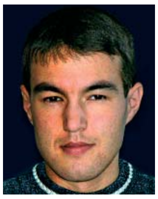
Stig Frode Mjølsnes received his Siv.ing. degree in Physical Electronics in 1980, and Dr.Ing. degree in Telecommunications in 1990, both at the Norwegian Institute of Technology, Trondheim. He has on several occasions served as security technology expert on Norwegian governmental committees in security and privacy, such as health privacy, and cryptographic policy. Just recently, he was asked to write the initial draft of a new national research programme in information security. He was called as expert witness in several legal cases, including the rather famous DVD DeCSS prosecution. He has worked on engineering analysis and design with respect to conditional access control in digital satellite television broadcast. Scientifically, he has maintained an interest in the technical approach of cryptographic protocols to help solve parts of the challenges of commercial access rights to digital content. Starting 2003, he holds a full professorship in information security at NTNU (Department of Telematics).

Stig Frode Mjølsnes is appointed committee executive manager of "NTNU Research Programme for Information Security" in the strategic focus area of ICT, and is also the manager of the research project "Time Stamps in Digital Forensics".

email: sfm@item.ntnu.no

Security models for electronic medical record

DAVRONDZHON GAFUROV, KIRSI HELKALA AND NILS KALSTAD SVENDSEN



Davrondzhon Gafurov is a PhD student at Gjøvik University College

Kirsi Helkala is a PhD student at Gjøvik University College

Nils Kalstad Svendsen is a PhD student at Gjøvik University College

In this article we give a definition of the electronic medical record (EMR) and summarize the main legal matters related to it. Further, we propose an information flow model of the current Norwegian EMR. This model is used to show that current access rights to the EMR are too general, and that requirements on accountability are not fulfilled. Finally we show how the use of dynamic access rights and non-repudiation can be used to improve the current situation.

1 Introduction

Over the last ten years our society has undergone an electronic revolution, and literally no segment of society in the industrialized countries is left untouched by the possibilities given by a multitude of computer applications and network services. The health sector is no exception, however the following characteristics differentiate this sector from the others:

1 The sensitivity of information treated in the system, emphasized by strict legal restrictions;

2 The complexity due to the wide range of systems and the large number of users;

3 The tight connection between private and public health sector.

Based on a general definition of the medical record and on the legal framework, we seek a classification of information sharing among health personnel working in the same organization, and communication between personnel from different organizations. Further, we compare the Norwegian model with the British Medical Association model (see Anderson [4]). As an extension of this model, we show how to model the EMR as a database with dynamic access rights, such that the requirements of confidentiality, integrity, availability and accountability are met. Finally we give an example of how the above framework can be used to achieve non-repudiation in message exchanges among health personnel.

2 What is an electronic medical record?

Based on the Health Personnel Act [12] and the Patients' Rights Act [14], KITH[1] defines an electronic patient record as: "An electronic collection of registered information on a patient related to health care". We find this definition and the term "electronic patient record" limiting. We consider Lærum's following discussion on the term electronic medical record (EMR) in [16] more appropriate in this context:

> *The EMR in its simplest form may be regarded as an electronic version of the paper-based medical record. It is the repository of clinical information on which health personnel base their decisions regarding health care of the individual patient. However, its content is not universally defined in the literature, and consequently, the concept is named in a multitude of ways.*

Lærum gives an overview of different definitions of EMR. For our purpose, to give general considerations of how to manage access rights to the database to ensure that requirements to confidentiality, integrity, availability and accountability are met, the general definition from [17] is appealing. Here the EMR is defined as a database containing data from various sources as shown in Figure 1. To our knowledge, a system like this is currently not implemented in any Norwegian hospital. Most hospitals use DIPS[2],

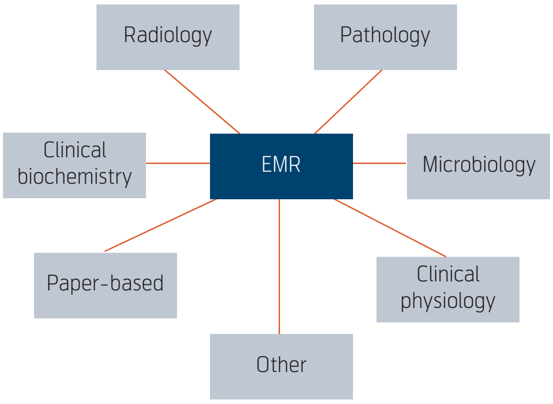


Figure 1 The EMR database

[1] *Kompetansesenter for IT i Helse- og sosialsektoren AS (Norwegian Centre for Informatics in Health and Social Care)*

[2] *www.dips.no*

ISSN 0085-7130 © Telenor ASA 2005

DocuLive EPR[3] or InfoMedix[4] to process health data. The solutions from these vendors are especially adapted to each institution to fit into the local systems. As academics, we are not constrained by these practical obligations and have the liberty to work with an idealized scenario. We propose an information theoretic model of how access to the EMR database should be administrated. But first we give a brief summary of the legal framework.

3 Legal restrictions

The following Norwegian legislation regulates the EPR: The Health Personnel Act [12], the Patients' Rights Act [14], Personal Health Data Filing System Act [13], the Journal Act [7], Personal data regulations [11] and the Archives Act [10]. The main features in these laws related to information security are:

1 Only the data controller, the data processors and people working under the instruction of the controller or the processor may be granted access to personal health data. Access may only be granted if this is necessary for the work of the person concerned and in accordance with the rules that apply regarding the duty of secrecy.

2 The health care provider is obliged to enter or record information in a patient record for the individual patient. The duty to keep patient records does not apply to co-operating personnel providing care in accordance with instructions or guidance from other health personnel.

3 Health institutions are obliged to designate one person with superior responsibility for the individual patient record including making decisions relating to what information is to be entered into the patient record.

4 It must be evident from the records who has entered the information into the patient records.

5 Corrections must be carried out through re-entering the information of the patient records, or by adding a dated correction in the records. Corrections must not be made by deleting information or comments.

6 Upon demand from the person whom the information in the patient record relates to, or of their own accord, health personnel can under certain conditions delete information or comments in the patient record.

7 Patient records may be kept electronically.

8 The data controller and the data processor must by means of planned, systematic measures ensure satisfactory data security with regard to confidentiality, integrity, quality and accessibility in connection with the processing of personal health data.

9 In legal or administrative cases relating to the professional conduct of health personnel, notes recorded in patient records, patient records and patient record material may be required for the purpose of being presented as evidence, either as originals, or as certified photocopies or printouts.

10 The patient is entitled to access to his or her medical records with enclosures and upon special request is entitled to a copy. Under certain circumstances the patient may be denied access to information in his or her medical records.

11 The patient is entitled to object to the disclosure of his or her medical records or information in the records. Furthermore, the information may not be disclosed if there is reason to believe that the patient would have objected to this if asked.

This legal framework poses strong requirements on availability, confidentiality, integrity and accountability, and states that patient approval is often necessary. Our belief is that an implementation of dynamic access rights and PKI is necessary to fulfill these requirements. We also note that the most common operations are read and write. Deletion and overwriting are relatively rare operations, and may be avoided by adding a deletion tag to an element instead of deleting it. In this report, we therefore focus on the management of read and write access rights.

4 Classification

The KITH report [2], which gives an overview of the need for PKI in the health sector, divides the PKI applications in the health sector into the following three groups:

- Communication within public administration/ health service;

- Communication between public administration/ health service and clients;

[3] *www2.siemens.no/med/*
[4] *www.infomedica.no*

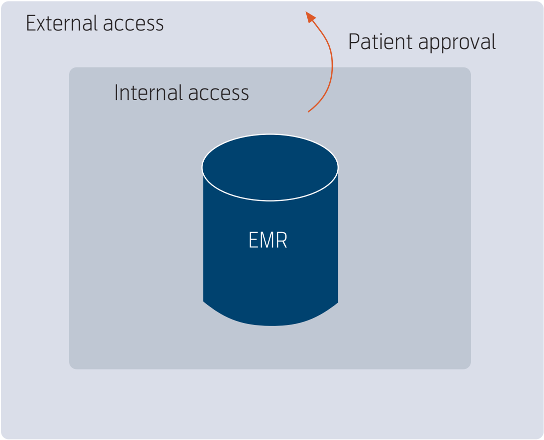


Figure 2 Simplistic view of ownership and access right to the EMR

- Communication between public administration and industry.

This focus is slightly different from ours; as we are concerned about who owns a medical record, who does not own a medical record, and how do these communicate between each other. This point of view is based on the first legal aspect of Section 3. Internally, the local administrator grants access rights, whereas access for external health personnel requires patients' agreement. Figure 2 illustrates the separation between these two cases.

4.1 Internal access policy

Internal information exchange takes place inside an organization. The organization has one manager who is responsible for all the employees. An organization might for example be the practice of a general practitioner, a hospital or united hospitals. From a legal aspect, the local data controller is the one who internally grants access to the data and ensures satisfactory information security with regard to confidentiality, availability, integrity and quality. A description of how this is done today in one of the Norwegian health regions is given in Section 5.

4.2 External access policy

External information exchange takes place when health information is transferred between working units, which have different managers. Scenarios of these situations are communication between a general practitioner and a specialist in a hospital or between a private laboratory and a general practitioner.

While the KITH report [2] focuses on all communications with a need for PKI within the health sector, we have tried to identify the communications where direct access to the EMR is necessary for at least one of the involved parties. These scenarios are summarized in Table 1. In Figure 3 we show the parties involved in this communication and we emphasize those who are in possession of an EMR. In Section 6 we treat the problem of reading and writing information to the different parts of the EMR without violating the legal framework. We especially focus on the case where laboratory results are sent from the laboratory to a doctor, and how the laboratory can be in the position to prove that the doctor was in possession of the result at a given time.

5 Models of Norwegian health service

Having classified different communication categories, this section introduces models of the information flow in Norwegian health service. The internal access model is based on the idea that there are two layers in every health institution. The outer layer consists of different departments and the sub-model

Communication	Authorization	Non-repudiation	No. of actors involved
Referral	X	X	2'
Case summary	X	X	2
Prescription	X	X	3
Journal transfer	X	X	2'
Lab answers		X	2
Sick leave		X	3
Medical certificate		X	3

Table 1 External communications where at least one of the parties needs direct access to EMR. This table is an abstract of the table given in the KITH report [2]. All communications require authentication, encryption and integrity. Non-repudiation for lab answers is not in the KITH report, but has been added to facilitate the use of the EMR as legal evidence. When the number of actors involved is marked with " ' " it means that patient approval could be included in the communication

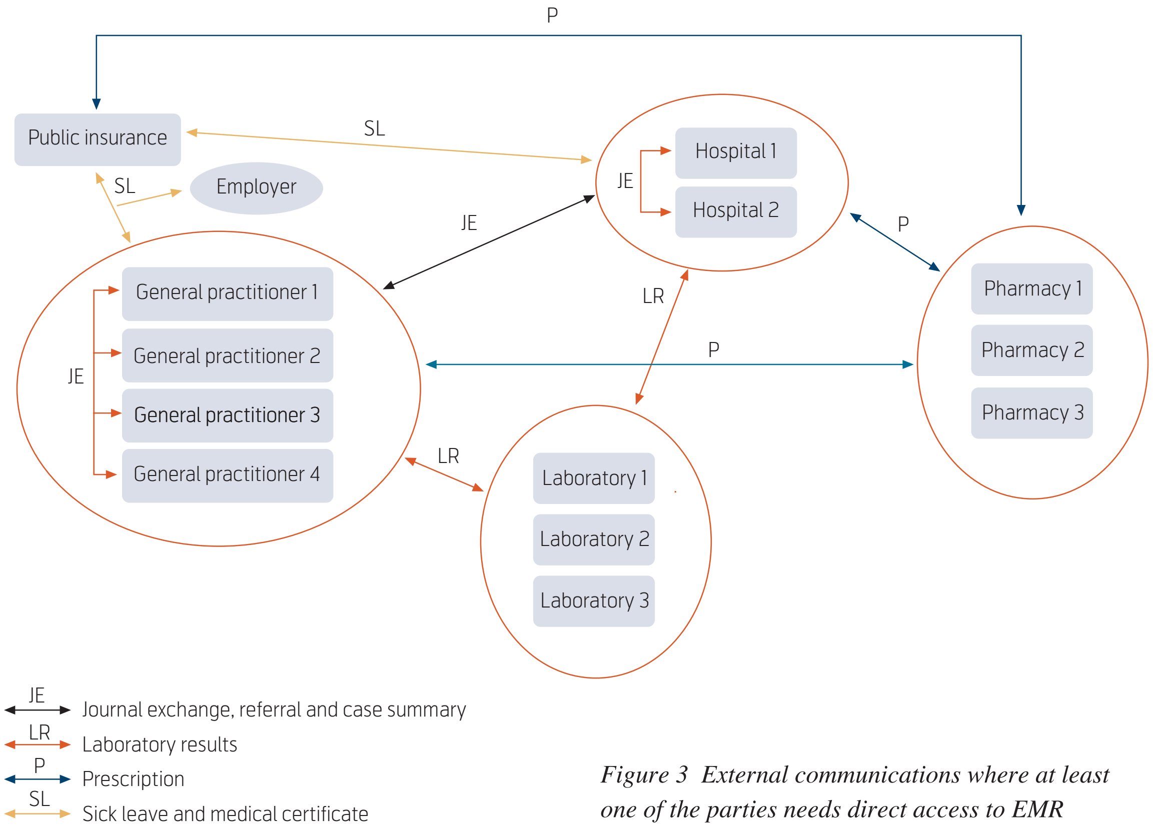


Figure 3 External communications where at least one of the parties needs direct access to EMR

explains the information flow between the departments. The inner layer consists of only one department and a sub-model is created to describe the information flow within the department. The majority external information flow is still based on the exchange of paper documents and use of fax machines.

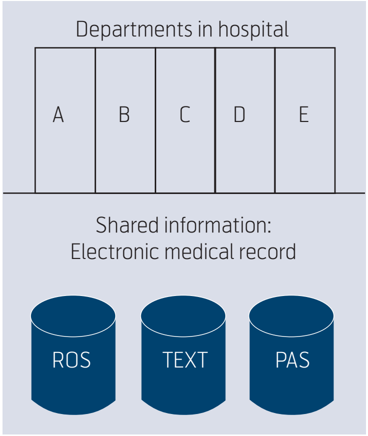


Figure 4 Multilateral policy in hospitals. ROS = Result of samples, TEXT = Text part of the patient record, often called the journal, PAS = patient administration system. Note: The EMR can contain more elements than shown in the figure

5.1 Internal access policy

A hospital can be divided organizationally into different departments as shown in Figure 4. This model can be generalized. If we were modeling a clinic with one doctor the model would contain only one department. Figure 4 also shows how different departments have access to the same EMR. This is the outer layer of the information flow, and can be modeled by a multilateral security policy.

BMA security policy. Multilateral security policy, shown in Figure 5, prevents information flowing across departments. One of the multilateral security policies is The British Medical Association security

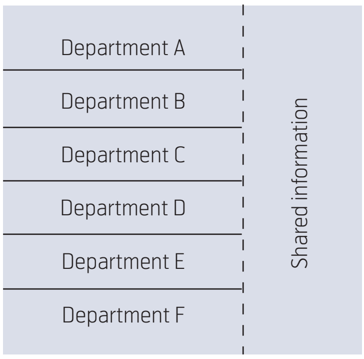


Figure 5 Multilateral security policy

ISSN 0085-7130 © Telenor ASA 2005

policy BMA proposed by Anderson [3]. It was designed specifically for the needs of the health service. BMA security policy consists of nine principal elements: access control, record opening, control, consent and notification, persistence, attribution, information flow, aggregation control, and trusted computing base.

A comparison of the information given by [3, 9, 15] shows that Norwegian electronic patient record system contains most of the principal elements of the BMA model, and can therefore be considered as similar to the BMA model.

Access control:

- In the BMA model, records are marked with an access control list naming the people or groups who may read them and append data to them.

- The Norwegian system is similar because health workers are divided into certain user groups based on their position in the hospital and these groups have different access rights to the records.

Record opening:

- In the BMA model, the patient might have multiple records with different sensitivity level. A clinician may open a new record with the patient. Where the patient is referred, the clinician may open a record with the patient and the referring clinicians.

- In the Norwegian system, records can be opened by anyone who has the right and obligation to do so. Patients are allowed to see their record in the presence of doctors.

Control:

- In the BMA model, one of the authorized clinicians may alter the access control list and add the health care professionals to it.

- In the Norwegian system, access privileges are given by the IT department or administrative department based on the recommendation given by the system owner. In the case of the internal information exchange the system owner is the manager of the united hospitals.

Consent and notification:

- In the BMA model, the responsible clinician must notify the patient whenever his patient record's access control list has been altered.

- In Norway, patients' permission will be asked when patient records are sent from one clinic to another, but inside the hospital this is not done. The reason for this is that a hospital is considered to be one clinic where all the workers already have previously mentioned rights to see the records.

Persistence:

- In the BMA model, no one should be able to delete the patient records before the expiration date.

- In the Norwegian system, normal users are not allowed to delete records and even deleting a sentence in the records is denied. Users are able to correct information by adding new information. If deleting is considered to be necessary then there is a certain protocol to follow. Request has to be made for the system owner and depending on the answer the IT department can perform the deletion.

Attribution:

- In the BMA model, all accesses to patient records must be marked on the record.

- The Norwegian system is similar because every access to the patient records will be logged (username and time, no digital signature on added information).

Information flow:

- In the BMA model, information from a less sensitive record can be added to a more sensitive record, but not the other way round.

- This is also true in Norway, but it is not relevant in our model because we consider only one medical record at a time.

Aggregation control:

- In the BMA model, patients must receive special notification if a person who has access to patient records on a large number of people, is proposed to be added to their access control list.

- There is no such requirement in the Norwegian laws.

Trusted computing base:

- In the BMA model, computer systems that handle personal health information must have a subsystem that enforces the above principals in an effective way.

- In Norwegian hospitals, there are three different previously mentioned EMR applications, which are based mostly on these principals.

Inside the hospital wards, the policy is not multilateral, but multilevel, because people working in the hospital, even in the same ward, have different clearance levels. These clearance levels, and therefore the access rights to the patient records depend on a user's position and working place in the hospital. Doctors have read and write access, nurses have read and limited write access and the secretaries in the wards have limited read and write access to the records. This indicates that the information flow inside the wards can be modeled by using the Bell-La Padula model [4].

Bell-La Padula confidentiality model. According to [4], the Bell-La Padula model belongs to the class of multilevel security policies meaning that it handles data at different sensitivity levels and prevents information flowing down in the hierarchy. Documents and users are given clearance levels. Users are able to read information written by users having the same or a lower clearance level than themselves and write to documents written by users having the same or superior clearance levels. The only one who can write down is the one who has no access to upper clearance level data. These principles are shown in Figure 6.

Some examples of the current situation inside the wards are given in Figures 7, 8 and 9. Figure 7 shows the information flow in a typical clinical ward. Nurses have limited write access to the medical record's text part (mentioned in the figure as NTEXT), where they write down medication and treatment they have given to the patient. However they are not allowed to see the whole text journal. As shown in Figure 7, doctors currently have this right. Figures 8 and 9 show information flows from radiography and laboratory departments. The full model of internal information exchange consists of two previously mentioned sub-models, Bell-La Padula and BMA.

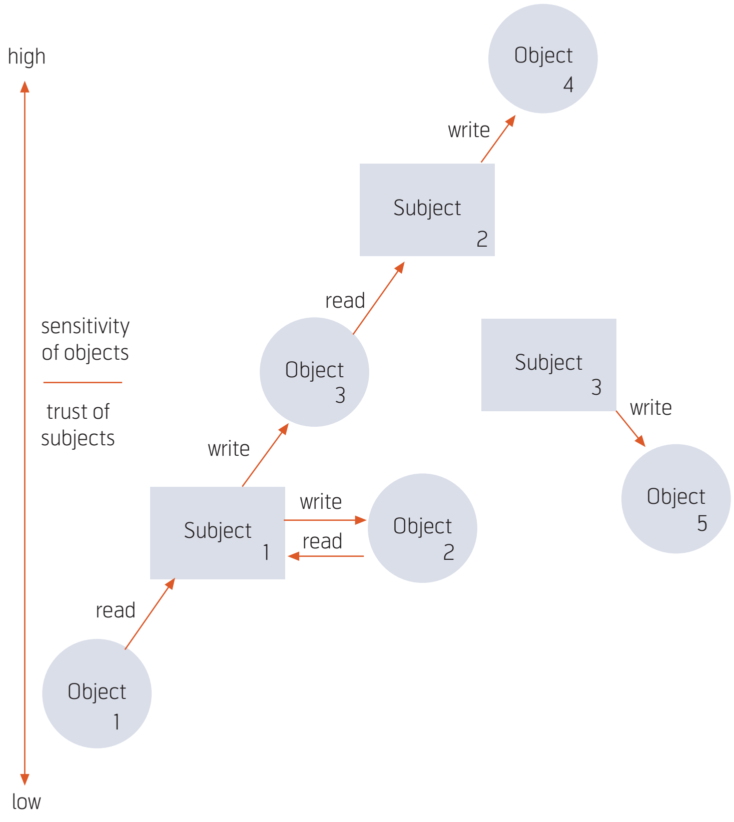


Figure 6 Bell-La Padula Confidentiality Model

5.2 External access policy

Today's situation is more or less paper based information exchange. For example, all medical prescriptions are written on a piece of paper, which is taken to the pharmacy by the patient. General practitioners have the possibility to send and receive referrals, case summaries and laboratory results by using electronic

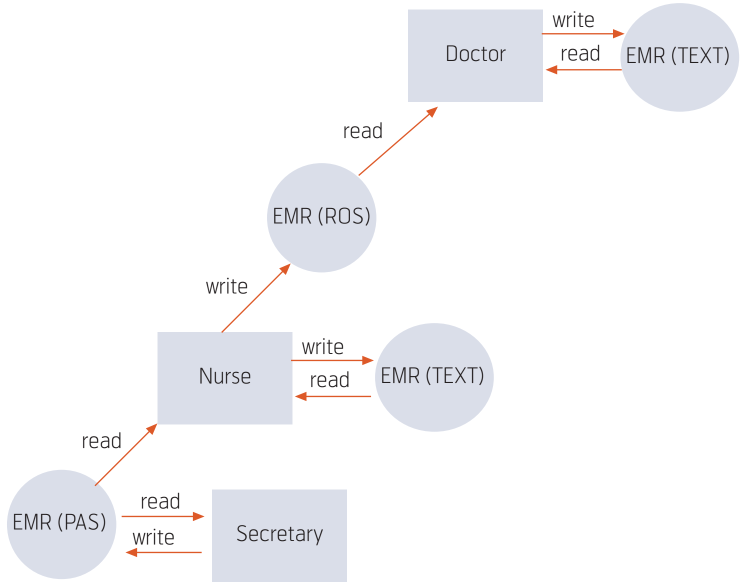


Figure 7 Information flow in the clinical ward

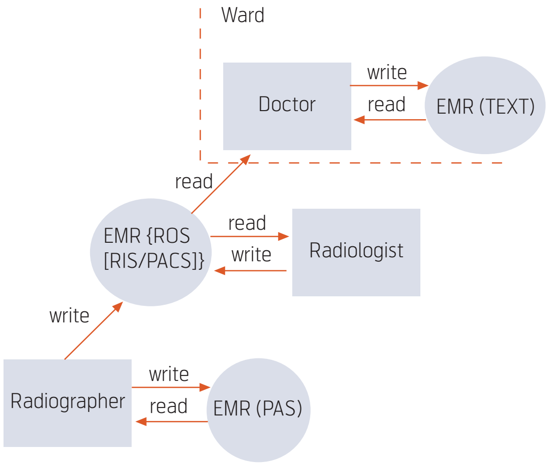


Figure 8 Information flow in the radiology department. RIS is the roentgen information system and PACS is the picture archive and communication system

ISSN 0085-7130 © Telenor ASA 2005

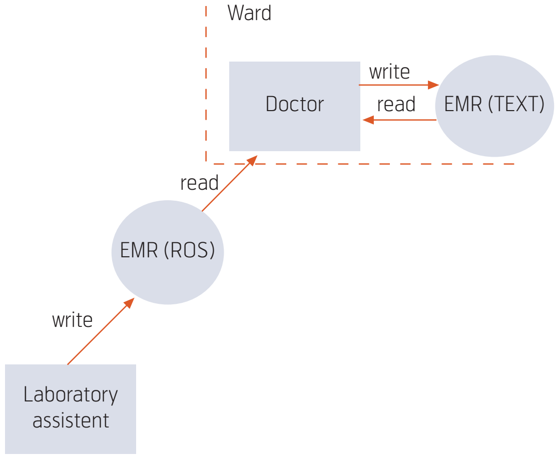


Figure 9 Laboratory department

secure envelope, which uses xml-protocol, but not all of them do so. Even if the document is in electronic form, it has to be stored in the health record by the general practitioner him- or herself. If the document is in paper form, the storing is time consuming because the practitioner has to scan it to the records. According to the Health Personnel Act (see Section 3) the private service providers or the hospitals do not have access rights to the medical records kept by the general practitioner, or vice versa. If the whole medical record needs to be transferred from one place to another, permission has to be asked from the patient and the medical record is sent as a paper document after receiving permission.

5.3 Weaknesses in the current system

Our models show that doctors working in the same department have the same access rights to patient records. This means that a doctor can access record information on a patient without being directly involved in the treatment of the patient. This is a violation of the first point in the list of legal restrictions stating that access to personal health data should only be granted if it is necessary for the work of the person concerned.

Access to the EMR is logged, but there is no use of digital signatures. This is a weakness if the EMR is to be used as evidences in a legal case as it can be claimed that the password had been stolen, the computer had been left open and so on. The use of a digital signature would make such statements less plausible.

Finally the external information exchange might be the greatest weakness of the system. The use of post or a courier to transfer paper files is an outdated method. This can be done quicker and safer by the use of a non-repudiation protocol to negotiate and carry out the exchange.

Motivated by the observations above, in Section 6 we propose models for administrating the access rights to the EMR such that:

1 Access to the EMR is more restricted than before and limited to only personnel currently working with a patient.

2 Accountability of actions is assured to a larger extent than before.

The proposed solutions are based on the existence of an internal and external PKI framework.

6 Models for the future

As pointed out, one may have the impression that compared to the legal constraints on patient data, access to EMR is too general and that the requirement of accountability may not be fulfilled. In this section we show how role-based access control (RBAC) and non-repudiation can be used to improve this situation.

6.1 Role-based internal information exchange

Motivation. The EMR is distributed among various places, e.g. laboratories and hospitals, and accessed by different users, e.g. doctors and nurses. The users of the system possess different access policies. Moreover, the nature of these access policies is dynamic. Dynamic access policy means that granting of operation on objects is done based not only on user's predefined set of functionality, but also based on some other contextual information (e.g. time or location). For example, a doctor in an operating theatre should have full access to patients' records but on other hand, from his home PC he should have limited access to patients' EMR.

A role-based access control is a promising access policy that can address dynamic policies. A core feature of the RBAC is that permissions (or rights or access rights) are assigned to roles, and users are members of the appropriate roles. Moreover, RBAC supports three well-known security principles: least privileges, separation of duties, and data abstraction [18]. Assignment of permissions to roles rather than to users seems a more natural way from the organizational point of view because roles are associated with work positions and duties of users within the organization. A study [6] shows that permissions assigned to roles tend to change relatively slowly compared to changes in the user membership of roles.

Role-based access control model. RBAC is defined in terms of four model components: Core RBAC, Hierarchical RBAC, Static Separation of Duty Relations and Dynamic Separation of Duty Relations [5]. We consider only a core RBAC with static and dynamic separation of duties. A core RBAC reference model proposed by NIST is depicted in Figure 10. The elements in the RBAC model are the set of users (U), roles (R), permissions (P) and their relationships that are user-to-role assignment (UA) and role-to-permission assignment (PA). A *role* is a job position or work title associated with a set of functions or responsibilities within the organization, for example nurse, laboratory assistant, etc. A *permission* is a predefined type of *operation* on an *object* within the system, for example permission to read an X-ray of a patient where read is an operation and *X-ray* is an object. A *session* is a mapping of a user to the activated subset of the roles assigned to the user. The functions *session_roles* and *user_session* returns the roles activated by session and the set of sessions associated with a user, respectively. The user assignment and permission assignment are many-to-many relationships. That is, one user can be assigned to several roles and one role to several users; and one role can be granted several permissions and one permission can be granted to several roles.

An RBAC with static separation of duties and dynamic separation of duties is an extension of the core RBAC model, see Figure 11. Static separation of duties (SSD) places constraints on the assignment of users to roles. Dynamic separation of duties (DSD) specifies constraints on permissions of the user to activate assigned roles. From another point of view, the difference between SSD and DSD is that SSD is set during design time while DSD is set during runtime.

RBAC model for EMR. We can show an example of RBAC for a model of internal information access being described in Section 5.1. We have the following sets:

- $U = \{User\_1, User\_2, \ldots, User\_n\}$;

- $R = \{$*Doctor, Nurse, Secretary, Radiologist, Radiographer, LaboratoryAssistant, Patient*$\}$;

- $P = \{$*ReadText*[5], *WriteText, ReadROS*[6], *WriteROS, ReadPAS*[7], *Write-PAS, ReadRIS*[8]*/PACS*[9], *WriteRIS/PACS*$\}$.

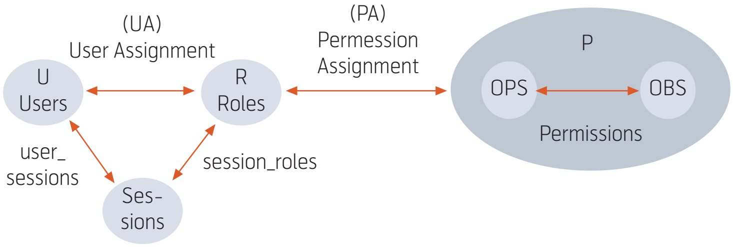


Figure 10 Core RBAC

User$_i$ can be named by using SSN[10] or some other identifier of the particular doctor, nurse and other users of the system[11]. And also, instead of dividing objects and operations on them, we have defined them together as a set of permissions P. We define user assignment as the function *user_roles*, i.e. a set of roles for the given user.

- *user_roles*(*User_1*) = {*Doctor*}
- *user_roles*(*User_2*) = {*Nurse*}
- *user_roles*(*User_3*) = {*Secretary*}
- *user_roles*(*User_4*) = {*Radiologist*}
- *user_roles*(*User_5*) = {*Radiographer*}
- *user_roles*(*User_6*) = {*LaboratoryAssistant*}
- *user_roles*(*User_7*) = {*Patient*}

We also define permission assignment as a function *role_permissions* that gives us a set of permissions for the given role:

- *role_permissions*(*Doctor*) = {*ReadText,WriteText,ReadROS,ReadRIS/PACS*}

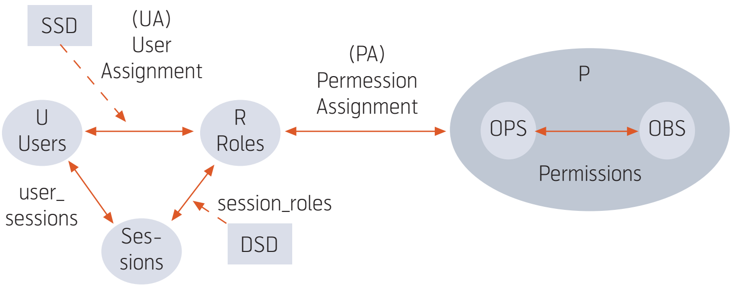


Figure 11 RBAC with static and dynamic separation of duties

[5] *Text is the text part of the patient record*
[6] *ROS is result of samples*
[7] *PAS is the patient administration system*
[8] *RIS is the roentgen information system*
[9] *PACS is the picture archive and communication system*
[10] *Social security number*
[11] *Such as the identifyer in the Norwegian health personnel register*

- *role_permissions*(*Nurse*) = {*ReadPAS, WriteROS*}

- *role_permissions*(*Secretary*) = {*ReadPAS, WritePAS*}

- *role_permissions*(*Radiologist*) = {*ReadRIS/PACS, WriteRIS/PACS*}

- *role_permissions*(*Radiographer*) = {*WriteRIS/PACS, ReadPAS, Write-PAS*}

- *role_permissions*(*LaboratoryAssitent*) = {*WriteROS*}

As mentioned earlier, rules regulating SSD can be set during design time, and constraints on user's role assignments can be specified. For example, a user cannot be assigned to both of the roles *Doctor* and *Pharmacist*. On the other hand, DSD puts constraints on activating assigned roles. For example, *Radiographer*'s permission to write to EPR (RIS/PACS) is activated when the doctor requests and sends the patient to take X-ray. Until that time a radiographer is not allowed to write RIS/PACS although he has such a right. Thus, for our purpose we define dynamic access policy as granting permission to a role based on some available runtime information. To address the dynamism we define the following abstract function, which serves as a precondition for activating roles:

can_activate(r, p, t)

This function returns true if the user with the role r can perform permission p based on some information available at time t, and false otherwise. In fact, the parameter t may encapsulate not only time, but other information like location as well. In [8], a spatial role-based access control framework applied to health care is presented. The framework utilizes *location* information in access control decisions, in order to determine the permissions a role encompasses at given location.

Accountability RBAC. In order to be compatible with governing rules and laws in our model, we assume that a local Trusted Third Party[12] (TTP) is running. The aim of TTP is to collect evidence of the users' activities by recording their operation on EMR. In its simplest form, the TTP can store data in log files.

6.2 External information exchange

Motivation. The external exchanges shown in Table 1 can be divided into two cases; communication between two main actors or communication among three main actors. Sending case summaries, referrals, records and laboratory results belongs to the model where there are only two main actors. The rest, prescription, sick leaves and medical certificates belong to the model where there are three main actors. These are shown in Figure 3. In this section we focus on how non-repudiation can be used to improve the accountability of these communications. We first give a brief introduction to non-repudiation.

Non-repudiation. Anyone slightly familiar with information security can list four of the main categories of security services defined by ISO 7498-2 [1]: Authentication, access control, confidentiality and data integrity. Fewer are aware of the fifth category, namely non-repudiation. Non-repudiation is related to authentication and data integrity, but has stronger proof requirements and is a protection against false denial of having been involved in a communication. The general goal of a non-repudiation service is to collect, maintain, make available, and validate irrefutable evidences concerning a claimed event or action and to resolve disputes about the occurrence or the non-occurrence of that event or action. Typical conflicts that can be solved by non-repudiation are the following:

- A claims having sent M to B while B denies having received M.

- B claims having received M from A while A denies having sent M.

- A claims having sent M before time T while B denies receiving it before T.

The main idea is that the parties in a communication should be in possession of a receipt when the communication is terminated. This receipt should be generated in such a way that both parties have the same opportunity to cheat, meaning that even though the computational power of one of the parties is superior to the other party, it should not have better possibilities of creating false evidence.

There are usually three parties involved in a transaction assuring non-repudiation: Originator, recipient and trusted third party (TTP). In a non-repudiation protocol, the TTP can have different degrees of involvement in the communication (inline, online or offline) and has to play different roles (certification authority, notary, delivery authority, time stamping authority and adjunctor). For more on non-repudiation and the role of TTPs see [19]. The evidences,

[12] *The role of the TTP is defined in Section 6.2.*

or receipts, can mainly be generated by two types of security mechanisms:

- Secure envelopes generated by TTPs using symmetric cryptographic techniques;

- Digital signatures generated by any party using asymmetric cryptographic techniques.

The use of secure envelopes requires an unconditional trust of the TTP, which in many cases is unrealistic. Therefore many users prefer digital signature based on evidences where the trust in the TTP as a certification authority can be reduced by appropriate registration and certification procedures.

Example. Let us now take a look at the two actors model and consider a communication between a general practitioner's (GP's) office and a laboratory with the purpose to analyze a patient's sample. In a real life situation, the main part of the communication is handled by the secretaries or assistants, as Figure 12 shows. Our model takes this into account. Our model also obeys the following privileges given to different roles: nurses and secretaries have limited access rights to the EMR, while the doctors have full reading and writing access. Limited access gives full rights to handle patient administration data but not the medical data. Results of the samples and diagnoses of the sickness are private issue between doctor and the patient.

The non-repudiation protocols between actors can be used for achieving confidentiality and integrity of the information exchange. Internal PKI can be implemented in the laboratory and at the doctor's office to achieve non-repudiation in that part of the communication too, but is not treated in this article.

Protocol. Before stating the protocol for the exchange, we introduce the necessary general notation.

- A, B and U: actors of the communication. U can be both A or B.

- M: message sent from A to B.

- K: message key defined by A.

- C = eK(M): commitment (cipher text) for message M.

- L = H(M,K): a unique label linking C and K.

- f_i (i = 1, 2, ...): flags indicating the intended purpose of a signed message.

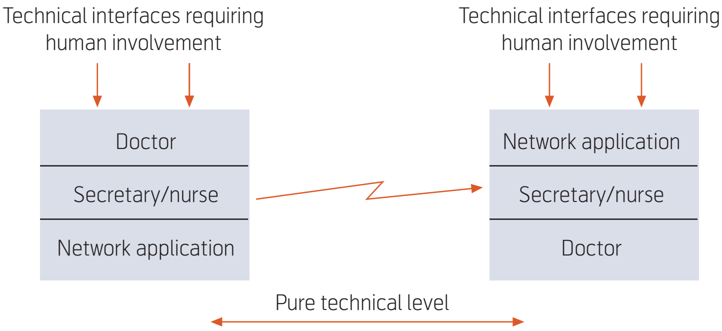


Figure 12 Three layers communication

- sub_K = sS_A (f_5, B, L, K, H(C)): authenticator of K provided by A.

- con_K = sS_{TTP} (f_6, A, B, L, K): evidence of confirmation of K issued by the TTP.

- abort = sS_{TTP} (f_8, A, B, L): evidence of abortion of a transaction issued by the TTP.

- T_{sub} is the deadline that A should either send a message key or submit it to the TTP.

The first part of the protocol handles booking of an appointment at a laboratory. The general practitioner's (GP's) secretary (GPS) sends a request to the laboratory. The receiver in the laboratory is also a secretary (LABS) who handles the first part of the communication. Based on the NR2 protocol described in [19], the communication goes as in Figure 13. The message contains the patient information and a description of what the laboratory should do. After receiving the final answer from the laboratory, the patient or the sample can be sent for testing. Using for example a courier to send the sample will ensure non-repudiation in this phase. Conflict resolution for this protocol is described in [19].

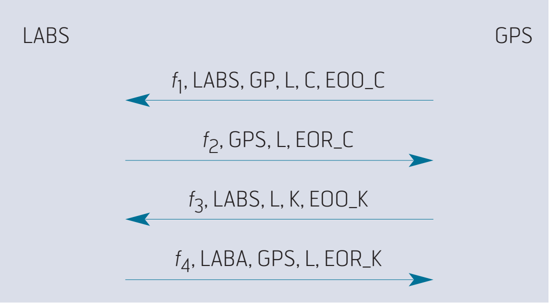


Figure 13 The NR2 protocol. Where EOO_C = sS_{GPS}(f_1, LABS, GP, L, C), EOR C = sS_{LABS}(f_2, GPS, L, C, T_{sub}), EOO_K = sS_{GPS}(f_3, LABS, L, K) and EOR_K = sS_{GPS}(f_4, LABA, GPS, L, K)

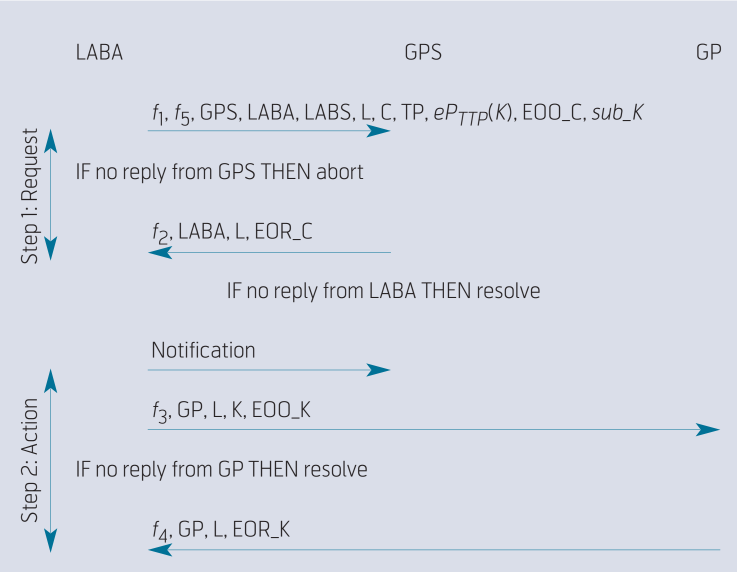

LABA
GPS
GP
Step 1: Request
f_1, f_5, GPS, LABA, LABS, L, C, TP, $eP_{TTP}(K)$, EOO_C, sub_K
IF no reply from GPS THEN abort
f_2, LABA, L, EOR_C
IF no reply from LABA THEN resolve
Notification
Step 2: Action
f_3, GP, L, K, EOO_K
IF no reply from GP THEN resolve
f_4, GP, L, EOR_K

Figure 14 The main protocol, where EOO_C = $sS_{LABA}(f_1, GP, GPS, LABA, LABS, L, C)$, *EOR_C =* $sS_{GPS}(f_2, LABA, GP, L, C, T_{sub})$, *EOO_K =* $sS_{LABA}(f_3, GP, L, K)$ *and EOR_K =* $sS_{GP}(f_4, LABA, L, K)$

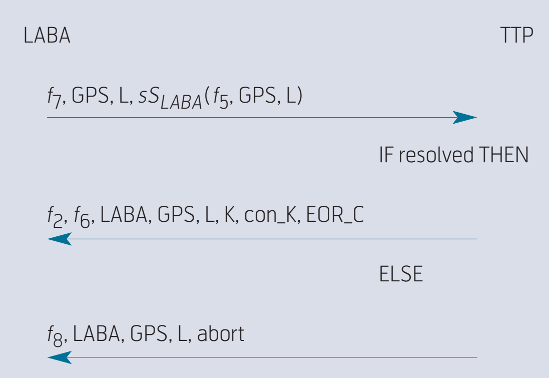

LABA
TTP
f_7, GPS, L, $sS_{LABA}(f_5$, GPS, L)
IF resolved THEN
f_2, f_6, LABA, GPS, L, K, con_K, EOR_C
ELSE
f_8, LABA, GPS, L, abort

Figure 15 The abort subprotocol

When the analysis is ready, the second part of the protocol can be initiated. In this step, there are two main actors but three parties. The purpose of this is to separate the secretary's and the doctor's duties. The laboratory assistant sends the encrypted results to the GP's secretary using the protocol described in Figure 14 (which is based on the NR3 protocol described in [19]). After Step 1, the secretary has an encrypted laboratory result, which she can save to the EMR. The secretary cannot read the results because she does not have the key. This is sent to the doctor in the next step after the secretary has sent her confirmation back to the laboratory assistant.

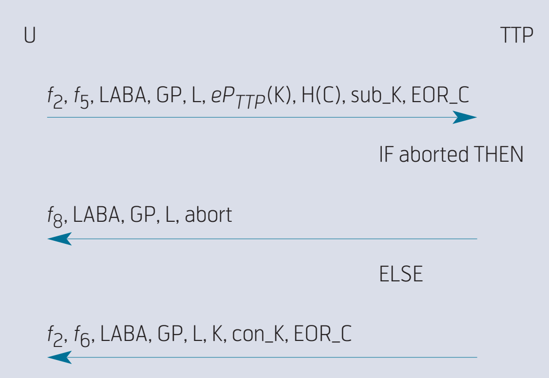

U
TTP
f_2, f_5, LABA, GP, L, eP_{TTP}(K), H(C), sub_K, EOR_C
IF aborted THEN
f_8, LABA, GP, L, abort
ELSE
f_2, f_6, LABA, GP, L, K, con_K, EOR_C

Figure 16 The resolve sub protocol

After having received the message from LABA in Step 2, the doctor has the key and can decrypt the result in the EMR. After receiving the final confirmation from the doctor, the laboratory assistant knows that the doctor has received the last result. The abort and resolve subprotocols are given in Figure 15 and Figure 16.

Problem places. In Step 1, the protocol will be aborted after the first message, if the laboratory does not receive the answer within a certain time period. Nobody has gained anything. The secretary then has four options:

1 The secretary saves the document and sends an answer: The protocol continues as normal and the doctor will receive the key eventually.

2 The secretary saves the document and does not answer: The protocol will be aborted by the laboratory. This protocol is described in Figure 15 and includes a TTP. The laboratory then has to send the first message again.

3 The secretary does not save the document but answers: The protocol will continue normally, but then the doctor has nothing to read later. In this case, the laboratory will be able to prove that the secretary has received the document. If the doctor locally does not trust his secretary to save it, a local PKI may be implemented to ensure non-repudiation.

4 The secretary does not save the document and does not answer: The protocol will be aborted and laboratory has to send the first message again.

If the GPS does not receive a notification from the LABA, the GPS can assume that the doctor has not received the decryption key and will, if time runs out, call for a resolution by using the protocol presented in Figure 16. It is therefore not in LABA's interest not to send the notification. On the other hand, the secretary can deny having received the notification even if LABA has sent it to her. This is the weakest point in our protocol. An easy, but relatively expensive, way to solve this problem is that the notification is sent through an online TTP. In this way the secretary cannot deny having received the notification.

After the message is sent from LABA to the GP, the doctor knows that the result has been sent earlier and he should be capable of retrieving the message and decrypting it in order to read the result. The doctor now has six options to proceed.

1 If the doctor finds the document from EMR, he reads it and answers: The laboratory will get the confirmation that the doctor has got the result.

2 If the doctor does not find the document from the EMR, he has to ask the secretary to save it to the EMR. He then proceeds as in case 1 and the laboratory will get the confirmation.

3 If the doctor does not find the document and the secretary cannot find it either, then the doctor knows that the secretary has made a mistake. To get the results, the communication has to be started again.

4 If the doctor does not read the document but answers, then the laboratory will get the confirmation that the doctor has the results.

5 If the doctor reads the document but does not answer, then the LABA calls for a resolution and will get evidence that it has sent the decryption key.

6 If the doctor does not read the document and does not answer, the LABA calls for a resolution and will get evidence that it has sent the decryption key.

In the case of a resolution the TTP generates a status report, gives evidences of what has happened and gives the message key by request. This protocol can be used for any information exchange between two main actors and their assistants. The protocol for the three main actors is to be defined in a future work. It can also be noted that in situations where a patient approval is necessary, the patient can be included as a party in a three actor protocol.

7 Conclusion

Based on a general definition of the electronic medical record as well as the legal framework for patient information, we have evaluated the current information flow in Norwegian health institutions. We have shown three examples of how the current methods do not meet the legal restrictions, and we have proposed a solution for two of them. Role based access control is proposed to limit the access to medical records in internal information exchange, and non-repudiation to achieve accountability for external information exchange.

Both solutions require the implementation of a public key infrastructure in the health sector. The proposed solution for external communication is an alternative to the propositions of a centralized health register. Our solution leads to effective communication among institutions, while avoiding the vulnerabilities introduced in collecting all the information in one database.

References

1 ISO. *Information processing systems – open systems interconnection – basic reference model – part 2: Security architecture*. International Organization for Standardization, 1989. ISO 7498-2.

2 Aksnes, B, Vestad, A, Henriksen, E, Skipnes, E, Kvaase, I E. *Forprosjekt for pki i helsenett, forprosjektrapport*. KITH, Sukkerhuset, 7489 Trondheim, January 2002. Technical Report R 3/02.

3 Anderson, R. A security policy model for clinical information systems. In: *IEEE symposium on Security and Privacy*, 1996.

4 Anderson, R. *Security engineering: A guide to building dependable distributed systems*. Wiley, 2001.

5 Ferraiolo, D, Sandhu, R, Gavrila, S, Chandramouli, R, Kuhn, D R. Proposed nist standard for role based access control. In: *ACM Transactions on Information and System Security*, 224–274, August 2001.

6 Ferraiolo, D F, Gilbert, D M, Lynch, N. An examination of federal and commercial access control policy needs. In: *Proceedings of the 16th NIST-NSA National Computer Security Conference*, September 1993.

7 *Forskrift om patientjournal*. Available from http://www.lovdata.no.

8 Hansen, F, Oleshchuk, V. Application of role-based access control in wireless healthcare information systems. In: *Scandinavian Conference in Health Informatics*, 30–33, June 2003.

9 Interview with J G Bronch, head of IT department in Sykehuset Innlandet 18.03.2005.

10 *Lov om arkiv (arkivloven)*. Available from http://www.lovdata.no.

ISSN 0085-7130 © Telenor ASA 2005

11 *Lov om behandling av personopplysninger (personopplysningsloven)*. Available from http://www.lovdata.no.

12 *Lov om helsepersonell m.v. (helsepersonelloven)*. Available from http://www.lovdata.no.

13 *Lov om helseregistre og behandling av helseopplysninger (helseregisterloven)*. Available from http://www.lovdata.no.

14 *Lov om pasientrettigheter (pasientrettighetsloven)*. Available from http://www.lovdata.no.

15 Lærum, H. *What is an electronic medical record and how should it be evaluated?* Trondheim, NTNU.

16 Lærum, H. *Evaluation of electronic medical record – A clinical task perspective*. Norwegian University of Science and Technology, 2004. PhD thesis.

17 Nystadnes, T. *What is an electronic medical record and how should it be evaluated*. Dissertation lecture available at http://kvalis.ntnu.no/PublicDocs/2004-03-18-laerum-dissertation-lecture.ppt.

18 Sandhu, R S, Coyne, E J, Feinstein, H L, Youman, C E. Role-based access control models. In: *IEEE Computer*, 29 (2), 1996.

19 Zhou, J. *Non-repudiation in Electronic Commerce*. Computer Security Series. Artech House, first edition, 2001.

Davrondzhon Gafurov received his MSc in Computer Engineering from the Technological University of Tajikistan (TUT), Khujand, Tajikistan in 2000. From 2000 to 2004 he was Research Assistant at the Department of Programming and Information Technology at TUT working on NLP project. He is currently pursuing the PhD degree in information security at Gjøvik University College, Gjøvik, Norway. His current research interests focus on information security.

email: davrondzhon.gafurov@hig.no

Kirsi Helkala received her MSc in Mathematics from the University of Joensuu, Finland, in 2001. After graduating she worked as assistant in the University of Joensuu and then as a mathematics, physics, and chemistry teacher in Kesämäenrinteen Koulu, Lappeenranta, Finland. She is currently a PhD student at Gjøvik University College. Her work interest is personnel authentication and the title of her PhD project is "Authentication in a health service context".

email: kirsi.helkala@hig.no

For a presentation of Nils Kalstad Svendsen, please turn to page 64.

Special

Enhanced UMTS services and applications characterisation

JAIME FERREIRA AND FERNANDO J. VELEZ

Jaime Ferreira is a Researcher at Portugal Telecom Inovação

Fernando J. Velez is an Assistant Professor at the University of Beira Interior, Portugal

Different points of view are compared for the classification of Enhanced UMTS (E-UMTS) applications, namely ITU-T I.211, UMTS Forum and 3GPP ones. A set of characterisation parameters (e.g. data rate, tolerance to delay and error, and session duration) is described, and a first assignment of parameter values for these services is proposed. An overview of E-UMTS deployment scenarios and supported services is then presented, based on the views of nowadays-relevant players. The influence of mobility in E-UMTS is discussed, and mobility models and scenarios are given. Deployment and mobility scenarios include expected population density, mobility characteristics, and usage of service mix, for each environment. A number of nearly thirty applications are considered. However, a reduced set of applications is needed for simulations purposes, and scenarios were defined with a selection of the most relevant applications. Finally, E-UMTS traffic generation and activity models, based on population and service penetration values, are described and characterised. ON/OFF states have been characterised by appropriate statistical distributions and parameters.

1 Introduction

Enhanced UMTS (E-UMTS) is a UMTS evolution step, which provides bit rates higher than 2 Mbit/s in the uplink and downlink directions over a 5 MHz frequency carrier, Figure 1.

It enables the provision of new wideband services and a significant reduction of the price per bit, running over flexible QoS enabled IP based access and core networks, and making possible an effective end-to-end packet based transmission. European projects (e.g. IST-SEACORN [1]) proposes a set of enhancements to UMTS, which include, among others, advanced modulation and radio transmission techniques, improved strategies for IP routing and QoS assurance.

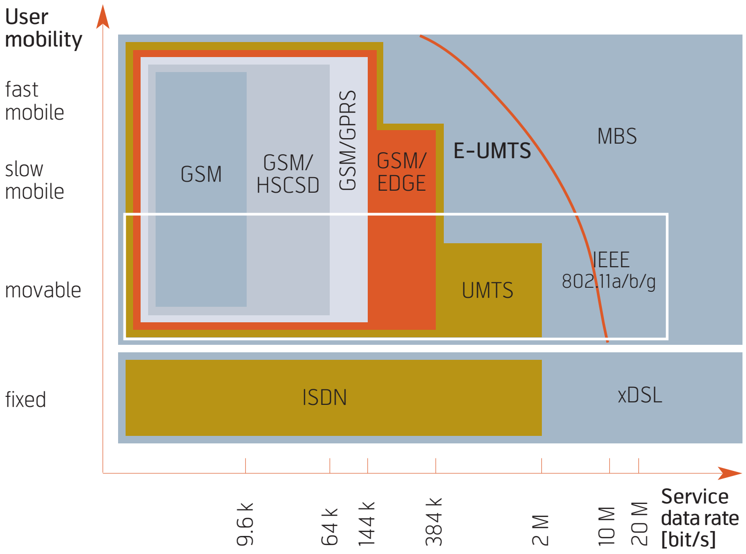


Figure 1 Enhanced UMTS concept. Enhanced UMTS (E-UMTS) is a UMTS evolution step, which provides bit rates higher than 2 Mbit/s in the uplink and downlink directions over a 5 MHz frequency carrier

Unlike HSDPA (high-speed downlink packet access), which will mostly extend UMTS maximum achieved data rates for the downlink, E-UMTS will allow for expansion in both down- and uplink directions. Hence, it will support wideband real-time/time-based mobile applications with a very high system capacity, and will set the ground for an initial introduction of actual broadband mobile applications, an important step towards 4G. E-UMTS will be a first step to achieve the goal of ubiquitous and seamless communications [2], scaling system capacity for mass-market services, which implies a capacity of the order of Gbps/km^2 will be available, and starting the emergence of adaptive services and new network modes, e.g. multicast, multihop and peer-to-peer. While in the Wireless LAN (WLAN) domain it is becoming possible with IEEE 802.11a, b, g, etc., in the mobile communications domain E-UMTS will be a first step to achieve this goal. Furthermore, with the use of multiple devices with various radio interfaces, E-UMTS will be the first 3.5G system to support values of capacity comparable to the ones needed, and will allow for the introduction of the ABC (Always Best Connected) concept [3], even before the introduction of OFDM/WCDMA and UWB systems for 4G. In this context, instead of being a competing technology, E-UMTS will be complementary to the various types of WLANs (and other radio interfaces and access technologies).

In this paper, the available data about mobile applications characterisation parameters is put together, enabling some insight into new approaches of performance analysis in E-UMTS. In this context, parameters are divided into six different types: service, traffic, communications, session and activity, service components, and operation environments.

ISSN 0085-7130 © Telenor ASA 2005

The effect of the proposed enhancements needs to be evaluated by means of simulation. For that purpose a set of services need to be used, in order to create a complete and realistic simulation framework, impacting directly on traffic generation models. These services will be accessed from a variety of operation environments, each with its distinctive set of service preferences, usage patterns, and associated mobility profiles. As an answer to these services and environmental conditions a matching set of deployment strategies need to be studied, adapted, and simulated.

Terminal mobility has a great influence in most UMTS communication aspects involving either performance or traffic generation as a result of handover. Issues such as radio resource management, location management and QoS, as well as traffic handling capacity, are directly affected by mobility. The purpose of mobility models is to describe typical terminal movement so that performance analysis may be made. On the one hand, many link level simulations will require the knowledge of detailed terminal position so that the effectiveness of studied techniques such as link adaptation (adaptive coding and modulation) transmission diversity and beam forming may be evaluated. On the other, some E-UMTS techniques, such as IP transport resource management, may not always require a detailed knowledge of individual terminal movement and positioning. In these cases, a mobility model that describes the average rate of handovers and cell cross-over velocity may be more convenient for traffic simulation purposes.

The *usage* of each application, i.e. the duration of connections of a given application relative to the duration of connections for all applications during the busy hour, is one of the most important aspects to be determined. These data will be essential for multi-service traffic engineering purposes, and it is the main motivation for the realisation of this study. Although there are nowadays few forecast results available for mobile communications, some estimations have already been made for narrow-, wide- and broadband applications in the residential market of fixed networks [4], [5], as well as for wireless [6], and mobile communications, e.g. the UMTS Forum [7] and RACE-MBS [8].

To collect applications data for SEACORN simulation purposes, a pragmatic approach was followed with the intention of extracting a service framework that is possible to implement, with low complexity, and maintaining a realistic and coherent model. At the same time, a further step is taken in traffic modelling by introducing Long Range Dependence models, better adapted to wireless IP communications. These are based on recent research on field data that has concluded that traditional traffic models produce too optimistic results when used to model some data traffic. This may result in networks that were under-dimensioned and therefore under-performing with respect to theoretical expectations.

The number of services and environments that have been considered initially has been integrated and condensed into a smaller set of services and service environments to meet the technical demands of the simulation tools. Without this summarisation effort the framework proposed would pose considerable difficulties to implement due to the amount of computing resources it would require from simulators. A subset of services and environments are therefore selected. Three operation environments, and four or five services result, depending on environment.

This paper is organized as follows. Different types of classification are presented in Section 2. Section 3 describes characterisation parameters, while Section 4 covers the range of variation of the parameters. Section 5 presents an overview of the UMTS Forum, ETSI, RACE-MBS and RACE-TITAN projects' views on service mixtures and deployment scenarios. They complement each other in supplying a background basis for the environment options to be selected for E-UMTS trials. The influence of mobility is discussed in Section 6, and a set of mobility models is given. Three models correspond to particular operation environments and describe the behaviour of individual terminals, while the fourth looks at average tele-traffic behaviour. Section 7 presents a set of scenarios that put together all components of E-UMTS deployment situations. These include expected population density, mobility characteristics and expected usage of service mix for each environment. Section 8 describes the selected set of services, environments and source traffic models to be used in E-UMTS simulations. They cover voice, interactive data (Multimedia Web Browsing, Instant Messaging for Multimedia) and streaming based services (Video-telephony, HD Video-telephony). Conclusions are drawn in Section 9.

2 Classifications

The characterisation of communications services and applications is a complex task that requires a strong effort of systematisation. It includes the organization of services and applications into classes and comprises several steps, from classification issues to the identification of the actual range of variation of the parameters. In the field of telecommunications, an application is defined as a task that requires communication of one or more information streams between two or more parties that are geographically separated,

Service hierarchies		Type of Information	Examples of broadband services	Examples of applications	3G framework	
					Market	QoS
Interactive	Conversational	Sound	Sound	Voice	RV	CONV
				Voice over IP	RV	CONV
		Moving pictures and sound	Video-telephony	Various purposes	RV	CONV
			HD Video-telephony	Tele-education	RV	CONV
			HIMM Videoconference	Various purposes	RV	CONV
			Video-conference	Tele-advertising	CI	CONV
			Video Surveillance	Mobile Video surveillance	MMS	CONV
		Data	High Volume File Transfer	Data File Transfer (FTP)	MIA	INTR
		Document (multimedia)	Mixed Document Communications Service	Multimedia (MM) Web Browsing	MIA	INTR
				Collaborative Working	MIEA	CONV
				Mobile Tele-working	MIEA	CONV/INTR
			High-resolution Image Service	Interactive Remote Games	CI	CONV
				Still Images Communication	CI/MMS	INTR
	Messaging	Mixed document	Multimedia Mail	Instant Messaging for MM	MMS	BACK/INTR
	Retrieval	Text, data, graphics, sound, still Images, moving pictures	Broadband Videotext	Audio Streaming	RV	STR
				E-commerce	MIA	STR
				Tourist Information	LBS	INTR
			Data Retrieval Service	Remote Procedure Call	MIEA	INTR
			Multimedia Retrieval Service	Urban Guidance	CI	INTR
				Mobile Portal (Content/commerce)	CI	INTR
				Assistance in Travel	LBS	INTR
Distribution	Broadcast	Moving pictures and sound	Video Distribution Service	Micro Movies (including video clips)	CI	STR
	Cyclical	Text, graphics, sound & still images	Full Channel Broadcast Videography	E-newspaper	CI	INTR

Table 1 Correspondence between services and applications
MIA – Mobile Internet Access; MIEA – Mobile Intranet/Extranet Access; CI – Customised Infotainment; MMS – Multimedia Messaging Service; LBS – Location-Based Services; RV – Rich Voice.

being characterised by the service attributes, and also by traffic and communications characteristics. A service is defined as a generic set of applications with similar characteristics, or a single application that is significant on itself. As a consequence an application is an instance of a service with specific characteristics. According to ITU-T I.211, applications and services can be divided into the following different groups: interactive (conversational, messaging, and retrieval) and distribution (broadcast, and cyclical) [9], Table 1.

UMTS attempts to fulfil the Quality of Service, QoS, request from the application or the user. In order to reach this the UMTS Forum and 3GPP use the following traffic classes [10]: (i) conversational, CONV, (ii) streaming, STR, (iii) interactive, INTR, and (iv) background, BACK.

The distinguishing factor of these classes is how delay-sensitive traffic is. While the conversational class is the most delay-sensitive, the background class is the less one.

The following relation can be established between 3GPP and ITU.T I.211 classes:

- Conversational class is the same in both classifications. There is a bi-directional dialogue between live end-users.

- Multimedia streaming has to correspond to retrieval and broadcast classes. Information is given in a continuous flow or stream.

- Interactive class can be connected with almost every I.211 class, except broadcast. User requests data from remote equipment.

- Background class treats communications like mail interchange, and is related with I.211 messaging class.

From the market perspective [11] services are the portfolio of choices offered by service providers to users, and can be charged separately. The study presented in [11] identifies six service categories that represent the majority of the demand for 3G services over the next five years. Besides voice, applications are classified into content connectivity (Internet) and mobile ones. While the content connectivity applications can be either mobile Internet access or mobile intranet/extranet access, mobile applications are divided into three categories: customised infotainment, multimedia messaging services (MMS) and location based. A classification according to Market and QoS points of view is also shown in Table 1.

3 Characterisation parameters

Characterisation parameters are necessary to characterise end-to-end services and their requirements, and it is useful to distinguish them into service, traffic, communication, service components, activity model, and operation environments as well. From the operation environments, in the mobile domain, terminal mobility is also of key relevance.

Initially, a set of 16 applications operating in an 'E-UMTS alone' environment was identified in [12], [13]. However, besides these applications, E-UMTS can still support UMTS specific ones, like voice, audio streaming, voice over IP, video clip transfer, or even others, with high foreseen demand, e.g. still image transfer, mobile portals or interactive games [14].

3.1 Service characteristics

The service characteristics are the following [12]:

- Intrinsic time dependency: time-based, TB, or non-TB, NTB;

- Delivery requirements: real-time, RT or non-RT, NRT;

- Directionality: unidirectional, Und, or bi-directional, Bid;

- Symmetry of the connection: Symmetric, Sym, or Asymmetric, Asy, and the respective asymmetry factor;

- Interactivity: existence or not;

- Number of parties: one-to-one, 1-1, one-to-many, 1-m, or multi-party (multi);

3.2 Traffic characteristics

On the one hand, it is important to characterise the generation process for applications:

Generation process – It describes the statistical distribution of session inter-arrivals (e.g. Poisson).

Distribution of the duration – If a negative exponential distribution is used to model call duration with service rate μ, the parameter corresponds to the mean value $1/\mu$. Other distributions such as the lognormal are also used for this purpose [15].

Average duration of connections – It refers to the connection holding time. On the other hand, it is important to describe accurately assumptions on latency/delay and channel hierarchy/bandwidth; i.e. data rates.

Latency/end-to-end delay – Absolute delay is one of the key QoS performance parameters that must be satisfied by the broadband network [16]. In the context of service characterisation, it is the maximum transfer time (in one way) that is tolerated by the service. To provide interactive response to viewers the response time between a user action and its effect should be less than 100 ms.

Data rate – It is important to define the average data rate associated with services. Specifying the average bandwidth requirement for a 'bursty' application is a challenge, because it varies according to the duration for which the average is taken. Furthermore, the values obtained vary widely across different users (such as the ones from image browsing), even for the same applications, because everyone has a unique usage pattern.

3.3 Communication characteristics

The communications characteristics consist of burstiness, service classes, BER (bit error rate), and protocol.

Burstiness – It is defined as the ratio between the peak and the average bit rates [17], several types of communication being highly 'bursty' in nature.

Service Classes – To support broadband applications, and based on QoS parameters, three classes of services must be supported [18]: best-effort delivery (ATM Forum ABR class of service), real-time (RT) delivery of time-based information (CBR or VBR), real-time delivery of non-time based information. ISO and NISO stand for isochronous & non-isochronous traffic, respectively [9].

Bit error rate (BER) – It is a non-dimensional variable that expresses service tolerance to uncorrected errors in the bearer service, including non-delivered information. It is calculated as the ratio between bits received with error or omitted, and the overall received bits.

Communication Protocol – The most common communication protocols [19] are User Datagram Protocol (UDP) and Transmission Control Protocol (TCP). While TCP coordinates the confirmation and retransmission of packets that are lost during a communication session, UDP does not provide any guarantees to data delivery.

3.4 Session and activity model

Detailed parameters describe service behaviour in terms of traffic generation. Two levels of behaviour may be distinguished: call/session, representing traffic generation process (birth and death), and activity models. Call and session parameters follow:

- BHCA and inter-arrival time
- Arrival distribution.

Activity models describe the relationship between active and silent periods in either direction. They are different depending on type of service, and the purpose of the model, and are the following:

- Duration and its distribution
- Average active/inactive time
- Active/inactive time distributions (e.g. exponential).

Applications	Intrinsic time dependency	Delivery requirements	Directionality	Symmetry/ Asymmetry	Asymmetry factor	Nb. of parties
Voice	TB	RT	Bid	Sym	1UL-1DL	1-1, 1-m
Voice over IP					1UL-1DL	1-m
Video-telephony (various purposes)				Sym/Asy	1UL-1DL	1-1
HD Video-telephony (Tele-education)					1UL-1DL/low	1-1
HIMM Videoconference					1UL-1DL	multi
Videoconference (Tele-advertising)					low	multi
Mobile Video Surveillance				Asy	5UL-0.001DL	1-1
Data File Transfer (FTP)	NTB				4UL-1DL	1-1
Multimedia (MM) Web Browsing	TB				1UL-5DL	1-1
Collaborative Working				Sym	1UL-1DL	1-1, 1-m
Mobile Tele-working					1UL-1DL	1-1
Interactive Remote Games	TB/NTB			Asy	26UL-1000DL	1-1, 1-m
Still Images Communication	NTB				26UL-1000DL	1-1
Instant Messaging MM	TB	NRT			4UL-100DL	1-1
Audio Streaming		RT			4UL-100DL	1-1, 1-m
E-commerce					4UL-100DL	1-m
Tourist Information	NTB				4UL-100DL	1-m
Remote Procedure Call		NRT/RT			4UL-100DL	1-m
Urban Guidance	TB				5UL-1e3DL	1-m
Mobile Portal (Content/commerce)	TB/NTB	RT			1UL-5DL	1-m
Assistance in Travel	TB	NRT/RT			5UL-1e3DL	1-1
Micro Movies (including video clips)	TB/NTB	RT	Und (with commands in the reverse link)		26UL-1000DL	1-m
E-newspaper	NTB	NRT	Bid		1UL-5DL	1-m

Table 2 Enhanced UMTS service characteristics

ISSN 0085-7130 © Telenor ASA 2005

Applications	Traffic characteristics			Communication characteristics			
	R_b [kb/s]	Avg. duration [min]	Latency/ delay [ms]	Burstiness	Service class	BER	Protocol
Voice	4-25	3	150	1	ISO&CBR/ RT-VBR	10-4	UDP
Voice over IP	4-25	3	150	1	ISO&CBR/ RT-VBR	10^{-4}	UDP
Video-telephony (various purposes)	32-384	5	200	1-5	ISO&CBR/ RT-VBR	10^{-4}	UDP
HD Video-telephony (Tele-education)	2000	30	200	1-5	ISO&CBR/ RT-VBR	10^{-4}	UDP
HIMM Videoconference	32-384	30	200	1-5	ISO&CBR/ RT-VBR	10^{-4}	UDP
Videoconference (Tele-advertising)	384-2000	30	200	1-5	ISO&CBR/ RT-VBR	10^{-4}	UDP
Mobile Video Surveillance	32-384	10-120	200	1-5	NISO&CBR	10^{-4}	UDP, TCP
Data File Transfer (FTP)	64-2000	1-5 s	10 s	1-50	NISO&CBR	10^{-6}	TCP
MM Web Browsing	384-2000	1-15	few sec.	1-20	ISO&RT-VBR	10^{-6}	TCP
Collaborative Working	64-2000	15-50	500	1-20	ISO&CBR(VBR)	10^{-6}	TCP
Mobile Tele-working	384-2000	15-50	200	1-20	ISO&CBR(VBR)	10^{-6}	TCP
Interactive Remote Games	64-1000	10-30	50	1-30	ISO&CBR(VBR)	10^{-7}-10^{-6}	UDP, TCP
Still Images Communication	64-1000	1-10	1000	1-20	ISO&CBR(VBR)	10^{-7}-10^{-6}	TCP
Instant Messaging for MM	1000-4000	0.1-15	15 s	1-20	NISO&UBR	10^{-6}	TCP
Audio Streaming	12-128	3-60	10 s	1-5	ISO&CBR/ RT-VBR	10^{-6}	UDP, TCP
E-commerce	64-1000	5	500	1-20	ISO&RT-VBR	10^{-6}	TCP
Tourist Information	64-1000	10-15	500	1-20	ISO&RT-VBR	10^{-6}	UDP
Remote Procedure Call	64-1000	5	250	1-50	NISO&ABR	10^{-6}-10^{-4}	TCP
Urban Guidance	128-4000	5-10	1-5 s	1-5	NISO&CBR	10^{-6}	TCP
Mobile Portal (Content/commerce)	64-2000	5-15	1-5 s	1-50	ISO&RT-VBR	10^{-6}	TCP
Assistance in Travel	128-4000	20-360	500	1-5	ISO&CBT & RT-VBR	10^{-6}	TCP
Micro Movies (including video clips)	64-384	3-5	10 s	1	NISO&CBR	10^{-6}	UDP
E-newspaper	1000-2000	20	500	1	ISO&RT-VBR	10^{-6}	TCP

Table 3 Communication and traffic parameters

In E-UMTS, both data and conversational communications are similarly supported as Packet Data Protocol (PDP) context establishment. In general, up- and downlink parameters have different values.

3.5 Service components

A given application can be supported by different services, having, as a consequence, different characteristics in terms of type of information, which are the following: sound, moving pictures or video, document (multimedia), data, text, graphics and still images. If applications have access to more than one type of information simultaneously, a simple activity/inactivity model is not enough, and a model for service components (that maps into the types of information) has to be defined for multi-service purposes. Details related to the service components (sound, data and video), and their correspondence with applications [20], were left for further work.

4 Range of variation

In Table 2 the range of variation of the parameters is identified. Table 3 presents communications and traffic parameters. All applications are interactive.

5 Hypothesis for 3G and beyond

5.1 UMTS Forum perspective

The UMTS Forum has identified six operational environments [7]: i) CBD – City business district (in building), ii) Suburban (in building or on street), iii) Home (in building), iv) Urban (pedestrian), v) Urban (vehicular), and vi) Rural in- & outdoor. The density of potential users per km^2 and the foreseen cell types have also been identified in low (pedestrian), medium and full (high) mobility scenarios.

UMTS penetration figures for years 2005 and 2010 in each operating environment can be extracted from these reports for each service class, Table 3.3 of [7]. These figures are based on extensive market research within Europe and represent the fraction of the density of potential users for each of the operation environments given in Table 3.1 of [7]. In order to achieve the number of active users it is necessary to know the busy hour connection attempt, BHCA, defined as the ratio between the total number of connections and the total number of subscribers in the considered area during the busy hour [7].

New forecasts are available in [21] for the categories of voice, location based services, business MM messaging service, mobile Internet access, consumer MM messaging service, mobile Intranet/Extranet access, and customised infotainment.

5.2 ETSI perspective

ETSI has identified the following three deployment scenarios for Wireless LANs [6]: Office, Industry and Studio (TV, radio or recording). The usage of applications deployed in those scenarios is presented in Tables 9–11 of [6], as well as their average data rate; different sets of applications exist in each of the scenarios. In this work, these values are the basis for the cases of E-UMTS deployment scenarios with movable or low mobility terminals.

5.3 MBS perspective

In the perspective of the RACE-MBS project, mobile applications can be divided into movable, slow (< 36 km/h) and fast mobile, each of them having different associated data rates, Table 4.1 of [8]. The fast mobile ones are: City Guidance, Freight and Fleet Management, Emergency, Pictorial Data for Travel, Public Transport Information, Electronic Newspaper, Traffic Advice, HDTV (High Definition Television) Contribution, Audio-visual Library and Surveillance of Property. The ones associated with slow mobility are: Access to Banking Services, Special Needs (health), Repair Assistance, CAD (Computer Added Design) Interconnection and HD Videophone; the movable ones are Tele-consultation and Wireless LAN (WLAN) Interconnection. User groups have been identified, and related with the following five geographical areas: Primary roads, City centres, Residential areas, Industrial areas and Hotspots. Estimations have been made for mature MBS on the busy hour rate; i.e. the percentage of total potential users that are active during the busiest hour [8]. These values are used here as a basis for the definition of various scenarios in a geographical area. Different notations are adopted in [6], [7], [8] for the usage. In order to have a common notation some definitions are presented in [9].

Demand as a percentage of the market		Residential	E-UMTS
Voice	-	55 %	57%
High interactive MM	< 144 kb/s	15 %	16 %
Narrowband	[144,384] kb/s	10 %	11 %
Wideband	]384, 2 048] kb/s	15 %	16 %
Broadband	> 2 Mb/s	5 %	-

Table 4 E-UMTS possible scenario (high mobility)

5.4 RACE-TITAN forecasts

From the RACE-TITAN project [4], [5] one extracted forecasts for 2010, defined as a percentage of the total residential market. Table 4 presents the adaptation of these forecasts to E-UMTS.

Here, the 144 kb/s limit of TITAN-ISDN applications was extended to 384 kb/s. Applications with data rates in the range [144, 384] kb/s were designated by narrowband ones, whereas the ones with data rate lower than 144 kb/s are called high interactive MM (multimedia) ones.

6 Mobility models

The purpose of mobility models is to describe typical terminal movement so that performance analysis may be made. Two basic types of models are covered: individual terminal and tele-traffic mobility models. Individual terminal models cover indoor (office), city centre and road mobility types, while the tele-traffic mobility model deals with average values, and may be parameterised from pedestrian to highway velocities, thus covering all types of environments.

6.1 Office environments

From the mobility perspective, indoor environments are characterised by slow speeds (pedestrian or slow vehicular) and relatively well defined mobility paths, determined by architectural topology and activity patterns. Each particular environment may however exhibit its own distinctive features. Among others, the following may be enumerated: home, office environments, airport and train stations, commercial zones, theatres/public diversion, and parking zones.

The particular case of office environments is characterised by a “boxy” topology, where office rooms are interconnected by passage areas, or corridors. Users will spend some considerable time stationary at a desk and when in motion will move towards a particular destination using a given path. Destinations may be chosen randomly, using a uniform distribution. Two cases may be considered, concerning the nature of movements. In the first one, both source and destination are an office room, while in the other, either source or destination is a corridor position.

The important parameters are the mean ratio of room-to-corridor situated mobile terminals, $r = 85\ \%$, the average time in office room, $T_r = 10$ min and the mobile speed, $v_m = 3$ km/h. Their values can be adjusted to particular environment conditions, especially the average office time and ratio ones.

6.2 Outdoor-to-indoor pedestrian – Business city centres

As the name implies this mobility model is associated with a business city centre environment where base stations are placed outdoors but also cover internal building areas [22]. To represent city centres, the Manhattan Model, i.e. a rectangular grid of intersecting streets, is normally used. Homogeneous squared buildings, with 200 m side and 30 m wide streets could characterise the environment. This sort of environment is characterised by small areas with high buildings, high user density and pedestrian mobility.

The urban mobility model is highly related to the Manhattan-like structure defined above. In such a structure, mobiles move along streets and may turn across streets with a given probability. The mobile's position is updated every five metres, and speed can be changed at each position update according to a given probability. The mobility model is described by the following parameters:

- Mean speed: 3 km/h
- Minimum speed: 0 km/h
- Standard deviation for speed (normal distribution): 0.3 km/h
- Probability to change speed at position update: 0.2
- Probability to turn at cross streets: 0.5

Mobiles are uniformly distributed in the street and their direction is randomly chosen at initialisation.

6.3 Vehicular environment

Vehicular environment applies to scenarios in urban and suburban areas outside the rise core, where the buildings have nearly uniform height, and are characterised by larger cells and higher transmit power.

The vehicular reference mobility model uses a pseudo-random process with semi-direct trajectories. The mobile position gets updated according to the de-correlation length and direction can change at each position update following a given probability within a sector. For reference example, it can be assumed that the mobile's speed is constant, and that the mobility model is defined by the following parameters [22]:

- Speed: 36 – 81 km/h
- Probability to change direction at update: 0.2
- Maximum angle for direction update: 45°
- De-correlation length: 20 metres

Mobiles are uniformly distributed on the map, and their direction is randomly chosen. The cell radius is 2000 m for services up to 144 kb/s, and 500 m for data rates above 144 kb/s. The base station antenna height must be 15 metres above the average rooftop height. The deployment scheme is a hexagonal cell layout with distances between base stations equal to 6 km. Tri-sectored cells should be used.

6.4 Flow equilibrium model for traffic

The high mobility associated with E-UMTS calls for a tele-traffic analysis, where both new connections and handover connections traffic must be considered simultaneously. In a first approach, linear coverage geometries, where mobiles travel randomly through cells located end-to-end, can be considered. In a linear coverage geometry, cells are placed end-to-end, and mobiles can handover only from a cell to one of the two adjacent ones; a connection comprises successive sessions τ_1, τ_2, τ_3, ... in cells traversed by a mobile terminal, and its duration τ follows an exponential distribution whose mean is [23], where μ is the service rate. The channel occupancy time τ_c is the time spent by a user in communication prior to handover (or subsequent to handover) or connection completion, and also follows an exponential distribution with reasonable accuracy [24].

The cell dwell time τ_h is the residing time of a mobile within a cell. Further assuming that the dwell time is exponentially distributed with mean $\bar{\tau}_c = \mu + \eta$, the channel occupancy time is $\tau_c = \min\{\tau, \tau_h\}$; i.e. it is either the time spent in a cell before crossing the cell boundary if the connection continues or the time until the channel is relinquished [23]. As the minimum of two exponential random variables is also exponentially distributed with parameter $\mu_c = \mu + \eta$, the mean channel occupancy time is given by

$$\bar{\tau}_c = \frac{1}{\mu_c} = \frac{1}{\mu + \eta} \tag{1}$$

The cross-over rate, η, can be derived by knowing the average speed of terminals. As a service parameter it may be relevant if there is a typical mobility pattern associated with the service. E-UMTS scenarios of mobility are characterised by a triangular distribution for the velocity with average velocity, $V_{av} = 1$, 10, 15 and 22.5 $\mathrm{m \cdot s^{-1}}$, for the pedestrian (PD), urban (UB), main roads (MR) and highways (HW) scenarios, respectively, Table 5.

Scenario	V_{av} [m · s^{-1}]	Δ [m · s^{-1}]
Static	0	0
Pedestrian	1	1
Urban	10	10
Main roads	15	15
Highways	22.5	12.5

Table 5 Characteristics of the scenarios of mobility

Services	Data rates	Demand, % of the market					
	[kb/s]	Residential		Mixed		Business	
Sound	–	57		42		27	
High Interact. MM	£ 144	16		16		16	
Narrowband	]144,384]	11	27	18.5	42	26	57
Wideband	]384,2 048]	16		23.5		31	

Table 6 Assumptions for usage in various markets

For these scenarios, a triangular distribution is considered for the velocity [25], with average $V_{av} = (V_{\max} + V_{\min}) / 2$, and deviation $\Delta = (V_{\max} - V_{\min}) / 2$. This leads to the following cross-over rate

$$\eta = \left\{ \frac{2R}{\Delta^2} \left[(V_{av} + \Delta) \cdot \ln \left(\frac{V_{av} + \Delta}{V_{av}} \right) \right. \right. \\ \left. \left. -(V_{av} - \Delta) \cdot \ln \left(\frac{V_{av}}{V_{av} - \Delta} \right) \right] \right\}^{-1} \qquad (2)$$

when $V_{\min}$, $\Delta > 0$, and when $V_{\min} = 0$, to the limit

$$\eta = \frac{V_{av}}{2 \cdot \ln(2)} \cdot \frac{1}{(2R)} \qquad (3)$$

Different types of mobility are assumed for each application in each of the scenarios.

7 E-UMTS scenarios definition

It is still difficult the have a clear view of the deployment scenarios in E-UMTS. However, it is already possible to clearly distinguish the following eight deployment scenarios [12]: i) business city centre, BCC (vehicular or pedestrian), ii) urban residential, URB (vehicular or pedestrian), iii) primary roads, ROA, iv) trains, TRA, v) commercial zones – COM (e.g. airports, railway stations, hospitals, commercial centres, universities), vi) offices, OFF (buildings), vii) industry, IND (large factories plant), and viii) home, HOM.

From the data available for UMTS (data rate from 144 kb/s to 2 Mb/s), MBS and Wireless LANs (> 2 Mb/s), it is possible to perform an updated extrapolation for E-UMTS communications. Broadband applications (> 2 Mb/s) will only be supported in scenarios without relevant mobility; hence, the offices and industry scenarios will be defined separately. Because the data from the RACE-TITAN project is for the residential market, some changes have to be made for business and mixed (half-business / half-residential) ones, Table 6. One considers an increase of 15 % on the demand of narrow- plus wideband applications from the residential market to the mixed one, and also from the mixed to the business market, corresponding to a decrease in traditional markets, i.e. sound and voice.

Values for offices and industrial markets [22] have slight differences compared to the business market, Table 7. In the context of these scenarios, new applications, like Control data, Monitoring [6], TV Programme Distribution (MPEG2-4), WLAN Interconnection and Professional Images, are considered. One further assumes that there is a correspondence between deployment scenarios and envisaged markets:

- Residential: URB, and HOM
- Mixed: ROA, TRA, and COM
- Business: BCC, OFF, and IND

Values are proposed for the usage of applications at the eight scenarios, Table 8. The set of applications considered are the ones from [12] plus Broadband, and Sound [14]. In Table 8, besides values for usage, the envisaged approximated maximum data rates, R_b, are introduced for all applications, in order to establish the service class (Sound, High Interactive Multimedia, Narrow-, Wide- or Broadband). These data rates refer to the link with higher bit rate (either the up- or the downlink). Asymmetric applications (e.g. FTP) will only need such high bit rates in one of the ways. The density factors for each of the scenarios are presented as well [8]. In the business market there is a fundamental difference between the BCC scenario, and the OFF and IND ones: applications are movable (not

Services	Data rates	Demand [%]	
	[kb/s]	Offices	Industry
Sound	-	25	15
High Interact. MM	£ 144	15	10
Narrowband	]144,384]	20	20
Wideband	]384, 2 048]	25	40
Broadband	> 2048	15	15

Table 7 Assumptions for offices and industry

ISSN 0085-7130 © Telenor ASA 2005

Applications Usage [%]	R_b [kb/s]	BCC	URB	ROA	TRA	COM	OFF	IND	HOM
Sound									
Voice	12	14.0	29.0	21.5	21.5	21.5	13.0	7.5	29.0
Voice over IP	12	10.0	21.0	15.5	15.5	15.5	9.0	6.0	21.0
Audio Streaming	64	3.0	7.0	5.0	5.0	5.0	3.0	1.5	7.0
Total		**27.0**	**57.0**	**42.0**	**42.0**	**42.0**	**25.0**	**15.0**	**57.0**
High Interactive Multimedia (HIMM)									
Interactive Remote Games	128	1.5	1.5	1.5	1.5	1.5	1.0	0.0	1.5
Still Images Communication	128	3.0	3.0	3.0	3.0	3.0	3.0	2.5	3.0
Mobile Portal	128	3.0	3.0	3.0	3.0	3.0	3.0	2.5	3.0
Micro-movies	128	1.5	1.5	1.5	1.5	1.5	1.0	0.0	1.5
Video-telephony	128	3.0	3.0	3.0	3.0	3.0	3.0	2.0	3.0
HIMM Videoconference (various purposes)	128	2.0	2.0	2.0	2.0	2.0	2.0	1.5	2.0
Collaborative Working (& tele-presence)	128	2.0	2.0	2.0	2.0	2.0	2.0	1.5	2.0
Total		**16.0**	**16.0**	**16.0**	**16.0**	**16.0**	**15.0**	**10.0**	**16.0**
Narrowband									
Videoconference (Tele-advertising)	384	3.2	1.5	2.2	2.2	2.2	3.0	0.0	1.5
Data File Transfer (FTP)	384	5.5	2.3	3.9	3.9	3.9	4.0	6.0	2.3
Multimedia Web Browsing	384	11.8	4.9	8.4	8.4	8.4	9.0	14.0	4.9
Broadband Videotex (E-commerce)	384	5.5	2.3	4.0	4.0	4.0	4.0	0.0	2.3
Total		**26.0**	**11.0**	**18.5**	**18.5**	**18.5**	**20.0**	**20.0**	**11.0**
Wideband									
Monitoring	480	-	-	-	-	-	-	5.0	-
Instant Messaging for Multimedia	1024	2.2	1.0	1.2	1.5	1.9	8.0	4.0	1.7
Remote Procedure Call	1024	2.2	2.5	1.8	2.0	4.0	2.0	8.0	2.5
Mobile Tele-working	1536	5.4	0.7	2.0	1.5	2.0	2.5	2.5	3.1
Assistance in Travel	1536	2.7	3.5	6.0	2.0	3.0	1.5	1.5	1.0
Urban Guidance	1536	0.8	1.1	2.0	1.5	1.9	0.5	0.5	0.5
Mobile Video Surveillance	1536	0.3	0.2	0.1	-	0.2	0.3	10.0	0.2
Tourist Information	1536	2.7	0.3	1.4	2.5	3.5	0.5	0.5	0.3
E-newspaper	1536	3.7	3.2	3.0	8.0	4.0	4.7	5.0	3.2
HD Videotelephony (Tele-education)	2048	11.0	3.5	6.0	4.5	5.0	5.0	3.0	3.5
Total		**31.0**	**16.0**	**23.5**	**23.5**	**23.5**	**25.0**	**40.0**	**16.0**
Broadband									
Control Data	3840	-	-	-	-	-		10.0	-
TV Programme Distribution (MPEG2-4)	12780	-	-	-	-	-	4.5	0.0	-
WLAN Interconnection	12780	-	-	-	-	-	6.5	1.5	-
Professional Images	12780	-	-	-	-	-	4.0	3.5	-
Total							**15.0**	**15.0**	
Density Factor (Number of users / m^2)		0.031	0.012	0.012	0.111	0.150	0.150	0.004	0.015

Table 8 Proposal for applications usage in each deployment scenario

mobile) in the latter, hence, broadband applications, with data rates up to 8 Mb/s, can be supported.

As an example, the following assumptions have been made for the business market (including BCC, OFF and IND scenarios):

- For sound applications, while the usage is 27 % in BCC it is 25 % in OFF, and 15 % in IND; i.e. slight differences exist among them. As an example, the values for application usage in BCC are the following: Voice, 14 %, Voice over IP, 10 %, and Audio streaming, 3 %.

- High interactive multimedia (HIMM) applications have an overall usage of 16 % in BCC, 15 % in OFF, and 10 % in IND. In the case of BCC, the usage is distributed in the following way: Interactive remote games, 1.5 %, Still Images Communication, 3 %, Mobile Portal, 3 %, Micro-movies, 1.5 %, Video-telephony, 3 %, Video-conference, 2 % and Collaborative Working (& tele-presence), 2 %. The slight difference between OFF and IND comes from the lower usage of entertainment applications (Interactive Remote Games and Micro-movies).

- Narrowband applications have a usage of 26 % in BCC and 20 % both in OFF and IND. As an example, in the BCC scenarios the values for the usage are distributed in the following way: Video-conference for Tele-advertising, 3.2 %, Data File Transfer (ftp), 5.5 %, Multimedia Web Browsing, 11.8 %, and Broadband Videotex for *e*Commerce, 5.5. %. While the differences from the OFF scenario are slight, there are important differences from the IND scenario [4]: one only considers 6 % of FTP usage and 14 % of Web browsing usage, while the other applications are not used.

- For wideband applications the usage is 31 % in BCC, 25 % in OFF, and 40 % in IND. The data from the MBS project, Table 4.1 of [8], and from Wireless LANs Tables 9 and 10 of [6] were used, except for Instant Messaging for Multimedia, *e*Newspaper and Remote Procedure Call, since data were not available. An example follows on the way that parameters have been obtained for this class of service in the BCC scenario: one considers a usage of 2.2 % of Instant Messaging for Multimedia, 3.7 % for *e*Newspaper and 2.2 % for Remote Procedure Call; next, the sum of these values was subtracted from the 31 % of usage of Wideband applications, a value of 22.9 % being obtained; finally, this usage was distributed by the remaining applications in the Wideband service class proportionally to the values for the usage extracted from Table 4.1 of [8], which are used as weights.

- The usage of broadband applications is 15 % in OFF and IND. In OFF, one considers the following values for the usage: TV Programmes Distribution (MPEG2-4), 4.5 %, WLAN Interconnection, 6.5 %, and Professional Images, 4 %. In IND one does not consider TV Programme Distribution because entertainment is less likely to occur. Instead one introduces an important application: Control Data, with a usage of 10 %.

In other types of market, residential (URB and HOM) and mixed (ROA, TRA and COM) the procedure followed to determine the values of the usage was similar, and data for wide- and broadband applications were extracted from data for Wireless LANs [6] and from MBS [8] as well. The names of the deployment scenarios are approximately the same as in [8], and the data from hotspots have been considered for the Train and Commercial Zones deployment scenarios; the Home scenario was considered as being similar to the Urban one, with slight changes in the usage, except for the one of Tele-working (higher usage at home), Assistance in Travel and Urban Guidance (lower usage at home).

Finally, it is worth noting that the values presented for the maximum data rates are approximate, and refer to the link with higher bit rate (either the up- or the downlink). Asymmetric applications (e.g. FTP) will only need such high bit rates in one of the ways, whereas for bursty VBR applications (e.g. Desktop MM) the average bit rate can be much lower leading to an improvement of the resource usage, and a statistical multiplexing gain occurs.

For a more complete description of the operation environments, the definition of scenarios from mobility is needed. Details are given in [26].

8 Source traffic and simulations

To meet the technical demands of the simulation tools a reduced set of services and environments are therefore selected. Without this summarisation effort the framework proposed would pose considerable difficulties to implement due to the amount of computing resources it would require from simulators. As a result we have three operation environments with four or five services each depending on the environment.

Proposed applications and their relative usage are shown in Table 9. This table is obtained from Table 8 by assuming, as a simplification, that the most signif-

Applications Usage [%]	Data rate [kb/s]	OFF	BCC	VEH
Sound				
Voice (VOI)	12.2	25.0	27.0	42.0
High Interactive Multimedia				
Video-telephony (VTE)	128	15.0	16.0	16.0
Narrowband				
Multimedia Web Browsing (MWB)	384	20.0	26.0	18.5
Wideband				
Instant Messaging for Multimedia (IMM)	1024	25.0		
Assistance in Travel (ATR)	1660			23.5
HD Video telephony (HDT)	2048		31.0	
Broadband				
WLAN Interconnection (WLI)	12780	15.0	-	-
Density factor (users/m^2)		**0.150**	**0.031**	**0.012**

Table 9 Proposal for applications usage in each of the SEACORN simulation deployment scenarios

icant application accounts for all the traffic usage in that service group. The envisaged approximate data rates are introduced for all applications in accordance with the service class associated with the application (Sound, High Interactive Multimedia, narrowband, wideband or broadband). Here, the ROA scenario was called vehicular (VEH). The population density factors for each of the scenarios are the ones from Table 8. Data rates are aligned to existing standard values in UMTS and HSDPA.

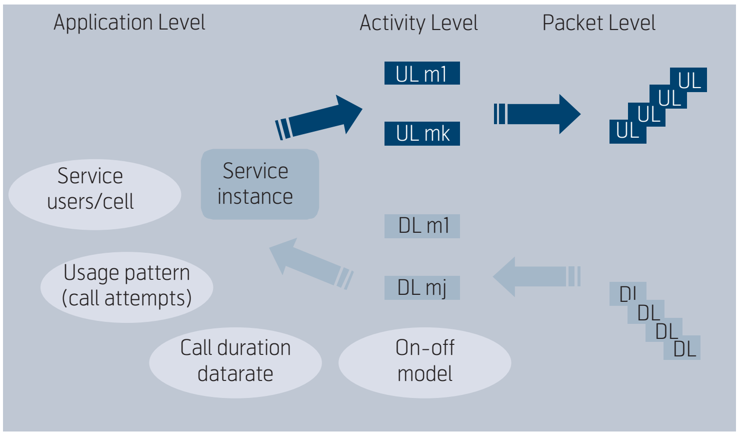


Figure 2 Application traffic. The amount of generated calls is dependent on the number of potential users and the session arrival rate per user. Call duration, and activity pattern determine traffic behaviour

Applications	OFF	BCC	VEH
Voice	0.3	0.3	0.3
Video-telephony	0.2	0.2	0.2
Multimedia Web Browsing	0.15	0.15	0.12
Instant Messaging for Multimedia	0.2	-	-
Assistance in Travel	-	-	0.2
HD Video-telephony	-	0.2	-
WLAN Interconnection	0.15	-	-

Table 10 E-UMTS penetration rates (2010)

8.1 Application traffic

The amount of generated calls is dependent on the number of users and the session arrival rate per user that characterises each service and environment. Call duration and activity pattern determine traffic behaviour, Figure 2.

User density per environment is defined in Table 9. Penetration rates of E-UMTS services are used to account for different adoption rates between services in each year and evolution of service take-up along the years. Service penetration rates for the year 2010 are estimated, extrapolating and adapting those made for UMTS [7], Table 10.

The total number of application j subscribers M_j is given as a function of the penetration P_j by

$$M_j = P_j \cdot M_T \qquad (4)$$

where M_T is the total population of users in the cell.

8.2 Busy hour call attempt

From the values of usage one is considering in SEACORN for OFF, BCC, and VEH scenarios it is important to obtain, for each considered service, the values of the busy hour call attempts to be used in simulations. Busy hour call attempt represents in this case the total number of call attempts by all users considered in one simulation. They will correspond to the users covered by a radio cell or part of a cell.

$$BHCA_j = \frac{Usage_j}{\bar{\tau}_j} \cdot M_T \cdot \bar{f} \qquad (5)$$

where M_T is the number of users in the cell, $\bar{\tau}_j$ is the average call duration and $\bar{f}$ is the average traffic per user, which can vary from 0 (no activity) to 1 (user permanently busy). From the values for usage from Table 9 and considering a user population of M_T = 100, one can obtain, as an example, results for the *BHCA* as a function of $\bar{f}$ for the office environment, Figure 3.

8.3 Call generation and traffic parameters

Call generation and traffic parameters describe service behaviour in terms of traffic generation. They include average data-rate, session arrival rate per user during the busy hour, average call duration, burstiness and asymmetry factor. The term call usually refers to conversational services while session refers to data connections. Because model parameters for each type may be different it is useful to have different designations even though the concept is basically the same, since in E-UMTS both entities are similarly supported on PDP context establishment. It must be referred that since uplink and downlink data rates are not necessarily equal, the data rate parameter value refers to the higher value between the two, which in all exemplified cases is the downlink.

The set of parameters that describe these services from the traffic modelling perspective are defined in Table 11. Detailed parameters describe service behaviour in terms of traffic generation. Two levels of behaviour may be distinguished: call/session representing traffic generation process, and activity models that describe how a session behaves in terms of idle and active periods.

Call and session related parameters are used to model the birth and death of calls and sessions. Session arrival rate represents the average number of calls generated per service subscriber during the busy hour. The Poisson process is used to model session arrivals. Values for session arrival rate are derived from service penetration and session duration values, considering $\bar{f} = 0.03$ (here $M_T = 1$). These parameters are tightly associated for each application and deployment scenario, as formulated below. Usage, U_j, is expressed as a ratio between the traffic for service j (derived from session arrival rate, duration and number of subscribers) given by the total average duration of connections of application j and the total produced traffic during the busy hour (6). The conversion to *BHCA* can be done by using (5).

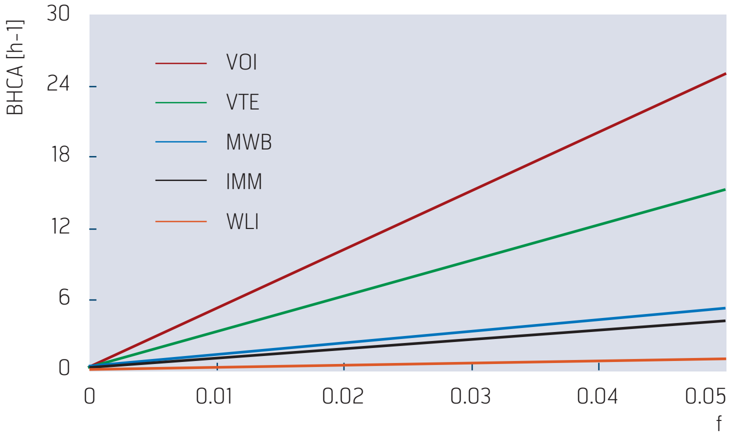


Figure 3 BHCA as a function of $\bar{f}$ for the office environment (M_T = 100). It depends on the demand and average duration of each application, and is a linearly increasing function. Typically $\bar{f}$ varies up to 4 – 5 %

$$U_j = \frac{SessArrRate_j \cdot P_j \cdot \tau_j}{\sum_i SessArrRate_i \cdot P_i \cdot \tau_i} \qquad (6)$$

8.4 Session activity parameters

Session activity parameters describe the detailed aspects of traffic within a call. This is accomplished by means of an alternating active/inactive state model (ON/OFF). The activity within a call can be modelled by defining an average duration of each period, together with an adequate statistical distribution. Video telephony applications are always active in both directions and so do not have OFF periods.

The basic model for application data normally uses a web session as a paradigm, although the model may be used for all types of data. A session is composed of a set of active periods made of packet sequences (packet calls) separated by inactivity periods. A

Applications	Data rate	Session arrival rate [h-1]			Avg.duration	Burstiness	Symmetry
	[kb/s]	OFF	BCC	VEH	$\bar{\tau}$ [min]		UL/DL
Voice	12	0.50	0.54	0.84	3	1	1
Video-telephony	128	0.45	0.48	0.48	3	1-5	1
Multimedia web browsing	384	0.16	0.21	0.19	15	1-20	0.25
Instant Messaging for Multimedia	1024	0.15	-	-	15	1-50	0.05
Assistance in Travel	1536	-	-	0.11	20	1-5	0.07
HD Video-telephony	2048	-	0.09	-	30	1-5	1
Wireless LAN Interconnection	12780	0.03	-	-	60	1-20	0.25

Table 11 Call generation & traffic parameters

Applications	Active state (ON)			Inactive state (OFF)	
	Average [s]	File size [kB]	Distribution	Average [s]	Distribution
Voice	1.4	2.14	EXP	1.7	Exponential
Video-telephony	τ	-	-	0	-
Multimedia Web Browsing	5	240	Pareto	13	Pareto
Instant Messaging for Multimedia	5	640	Weibull	90	Pareto
Assistance in Travel	60	11520	Weibull	14	Pareto
HD Video-telephony	τ	-	-	0	-
Wireless LAN Interconnection	5	7988	Weibull	1	Pareto

Table 12 Application activity parameters

packet call is a sequence or burst of packets, corresponding, e.g. to a Web page or other data item. Inactivity periods between packet call arrivals are often called reading or inactivity time.

Table 12 describes average active versus inactive durations, the corresponding file sizes of activity packet call periods and statistical distributions of the active and inactive durations.

It may be noted that streaming services do not exhibit inactive states, therefore the ON state is equal to call duration and OFF state duration is zero.

Entity	Random variable	Parameters
Session arrival [h^{-1}]	Exponential	Mean = 0.16~0.21
Session duration [min]	Exponential	Mean = 15
Packet calls size [kB]	Pareto	α = 1.1, k = 22 kB Mean = 240 kB)
Inactive time distribution [s]	Pareto	α = 1.5, k = 3s (Mean = 13 s)

Table 13 Multimedia web browsing traffic model

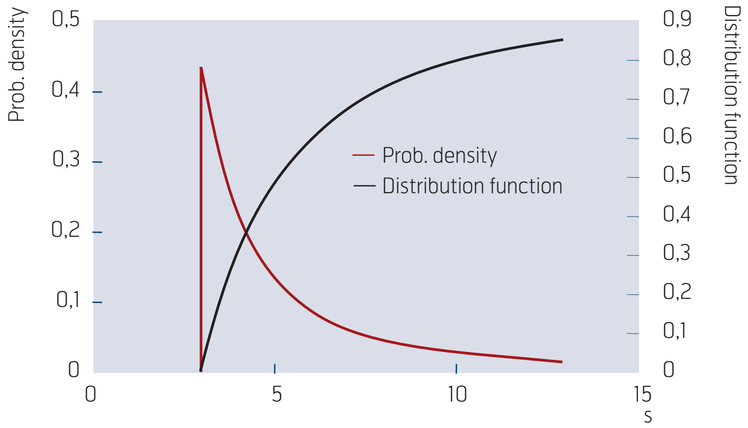


Figure 4 Multimedia Web browsing inactive time distribution. Inactive time (or reading time) is modelled by a Pareto distribution with a minimum of 3 s, and average of 13 s

8.5 Speech source model

For simulation, speech sources are based on adaptive multi rate (AMR) codecs with only the 12.2 kb/s mode. Each AMR 12.2 kb/s source is modelled with an ON/OFF model for discontinuous transmission (DTX) as follows:

- Voice Call Duration Distribution: Exponential, mean: 180 sec
- Duration of On-state Distribution: Exponential, mean: 1.4 sec
- Duration of Off-state Distribution: Exponential, mean: 1.7 sec.

During talk spurts the source generates 32-byte speech payload at 20 ms intervals, while due to DTX during silence periods a 7-byte payload carrying Silence Descriptor (SID) frame at 160 ms intervals is generated [27] [28]. Mapping of AMR data into MAC PDUs is done according to the description in [26]. If we assume the typical VoIP protocol stack employing Real-Time Transport Protocol (RTP) encapsulated in User Datagram Protocol (UDP), which is further carried by the IP the combination of these protocols introduces a total of 40 bytes header data when using IP version 4 (IPv4), and 60 bytes header if IPv6 is used [29].

8.6 Multimedia Web Browsing

This data source model is based on source models for web browsing but considers higher average file sizes due to the multimedia nature of the service. Multimedia traffic exhibits potentially large file sizes with heavy tail distribution patterns. Each session is modelled as a WWW application, consisting of a sequence of packet calls corresponding to file downloads having the statistics shown in Table 13. It must be noted that the reading time (or time between packet calls) may be adjusted to produce different data rates and simulate multiple users. The packet call size corresponds to the file size parameter stated in Table 12. Inactive time (or reading time) is modelled by a Pareto distribution with a minimum of 3 s, and average of 13 s (Figure 4).

8.7 Instant Messaging for Multimedia (IMM)

Real-time Multimedia messages will be exchanged between users via an MM Instant Messaging server. There will be a pattern of exchange/read sequences as in MM Web browsing. It is considered that messages will not have a minimum size, and therefore the Weibull distribution is used to model packet call size, Table 14.

A graphical representation of the distribution is shown in Figure 5.

8.8 Assistance in Travel

This service consists of

a City Guidance – a tourist can have assistance to meet a given location through this application, choosing the best path and informing him/her the average time it takes to get there. Additionally, virtual reality images may be shown offering a virtual tour or guidance through super-imposed imaging by means of an adequate visualisation device such as eyeglasses.

b Traffic Advice and Road Conditions – this component of the application is intended to provide user-oriented information (data, video and audio comments) about how to get to a destination and the traffic conditions. A user can decide on the more convenient road to reach a specific destination, get information about parking, traffic flow and traffic forecast. Images of the path and critical spots may be shown (e.g. on a built-in car device such as the windscreen or ergonomic terminal).

The service combines location based services with personal assistance and may use interactive video streaming for virtual reality sessions. The Weibull distribution is used, Table 15.

8.9 WLAN Interconnection

WLAN interconnection is expected to follow a behaviour similar to LAN data traffic, Table 16. Accounting for a slight levelling due to source traffic aggregation, a Weibull distribution with a shape parameter of $\alpha = 1.5$ is used to produce a less pronounced distribution tail, as shown in Figure 6.

8.10 MPEG-4 Video Model

MPEG-4 was developed for Internet and mobile applications and will be extensively used in UMTS, together with H.263, which has similar characteristics.

Entity	Random variable	Parameters
Session arrival [h^{-1}]	Exponential	Mean =0.15
Session duration [min]	Exponential	Mean = 15
Packet calls size [kB]	Weibull	α = 1, β = 640 (Mean = 640 kB)
Inactive time distribution [s]	Pareto	α = 1.5, k = 30 s (Mean = 90 s)

Table 14 IMM traffic model parameters

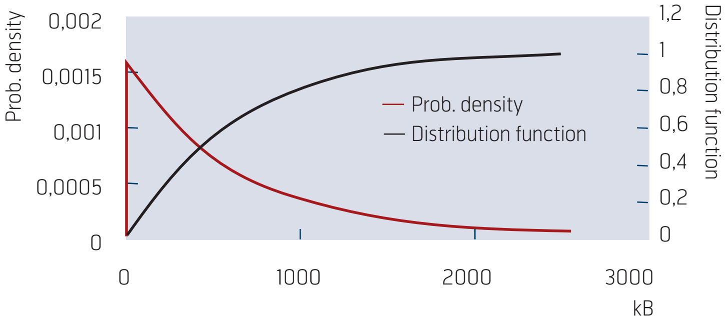


Figure 5 Instant Messaging for Multimedia Packet Call Size Distribution. It is considered that messages will not have a minimum size, and therefore the Weibull distribution is used to model packet call size

Entity	Random variable	Parameters
Session arrival [h^{-1}]	Exponential	Mean = 0.11
Session duration [min]	Exponential	Mean = 20
Packet calls size [kB]	Weibull	α = 1.2, β = 12246 (Mean = 11520kB)
Time between packet calls [s]	Pareto	α = 1.5, k = 3 s, (Mean = 13 s)

Table 15 Assistance in travel traffic model parameters

Entity	Random variable	Parameters
Session arrival [h^{-1}]	Exponential	Mean = 0.03
Session duration [min]	Exponential	Mean = 60
Packet calls size [kB]	Weibull	α = 1.5, β = 8848 (Mean = 7988 kB)
Time between packet calls [s]	Pareto	α = 2, k = 0.5 s (Mean = 1 s)

Table 16 WLAN Interconnection traffic model parameters

MPEG-4 encoders generate three types of frames: I frames, P frames and B frames. Several methods to model and simulate MPEG video traffic have been presented [30], [31]. These models describe an MPEG video stream, and are suitable for streaming and real

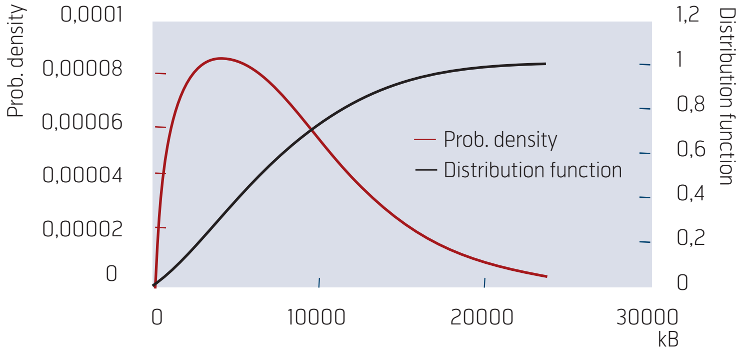


Figure 6 WLAN Interconnection file size distribution. It is expected to follow a behaviour similar to LAN data traffic. Accounting for a slight levelling due to source traffic aggregation, a Weibull distribution with a shape parameter of α = 1.5 is used

time videotelephony simulations. The Gamma, Weibull and Lognormal distributions are almost equally suitable for this purpose. In [30] a set of statistical data that captures a set of MPEG samples is analysed, and details on the parameters are given.

However, the main result from the ongoing research on MPEG-4 traffic simulation is that statistics can vary widely even in the case of similar material (e.g. coded movies). The only way of reaching an optimal fit involves the study of the particular source in question.

9 Conclusions

Future 3.5G systems have to be able to support current applications and new ones with different capacities and requirements. Hence, new classifications and characterisation parameters of these applications are needed, including a first assignment of the range of variation of the parameters. Parameters include data rate, tolerance to delay and error, busy hour call attempt, session duration and activity/inactivity models, among others. This is an important step in order to define realistic simulation traffic scenarios and models.

After presenting current views on operation environments, mobility models were defined for specific E-UMTS environments. A set of scenarios was drawn by associating values of service usage with each of the eight deployment scenarios. Nearly thirty services were considered, grouped into Sound, High Interactive Multimedia, Narrow-, Wide- and Broadband. They are examples of a mixture of applications that may exist in E-UMTS. However, a reduced set of applications is needed for simulations purposes. A selection of the most relevant service environments and applications was made in response to the need to diminish the burden on simulation work.

Then, a traffic generation model was described in order to allow quantification and description of traffic offered to the E-UMTS network. This model is based on population and service penetration values in order to determine call generation rates for the constituent services within each of the selected scenarios.

For each service an activity model was described, and the ON and OFF states were characterised by appropriate statistical distributions. Concerning streaming services, a model for streaming video was outlined, based on MPEG-4 trace statistics. This completes the basic output to the SEACORN simulation work, whose objective is to determine the aggregate traffic, packet error rates, detailed loss rates, and other parameters.

10 Acknowledgements

This work was partially funded by MULTIPLAN (a Portuguese Foundation for Science and Technology POSI project with FEDER funding), SAMURAI (a PRAI-Centro project with FEDER funding), and IST-SEACORN. The authors acknowledge the fruitful contributions from all IST-SEACORN partners.

References

1 *Seacorn*. April 29, 2005 [online] – URL: http://seacorn.ptinovacao.pt

2 Raychaudhuri, D. Topics in 4G Wireless Networks: Ad-Hoc Nets, Adaptive Services & QoS. In: *Proc. of Wireless, Mobile and Always Best Connected 1st International ANWIRE Workshop*, Glasgow, UK, April 2003.

3 Gustafson, E, Jonsson, A. Always Best Connected. *IEEE Wireless Communications*, 10 (1), 49–55, 2003.

4 Stordahl, K, Murphy, E. Forecasting Long-term Demand for services in the Residential Market. *IEEE Communications Magazine*, 33 (2), 44–49, 1995.

5 Olsen, B T et al. Techno-Economic Evaluation of Narrowband and Broadband Access Network Alternatives and Evolution Scenario Assessment. *IEEE Journal on Selected Areas in Communications*, 14 (6), 1184–1203, 1996.

6 ETSI. *Radio Equipment and Systems (RES); High Performance Radio Local Area Networks (HIPERLAN); Requirements and architectures for Wireless ATM Access and Interconnection.*

Sophia Antipolis, France, 1997. (ETSI TR 101.031 v.1.1.1)

7 UMTS. *UMTS/IMT-2000 Assessing Global Requirements for the Next Century*. London, UK, UMTS Forum, 1999. (Report No. 6)

8 Rokitansky, C H, Scheibenborgen, M (eds.). *Updated Version of SDD*. Brussels, Belgium, RACE Central Office, 1994. (RACE MBS Deliverable R2067/UA/WP 2.1.5/DS/P/68.b1)

9 Velez, F J, Correia, L M. Mobile Broadband Services: Classification, Characterisation and Deployment Scenarios. *IEEE Communications Magazine*, 40 (4), 2002.

10 Holma, H, Toskala, A. *WCDMA for UMTS*. Chichester, UK, John Wiley, 2001.

11 UMTS. *The UMTS Third Generation Market – Structuring the Service Revenues Opportunities*. London, UK, UMTS Forum, Sep. 2000. (Report No. 9)

12 San José, E R, Velez, F J. Enhanced UMTS Services and Applications: a perspective beyond 3G. In: *Proc. of EPMCC'2003 – 5th European Personal Mobile Communications Conference*, Glasgow, Scotland, April 2003.

13 Ferreira, J (ed.). *Classification of Mobile Multimedia Services*. Brussels, Belgium, IST Central Office, 2002. (IST SEACORN CEC deliverable 34900/PTIN/DS/ 011/b8)

14 Ahonen, T T, Barret, J. *Services for UMTS – Creating the Killer Application*. Chichester, UK, John Wiley, 2002.

15 Jedrzycki, C, Leung, V. Probability distributions of channel holding time in cellular telephony systems. In: *Proc. of VTC'96 – 46th Vehicular Technology Conference*, Atlanta, Georgia, USA, May 1996.

16 Kwok, T C. Residential Broadband Internet Services and Applications Requirements. *IEEE Communications Magazine*, 35 (6), 76–83, 1997.

17 Händel, R, Anber, M, Schröder, S. *ATM Networks, Concepts, Protocols, Applications*. New York, NY, Addison-Wesley, 1996.

18 Kwok, T C. A Vision for Residential Broadband Services: ATM-to-the-Home. *IEEE Network*, 9 (5), 14–28, 1995.

19 Laiho, J, Wacker, A, Novosad, T. *Radio Network Planning and Optimisation for UMTS*. Chichester, UK, John Wiley, 2002.

20 Ashby, S et al. *Final Report on MBS Applications and Services*. Brussels, Belgium, RACE Central Office, 1994. (RACE MBS Deliverable R2067/BTL/1.2.2/DS/R/056.b1)

21 UMTS. *The UMTS Third Generation Market – Phase II: Structuring the Service Revenue Opportunities*. London, UK, UMTS Forum, 2001. (Report No. 13)

22 ETSI. *Technical Report Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0)*. Sophia Antipolis, France, April 1998. (ETSI TR 101 112 V3.2.0)

23 Jabbari, B. Teletraffic Aspects of Evolving and Next-generation Wireless Communication Networks. *IEEE Personal Communications Magazine*, 3 (6), 4–9, 1996.

24 Guérin, R. Channel Occupancy Time Distribution in a Cellular Radio System. *IEEE Transactions on Vehicular Technology*, 36 (3), 89–99, 1987.

25 Chlebus, E, Ludwin, W. Is handoff traffic really Poissonean? In: *Proc. of IEEE ICUPC'95 – IEEE International Conference on Universal Personal Communications*, Tokyo, Japan, Nov. 1995.

26 Ferreira, J (ed.). *Enhanced UMTS deployment and mobility scenarios*. Brussels, Belgium, IST Central Office, 2002. (IST SEACORN CEC Deliverable 34900/PTIN/DS/013/b1)

27 3rd Generation Partnership Project TS 26.093: *Source Controlled Rate operation (Release 5).*

28 3rd Generation Partnership Project TS 26.101: *AMR Speech Codec Frame structure (Release 5).*

29 Cuny, R, Lakaniemi, A. VoIP in 3G Networks: An End-to-End Quality of Service Analysis. *IEEE VTC 2003 Spring – The 57th IEEE Semiannual Vehicular Technology Conference*, Jeju, Korea, April 2003.

30 Krunz, M, Hughes, H. A Traffic Model for MPEG-Coded VBR Streams. *Performance Evaluation Review* (In: *Proc. of the ACM SIGMETRICS'95*), 47–55, 1995.

31 Matrawy, A, Lambadaris, I, Huang, C. MPEG4 Traffic Modeling Using The Transform Expand Sample Methodology. In: *Proc. of IWNA4 – The 4th IEEE International Workshop on Networked Appliances*, Gaithersburg, MD, Jan. 2002.

List of acronyms

3GPP	3rd Generation Partnership Project
ABC	Always Best Connected
ABR	Available bit rate
AMR	Adaptive multi rate
Asy	Asymmetric
ATM	Asynchronous transfer mode
ATR	Assistance in Travel
BACK	Background
BCC	Business city centre
BER	Bit error rate
BHCA	Busy hour call attempt
Bid	Bi-directional
CBD	City business district
CBR	Constant bit rate
CI	Customized Infotainment
COM	Commercial zones
CONV	Conversational
DTX	Discontinuous transmission
ETSI-RES	European Telecommunications Standard Institute – Radio Equipment and Systems
E-UMTS	Enhanced UMTS
FTP	File Transfer Protocol
HD	High definition
HDT	HD Video-telephony
HDTV	HD Television
HIMM	High Interactive MM
HOM	Home
HSDPA	High-speed downlink packet access
IMM	Instant Messaging for MM
IND	Industry
INTR	Interactive
IP	Internet protocol
IPv4	Internet protocol – version 4
IPv6	Internet protocol – version 6
ISO	Isochronous
IST-SEACORN	Simulation of Enhanced UMTS Access and Core Networks
ITU-T	International Telecommunications Union – Telecommunications sector
LBS	Location Based Services
MAC	Medium access and control
MIA	Mobile Internet Access
MIEA	Mobile Internet/Extranet Access
MM	Multimedia
MMS	MM Message Service
MPEG	Moving pictures expert group
MWB	MM Web Browsing
NISO	Non-ISO
NRT	Non-RT
NTB	non-TB
OFDM	Orthogonal Frequency Division Multiplexing
OFF	Offices
PDP	Packet Data Protocol
PDU	Packet data unit
QoS	Quality of Service
RACE	Research and Development on Advanced Communications Technologies in Europe
RACE-MBS	RACE Mobile Broadband System
RACE-TITAN	RACE Tool for Introduction Scenarios and Techno-Economic Evaluation of the Access
ROA	Roads
RT	Real-time
RTP	Real-time Transport Protocol
RV	Rich voice
SID	Silence descriptor
STR	Streaming
Sym	Symmetric
TB	Time-based
TCP	Transmission Control Protocol
TRA	Trains
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
Und	Unidirectional
URB	Urban
UWB	Ultra Wideband
VEH	Vehicular
VOI	Voice
VoIP	Voice over IP
VTE	Video-telephony
WCDMA	Wideband CDMA, Code Division Multiple Access
WLAN	Wireless LAN, Local Area Network
WLI	WLAN Interconnection

Jaime Ferreira (51) is a Researcher at Portugal Telecom Inovação. He graduated in electrical engineering from Universidade do Porto in 1975. His current focus is on service design, evolution of service architecture and models for wireless traffic. He represented Portugal Telecom in 3GPP and authored several papers and presentations for IEE, IST and EURESCOM events. Mr. Ferreira contributed to operator initiatives "4G, the next frontier" and "The operator's vision on systems beyond 3G". He was contributor to Enhanced UMTS service aspects of the SEACORN project and is currently leading the EURESCOM project TRAWIS, "Traffic models for the new wireless services" and TIMES, "The inter-operator IM and mobile IM service".

email: jaime@ptinovacao.pt

Fernando J. Velez (35) received the Licenciado, MSc and PhD degrees in Electrical and Computer Engineering from Instituto Superior Técnico, Technical University of Lisbon in 1993, 1996 and 2001, respectively. Since 1995 he has been with the Department of Electromechanical Engineering of University of Beira Interior, Covilhã, Portugal, where he is assistant professor. He is also researcher at Instituto de Telecomunicações, Lisbon. He made or makes part of the teams of RACE/MBS, ACTS/SAMBA, COST 259, COST 273, COST 290 and IST/SEACORN European projects, and he is the coordinator of four Portuguese projects: SAMURAI, MULTIPLAN, CROSSNET, and MobileMAN. He has authored more than thirty papers and communications in international journals and conferences and is a member of IEEE. His main research areas are cellular planning tools, traffic from mobility, multi-service traffic and cost/revenue performance of advanced mobile communication systems.

email: fjv@ubi.pt

Analysis of QoS in WLAN

PAAL E. ENGELSTAD AND OLAV N. ØSTERBØ

Paal E. Engelstad is Research Scientist in Telenor R&D

Olav N Østerbø is Senior Research Scientist in Telenor R&D

An analytical model is proposed to describe the priority schemes of the EDCA mechanism of the IEEE 802.11e standard. EDCA provides class-based differentiated QoS to IEEE 802.11 WLANs. The main contribution of this paper as opposed to other works, is that the model predicts the throughput, delay and frame dropping probabilities of the different traffic classes in the whole range from a lightly loaded, non-saturated channel to a heavily congested, saturated medium. Furthermore, the model describes differentiation based on different AIFS-values, in addition to the other adjustable parameters (i.e. window sizes, retransmission limits, TXOP lengths, etc.) also encompassed by previous models. AIFS differentiation is described by a simple equation that enables access points to predict at which traffic loads starvation of a traffic class will occur. Moreover, virtual collision handling is included into the model. We show how this part of the model can also be used to model the performance of the Virtual Collision Handler. The model is calculated numerically and validated against simulation results. We observed a good match between the analytical model and simulations.

1 Introduction

During recent years the IEEE 802.11 WLAN standard [1] has been widely deployed as the most preferred wireless access technology in office environments, in public hot-spots and in the homes. Due to the inherent capacity limitations of wireless technologies, the 802.11 WLAN easily becomes a bottleneck for communication. In these cases, the QoS features of the 802.11e standard will be beneficial to prioritize for example voice and video traffic over more elastic data traffic.

The IEEE 802.11 medium access control (MAC) comprises the mandatory Distributed Coordination Function (DCF) as a contention-based access scheme, and the optional Point Coordination Function (PCF) as a centrally controlled polling scheme. However, PCF is hardly implemented in any products, and DCF represents the commonly used MAC mechanism of 802.11. DCF adopts carrier sense multiple access ("listen-before-talk") with collision avoidance (CSMA/CA) and uses binary exponential backoff. A station not only goes into backoff upon collision. It also carries out a "post-backoff" after having transmitted a packet, to allow other stations to access the channel before it transmits the next packet.

The IEEE 802.11e standard [2] works as an extension to the 802.11 standard, and the Hybrid Coordination Function (HCF) is used for medium access control. HCF comprises the contention-based Enhanced Distributed Channel Access (EDCA) as an extension for DCF, and the centrally controlled Hybrid Coordinated Channel Access (HCCA) as a replacement for PCF. EDCA has received most attention recently, and it seems that this is the WLAN QoS mechanism that will be promoted by the majority of vendors. EDCA is therefore the area of interest of this paper, and HCCA will not be discussed any further here.

EDCA enhances DCF by allowing four different access categories (ACs) at each station and a transmission queue associated with each AC. Each AC at a station has a conceptual module responsible for channel access for each AC and in this paper the module is referred to as a "backoff instance". Hence each of the four transmission queues (and the associated ACs) on a station is represented by one backoff instance. The channel access between different backoff instances on a station is not completely independent due to the virtual collision handling between the queues on the station. If two or more backoff instances on the same station try to access the channel in the same timeslot, the station attempts to transmit the frame of the highest priority AC, while the lower priority frames will go through backoff.

The traffic class differentiation of EDCA is based on assigning different access parameters to different ACs. First and foremost, a high-priority AC, *i*, is assigned a minimum contention window, $W_{0,i}$, that is lower than (or at worst equal to) that of a lower-priority AC. At a highly loaded (or "saturated") medium, the post-backoff of the high-priority AC will normally be smaller than the post-backoff of a low-priority AC. This results in an average higher share of the channel capacity, because the high-priority AC will on average have to refrain from the channel for a shorter period of time than the low priority AC.

Another important parameter setting is the AIFS value, measured as a Short Interframe Space (SIFS) pluss an AIFSN number of timeslots. A high-priority AC is assigned an AIFSN that is lower than (or at

ISSN 0085-7130 © Telenor ASA 2005

worst equal to) the AIFSN of a lower-priority AC. The most important effect of the AIFSN setting is that the high-priority AC normally will be able to start earlier than a low priority AC to decrement the backoff counter after having been interrupted by a transmission on the channel. At a highly loaded channel where the decrementing of the backoff counter will be interrupted by packet transmissions a large number of times, the backoff countdown of the high-priority AC will occur at a higher average speed than that of the lower-priority AC. As the wireless medium gets more and more congested, the average number of empty timeslots between the frames transmitted by the higher-priority ACs might be lower than the AIFSN value of the low-priority AC. At this point, the AC will not be able to decrement its backoff counter, and all packets will finally be dropped instead of being transmitted. This is referred to as "starvation".

Other differentiation parameters that may be adjusted in 802.11e (and which are also explicitly or implicitly included in the model proposed below) are the retry limit, L_i (of short and long packets), the maximum contention window W_i, L_i and the TXOP-limit of each AC, i.

Most of the recent analytical work on the performance 802.11e EDCA stems from the simple and fairly accurate model proposed by Bianchi [3] to calculate saturation throughput of 802.11 DCF. Later, Ziouva and Antonakopoulos [4] improved the model to find saturation delays, however, still of the undifferentiated DCF. They also improved the model by stopping the backoff counter during busy slots, which is more consistent with the IEEE 802.11 standard.

Based on this work, Xiao [5] extended the model to the prioritized schemes provided by 802.11e by introducing multiple ACs with distinct parameter settings, such as the minimum and maximum contention window. It is also straightforward to extend the model for the use of different TXOP sizes of different classes. Furthermore, this prioritized model also introduced a finite retry limit. This additional differentiation parameter leads to more accurate results than previous models. (A list of references for other relevant efforts and model improvements of DCF can also be found in [5].)

A differentiation parameter lacking in Xiao's model, however, is the important AIFS parameter. Xiao assumed equal AIFSN of all traffic classes. (In fact, a situation with only two different ACs was analyzed in the work.)

Furthermore, the model does not correctly capture starvation. In many cases the QoS-enabled Access Point (QAP) would need to predict when the starvation of an AC will occur, so that it will be able to know when to change the current parameter settings, e.g. to avoid that any AC is completely starved. The main problem is that the QAP cannot know that an AC is starved simply by observing that it does not receive any traffic of that AC. The reason for not receiving the traffic might just as well be that the other stations, QSTAs, are accidentally not transmitting any traffic of that AC. By having starvation correctly descibed by the model, the QAP has a means to predict at which traffic loads starvation of an AC will occur. It can predict starvation simply by measuring the traffic load on the channel. The QAP might also be interested in predicting at which traffic levels the Virtual Collision Handler (VCH) will start to starve a traffic class.

The model is also limited, as a fully saturated channel is assumed. Due to the bursty characteristics of many types of data traffic, it is unlikely that the channel will be fully saturated all the time. The rate adaptation of TCP, for example, will often ensure that the total channel load will not be fully saturated. Hence, in many cases an access point will be more interested in knowing how to set parameters for a lightly saturated channel, and to adjust these parameters dynamically in this region. An analytical model that covers the full range from a non-saturated to a fully saturated channel would be more useful.

This paper extends Xiao's model to enhance it in a number of ways:

- The presented model predicts the performance not only in the saturated case, but on the whole range from an unsaturated medium to a fully saturated channel. (Some works, such as [6] and [7] have explored unsaturated conditions, however, only of the one-class 802.11. They are also primarily focussed on the non-saturation part instead of finding a good descriptive solution for the whole range.)

- In the non-saturation situation, our model accounts for "post-backoff" of an AC, although the queue is empty, according to the IEEE 802.11 standard. If the packet arrives in the queue after the "post-backoff" is completed, the listen-before-talk (or CSMA) feature of 802.11 is also incorporated in the model.

- Our model describes the use of AIFSN as a differentiating parameter, in addition to the other differentiation parameters encompassed by Xiao's efforts and other works.

- A simple closed-form equation that predicts with satisfactory accuracy the starvation point (or “freeze point”) of each traffic class is provided. The only prerequisite for a station (e.g. an access point) to determine that starvation of an AC has occurred, is to know the AIFSN value of the AC and to monitor the traffic load on the channel.

- Virtual Collision Handling is treated in the model.

The model is validated against simulation. Both uplink and downlink traffic scenarios are used. The majority of other works that do analytical performance evaluations, empirical simulations and/or validations between analytical numerical results and simulations, seem to focus only on the uplink traffic problem. They present results with a number of stations contending for the channel, and with fairly equal shares of traffic allocated to each station and to each AC. In this paper, we do similar uplink analysis as in earlier works. However, in addition we look at downlink traffic scenarios, which might be important in many real-life scenarios.

The remaining part of the paper is organized as follows: The next section summarizes the differentiation parameters of 802.11e and provides the basis for understanding the analytical model. Section III presents the analytical model with virtual collision handling and AIFS differentiation. Expressions for delays and estimation of delay-dependent traffic parameters are calculated in Section IV. In Section V, expressions for the throughput (with or without virtual collision handling) are found. Then, a section is allocated to the validation of the model against simulations. Both uplink and downlink traffic scenarios are considered. Our findings are finally summarized in our concluding remarks.

II Important differentiation parameters of 802.11E

A Priority based on Contention Windows (CWs) and Exponential Backoff

For each AC, $i (i = 0, ..., 3)$, we let $W_{i,j}$ denote the contention window size in the j-th backoff stage, i.e. after the j-th unsuccessful transmission; hence $W_{i,0} = CW_{i,\min} + 1$, where the recommended values for $CW_{i,\min}$ are listed in Table 1. We also denote $j = m_i$ as the j-th backoff stage where the contention window has reached $CW_{i,\max} + 1$; the window will no longer be increased in the subsequent backoff stages. Hence, $m_i = \log_2((CW_{i,\max} + 1) / CW_{i,\min} + 1)$. Finally, we let L_i denote the retry limit of the retry counter; if the transmission is unsuccessful after the L_i-th backoff stage, the packet will be dropped. The 802.11 specification allows for different retry limits, Dot11ShortRetryLimit and Dot11LongRetryLimit, for packets that are longer than and shorter than the Dot11RTSThreshold, respectively. In this paper, however, we will assume that the sizes of all packets of a class i are either below or above the Dot11RTSTreshold-parameter, so that L_i is equal for all packets belonging to the same class.

	AC[3]	AC[2]	AC[1]	AC[0]
AIFSN	2	2	3	7
CWmin	3	7	15	15
CWmax	15	31	1023	1023
Retry Limit (long/short)	7/4	7/4	7/4	7/4

Table 1 Recommended (default) parameter settings for 802.11e

$$W_{i,j} = \begin{cases} 2^j W_{i,0}; & j = 0, 1, .., m_i - 1 \\ 2^{m_i} W_{i,0} = CW_{i,\max}; & j = m_i, ..., L_i \end{cases} \quad (1)$$

In the special case where $m_i \leq L_i$, Eq. (1) is reduced to $W_{i,j} = 2^j W_{i,0}$ for $j = 0, 1, ..., L_i$.

B Priority based on Inter-Frame Spaces (IFSs)

When a backoff instance senses that the channel is idle after a packet transmission, it normally waits a guard time called the Distributed Inter-Frame Space (DIFS) during which it is not allowed to transmit packets or do backoff countdown. The duration of DIFS is the sum of a SIFS and two time slots. The two time slots of DIFS allow the Hybrid Coordinator (HC) on the QAP (or Point Coordinator with only plain 802.11) to access the channel with higher priority. The HC is allowed to enter the channel after only waiting one time slot (in addition to SIFS) and it does not need to go through “post-backoff” before accessing the channel. Moreover, certain packets – including the Clear-To-Send (CTS) and Acknowledgement (ACK) packets – can be sent after only waiting SIFS. This gives maximum priority to this traffic, and ensures that a data exchange (such as a data transmission followed by an ACK) can be considered nearly an atomic transaction. Instead of using DIFS, each AC[i] of 802.11e uses an Arbitration Inter-Frame Space (AIFS[i]) that consists of a SIFS and an AIFSN[i] number of additional time slots. In this paper we define A_i as:

$$A_i = AIFS[i] - \min(AIFS[i]) \geq 0 \quad i = 0, ..., N - 1 \quad (2)$$

where N is the number of different ACs (i.e. normally four). The 802.11e standard mandates that $AIFSN[i] \geq 2$, where the minimum limit of 2 slots corresponds to the DIFS interval. The use of AIFSN to differentiate between ACs has two consequences. Assume two backoff instances with different AIFSNs. Both have a packet to send, but the channel is busy, so they have to wait. The first effect occurs when neither of the two backoff instances is in backoff; i.e. they are not in binary backoff and the post-backoff is completed (or the post-backoff is not necessary because it is the first packet ever to be sent). In this case, when the channel is sensed idle, the backoff instance with the lowest AIFSN is the first to be allowed to make a transmission attempt. The other backoff instance will sense that the channel is busy during this time slot. It will have to wait until the transmission attempt of the first backoff instance is completed before it is allowed to transmit the packet.

The second effect occurs when both backoff instances are in backoff. To describe the effect, assume that the backoff counters of both backoff instances are equal. Each time a packet transmission is completed, the backoff continues the countdown of idle time slots as soon as the AIFS interval is completed. However, after the packet transmission, the backoff instance with the lowest AIFSN will start counting down idle time slots before the other backoff instance is allowed to enter the count down procedure. Thus, the backoff instance with the lowest AIFSN will be able to count down the backoff window faster than the backoff instance with higher AIFSN.

B Priority based on Transmission Opportunities (TXOPs)

Priority based on differentiated Transmission Opportunities is not treated explicitly in this paper. For simplicity and to keep focus on the most important issues, we have assumed that all traffic classes send packets of equal lengths (i.e. of 1024 bytes), and that each packet fits perfectly into one TXOP. Calculating the model with respect to different packet lengths is easy, as shown by Xiao [5]. It is also easy to extend our analysis to contention-free bursting (CFB) within a TXOP by summing up the different SIFS occurring between subsequent packets of the burst.

III Analytical model

A Priority based on Transmission Opportunities (TXOPs)

Figure 1 illustrates the Markov chain for the transmission process of a backoff instance of priority class i. The state changes in the figure occur only when the backoff instance is able to contend for the channel. If one or more stations transmit in a time slot, that slot is sensed busy by the backoff instances. For the subsequent time that the packet is under transmission, the backoff instance remains in the same state. The Markov process finally resumes at the first slot where the channel again is open for contention, i.e. after the transmission (or collision) is completed and after the real-time duration of an additional DIFS. Hence the (virtual) time scale is discrete and integral, and a backoff slot that is open for contention triggers each “clock tick”. Measured in real-time the duration of a slot that leads to transmission is equal to the length of the transmission including the RTS and CTS (if being used), the data packet and the ACK as well as all the associated inter-frame spaces. The real-time duration of an empty slot is equal to the duration of a regular backoff slot.

In the Markov chain, the utilization factor, ρ_i, represents the probability that there is a packet waiting in the transmission queue of the backoff instance of AC i at the time a transmission is completed (or a packet dropped). Now, the backoff selects a backoff interval k at random and goes into post-backoff. If the queue is empty, at a probability $1 - \rho_i$, the post-backoff is started by entering the state $(i, 0, k, e)$. If the queue on the other hand is non-empty, the post-backoff is started by entering the state $(i, 0, k)$. Hence, ρ_i balances the fully non-saturated situation with the fully saturated situation and therefore plays a role to model the behaviour of the intermediate semi-saturated situation. We see that when ρ_i –> 1 the Markov chain behaviour approaches that of the non-saturation case similar to the one presented by Xiao [5]. On the other hand, when ρ_i –> 0 the Markov chain models a stochastic process where the backoff instance after transmission always goes into “post-backoff” without a packet to send.

As mentioned above, the states $(i, 0, k, e)$ in the upper row are entered when the channel is not fully saturated and when the queue of a backoff instance is empty at the time a transmission is completed (or a packet dropped). The states $(i, 0, 1, e)$, ..., $(i, 0, W_{i,0} - 1, e)$ represent a situation where the transmission queue is empty, but the station is counting down backoff slots. The probability that a backoff instance of AC i is sensing the channel busy and is thus unable to count down the backoff slot from one timeslot to the other is denoted by the probability p_i^*. (We use the asterisk to indicate that this is a probability related to the countdown process.) It undertakes a countdown at probability $1 - p_i^*$and moves to another state. If it has received a packet while in the previous state at a probability q_i^*, it moves to a corresponding state in the second row with a packet waiting for transmission. Otherwise, it remains in the first row with no packets waiting for transmission.

While in the state $(i, 0, 0, e)$, on the contrary, the backoff instance has completed post-backoff and is only waiting for a packet to arrive in the queue. If it

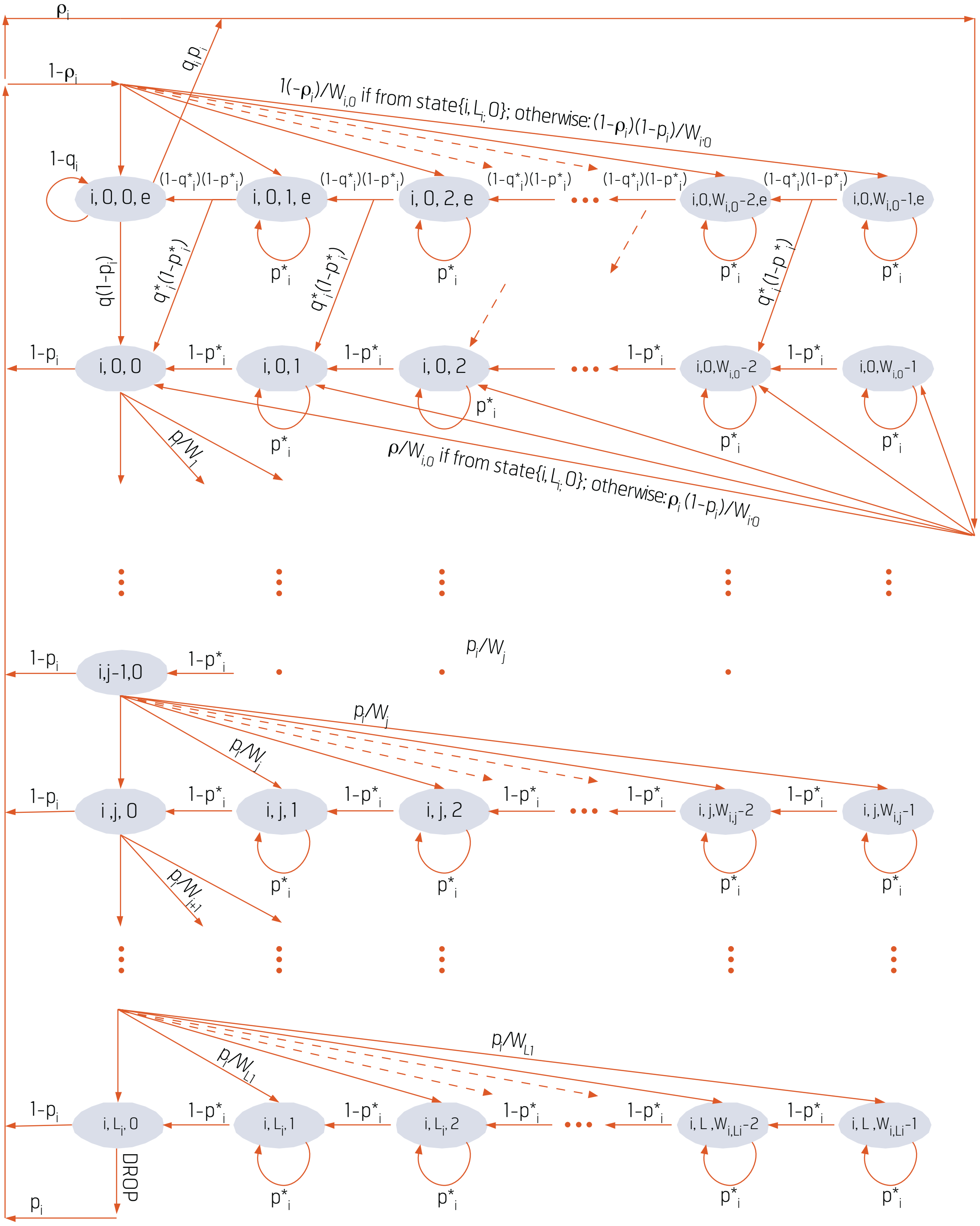


Figure 1 Markov Chain (both saturation and non-saturation)

receives a packet during a timeslot at a probability q_i, it does a "listen-before-talk" channel sensing and moves to a new state in the second row, since a packet is now ready to be sent. If the backoff instance senses the channel busy, at a probability p_i, it performs a new backoff. Otherwise, it moves to state $(i, 0, 0, e)$ to do a transmission attempt. The transmission succeeds at a probability $1 - p_i$. Otherwise, it doubles the contention window and goes into another backoff.

All other rows apart from the first row in the figure illustrate a situation with at least one packet in the system. Indeed, only these states are entered in the extreme case of a fully saturated channel and a non-zero traffic load on each AC, so that the transmission queue is always full. Hence, after successful transmission or after a packet has been dropped, the backoff instance proceeds directly into one of the post-backoff states $(i, 0, k)$ (for $k = 0, 1, ..., W_{i,j} - 1$). For each unsuccessful transmission attempt, the backoff instance moves to a state in a row below at a probability p_i. However, if the packet has not been successfully transmitted after $L_i + 1$ attempts, the packet is dropped. Hence, the accumulated frame dropping probability, $P_{i,drop}$, can be estimated as:

$$P_{i,drop} = p_i^{L_i+1} \tag{3}$$

Let $b_{i,j,k}$ denote the state distributions. Since the probability of a transmission attempt entering stage j (where $j = 0, 1, ..., L_i$) is p_i^j, chain regularities yield:

$$b_{i,j,0} = p_i^j \, b_{i,0,0}\,; \; j = 0, 1, ..., L_i \tag{4}$$

Furthermore, we observe that a backoff instance transmits when it is in any of the states $(i, j, 0)$ where $j = 0, 1, ..., L_i$. Hence, if we let τ_i denote the transmission probability (i.e. the probability that a backoff instance in priority class i transmits during a generic slot time, independent on whether the transmission results in a collision or not), we have:

$$\tau_i = \sum_{j=0}^{L_i} b_{i,j,0} = b_{i,0,0} \frac{1 - p_i^{L_i+1}}{1 - p_i} \tag{5}$$

In the following subsection we will find ways to express $b_{i,j,0}$ and p_i in terms of τ_i. Hence, a complete description of the system can be found by solving the above set of equations (one equation per AC i).

B Markov chain analysis

First we will look at the post-backoff stage of the Markov chain for $j = 0$. From chain regularities, we observe that:

$$b_{i,0,0} = b_{i,L_i,0} + \sum_{j=0}^{L_i-1} (1 - p_i) b_{i,j,0} \tag{6}$$

By working recursively through the chain from right to left in the upper row, we get:

$$b_{i,0,k,e} = \frac{(1-\rho_i) b_{i,0,0}}{W_{i,0}(1 - p_i^*)} \frac{1 - (1 - q_i^*)^{W_{i,0}-k}}{q_i^*}; \tag{7}$$

$$k = 1, 2, ..., W_{i,0} - 1$$

Furthermore, we see that:

$$b_{i,0,0,e} = \frac{(1-\rho_i) b_{i,0,0}}{W_{i,0} q_i} \frac{1 - (1 - q_i^*)^{W_{i,0}}}{q_i^*} \tag{8}$$

From the upper left part of the Markov diagram we see that $b_{i,0,W_{i,0}\text{-}1} = b_{i,0,0} / ((1 - p_i^*)W_{i,0})$. By working recursively and horizontally through the chain we also observe that:

$$b_{i,0,k} = \frac{W_{i,k} - k}{W_{i,0}(1 - p_i^*)} (b_{i,0,0} + q_i p_i b_{i,0,k,e}) - b_{i,0,k,e} \tag{9}$$

$$\text{for} \;\; k = 1, 2, ..., W_{i,0} - 1$$

Undertaking the same analysis for the rest of the chain, we get:

$$b_{i,j,k} = \frac{W_{i,j} - k}{W_{i,j}(1 - p_i^*)} p_i^j b_{i,0,0}; \tag{10}$$

$$j = 1, ..., L_i \text{ and } k = 1, 2, ..., W_{i,0} - 1$$

Finally, the normalization requires that:

$$\sum_{k=0}^{W_{i,0}-1} b_{i,0,k,e} + \sum_{j=0}^{L_i} \sum_{k=0}^{W_{i,j}-1} b_{i,j,k} = 1 \tag{11}$$

This yields:

$$\frac{1}{b_{i,0,0}} = \sum_{j=0}^{L_i} \left[1 + \frac{1}{1 - p_i^*} \sum_{k=0}^{W_{i,j}-k} \frac{W_{i,j} - k}{W_{i,j}} \right] p_i^j$$

$$+ \frac{1 - \rho_i}{1 - q_i} \frac{1 - (1 - q_i^*)^{W_{i,0}}}{W_{i,0} q_i^*} \left(1 + \frac{(W_{i,0} - 1) q_i p_i}{2(1 - p_i)} \right) \tag{12}$$

The first sum in the equation above represents the saturation-part, while the second part is the dominant term under non-saturation. Hence, the expression provides a unified model encompassing all channel loads from a lightly loaded non-saturated channel, to a highly congested, saturated medium. This full-scale model will be validated below.

The non-saturation part of the expression might require further explanation. First, the factor $(1 - \rho_i)$ (where ρ_i is the utilisation factor of the backoff instance) represents the probability of having an empty queue after successful transmission. If this happens, it enters the empty-queue post-backoff procedure, which is represented by the states $(i, 0, 0, e)$, …, $(i, 0, W_{i,0} - 1, e)$. Second, the geometric sum $1 - (1 - q_i^*)^{W_{i,0}} / (W_{i,0}\, q_i^*)$ expresses the probability of not receiving any packets in the transmission queue while performing the complete empty-queue post-backoff. In other words, it is the probability of finally ending in the state $(i, 0, 0, e)$, instead of transitioning to any of the regular post-backoff states (i, 0, 0), …, $(i, 0, W_{i,0} - 2)$ where a packet is waiting to be transmitted. In the geometric sum, q_i^* is the probability that such a transition will take place between any of the countdowns. Hence, q_i^* is the probability of receiving a packet during the time-scale of one slot that is counted down. (The asterisk denote that the probability is associated with the countdown process.) Third, while waiting for a packet in the state $(i, 0, 0, e)$, q_i (without the asterisk) represents the traffic generation probability. With a lightly loaded channel, the factor $1 / q_i$ will be the dominant part of the equation. At low loads the factor ensures the typical non-saturation behaviour where successfully transmitted traffic equals the traffic entering the transmission queue. Finally, the factor $(1 + (W_{i,0} - 1)\, q_i p_i / 2(1 - p_i))$ appears as a consequence of the listen-before-talk test in state $(i, 0, 0, e)$ (which can be replaced with 1 if the test is not implemented).

We also note that the saturation part of the equation can be written out. Performing the summation, we write the first sum as eq. (13), see overleaf.

$$\sum_{j=0}^{L_i}\left[1+\frac{1}{1-p_i^*}\sum_{k=0}^{W_{i,j}-k}\frac{W_{i,j}-k}{W_{i,j}}\right]p_i^j=\begin{cases}\frac{(1-2p_i^*)(1-2p_i)(1-p_i^{L_i+1})+W_{i,0}(1-p_i)(1-(2p_i)^{L_i+1})}{2(1-p_i^*)(1-p_i)(1-2p_i)}; & m_i>L_i\\ \frac{(1-2p_i^*)(1-2p_i)(1-p_i^{L_i+1})+W_{i,0}\left[(1-p_i)(1-(2p_i)^{m_i})+(1-2p_i)(2p_i)^{m_i}(1-p_i^{L_i-m_i+1}\right]}{2(1-p_i^*)(1-p_i)(1-2p_i)}; & m_i\le L_i\end{cases}\qquad(13)$$

C Modelling Probabilities without Virtual Collision Handling

We let p_b denote the probability that the channel is busy. Since this means that at least one backoff instance transmits during a slot time, we have:

$$p_b=1-\prod_{i=0}^{N-1}(1-\tau_i)^{n_i}\qquad(14)$$

Here, n_i denotes the number of backoff instances contending for channel access in each priority class i, and N denotes the total number of classes.

The probability of unsuccessful transmission p_i from one specific backoff instance (as described in the Markov chain), requires that at least one of the other backoff instances does transmit in the same slot:

$$p_i=1-\prod_{c=0,c\neq i}^{N-1}(1-\tau_c)^{n_c}=1-\frac{1-p_b}{1-\tau_i}\qquad(15)$$

D Probabilities with Virtual Collision Handling

It is possible to make modifications to take virtual collisions into account in the analytical model. Consider for example a situation with n stations and four active transmission queues on each station. A backoff instance can transmit packets if other backoff instances do not transmit, *except* the backoff instances of the lower priority ACs *on the same QSTA*. The reason for this exception is that the virtual collision handling mechanism ensures that upon virtual collision the higher priority AC will be attempted for transmission while the colliding lower priority traffic goes into backoff. This can be generalized by the expression:

$$p_i=1-\frac{\prod_{c=0}^{N-1}(1-\tau_c)^{n_c}}{\prod_{c=0}^{i}(1-\tau_c)}\qquad(16)$$

If this expression is replaced with our original expression in (15), virtual collisions should be correctly incorporated in the model.

E Modelling probabilities of the Virtual Collision Handler itself

One may use exactly the same analytical model to study the behaviour of the Virtual Collision Handler (VCH). Here the VCH represents the channel, while there are only N (typically four) queues contending for access; i.e. one queue per AC i. Hence, one may model the throughput of the Virtual Collision Handler by setting $n_i = 1$ for all i. In this case, Eq. (14) is simply replaced by:

$$p_b=1-\prod_{i=0}^{N-1}(1-\tau_i)\qquad(17)$$

and Eq. (15) is replaced by:

$$p_i=1-\frac{(1-p_b)}{\prod_{c=0}^{i}(1-\tau_c)}=1-\prod_{c=i+1}^{N-1}(1-\tau_c)\qquad(18)$$

Here we note that the highest priority class will correctly have $p_{N-1} = 0$, which means that it is never blocked and never experiences a collision when it tries to access the channel for transmission.

Like before, N denotes the total number of classes. Note that the expressions typically describe the special case with only one station contending for the channel, which is the case when the channel refers to the "virtual" channel represented by the VCH. These expressions will be useful for analysis of downlink traffic scenarios later in this paper.

F The backoff countdown rate with AIFS differentiation

At this point, we are ready to find all transmission probabilities τ_i ($i = 0, 1, 2, 3$) for the saturation condition (i.e. by setting $\rho_i = 1$), without AIFS differentiation (by setting $p_i^* = p_i$). However, first we will incorporate AIFS differentiation in this subsection. Then, in the next subsection we will subsequently find expressions for ρ_i, q_i, in order to later be able to find transmission probabilities also under non-saturation conditions.

We see that the model presented so far does not encompass AIFS differentiation. In the following, however, we will include AIFS differentiation as an integral part of the count-down blocking probability, p_i^*.

The probability that a backoff instance senses a generic slot as idle is denoted p_i^* in the Markov chain. Without AIFS differentiation, it equals the probability that all other stations do not transmit, p_i:

$$p_{i,A_i=0}^*=p_i=1-\frac{1-p_b}{1-\tau_i}\qquad(19)$$

In this article, however, we argue that IFS-differentiation can be modelled with pretty good accuracy by adjusting the countdown blocking probability p_i^*. For the highest priority AC, AC[3], we set $p_3^* = p_3$. For lower priority ACs with a higher AIFS (for which $A_i \geq 1$) we reduce their countdown rate correspondingly. The additional A_i slots, where lower priority backoff instances of class i have to suspend the back-

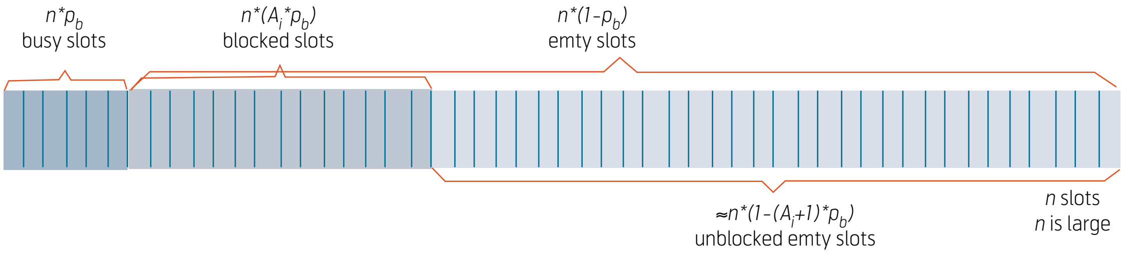


Figure 2 Simplified illustration of the principle of AIFS differentiation

off countdown, are modelled as being smeared out randomly and distributed uniformly over all slots. This assumption is required to be able to treat the A_i slots within the Markov model. (In reality, an A_i slot will only occur after another A_i slot, or after a successful or unsuccessful packet transmission.) Using this simple assumption, it is possible to "scale down" the probability of detecting an empty slot.

This down-scaling can be illustrated in Figure 2. Here a large number of n slots are grouped into three groups; busy slots (due to successful transmissions or collisions), empty slots that are blocked (due to the AIFSN setting of the AC in consideration) and other empty slots that are not blocked. Each busy slot leads to a proportional share of empty slots being blocked, where A_i yields the proportion of blocked slots for the AC in question.

By replacing p_b with the new "scaled" expression $(A_i + 1)p_b$, we derive a new "scaled" expression for p_i, which is denoted p_i^*:

$$(1 - p_i^*) = \frac{1 - (A_i + 1)p_b}{1 - \tau_i} \qquad (20)$$

However, to maintain consistency in the model, a minimum bound is introduced:

$$p_i^* = \min(1, p_i + \frac{A_i p_b}{1 - \tau_i}) \qquad (21)$$

Thus, starvation occurs when $p_i^* \geq 1$ in Eq. (17) or when $p_i^* = 1$ in Eq. (18). Since at this point $\tau_i \to 0$, starvation can be roughly predicted to occur when:

$$p_b \geq \frac{1}{1 + A_i} \qquad (22)$$

Although this approximate expression seems rather rough, its usefulness is striking. In the semi-saturated case, an AP is normally interested in adjusting channel access parameters (such as contention windows, AIFSN etc.) for each traffic class to control the share of the channel allocated to each AC. By means of the above expression, the AP can simply predict from the traffic load that it pours into the transmission queues, whether any AC will face starvation when the traffic is handled by the VCH.

An alternative approach is to include AIFS countdown explicitly in the Markov chain [11]. However, the model becomes all too complex to be useful except for the simplest settings of the AIFS parameters. For example, in [11] the lowest priority AC is configured with AIFSN = 3, while for all other ACs the AIFS is set equal to DIFS. More flexible configurations of the AIFSNs seem complicated within this framework.

IV Estimation of the traffic parameters ρ_i, q_i and q_i^*

A Delay and Service Time under saturation conditions and infinite queues

In order to be able to determine an expression for ρ_i we first need expressions for the packet delays under both saturated and non-saturated conditions. Under saturation conditions, the queue is always full of packets ready to be transmitted (i.e. the utilization, ρ, of the queue is equal to 1).

To study delay under these conditions, we first deal with the delay associated with counting down backoff slots for the packets to be transmitted. The probability of a successful transmission exactly in the j-th stage is $p_i^j(1 - p_i)$. In each stage, h, the average countdown delay is $T_c * (W_{i,h} - 1) / 2$, and the accumulated delay for a packet sent in the j-th stage is found by summing all h stages up till j. In summary, the expected countdown delay, $\bar{D}_i^{CD}$, is:

$$\overline{D}_i^{CD} = T_e \sum_{j=1}^{L_i} p_i^j (1 - p_i) \sum_{h=0}^{j} \frac{W_{i,h} - 1}{2} \qquad (23)$$

While the backoff instance is counting down, the probability of facing an empty slot is $1 - p_i^*$ while the probability of being blocked is p_i^*. Hence, $p_i^* / (1 - p_i^*)$ represents the share of slots where the countdown process is being blocked. While being blocked, the average delay is $p_s T_s + (p_b - p_s) T_c$. In summary, the expected blocking delay, $\bar{D}_i^B$, is:

$$\overline{D}_i^B = \overline{D}_i^{CD} \frac{p_i^*}{1 - p_i^*} \frac{\left(\frac{p_s}{p_b} T_s + \left(1 - \frac{p_s}{p_b}\right) T_c\right)}{T_e} \qquad (24)$$

ISSN 0085-7130 © Telenor ASA 2005

Furthermore, each time a backoff instance goes from the $(h-1)$-th to the h-th stage, there is a collision delay, $T_c^* = T_c + T_{TO}$, associated with the transmission attempt that a backoff instance must wait after experiencing a collision before it can sense the channel idle again. (Here, T_{TO} is simply defined as the difference between the collision delay, T_c^*, of a station that participates in a collision and that of a station not participating, T_c. When validating against simulations, its setting should reflect the implementation of the simulator.) The number of retry attempts is found by finding the average of j. In summary the expected retry delay, $\bar{D}_i^R$, is:

$$\overline{D}_i^R = T_c(1-p_i)\sum_{j=1}^{L_i} j p_i^j \tag{25}$$

Moreover, we must add the average transmission delay, T_s, associated with the successfully transmitted packets (occurring at a probability of $(1-p_i^{L_i+1})$, resulting in the delay, $\bar{D}_i^T$:

$$\overline{D}_i^T = T_s(1-p_i^{L_i+1}) \tag{26}$$

Finally, we must also take into account the delay $\bar{D}_i^{drop}$ associated with dropped packets (occurring at a probability of $p_i^{L_i+1}$). We have:

$$\overline{D}_i^{CD,drop} = T_e \sum_{h=0}^{L_i} \frac{W_{i,h}-1}{2}$$

$$\overline{D}_i^{B,drop} = \overline{D}_i^{CD,drop} \frac{p_i^*}{1-p_i^*} \frac{\left(\frac{p_s}{p_b}T_s + \left(1-\frac{p_s}{p_b}\right)T_c\right)}{T_e}$$

$$\overline{D}_i^{R,drop} = T_c^*(L_i+1) \tag{27}$$

$$\overline{D}_i^{drop} = p_i^{L_i+1}(\overline{D}_i^{CD,drop} + \overline{D}_i^{B,drop} + \overline{D}_i^{R,drop}$$

In conclusion, we find that the total delay under saturation conditions between packets that are either successfully transmitted or dropped, is:

$$\overline{D}_i^{SAT} = \overline{D}_i^{CD} + \overline{D}_i^{B} + \overline{D}_i^{R} + \overline{D}_i^{T} + \overline{D}_i^{drop} \tag{28}$$

B Delay and Service Time on a non-saturated channel

Under extreme non-saturation conditions, however, the post-backoff is completed before a packet arrives in the transmission queue to be transmitted. Thus, under these conditions the post-backoff will not add to the transmission delay, as it did when we calculated the saturation delays above. The easiest way to handle this is to subtract the post-backoff delay from the expressions above.

$$\overline{D}_i^{NON-SAT} = \overline{D}_i^{SAT} - \frac{W_{i,0}-1}{2}$$

$$\left(T_e + \frac{p_i^*}{(1-p_i^*)}\left(\overline{D}_i^{CD,drop}\left(\frac{p_s}{p_b}T_s + \left(1-\frac{p_s}{p_b}\right)T_c\right)\right)\right) \tag{29}$$

C Estimating ρ_i

First, for a G/G/1 queue, the probability that the queue is non-empty, ρ, is given by $\rho = \lambda\bar{x}$, where λ represents the traffic rate in terms of packets per second and $\bar{x}$ is the average service time.

For simplicity, we assume here that the traffic rate faced by all backoff instances of a class is the same on all stations and use λ_i to denote the traffic rate (in terms of packets per seconds) of traffic class i on one station. Then we have

$$\min\left(1, \lambda_i\overline{D}_i^{NON-SAT}\right) \geq \rho_i \geq \min\left(1, \lambda_i\overline{D}_i^{SAT}\right) \tag{30}$$

It is possible to use arguments to determine ρ_i with higher accuracy. In all scenarios we studied and validated, however, we did not experience any significant differences between setting $\rho_i = \min(1, \lambda_i\bar{D}_i^{SAT})$ and $\rho_i = \min(1, \lambda_i\bar{D}_i^{NON\text{-}SAT})$. Hence, elaborating on this issue is beyond the scope of this paper.

If we measure the delay parameters in terms of μs and the rate, R_i, in terms of *Mbps* and all packets are of 1024 bytes, we find that:

$$\lambda_i(\mu s) = \frac{R_i(Mbps)}{8 * 1024} \tag{31}$$

D Estimating q_i and q_i^*

To estimate q_i of the non-saturation model we assume that the traffic arriving in the transmission queue is Poisson distributed, i.e. that we have an M/G/1 queue. q_i is the probability that at least one packet will arrive in the transmission queue during the following generic time slot under the condition that the queue is empty at the beginning of the slot.

Starting out with the pdf of the length of a generic slot, $b(t)$, we have:

$$\begin{aligned} b(t) = p_s\delta(t-T_s) &+ (1-p_b)\delta(t-T_c) \\ &+ (p_b-p_s)\delta(t-T_c) \end{aligned} \tag{32}$$

Consequently, q_i is calculated as:

$$q_i = 1 - \int_0^\infty e^{-\lambda_i t}b(t)dt = 1-$$

$$\left(p_s e^{-\lambda_i T_s} + (1-p_b)e^{-\lambda_i T_e} + (p_b-p_s)e^{-\lambda_i T_c}\right) \tag{33}$$

Noting, however, that the model is approximate by nature, it is possible to find the probability based on the average length of the timeslot. Hence, as a simplification we might approximate:

$$q_i \approx 1 - e^{-\lambda_i((1-p_b)T_e + p_s T_s + (p_b-p_s)T_c)} \tag{34}$$

Both expressions for q_i were validated in a number of scenarios, and we did not observe any significant differences between either of them.

Ideally, q_i^* should be estimated differently, since it expresses the probability of receiving a packet in the timescale of the countdown of one backoff slot, in contrast to q_i^*. We tested a number of different expressions for this, but observed that setting q_i^* equal to q_i for simplicity worked as a good approximation in all the scenarios we explored.

V Throughput

A Throughput without Virtual Collision Handling

Let $p_{i,s}$ denote the probability that a packet from any of the backoff instances of class i is transmitted successfully in a time slot.

$$p_{i,s} = \frac{n_i \tau_i}{(1-\tau_i)} \prod_{c=0}^{N-1} (1-\tau_c)^{n_c} \tag{35}$$

Let also p_s denote the probability that a packet from any class i is transmitted successfully in a time slot.

$$p_s = \sum_{i=0}^{N-1} p_{i,s} = \sum_{i=0}^{N-1} \frac{n_i \tau_i}{(1-\tau_i)} \prod_{h=0}^{N-1} (1-\tau_h)^{n_h} \tag{36}$$

Then, the throughput of class i, S_i, can be written as the average real-time duration of successfully transmitted packets by the average real-time duration of a contention slot that follows the special time scale of our model:

$$S_i = \frac{p_{i,s} T_{i,MSDU}}{(1-p_b)T_e + p_s T_s + (p_b - p_s)T_c} \tag{37}$$

T_e denotes the real-time duration of an empty slot, while T_s, T_c denote the real-time duration of a slot containing a successfully transmitted packet and of a slot containing two or more colliding packets, respectively. The length of the longest colliding packet on the channel determines T_c. If all packets are of the same length, which we will consider in this paper, T_c = $T_{s.}$ (Otherwise refer to [12] to calculate T_c based on the average duration of the longest colliding data packet on the channel.) Finally, T_{MDSDU} denotes the average real-time required transmitting the MSDU part of a data packet.

First, we notice that the share s_i of the total data bandwidth (that is given by the current traffic load) allocated to a class i is given by:

$$s_i = \frac{\frac{n_i \tau_i}{(1-\tau_i)} T_{i,MSDU}}{\sum_{i=0}^{N-1} \frac{n_i \tau_i}{(1-\tau_i)} T_{i,MSDU}} \tag{38}$$

B Throughput with Virtual Collision Handling

If there is one transmission queue of each AC on each station, on the contrary, there will be Virtual Collision Handling between the queues on each station. Then, higher priority traffic does not need to take transmission of lower-priority queues on the same station. The probability of their transmissions will not affect the throughput of the higher priority AC. Thus, Eq. (35) above must be replaced by:

$$p_{i,s} = \frac{n_i \tau_i}{(1-\tau_i)} \frac{\prod_{c=0}^{N-1} (1-\tau_c)^{n_c}}{\prod_{c=0}^{i} (1-\tau_c)} \tag{39}$$

Using this expression for $p_{i,s}$, p_s is calculated, as before, by summing over all $p_{i,s}$, i.e. $p_s = \sum_{i=0}^{N-1} p_{i,s}$. S_i is also calculated as above.

C Throughput of the Virtual Collision Handler

One can use expression (39) to look at behaviour with only one station by setting $n_i = 1$ for all i. Then we get:

$$p_{i,s} = \tau_i (1 - p_i) \tag{40}$$

Here, we note that the expression describes the special case with only one station contending for the channel, which is the case when the channel refers to the "virtual" channel represented by the VCH.

Using this new expression for $p_{i,s}$, both p_s and S_i of the VCH are calculated as earlier.

VI Validations

A Parameters used for validations

For validations, we compared numerical computations in Mathematica of the model presented above with ns-2 simulations, using the TKN implementation of 802.11e in ns-2 [8].

In Table 1 we can see the parameter settings for 802.11a, 802.11b, 802.11g. Note however that parameters such as CWmin and CWmax are overridden by the use of 802.11e [2]. For our validations, we simply used the default 802.11e values summarized in Table 1. (Hence, a scenario where the HC adjusts these parameters dynamically, was not considered.)

The scenario selected for validations is 802.11b [9], since this is the most widely deployed configuration per se. A configuration with the mandatory long preamble was explored [9]. According to the standard the long preamble and physical PLCP header are

	802.11a	802.11b	802.11g
Nom. BW	54 Mb/s	11 Mb/s	54 Mb/s
SIFS	16 us	11 us	10 us (+ 6 us)
SLOT	9 us	20 us	9 us (11g-only) 20 us (legacy)
CWmin	15	31	15
CWmax	1023	1023	1023
Retry limit (long/short)	7/4	7/4	7/4
PHY-header	20 us	192 us (long) 96 us (short)	20 us
Retry limit (long/short)	7/4	7/4	7/4

Table 2 Parameter settings for 802.11a, 802.11b and 802.11g

always transmitted at 1 Mbps, and takes 192 μs in total. In our selected scenario, we also consider that all data payloads (i.e. MSDU) are of 1024 bytes of length and are transmitted at the maximum 802.11b rate of 11 Mbps. Furthermore, we consider a case with the basic transmission mechanism of sending a data packet followed by an acknowledgement (ACK) without the initiation RTS/CTS-mechanism. According to the standard, the MAC-part of the ACK shall be transmitted at the same rate as the proceeding frame, i.e. at 11 Mbps. However, in our scenario we consider an implementation where the MAC-part of the ACK is transmitted at 1 Mpbs. The reason that we make this choice is to match with the implementation of the ns-2 network simulator that is being used to validate our results.

B Determining the 802.11b parameters for numerical calculations

With a transmission range in the order of 30 m the propagation delay will be around 0.1 μs, and is neglected in the estimation of the parameters (which is often normal practice also with various simulator implementations). Conceptually, the propagation delay can be considered as an already included part of the value for the SIFS, eq. (41) – see below.

Here T_c denotes the time a non-colliding station has to wait when observing a collision on the channel, while T_c^* denotes the time a colliding station has to wait when experiencing collision. A non-colliding station has to wait for a period determined by the fixed EIFS parameter, while a colliding station has to wait by a period determined by the configurable AckTimeout interval. For simplicity, we have set the AckTimeout equal to EIFS (which is also a normal practice with many simulators, such as ns-2), such that T_c equals T_c^*, and $T_{TO} = 0$. (Xiao [5] sets T_{TO} = EIFS – DIFS = 314 μs.)

Note also that the calculation of T_c (and T_c^*) includes the element T_{1024} since all packets on the air are of the same size. In a system where there are packets of different length T_c (and T_c^*) should instead consider the transmission time of the longest packet, which is not difficult to estimate (e.g. see [12]). If we had considered transmission with the RTS/CTS mechanism, on the contrary, we would have had the changes as shown in Eq. (42) – see below.

$$\begin{aligned}
T_e &= 20\ \mu s \\[1ex]
T_{i,MSDU} &= T_{1024} = 8 * 1024 / 11\ \mu s = 520.4\ \mu s \\[1ex]
T_s &= (T_{PHY} + T_{MAC} + T_{1024}) + SIFS + (T_{PHY} + T_{ACK\text{-}MAC}) + \min(AIFS[0], ..., AIFS[3]) \\
&= (192\ \mu s + 8(24 + 4) / 11\ \mu s + 8 * 1024 / 11\ \mu s) + 10\ \mu s + (192\ \mu s + 8 * 14 / 1\ \mu s) + 50\ \mu s \\
&= (957.1\ \mu s) + 10\ \mu s + (304\ \mu s) + 50\ \mu s = 1321.1\ \mu s \\[1ex]
T_c &= (T_{PHY} + T_{MAC} + T_{1024}) + (EIFS) \\
&= (957.1\ \mu s) + (SIFS + (T_{PHY} + T_{ACK\text{-}MAC@1Mbps} + DIFS) \\
&= (957.1\ \mu s) + (10\ \mu s + 192\ \mu s + 8 * 14 / 1\ \mu s + 50\ \mu s) = 1321.1\ \mu s
\end{aligned} \tag{41}$$

$$\begin{aligned}
T_s^{RTS/CTS} &= (T_{PHY} + T_{RTS\text{-}MAC}) + SIFS + (T_{PHY} + T_{CTS\text{-}MAC}) + SIFS + T_s \\
&= (192\ \mu s + 8 * 20 / 2\ \mu s) + 10\ \mu s + (192\ \mu s + 8 * 14 / 2\ \mu s) + 10\ \mu s + (1265.1\ \mu s) \\
&= 1603.1\ \mu s \\[1ex]
T_c^{RTS/CTS} &= (T_{PHY} + T_{RTS\text{-}MAC}) + DIFS \\
&= (192\ \mu s + 8 * 20 / 2\ \mu s) + 50\ \mu s \\
&= 322\ \mu s \\[1ex]
T_c^{*,RTS/CTS} &= T_s^{RTS/CTS} + SIFS + T_{CTS\text{-}Timeout}
\end{aligned} \tag{42}$$

C Validation of the full-range model with starvation (Uplink Scenario)

First we look at a typical uplink scenario with a number of stations, QSTAs, contending for channel access (Figure 3). In this scenario, the role of the access point, QAP, is simply to acknowledge packets sent by the QSTAs.

This scenario corresponds to the uplink scenario presented in the frequently cited paper by Mangold et al. [10], except that here we consider 802.11b instead of 802.11a. Each station has four active queues and sends 250 kbps of traffic of each of the four ACs. For simplicity, the packets of all ACs have the same length of 1024 bytes (i.e. IP header and IP payload).

The throughput values of our ns-2 simulations were measured over three minutes of simulation time. The simulations were started with a 100 seconds transition period to let the system stabilize before the measurements were started. Each QSTA generated 250 kbps Poisson distributed traffic for each of the four ACs. The dot11RTSThreshold was set so high (e.g. 3000 bytes) that the optional RTS/CTS mechanism of 802.11e was not used.

Figure 4 compares numerical calculations of the analytical model with the actual simulation results. We observe that our full-range model, which models 802.11e on the full range from a non-saturated to a saturated medium, gives a qualitatively good match when compared with simulations. We also observe inaccuracies in the model in the semi-saturation part (middle part) of the figure. The numerical calculations of Mathematica have difficulty in converging in this region, for example for $n = 8$.

In Figure 4 we also observe that the starvation of AC[0] and AC[1] experienced with simulations is described with relatively good accuracy by the analytical model. We will take a closer look at the importance of the expression to the analytical model. Figure 5 shows how the probability of a busy slot on the channel, p_b, changes as a function of the traffic load. Here, the curve for p_b is taken from exactly the same numerical calculations as were drawn in Figure 4. The horizontal lines illustrate the starvation conditions of the two classes according to Eq. (19); i.e. when $p_b = 0.167$ and when $p_b = 0.5$. The vertical lines map these freeze points down onto the x-axis, and translate them into the corresponding traffic loads. Despite integer resolution of the x-axis, we have linearly interpolated down to non-integer points on the axis, for illustrative purposes only. Returning to Figure 4 above, we observe that the starvation points were predicted with satisfactory accuracy.

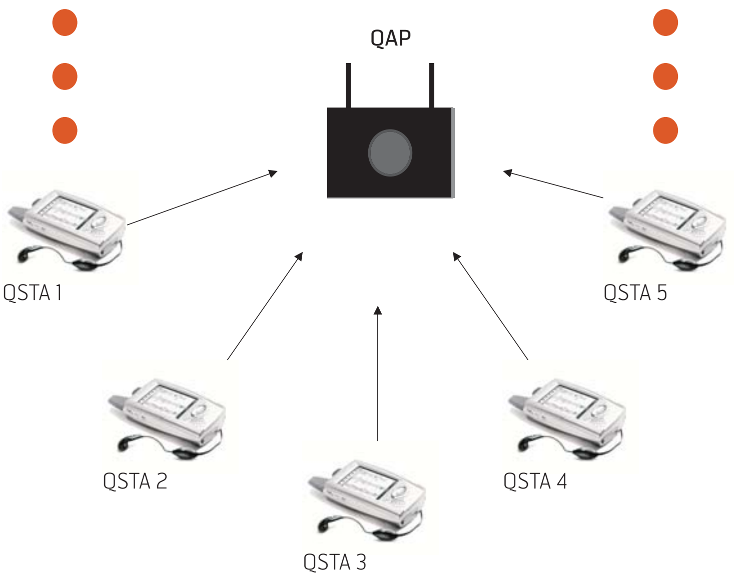


Figure 3 Simulation setup to validate numerical results of uplink traffic

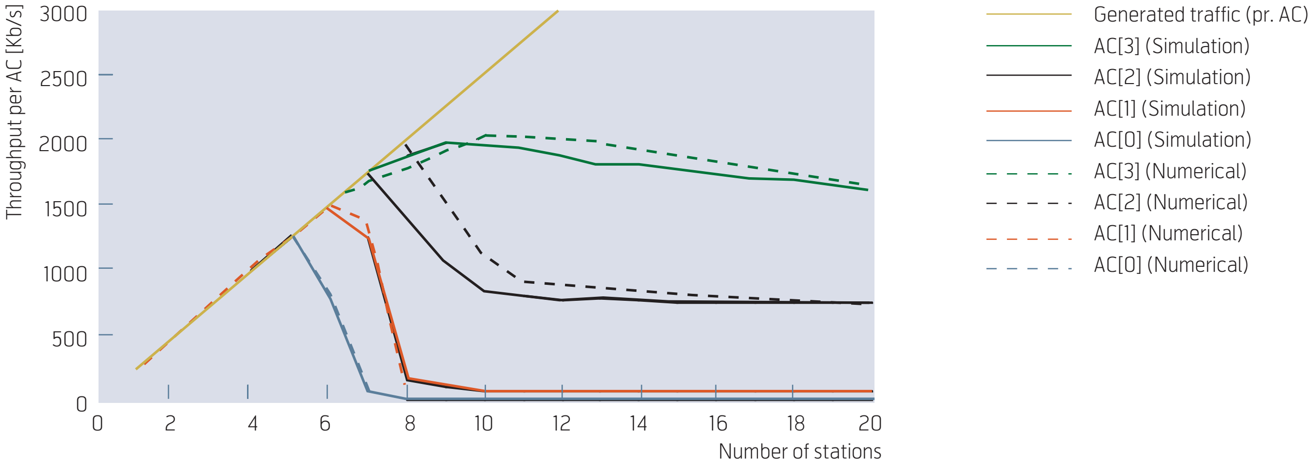


Figure 4 Comparing analysis (numerical) with simulations. Four ACs per station and 250 kbps per AC per station

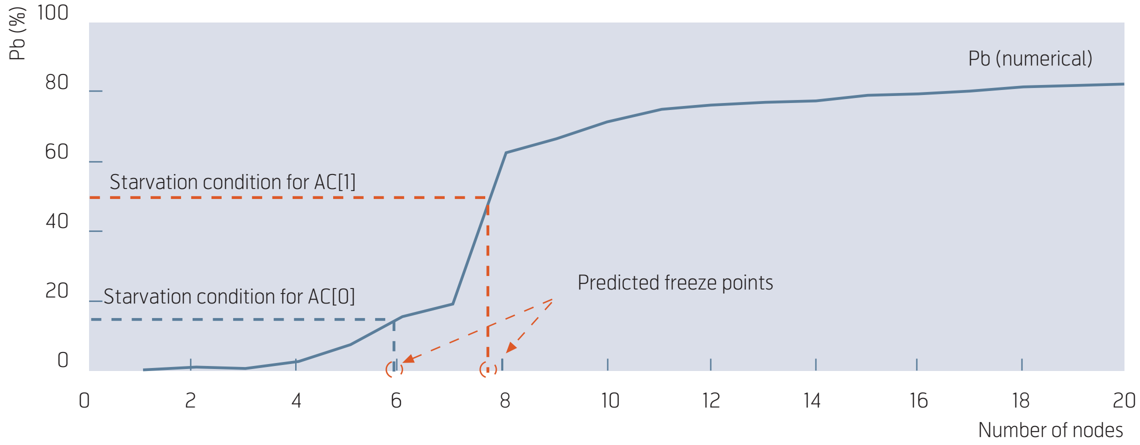


Figure 5 Using the analytical model to predict at which traffic loads starvation will occur

D Validation of the performance of the Virtual Collision Handler (Downlink Scenario)

Most related works on analytical modelling of 802.11e seem to focus on uplink scenarios when it comes to validations. However, here we argue that in daily life, 802.11 is mainly used for Internet Access or for access to a wired Local Area Network (LAN) infrastructure. In both cases, the wireless station (QSTA) is often a client that retrieves large amounts of information from the wired network. In other words, traffic patterns are normally asymmetric, with little uplink traffic from the STAs, but a large amount of downlink traffic from the access point (QAP). Indeed, many networks are designed and optimized with respect to this feature. Asymmetric Digital Subscriber Line (ADSL), for example, which connects the majority of households to the Internet, often allocates a significantly larger share of the available bandwidth for the downlink traffic, assuming that the uplink traffic will be limited. Ensuring quality of service and appropriate differentiation of the downlink traffic is therefore of utmost importance.

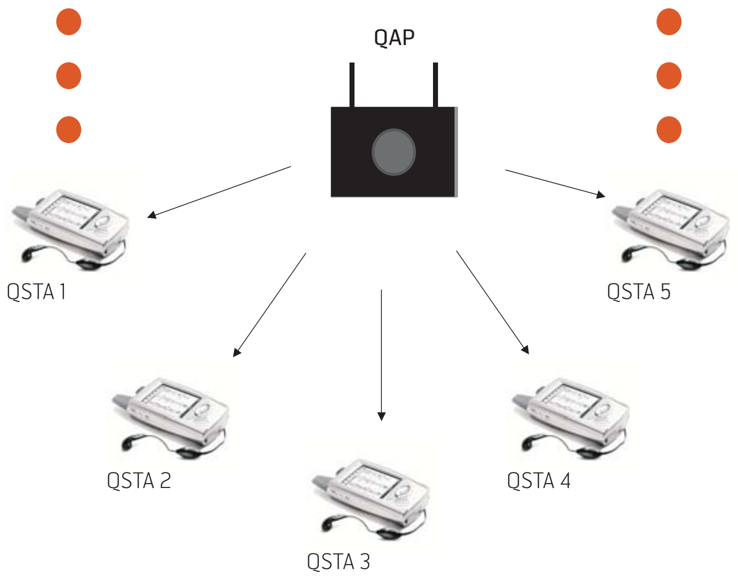


Figure 6 Simulation setup to validate numerical results of downlink traffic

In this section we look at a typical downlink scenario. Here we go to the extreme and assume that all traffic is downlink traffic. This assumption means that the QAP is always free to use the wireless channel and will not experience collision from any other station. This actually means that all traffic contention will occur in the Virtual Collision Handler (VCH), which will represent a "virtual" traffic channel. An underlying assumption is that the QAP uses EDCA for downlink traffic, and not HCCA.

For our validation, we consider a scenario with QAP, implementing a VCH and four transmission queues for the four possible ACs. This configuration is depicted in Figure 6. It corresponds to the downlink scenarios presented in the frequently cited paper by Mangold et al. [10], except that here we consider 802.11b instead of 802.11a.

Figure 6 shows that a number of different stations (QSTAs) might be listening for traffic on the radio channel in order to receive any downstream traffic from the access point (QAP). However, it is only the QAP that sends data traffic, while the QSTAs are not actively initiating traffic. Hence, the role of the QSTAs is only to acknowledge all MAC frames that the QAP successfully transmit on the channel.

Figure 7 compares numerical calculations of the analytical model with the actual simulation results. As before, we use Poisson distributed traffic consisting of 1024 byte packets sent without the optional RTS/CTS mechanism. Here we are only dealing with

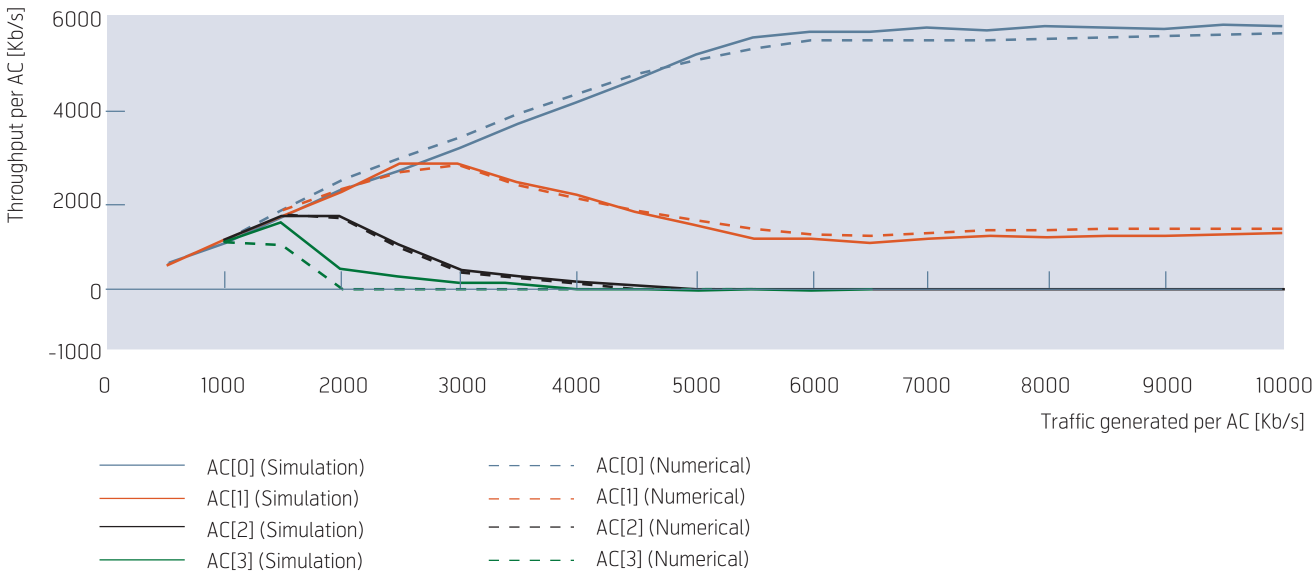


Figure 7 Comparison between analytical results and simulation results. (Recommended 802.11e parameter settings)

one node. Instead of increasing the number of nodes (as we did in the Uplink Scenario above) we increase the traffic generated per AC. For simplicity, we assumed that the QAP generated the same amount of downlink traffic for each of the four ACs. In Figure 7 we observe that our full-range model, which models 802.11e on the full range from a non-saturated (finite queue) to a saturated (infinite queue) system, gives a good match when compared to simulations.

In Figure 8 we repeat the validations using different values for the contention window. Here we have doubled all minimum and maximum contention windows compared to the recommended values given in Table 1. We have also shown the results on a larger scale (up to 20,000 kbps per AC) to illustrate the remarkably good accuracy between model and simulation results in the saturation part of the figure.

However, there are ranges of Figure 7 and Figure 8 where there are noticeable discrepancies between the curves. For Figure 8, this range is expanded and shown on a smaller scale in Figure 9. Here we observe that the model – probably the AIFS-approximation – is a little too rough on the lowest priority AC, AC[0]. Due to the fact that AC[0] and partly also AC[1] are underestimated here, the model incorrectly gives a throughput of AC[3] that exceeds the 1-to-1

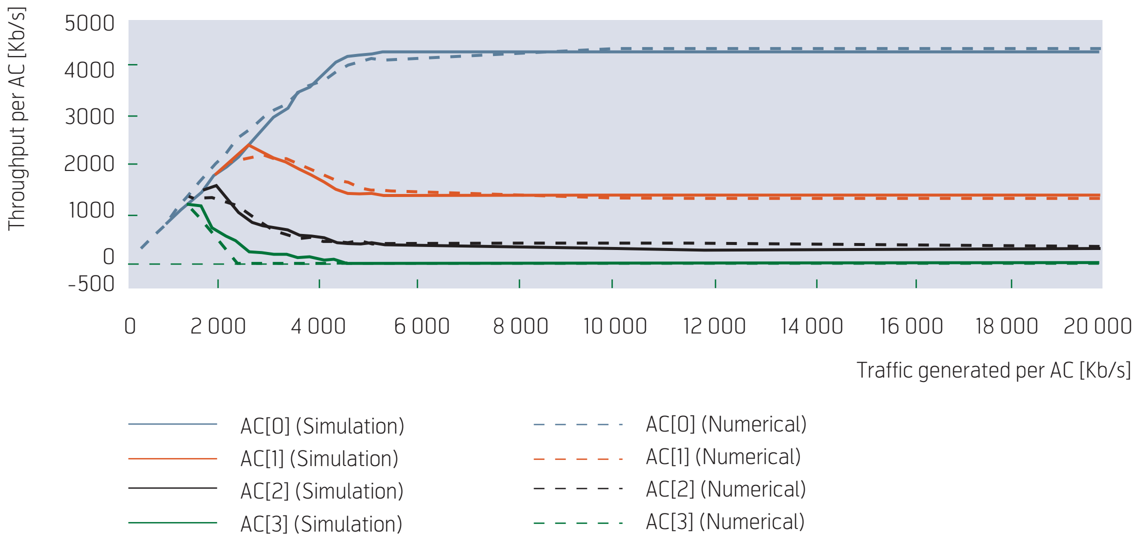


Figure 8 Comparison between analytical results and simulation results (doubled CW values)

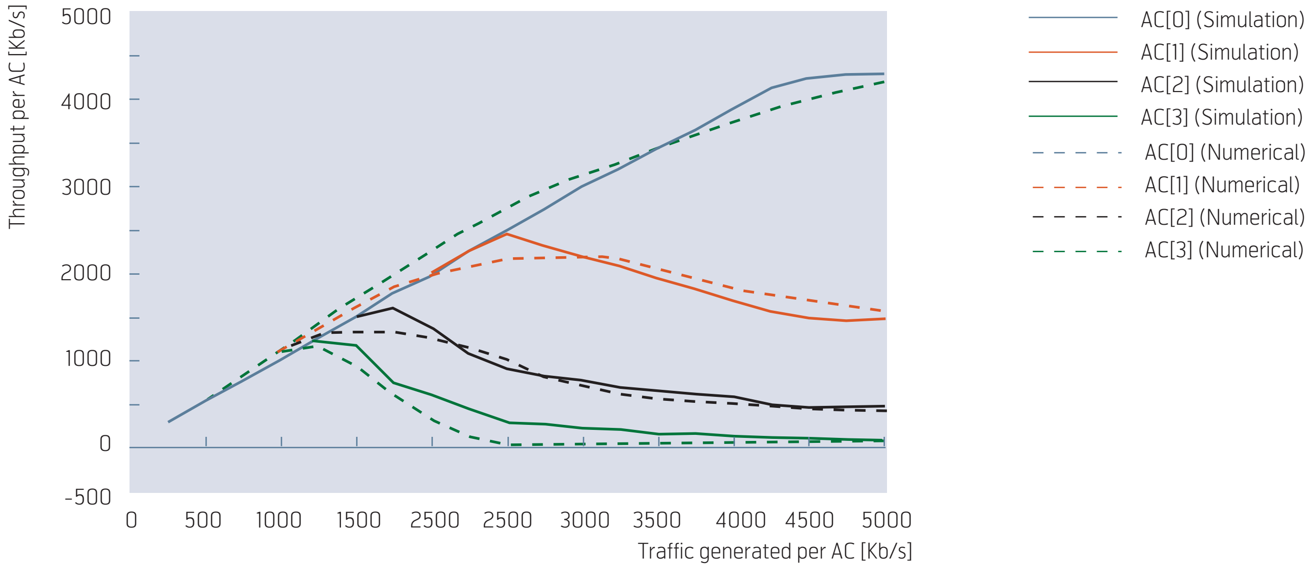


Figure 9 Comparison between analytical results and simulation results

linear line. This would mean that AC[3] transmits more traffic than it generates, which is obviously not correct. It is indeed possible to do some improvements of the model in this region, although one must keep in mind that the model is approximate, and a complete match might be difficult to find without adding considerable complexity to the model.

VI Conclusions

This paper shows how analytical saturation models can be extended to cover the full range from an unsaturated to a fully saturated channel.

Furthermore, AIFS differentiation was introduced in the model. It provides an approximate expression to determine the starvation point of different access categories (ACs) at a given traffic load and at given channel access parameters, such as the AIFSN assigned to each channel. (The other differentiation parameters also play a role in this expression as they indirectly influence the traffic load on the channel.) By measuring the channel load and by knowing the AIFSN assigned to each AC, the access point is able to tell when the starvation conditions are present for any of the ACs, independent of whether packets of these ACs are attempted for transmission.

Moreover, we show how virtual collision handling can be incorporated into the presented model. By encompassing virtual collision handling, we demonstrated that it is also possible to describe the behaviour of a Virtual Collision Handler. Hence, an access point that uses EDCA for massive downlink traffic is therefore able to predict the levels of QoS that the data traffic it is transmitting will obtain by the Virtual Collision Handler. In this way it is to a larger extent in control of the QoS of the traffic it is sending. (Needless to say, any station – whether it is an access point or not – may benefit from predicting the behaviour of the Virtual Collision Handler, although we anticipate that the model will be mostly appreciated by the access points.)

The model is calculated numerically and validated against simulations, using 802.11b and the default parameter settings for 802.11e. We observed that the expansion of the model to cover unsaturated conditions gave a relatively good match with simulations. AIFS differentiation and starvation did also match well. Also the analytical model for the behaviour of the Virtual Collision Handler corresponded well with our simulations.

Acknowledgements

The authors would like to acknowledge the OBAN project of the European Union’s 6th Framework Programme for supporting this work; our colleague, Einar Paul Edvardsen, for encouraging the work on this paper; and our Master student, Bjørn Selvig, for supporting us on the simulation part.

References

1 IEEE 802.11 WG, Part 11: *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specification*. IEEE 1999.

2 IEEE 802.11 WG, Draft Supplement to Part 11: *Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Medium*

Access Control (MAC) Enhancements for Quality of Service (QoS). IEEE 802.11e/D13.0, Jan. 2005.

3 Bianchi, G. Performance Analysis of the IEEE 802.11 distributed coordination function. *IEEE J. Select. Areas Commun.*, 18, 525–547, 2003.

4 Ziouva, E, Antonakopoulos, T. CSMA/CA performance under high traffic conditions: throughput and delay analysis. *Computer Communications*, 25, 313–321, Feb. 2002.

5 Xiao, Y. Performance analysis of IEEE 802.11e EDCF under saturation conditions. *Proceedings of ICC*, Paris, France, June 2004.

6 Malone, D W, Duffy, K, Leith, D J. Modelling the 802.11 Distributed Coordination Function with Heterogeneous Load. *Proceedings of Rawnet 2005*, Riva Del Garda, Italy, April 2005.

Paal E. Engelstad started work in Telenor R&D August 2000. He wrote his Master Thesis in Kyoto University (Japan), received his MSc degree in physics from NTNU and his bachelor degree in Computer Science from the Univ. of Oslo. Apart from working with science and technology, Engelstad has four years work experience as a management trainee in Orkla asa, where he worked as a product manager (marketing), technical advisor (engineering), and project manager (business development).

email: paal.engelstad@telenor.com

Olav N. Østerbø received his MSc in Applied Mathematics from the University of Bergen 1980 and his PhD from the Norwegian University of Science and Technology in 2004. He joined Telenor R&D in 1980. His main interests include teletraffic modelling and performance analysis of various aspects of telecom networks. Activities in recent years have been related to dimensioning and performance analysis of ATM networks and now moving towards IP networks, where the main focus is on modelling and control of different parts of next generation IP-based networks.

email: olav-norvald.osterbo@telenor.com

ISSN 0085-7130 © Telenor ASA 2005

Status

Introduction

PER HJALMAR LEHNE

Per Hjalmar Lehne is Researcher at Telenor R&D and Editor-in-Chief of Telektronikk

In this issue of *Telektronikk*'s status section, we give a report from last years *World Telecommunication Standardization Assembly – WTSA-04*. *Anne Lise Lillebø*, Senior Adviser in Telenor Corporate Communications, has produced a comprehensive report from the Assembly, which was held in Florianópolis, Brazil in October 2004.

The WTSA is the supreme organ of the *International Telecommunication Union Standardization Sector (ITU-T)*, and meets every four years. Telenor is a sector member of ITU-T and Ms Lillebø has participated on behalf of Telenor since the first Assembly in 1993. WTSA-04 was the first Assembly held in South America.

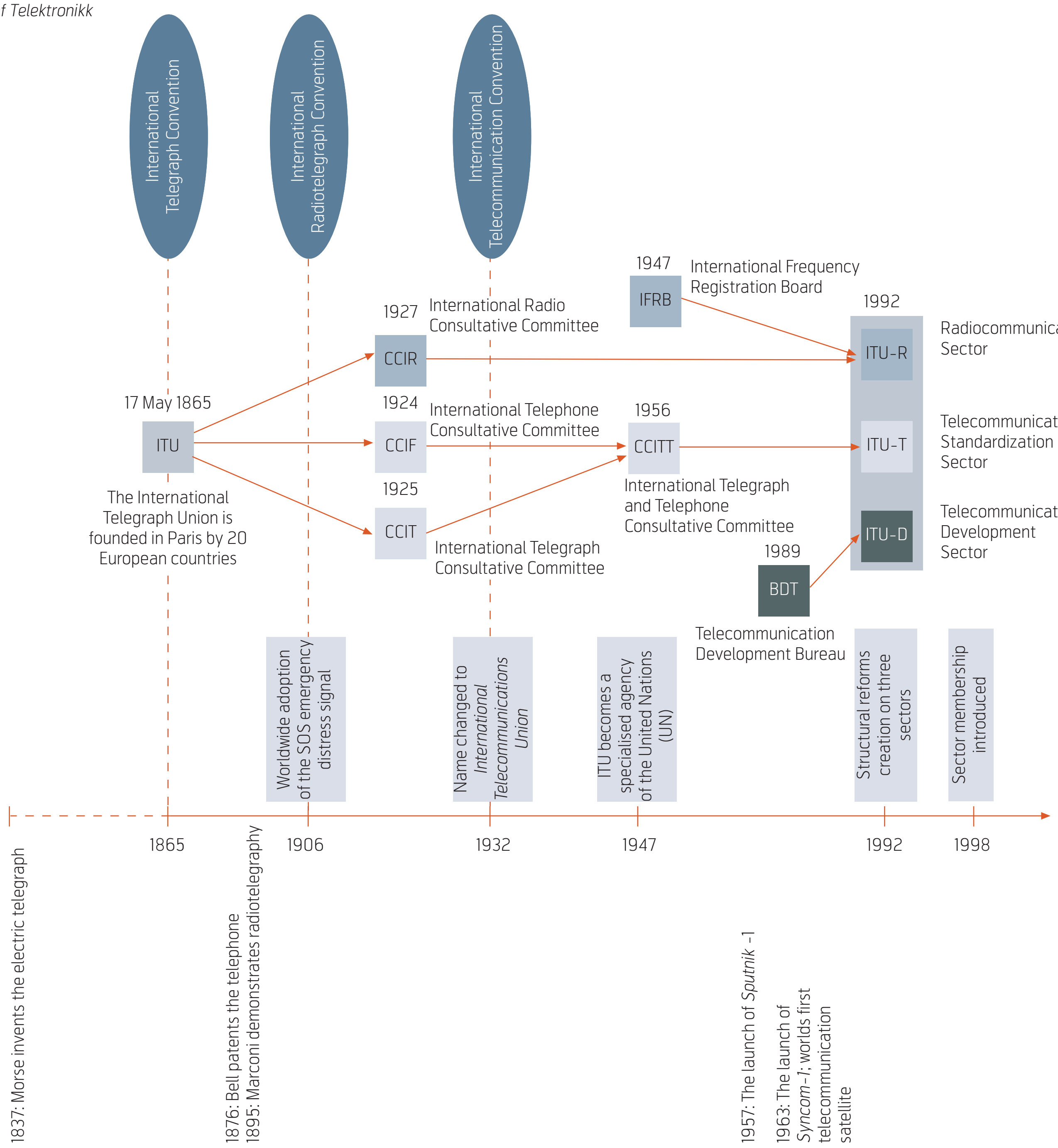

International Telegraph Convention
International Radiotelegraph Convention
International Telecommunication Convention
1947
IFRB
International Frequency Registration Board
1992
1927
International Radio Consultative Committee
CCIR
ITU-R
Radiocommunication Sector
17 May 1865
ITU
1924
International Telephone Consultative Committee
CCIF
1956
CCITT
ITU-T
Telecommunication Standardization Sector
The International Telegraph Union is founded in Paris by 20 European countries
1925
CCIT
International Telegraph and Telephone Consultative Committee
International Telegraph Consultative Committee
1989
BDT
ITU-D
Telecommunication Development Sector
Telecommunication Development Bureau
Worldwide adoption of the SOS emergency distress signal
Name changed to *International Telecommunications Union*
ITU becomes a specialised agency of the United Nations (UN)
Structural reforms creation on three sectors
Sector membership introduced
1865
1906
1932
1947
1992
1998
1837: Morse invents the electric telegraph
1876: Bell patents the telephone
1895: Marconi demonstrates radiotelegraphy
1957: The launch of *Sputnik -1*
1963: The launch of *Syncom-1*; worlds first telecommunication satellite

Figure 1 The organisational development of the International Telecommunication Union (ITU) from 1865 until today

ISSN 0085-7130 © Telenor ASA 2005

The first *World Telecommunication Standardization Conference* (WTSC) started in 1993 after the reorganisation of ITU. In 1992 ITU was transformed and three sectors were formed: the *Radiocommunication Sector* (ITU-R), the *Telecommunication Standardization Sector* (ITU-T) and the *Telecommunication Development Sector* (ITU-D). ITU-T replaced the former *International Telegraph and Telephone Consultative Committee* (CCITT). In 1998, the ITU introduced a new member category of *Sector membership* to give wider rights and obligations to private sector members.

Telenor has participated in all WTSC/WTSA and several have been previously reported in *Telektronikk*. Table 1 gives an overview of previous major ITU events and reports in later years. *Telektronikk* has a long history in printing reports from ITU events since 1927. In addition, Telektronikk has printed several reports and results from different ITU study groups. Consult the web: *http://www.telektronikk.com* for an overview of reports since 1992.

World Telecommunication Standardization Conference – WTSC-93 – Helsinki, Finland	
World Telecommunication Standardization Conference – WTSC-96 – Geneva, Switzerland, 9–18 Oct 1996	*Arve Meisingset*, *Telektronikk*, 89 (2), 84–86, 1997
World Telecommunication Standardization Assembly – WTSA 2000 – Montreal, Canada, 27 Sep – 6 Oct 2000	*Anne Lise Lillebø*, *Telektronikk*, 97 (1), 138–143, 2001
ITU Plenipotentiary Conference PP-02, Marrakech, Morocco, 23 Sep – 18 Oct 2002	*Einar Utvik* and *Anne Lise Lillebø*, *Telektronikk*, 99 (1), 138–153, 2003
World Telecommunication Standardization Assembly – WTSA-04 – Florianópolis, Brazil, 5 – 14 Oct 2004	*Anne Lise Lillebø*, *Telektronikk*, 101 (1), 153–161, 2005 (this issue)

Table 1 Later ITU events and corresponding reports in Telektronikk

Per Hjalmar Lehne is Researcher at Telenor R&D and Editor-in-Chief of Telektronikk. He obtained his MSc from the Norwegian Institute of Science and Technology (NTH) in 1988. He has since been with Telenor R&D working with different aspects of terrestrial mobile communications. His work since 1993 has been in the area of radio propagation and access technology, especially on smart antennas for GSM and UMTS. He has participated in several RACE, ACTS and IST projects as well as COST actions in the field. His current interests are antennas and the use of MIMO technology in terrestrial mobile and wireless networks and on access network convergence.

email: per-hjalmar.lehne@telenor.com

WTSA-04 – World Telecommunication Standardization Assembly, Florianópolis, 5 – 14 October 2004 – An overview of main results

ANNE LISE LILLEBØ

Anne Lise Lillebø is Senior Adviser, Telenor Corporate Communications

This article gives an account of the results of WTSA-04 seen from a Telenor perspective and with emphasis on European preparations and European common proposals.

1 Introduction

1.1 Background

The World Telecommunication Standardization Assembly – WTSA – is the supreme organ of the Telecommunication Standardization Sector (ITU-T) of the International Telecommunication Union (ITU), the United Nations' specialised agency for telecommunications. The WTSA meets every four years to coordinate the development of global standards for telecommunications networks and services.

For the first time in ITU's history, the WTSA was held in South America – on the island of Santa Catarina, Florianópolis, Brazil, from 5 to 14 October 2004. The WTSA was hosted by the Brazilian regulator ANATEL (Agência Nacional de Telecomunicacôes) and was opened by Mr Pedro Jaime Ziller de Araujo, president of ANATEL.

The Assembly was preceded by a Symposium on Cybersecurity held on 4 October and which attracted some 300 participants from 50 countries and three regional telecommunications organisations. During the WTSA-04 there were technical briefing sessions on IDN and IPv6, NGN, telemedicine and telecommunications for disaster relief.

1.2 Duties of the WTSA

The duties and functions of the Assembly are laid down in Article 18 of the ITU Constitution (CS) and Article 13 of the ITU Convention (CV). The tasks of the WTSA are the following:

- adopt the work programme for the period up to the next WTSA normally referred to as the "study period";
- establish Study Groups that will carry out the work;
- appoint chairmen and vice-chairmen of the Study Groups;
- elaborate working methods;
- approve resolutions regarding the work of the ITU-T;
- approve Recommendations.

The main issues of the WTSA-04 in Florianópolis were:

- Adoption of the ITU-T work programme and new Study Group structure;

Florianópolis Convention Center

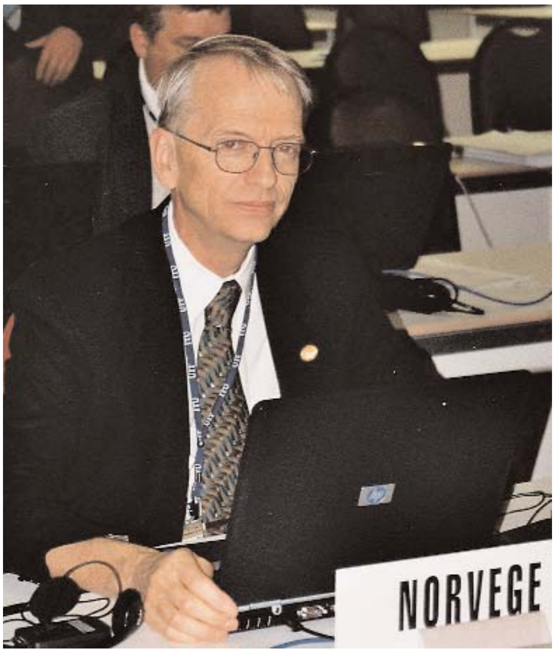
- Establishment of a new Study Group for NGN (Next Generation Networks);
- Appointment of chairmen and vice-chairmen of the Study Groups;
- Revision of Recommendation A.8 (Alternative Approval Procedure);
- Involving developing countries in standardisation activities.

Resolutions adopted by the Assembly will enter into force immediately after the closure of the WTSA.

1.3 Participation at the WTSA-04

The WTSA is open for direct participation from Member States of the ITU and Sector Members of the ITU-T Sector. The Norwegian Post and Telecommunications Authority has been appointed by the Ministry of Communications as Norway's ITU administration and takes care of the day to day business concerning ITU in Norway. At the WTSA, Norway was

ISSN 0085-7130 © Telenor ASA 2005



Knut Bryn, Head of the Norwegian delegation

Dominique Würges, France Télécom, representing ETNO

represented by Mr Knut Bryn, Technical Director of the Norwegian Post- and Telecommunications Authority.

Telenor is a Sector Member of the ITU-T and participated directly at the WTSA in its capacity as Sector Member. Telenor was represented by Mr Arve Meisingset, Telenor R&D, who was Norway's candidate for vice-chairmen of SG17 and Anne Lise Lillebø, Telenor ASA. Telenor's representatives were not members of the Norwegian delegation.

ETNO – the European Telecommunications Operators' Assocation – was represented by Mr Dominique Würges, France Télécom. Most of the major European telecommunications operators participated at the WTSA-04.

The European Commission was not present at the WTSA-04, and European Union coordination was entrusted to the Netherlands which held the EU presidency at the time of the Assembly.

Delegates from a total of 72 Member States and representatives from 43 Sector Members participated at the WTSA-04.

1.4 WTSA structure and management

The Assembly was chaired by Mr Savio Pinheiro from the Brazilian regulator ANATEL, supported by six vice-chairmen including Mr Anhony De Bono from Maltacom representing region B (Western Europe). The work was carried out in seven committees:

Committee 1:	Steering Committee
Committee 2:	Budget control
Committee 3:	Working methods of the ITU-T
Committee 4:	ITU-T work programme and organisation
Committee 5:	Telecommunication network infrastructure – reports by the Study Groups
Committee 6:	Telecommunication services and tariff issues – reports by the Study Groups
Committee 7:	Editorial Committee

Anthony De Bono, Maltacom, Malta, Vice-chairman

Committee 3 (Working methods) was chaired by Mr Fabio Bigi, Italy, and Committee 5 (Telecommunication network infrastructure: reports by the SGs) by Mr Jean-Yves Monfort, France.

1.5 European preparations for the WTSA-04

Joint European preparations for the WTSA-04 were organised by CEPT (Conférence Européenne des Postes et Télécommunicatons) with the establishment of a special project team (PT WTSA-04) chaired by Ms Marie-Thérèse Alajouanine from the French Administration. The PT WTSA-04 was open for direct participation from members of EICTA (European Information and Communications Technology Industry Association) and ETNO (European Telecommunications Network Operators' Association). A number of ETNO members including Telenor played an active role in the preparatory work of CEPT, focusing their action on two issues of high priority, i.e. the restructuring process including the creation of a Study Group on NGN and the support of the European Common Proposals (ECPs).

The chairman of the CEPT PT WTSA-04 conducted informal coordination with other regional organisations concerning individual topics such as NGN etc before and during the WTSA. An official coordination meeting under the chairmanship of CITEL (Inter-American Telecommunication Commission) was held on 4 October 2004. Regional coordination held during the WTSA helped to understand the different positions. However, a more extensive exchange of views between the various regions before the WTSA would have been very useful to clarify points.

Mr Dominique Würges, France Telecom, acted as the ETNO coordinator throughout the WTSA, and chaired an ETNO coordination meeting. ETNO members had not prepared any specific contribution for this Assembly, but sought close cooperation with their European administrations on priority matters such as restructuring and NGN.

For the first time CEPT had decided that ECPs (European Common Proposals) following the CEPT procedures should be developed as European input to the WTSA-04. An ECP is a common proposal among members of the CEPT. Sector Members in ITU-T cannot co-sign an ECP, but will have to coordinate with their respective administrations. At the WTSA 2000, CEPT members, Sector Members from ETNO and EICTA presented European proposals open for signature both from Member States and Sector Members. However, experience shows that an ECP is a stronger mechanism and is considered as a more formal position on behalf of CEPT.

The work of the CEPT PT WTSA-04 resulted in six ECPs:

- ECP 1: Appointment of SG Chairmen and Vice-Chairmen
- ECP 2: Other groups
- ECP 3: A Study Group on NGN
- ECP 4: Proposal for a revised Resolution 2 (numbering issues)
- ECP 5: Co-location of Study Group meetings
- ECP 6: Coordination role of TSAG (Telecommunication Standardization Advisory Group)

Telenor endorsed the ECPs and the Norwegian Post- and Telecommunications Authority co-signed the six ECPs on behalf of Norway.

2 Work programme and restructuring of Study Groups

In Committees 5 and 6 reports by Study Group chairmen presented an overview of the main achievements in the study period 2000 – 2004 and highlighted that a total of 982 Recommendations had been approved since the last Assembly in 2000 representing one Recommendation per working day.

The restructuring of the ITU-T Study Groups and the assignment of Questions into areas of study was one of the key issues at this WTSA. The main focus was on the management of work related to NGN (Next Generation Networks) and how WTSA-04 should define the long-term vision for NGN and migratory services and the strategy for achieving it.

Many European members were in favour of a reduced number of Study Groups (SGs), but after the completion of the WTSA-04, there are still 13 Study Groups in the ITU-T. The areas of work of the individual Study Groups were consolidated without any big changes. The former SG13 has been transformed into SG13 NGN and the former "Special Study Group on IMT-2000 and beyond" is now specifically devoted to mobility and has been converted into a "regular" Study Group on "Mobile telecommunication networks" (SG19).

WTSA-04 allocated a total of 165 Questions to the 13 Study Groups with the general areas of responsibility as shown in the above box.

New ITU-T Study Groups in the next four-year period 2005 – 2008:

Study Group	Mandate/Area of responsibility	Chairman
SG2	Operational aspects of service provision, networks and performance	Marie-Thérèse Alajouanine, France
SG3	Tariff and accounting principles including related telecommunication economic and policy issues	Ki-Shik Park, Republic of Korea
SG4	Telecommunication management	David Sidor, USA
SG5	Protection against electromagnetic environment effects	Roberto Pomponi, Italy
SG6	Outside Plant and Related Indoor Installations	Francesco Montalti, Italy
SG9	Integrated broadband cable networks and television and sound transmission	Richard R. Green, USA
SG11	Signalling requirements and protocols	Yukio Hiramatsu, Japan
SG12	Performance and quality of service	Jean-Yves Monfort, France
SG13	Next generation networks – architecture, evolution and convergence	Brian Moore, UK
SG15	Optical and other transport network infrastructures	Yoichi Maeda, Japan
SG16	Multimedia terminals, systems and applications	Pierre-André Probst, Switzerland
SG17	Security, languages and telecommunication software	Herbert Bertine, USA
SG19	Mobile telecommunication networks	John Visser, Canada
TSAG	Telecommunication Standardisation Advisory Group	Gary Fishman, USA

2.1 A Study Group for NGN

The management of NGN was considered by operators to be the most crucial at this WTSA. The NGN will provide fixed line and mobile users seamless communication and offer unrestricted access by users to different service providers in a multi-service, multi-protocol, multi-vendor environment. The operators want a converged multi-service platform to deliver all services. Considerable preparatory work – especially by operators – had been carried out in TSAG prior to the WTSA-04.

Telenor supported a Sector Member contribution on NGN to WTSA-04 developed by BT and France Telecom endorsed by 12 other Sector Members (i.a. TDC, TeliaSonera, Swisscom, Telefonica etc.). The Sector Member proposal aimed at creating a Study Group with sufficient critical mass to address NGN issues, advocating a single Study Group for core NGN studies where signalling requirements and protocols, naming and addressing and mobility were included in the same SG. Regrettably it was impossible to obtain the necessary support for this proposal from the ITU Member States.

CEPT had developed an ECP on NGN (ECP 3) promoting a compromise where elements from the former SG11, SG17 and SG13 were merged into a new SG on NGN to include architecture, signalling, protocols etc. The ECP was signed by 22 countries.

A successful compromise was reached on a new Study Group for NGN after substantial negotiations with countries such as Japan, Korea and the USA. The WTSA agreed to create a new Study Group for NGN comprising the responsibility of the former SG13 with the addition of areas related to architecture and frameworks. The new SG13 NGN will be the “lead Study Group” for all NGN studies and will be responsible for the coordination of all NGN related activities in the ITU-T Study Groups. In addition, the SG13 NGN will act as the “parent SG” for the NGN Focus Group that was created by the TSB (Telecommunication Standardization Bureau) Director before the WTSA-04 in order to speed up the ITU-T work on NGN. WTSA recommends that the SG13 NGN holds its meeting at the same place and during the same period as the meetings of SG11 and SG19 in cooperation with the management teams of these SGs.

3 Appointment of Study Group Chairmen and Vice-Chairmen

The candidacies for Study Group chairmen and vice-chairmen are discussed in the meetings of heads of delegation of Member States where Sector Members are not authorised to participate. The candidates for the chairmen and vice-chairmen are formally proposed by the WTSA.

This time there were far more candidates than posts, in particular for the posts of vice-chairmen. Despite the fact that CEPT had submitted a proposal aiming at limiting the number of vice-chairmen to the number required to do the job in the relevant Study Groups, there were numerous candidates for vice-chairmen from the two European administrative regions. The TSB Director had a very challenging task in trying to persuade Member States to withdraw their candidates for vice-chairmen, but this proved almost impossible.

There are now a total of 68 vice-chairmen distributed among the 13 Study Groups and TSAG – which means an average number of five vice-chairmen per Study Group! Both the new SG13 NGN and SG19 Mobile Telecommunication Networks each have eight vice-chairmen!

Mr Arve Meisingset, Telenor R&D, candidate from Telenor and supported by the Norwegian administration, was reappointed vice-chairman in SG17 for a new study period. Mr Meisingset is the only person from a Nordic country holding such a position within the ITU-T at present.

Marie-Thérèse Alajouanine, France, new Chairman of Study Group 2

Ms Marie-Thérèse Alajouanine, France, was appointed chairman of Study Group 2. Ms Alajouanine is the first woman ever to chair a Study Group in the ITU-T.

4 ECPs – European Common Proposals

ECP 1 (Appointment of Study Group Chairmen and Vice-Chairmen)

CEPT recommends that chairmen and vice-chairmen should be active in the field of the Study Group concerned and fully committed to the work of the Study Group. Appointed Vice-Chairmen having both technical and management competence in the area of the Study Group to which they have been appointed should be considered first in the appointment of working party chairmen. Only the number of vice-chairmen and working party chairmen should be appointed necessary for the efficient and effective management and functioning of the Study Group consistent with the work programme.

The Arab States had submitted a proposal for the amendment of ITU-T Resolution 35 (Appointment and maximum term of office for chairmen and vice-chairmen of ITU-T Study Groups and of TSAG), highlighting the importance of equitable geographical distribution of chairmen and vice-chairmen in the ITU-T Study Groups. The WTSA agreed on a compromise text taking into account the following: “In cases where there are two or more candidates with equal competence for the same chairman position, preference should be given to candidates from member States and Sector Members having the lowest number of designated study group chairmen”.

ECP 2 (Other groups)

The intention of this ECP was to update ITU-T Resolution 22 (Authorisation for TSAG to act between WTSAs) in order to reflect the provisions in ITU’s Convention 191A and 191B, adopted at the plenipotentiary conference in Marrakesh in 2002. In addition to creating Study Groups, the WTSA is also authorised to create “other groups” and setting their terms of reference. TSAG had agreed to amend Resolution 22 to take into account these provisions and had submitted a draft text to this effect. After extensive debate a compromise text was finally agreed emphasising that such “other groups” are not authorised to adopt Questions or Recommendations.

ECP 3 (A Study Group on NGN)

The objective of the CEPT was to prompt the WTSA to take an initiative to restructure the ITU-T Sector to address the next generation of telecommunications networks (NGN) in a proper manner in order to maintain ITU as the leading global standards body in telecommunications. CEPT proposed to establish a new NGN Study Group with sufficient critical mass of expertise to cover core aspects of the NGN effectively and to be a focal point for industry to participate and invest its resources for the development of the NGN.

The former SG13 will be transformed into a new SG 13 on NGN with work on NGN and Packet based networks transferred from the old SG11 and SG17. The new SG 13 NGN will be the "parent SG" for the NGN Focus Group and will also be responsible for coordination of all NGN related work in other study groups. The compromise is acceptable to CEPT.

ECP 4 (Proposal for a revised Resolution 2 – numbering issues)

The French Administration was concerned about numbering resources being misused in SG2 and proposed a revision of Resolution 20 (Procedures for allocation and management of international telecommunication numbering, naming, addressing and identification resources) to counter this. The revised agreed text in Resolution 20 covers the point raised by CEPT.

ECP 5 (Co-location of Study Group meetings)

CEPT proposed that co-location (meeting at the same time in the same place) of Study Group meetings could be a means to avoid the duplication and to improve the efficiency of the work. Co-location will enable delegates to participate in more than one Study Group, limit the need for liaison statements between Study Groups and reduce costs for ITU and its members. The ECP5 proposes additional text to the introductory part of ITU-T Resolution 2 (Study Group responsibility and mandates).

The principle of co-location of Study Group meetings was unanimously approved by the WTSA-04.

ECP 6 (Coordination role of TSAG)

The ECP proposed that the working methods of the ITU-T Study Groups should be improved to ensure better coordination of the work on topics spanning several Study Groups. This could be achieved by expanding the role of TSAG in its field of coordination and by requiring Lead Study Groups to support TSAG in carrying out this role.

WTSA agreed to amend Resolution 22, and sections 2.17 and 4.2 in Resolution 1 accordingly.

5 Working methods

5.1 Recommendation A.8 AAP

So-called "Technical Recommendations" which do not have regulatory or policy implications can be approved according to the AAP (Alternative Approval Procedure). Currently a Recommendation can be approved by a Study Group if it is not opposed. Both African and Asian countries had submitted proposals aiming at removing the possibility of vetoing a Draft Recommendation to be approved by AAP in a Study Group meeting. These countries proposed that the requirement for not adopting a Draft Recommendation would be that more than one Member State opposed the Recommendation in question, preferably three to five.

This issue proved to be very contentious, and there was no consensus in Committee 3 where this issue was handled. It was finally resolved in the plenary where a clear majority supported the requirement that more than one Member State must object to the adoption of a Draft Recommendation in order to reject the approval of the Recommendation. The Assembly tasked TSAG to review the matter if required. The text in Recommendation A.8 has been amended to reflect this change.

A number of Member States preferred to keep the status quo and were strongly against this change of principle in the ITU-T approval procedures. The CEPT Brief advised the European countries to support the change, but some CEPT members voted against any change. USA in particular favoured the status quo. Telenor welcomes this change that one Member State can no longer block the approval of a Recommendation. We consider this to be a step in the right direction regarding the approval procedure for "technical Recommendations", which are voluntary Recommendations.

6 Technology watch and Seminar coordination committee

6.1 Technology watch

WTSA-04 agreed that TSAG should establish an appropriate mechanism – a "technology watch" – to monitor the market for new topics and emerging technologies in order to identify new subjects for study by the ITU-T. This task has been included in the updated Resolution 22 "Authorization for TSAG to act between WTSAs".

6.2 Seminar and workshop coordination committee

The WTSA agreed to establish a new group "The seminar and workshop coordination committee – SCC" (Resolution 53) under the supervision of TSAG to monitor the technological evolution and to oversee the workshop programme of ITU-T. The SCC will also seek to encourage developing countries to be involved in these events. A more detailed mandate for the group will be developed by TSAG.

The acceptance of this new tool was controversial, and the opponents found that it might add more administrative work and bureaucracy for the experts responsible and limit their ability for an efficient action. Telenor and other ETNO members considered this to be micro-management and were in favour of entrusting this task to the SG chairmen and the TSB Director.

7 Regional preparations for WTSAs

The Assembly agreed that it would be of great advantage to all members to have the possibility of consolidating views at a regional level combined with the opportunity for interregional discussions prior to the WTSA and that this will facilitate consensus building.

Resolution 43 calls for the organisation of one regional preparatory meeting before a WTSA per region, followed by an informal meeting of the chairmen and the vice-chairmen of the regional preparatory meetings to be held not earlier than six months prior to the WTSA.

8 Internet related matters

There were extensive discussions on Internet matters, especially due to many controversial proposals from the Arab States. The EU coordination was conducted by the Netherlands (Mr W. Rullens). The result of the deliberations was satisfactory to European members.

The mandate of SG2 “Operational aspects of service provision, networks and performance” was not modified to introduce Internet. The debates on Internet were mainly focused on two technical aspects: security and naming and numbering issues. The Assembly adopted a number of resolutions and related measures on Internet matters acceptable for CEPT:

8.1 ITU-T’s contribution to WSIS

The issue of Internet Governance was discussed in the context of how to include ITU-T into the WSIS (World Summit on the Information Society) process. The proposed Internet Governance Group was not created, but the WTSA decided to create a short life group within ITU-T with the mandate of providing input on the technical aspects of the telecommunication networks used by the Internet to the first meeting of the ITU Council Working Group on WSIS in 2005.

8.2 Measures to combat spam

WTSA-04 approved a resolution (Res 52) tasking the ITU-T Study Groups to work with the IETF (Internet Engineering Task Force) and others to develop technical recommendations aimed at countering spam. TSAG will follow up work carried out in this area for further guidance.

Another resolution on combating spam (Res 51) instructs the TSB Director “to prepare urgently a report to the Council on relevant ITU and other international initiatives for countering spam, and to propose possible follow-up actions for consideration by the Council” and invites Member States to take the necessary steps within their national legal frameworks to ensure that appropriate measures are taken to combat spam.

8.3 Naming, numbering, addressing, routing and identification resources

Naming and numbering issues are part of the traditional scope of ITU-T work, mainly in its Study Group 2 responsible for the operational aspects of service provision, networks and performance. SG2 was tasked by the WTSA to examine naming, numbering, addressing and routing (NNAR) for NGN.

WTSA revised Resolution 20 on the allocation of international telecommunication numbering, naming, addressing and identification codes such as country code, signalling area and network codes, data country codes, mobile country codes etc. The text recognises ITU’s responsibility in the area of assignment of such international resources. As proposed by CEPT in its ECP4, the resolution requests the Study Groups to give guidance on these issues, especially regarding complaints about misuse of an international numbering resource.

8.4 ENUM

ENUM is an Internet telephone number mapping protocol. This protocol allows consumers to use a single number to access many types of terminals and services such as phone, e-mail, mobile phones, websites or any other service available through an Internet addressing scheme.

WTSA tasked SG2 to study how ITU could have administrative control over changes relating to international telecommunication resources such as naming, numbering, addressing and routing used for ENUM.

8.5 Country code top level domain names – ccTLD

WTSA adopted a new resolution (Resolution 44) instructing ITU-T SG2 to cooperate with governments and industry to review Member States’ ccTLD experiences.

8.6 Internationalised Domain Names (IDN)

A new Question on internationalised domain names was surprisingly assigned to Study Group 17. The Study Group will engage in an in-depth discussion of the political, economic, and technical issues related to IDN.

9 Resolution on Cybersecurity

Following the workshop on cybersecurity organised immediately before the WTSA, the Assembly adopted Resolution 50 asking ITU-T to evaluate its Recommendations, especially in the area of signalling and communications protocols in order to prevent potential exploitation by malicious parties. ITU-T is also tasked to raise awareness of the need to defend information and communications systems against the threat of cyber attack.

10 Call back

The existing Resolution 29 "Alternative calling procedures on international telecommunication networks" was amended and SG3 is requested to study the economic effect of call-back and other similar calling practices in developing countries and how they impact on their ability to develop their telecommunication networks and services.

11 Greater involvement of developing countries in standardisation activities

During this WTSA there was an increased emphasis on the role played by the developing countries in ITU-T. There were extensive debates on how to ensure the active participation and involvement of developing countries, least developed countries (LDCs) and countries with economies in transition in the work of the ITU-T and how initiatives could be taken with the aim of bridging the digital divide.

11.1 Bridging the standardisation gap

In Resolution 44 "Bridging the standardisation gap between developing and developed countries", the Assembly adopted an action plan based on five programmes:

Programme I: Strengthening standard-making capabilities

This programme includes the introduction of webcasting systems enabling developing-country experts to follow Study Group meetings from their office work-stations and consultancy projects to support developing countries in the development of standardisation plans, strategies, policies, etc.

Programme II: Assisting BDT in enhancing efforts in respect of standards application

The second programme will ensure that developing countries have a clear understanding of ITU-T Recommendations and will assist the BDT (Bureau de dévéloppement des télécommunications) i.a. in assessing whether existing national standards of developing countries are consistent and in accordance with the current ITU-T Recommendations.

Programme III: Human resource building

This will be achieved through the organisation of seminars, workshops and Study Group meetings in developing countries and through the establishment of a forum, moderated by a group of experts, to support and provide advice to standardisation bodies in developing countries.

Programme IV: Flagship groups for bridging the standardisation gap

This programme encourages direct voluntary support from developed countries in assisting groups of developing countries – so-called flagship groups – in their standardisation activities.

Programme VI: Fundraising for bridging the standardisation gap

Invitation to the public and private sector to fund the implementation of the action plan

11.2 Regional groups

Resolution 54 calls for the establishment of regional groups based on the model already used by SG3. Regional groups can serve as a forum where needs of the countries of the region are discussed and agreed for submission to the appropriate Study Group. Chairmen of regional groups act as a liaison with countries of their region to inform about the work of the Study Group.

12 Gender mainstreaming in ITU-T

For the first time mainstreaming of gender issues was discussed at a WTSA. The Assembly adopted Resolution 55 encouraging the inclusion of a gender perspective in the work of TSAG and the ITU-T Study Groups over the next four-year cycle and the mainstreaming of a gender perspective in the work of the TSB.

13 European coordination during the WTSA

Ms Marie Thérèse Alajouanine, France, the chairman of the CEPT PT WTSA-04, held several coordination meetings during the Assembly open for members both from EICTA and ETNO. An informal coordination meeting for ETNO members was organised by

Mr Dominique Würges, France Telecom, during the WTSA-04, providing the possibility to express points of interest to ETNO members, in particular highlighting key points such as NGN and the appointment of competent chairmen in the SGs.

Despite the extensive coordination efforts made by CEPT, certain difficulties were encountered during the course of the Assembly. In general, there were not enough countries to support the ECP and the coordinator was often left alone to fight the battle. In connection with the "show of hands" related to the dispute on the veto procedure in the AAP, some European countries voted against the CEPT position outlined in the CEPT Brief, instead of abstaining. Such difficulties can easily undermine a united CEPT approach and should be resolved for the next Assembly.

Abbreviations

AAP	Alternative Approval Procedure
BDT	Bureau de développement des télécommunications
ccTLD	Country code Top Level Domain Names
CEPT	Conférence européenne des administrations des postes et des telecommunications
CEPT PT-WTSA-04	CEPT Project Team on World Telecommunication Standardization Assembly 04
CITEL	Inter-American Telecommunication Commission
CV	ITU Convention
CS	ITU Constitution
CWG	Council working group
ECP	European common proposal
EICTA	European Information and Communications Technology Industry Association
ENUM	Electronic numbering
ETNO	European Telecommunications Network Operators' Association
IDN	Internationalised Domain Names
Ipv6	Internet Protocol addresses version 6
IETF	Internet Engineering Task Force
IMT-2000	International Mobile Communications 2000
IP	Internet Protocol
IT	Information technology
ITU	International Telecommunication Union
ITU-T	Telecommunication Standardization Sector
LDCs	Least Developed Countries
NGN	Next Generation Networks
NNAR	Naming, numbering, addressing and routing
PP02	Plenipotentiary Conference 2002
SCC	Seminar and workshop coordination committee
SDO	Standards Development Organisation
SG	Study Group
SSG	Special Study Group
TSAG	Telecommunication Standardization Advisory Group
TSB	Telecommunication Standardization Bureau
WSIS	World Summit on the Information Society
WTSA	World Telecommunication Standardization Assembly

Anne Lise Lillebø is Senior Adviser, Telenor Corporate Communications. Her main responsibilities include frequency coordination and policy matters related to international standardisation organisations. She participated in CCITT in 1988 and in the WTSAs in 1993, 1996, 2000 and 2004. She holds a Master of Arts degree from the University of Oslo.

email: anne-lise.lillebo@telenor.com

Terms and acronyms in Information Society and Security

AES	Advanced Encryption Standard	Is also known as Rijndael. In cryptography, it is a block cipher adopted as an encryption standard by the US government. It is expected to be used worldwide and is analysed extensively, as was the case with its predecessor, the Data Encryption Standard (DES). AES was adopted by National Institute of Standards and Technology (NIST) in November 2001 after a 5-year standardisation process. The cipher was developed by two Belgian cryptographers, Joan Daemen and Vincent Rijmen, and submitted to the AES selection process under the name "Rijndael", a blend comprising the names of the inventors. www.nist.gov
AP	Access Point	A point where users access the system/network, e.g. a base station in a wireless network.
ARP	Address Resolution Protocol	Protocol used in Ethernets and WLANs for requesting conversion of the IP address to the MAC address in order to send messages to the right terminal on the network.
ARPA	Advanced Research Project Agency	ARPA was established in 1958 in response to the Soviet launching of Sputnik, with the mission of keeping the US military technology ahead of its enemies. It was renamed DARPA (for Defense) in 1972, then back to ARPA in 1993, and then back to DARPA again on March 11, 1996. DARPA is independent from other more conventional military R&D and reports directly to senior Department of Defense management. DARPA has around 240 personnel (about 140 technical) directly managing a $2 billion budget. DARPA focuses on short-term (two to four year) projects run by small, purpose-built teams. ARPA was responsible for funding the development of ARPANET (which grew into the Internet), as well as the Berkeley version of Unix (BSD) and TCP/IP. www.darpa.mil
AS network	Autonomous System network	Term used for large national networks – the backbone of Internet
BAS	Beskyttelse Av Samfunnet (Eng: protection of society)	A series of projects run by the Norwegian Defence Research Establishment (FFI) in order to analyse the vulnerability and civil preparedness of the society. www.ffi.no
BGP	Border Gateway Protocol	The core routing protocol of the Internet. It works by maintaining a table of IP networks or 'prefixes' which designate network reachability between autonomous systems (AS). It is described as a path vector protocol. BGP does not use technical metrics, but makes routing decisions based on network policies or rules. The current version of BGP, BGP version 4, is specified in request for comment RFC 1771. http://www.ietf.org/rfc/rfc1771.txt
BMA	British Medical Association	The British Medical Association represents doctors from all branches of medicine all over the UK. It is a voluntary association with about 75 per cent of practising doctors in membership. It has a total membership of over 134,000. www.bma.org
CAIDA	Cooperative Association for Internet Data Analysis	The Cooperative Association for Internet Data Analysis (CAIDA) is a collaborative undertaking among organisations in the commercial, government, and research sectors aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global Internet infrastructure. www.caida.org
CART	Computer Analysis and Response Team	Division within the US Federal Bureau of Investigations – FBI. The Computer Analysis and Response Team provides assistance to FBI field offices in the search and seizure of computer evidence as well as forensic examinations and technical support for FBI investigations. http://www.fbi.gov/hq/lab/org/cart.htm

CD	Compact Disc	An optical disc used to store digital data, originally developed for storing digital audio. A standard compact disc, often known as an audio CD to differentiate it from later variants, stores audio data in a format compliant with the Red Book standard. An audio CD consists of several stereo tracks stored using 16-bit PCM coding at a sampling rate of 44.1 kHz. Standard compact discs have a diameter of 120 mm, though 80-mm versions exist in circular and "business-card" forms. The 120-mm discs can hold 74 minutes of audio, and versions holding 80 or even 90 minutes have been introduced. The 80-mm discs are used as "CD-singles" or novelty "business-card CDs". They hold about 20 minutes of audio. Compact disc technology was later adapted for use as a data storage device, known as a CD-ROM. The first edition of the Red Book was released in June 1980 by Philips and Sony; it was adopted by the Digital Audio Disc Committee and ratified as IEC 908. http://www.iec.ch
CNA	Computer Network Attack	Term used in electronic warfare.
CobiT	Control objectives for information and related Technology	A framework for information security. It provides managers, auditors, and IT users with a set of generally accepted information technology control objectives to assist them in maximising the benefits derived through the use of information technology and developing the appropriate IT governance and control in a company. In its 3rd edition, CobiT has 34 high level objectives that cover 318 control objectives categorised in four domains: Planning and Organization, Acquisition and Implementation, Delivery and Support, and Monitor. It comprises six elements: management guidelines, control objectives, CobiT framework, executive summary, audit guidelines and an implementation toolset. It was developed by the IT Governance Institute (ITGI) and the Information Systems Audit and Control Foundation (ISACA) in 1992 when the control objectives relevant to information technology were first identified. The first edition was published in 1996; the second edition in 1998; the third edition in 2000, and the on-line edition became available in 2003. It has more recently found favour due to external developments, especially the Enron scandal and the subsequent passage of the Sarbanes-Oxley Act. The CobiT mission is "to research, develop, publicize and promote an authoritative, up-to-date, international set of generally accepted information technology control objectives for day-to-day use by business managers and auditors". Managers, auditors and users benefit from the development of CobiT because it helps them understand their IT systems and decide the level of security and control that is necessary to protect their companies' assets through the development of an IT governance model. www.isaca.org, www.itgi.org
COSO	The Committee of Sponsoring Organizations of theTreadway Commission	Guidelines for internal control. COSO refers to a document providing a common definition of internal controls, standards, and criteria against which companies and organizations can assess their control systems. It is one of the most widely-accepted internal control frameworks for the audit of internal controls.
COTS	Commercial Off-The-Shelves	A product that is used "as-is". COTS products are designed to be easily installed and to interoperate with existing system components.
CPU	Central Processing Unit	The part of a computer that interprets and carries out the instructions contained in the software. The term processor usually refers to a CPU as well.
CSIRT	Computer Security Incident Response Team	Team set up by an organisation to handle computer security incidents.
CSNET	Computer Sciences Research Network	A network created by the US National Science Foundation in the early 1980s. It was a computer network linking academic Computer Science departments nationwide. It was an alternative to ARPANET, to which many Computer Science departments did not have the privilege of access. CSNET connected with ARPANET using TCP/IP and ran TCP/IP over X.25, but it also supported departments without sophisticated network connections, using automated dial-up mail exchange. It was a forerunner to NSFNET. CSNET operated autonomously until 1989, when it merged with Bitnet to form CREN (Corporation for Research and Educational Networking). By 1991 the growth of the Internet had made the CSNET services redundant, and CREN discontinued them.
DSD	Dynamic Separation of Duties	Constraints on permission to activate assigned roles in RBAC (Role Based Access Control).

DSS.1	Digital Subscriber Signalling System no. 1	Also known as Euro-ISDN or E-DSS1 (European DSS1), a digital signalling protocol (D channel protocol) used for ISDN.
DVD	Digital Versatile Disc	Formerly Digital Video Disc. Data storage format released in 1995. The discs have the same physical size as the CD, but the capacity is more that 7 times higher, approx. 4.7 GB on one side. The discs can have dual layers per side, thus a double-sided, dual-layer disc can store approx. 17 GB of data. Used for storing video, sound, computer software and data, games, etc. A single-sided, single-layer disc can store a typical feature film of 130 minutes with 8 different surround quality sound tracks. Available as read-only (DVD-Video, DVD-ROM), Once writable (DVD-R, DVD+R) and Re-writable (DVD-RW, DVD+RW, DVD-RAM). www.dvdforum.org
EMR	Electronic Medical Record	
ETSI	European Telecommunications Standards Institute	A non-profit membership organization founded in 1988. The aim is to produce telecommunications standards to be used throughout Europe. The efforts are coordinated with the ITU. Membership is open to any European organisation proving an interest in promoting European standards. It was e.g. responsible for the making of the GSM standard. The headquarters are situated in Sophia Antipolis, France. www.etsi.org
EVC	EigenVector Centrality	A structural measure of networks (graphs).
FACTS	Flexible AC/DC transmissions	
FBI	Federal Bureau of Investigations	US federal agency founded in 1908, comprising almost 30 000 employees in 2004. The mission is to "to protect and defend the United States against terrorist and foreign intelligence threats and to enforce the criminal laws of the United States". www.fbi.gov
FFI	Forsvarets Forskningsinstitutt (Eng: Norwegian Defence Research Establishment – NDRA)	Norwegian Defence Research Establishment is the prime institution responsible for defence-related research in Norway. The Establishment is also the chief adviser on defence-related science and technology to the Ministry of Defence and the Norwegian Armed Forces' military organisation. www.ffi.no
ICT	Information and Communication Technology	
IDS	Intrusion Detection System	A software/hardware tool used to detect unauthorised access to a computer system or network. This may take the form of attacks by skilled malicious hackers or Script kiddies using automated tools. An IDS is required to detect all types of malicious network traffic and computer usage. This includes network attacks against vulnerable services, data driven attacks on applications, host based attacks such as privilege escalation, unauthorised logins and access to sensitive files, and malware (viruses, Trojan horses, and worms).
IEEE	The Institute of Electrical and Electronic Engineers	USA based organisation open to engineers and researchers in the fields of electricity, electronics, computer science and telecommunications. Established in 1884. The aim is to promote research through journals and conferences and to produce standards in telecommunications and computer science. IEEE has produced more than 900 active standards and has more than 700 standards under development. Divided into different branches, or 'Societies'. Has daughter organisations, or 'chapters' in more than 175 countries worldwide. Headquarters in Piscataway, New Jersey, USA. www.ieee.org
IPS	Intrusion Prevention System	Any device which exercises access control to protect computers from exploitation. "Intrusion prevention" technology is considered by some to be an extension of intrusion detection (IDS) technology, but it is actually another form of access control.
ISDN	Integrated Services Digital Network	A digital telecommunications network that provides end-to-end digital connectivity to support a wide range of services, including voice and non-voice services, to which users have access by a limited set of standard multi-purpose user-network interfaces. The user is offered one or more 64 kb/s channels. www.itu.int

ISF	Information Security Forum	Non-for-profit organisation with more than 250 members, the ISF aims to deliver practical guidance and solutions to overcome wide-ranging security challenges impacting business information. http://www.securityforum.org/
ITU	International Telecommunications Union	On May 17, 1865, the first International Telegraph Convention was signed in Paris by the 20 founding members and the International Telegraph Union (ITU) was established to facilitate subsequent amendments to this initial agreement. It changed name to the International Telecommunications Union in 1934. From 1948 a UN body with approx. 200 member countries. It is the top forum for discussion and management of technical and administrative aspects of international telecommunications. www.itu.int
LAN	Local Area Network	A network shared by communicating devices, usually in a small geographical area. A system that links together electronic office equipment, such as computers and word processors, and forms a network within an office or building.
MAC	Medium Access Control	The lower of the two sub layers of the Data Link Layer. In general terms, MAC handles access to a shared medium and can be found within many different technologies. For example, MAC methodologies are employed within Ethernet, GPRS and UMTS.
MFENET	Magnetic Fusion Energy Network	A research network of the US Department of Energy
MTU	Main Terminal Unit	
NSFNET	National Science Foundation Network	A major part of early 1990s Internet backbone. Organisations connecting to the Internet in the early 1990s had to sign a usage agreement directly with NSFNET to gain access to large parts of the Public Internet, regardless of which Internet Service Provider they purchased Internet access from. From 1987 to 1995 the NSFNET was operated on behalf of the NSF by Merit Network, Inc, a non-profit corporation governed by public Universities. On April 30, 1995, the NSFNET Backbone Service was successfully transitioned to a new architecture, where traffic is exchanged at interconnection points called NAPs (Network Access Point).
NTNU	Norwegian University of Science and Technology	www.ntnu.no
OECD	Organization for Economic Cooperation and Development	The OECD groups 30 member countries sharing a commitment to democratic government and the market economy. With active relationships with some 70 other countries, NGOs and civil society, it has a global reach. Best known for its publications and its statistics, its work covers economic and social issues from macroeconomics to trade, education, development and science and innovation. The OECD produces internationally agreed instruments, decisions and recommendations to promote rules of the game in areas where multilateral agreement is necessary for individual countries to make progress in a globalised economy. www.oecd.org
OS	Operating System	
PDA	Personal Digital Assistant	Handheld device that combines computing, telephone/fax, Internet and networking features. A typical PDA can function as a cellular phone, fax sender, Web browser and personal organiser.
PDS	Politiets Datakrimsenter	The National Computer Crime Center of the Norwegian Police.
PKI	Public Key Infrastructure	An arrangement which provides for third-party vetting of, and vouching for user identities. It also allows binding of public keys to users. This is usually carried by software at a central location together with other coordinated software at distributed locations. The public keys are typically in certificates. The term is used to mean both the certificate authority and related arrangements as well as more broadly and somewhat confusingly to mean use of public key algorithms in electronic communications. The latter sense is erroneous since PKI methods are not required to use public key algorithms.
PTO	People Technology Organisation	

RBAC	Role Based Access Control	An approach to restricting system access to authorised users. http://csrc.nist.gov/rbac/
RC4	Rivest Cipher 4	One of several ciphers proprietary to the RSA Data Security Inc. Also called ARCFOUR, it is the most widely-used software stream cipher and is used in popular protocols such as Secure Sockets Layer (SSL) (to protect Internet traffic) and WEP (to secure wireless networks). RC4 falls short of the high standards of security set by cryptographers, and some ways of using RC4 lead to very insecure cryptosystems (including WEP). RC4 was designed by Ron Rivest of RSA Security in 1987; while it is officially termed "Rivest Cipher 4", the RC acronym is alternatively understood to stand for "Ron's Code". RC4 was initially a trade secret, but in September 1994 a description of it was anonymously posted to the Cypherpunks mailing list. It was soon posted on the sci.crypt newsgroup, and from there to many sites on the Internet. Because the algorithm is known, it is no longer a trade secret.
RFID	Radio Frequency IDentification	A method of storing and remotely retrieving data using devices called RFID tags or transponders. An RFID tag is a small object that can be attached to or incorporated into a product, animal, or person. RFID tags contain antennas to enable them to receive and respond to radio-frequency queries from an RFID transceiver. Passive tags require no internal power source, whereas active tags require a power source.
RTU	Remote Terminal Unit	
SCADA	Supervisory Control And Data Acquisition	Systems used in industrial and engineering applications to monitor and control distributed systems from a master location. SCADA is a very broad umbrella that describes solutions across a large variety of industries.
SI	Susceptible-Infected	Two-state model for spreading of malicious software where nodes may be susceptible to infection or infected.
SIR	Susceptible-Infected-Refractory	Model where nodes cure after a while and then stay immune for a period.
SIS	Susceptible-Infected-Susceptible	Alternative model where nodes stay infected only for a certain time and then again become susceptible.
SMART	Specific, Measurable, Attainable, Repeatable and Time dependent	Characteristics of a good metric.
SOX	Sarbanes-Oxley Act of 2002	Corporate governance rules, regulations and standards for specified public companies including SEC registrants (SEC: security – in this case a particular subcategory of companies certified for information security (revision, provision, consulting etc)).
SPAN	Space Physics Analysis Network	A research network for NASA space physicists. www.nasa.gov
SSD	Static Separation of Duties	Constraints on the assignment of roles to users in RBAC.
SSID	Service Set IDentifier	A code attached to all packets on a wireless network to identify each packet as part of that network. The code consists of a maximum of 32 alphanumeric characters. All wireless devices attempting to communicate with each other must share the same SSID. Apart from identifying each packet, SSID also serves to uniquely identify a group of wireless network devices used in a given "Service Set". There are two major variants of the SSID. Ad-hoc wireless networks that consist of client machines without an access point use the BSSID (Basic Service Set Identifier); whereas on an infrastructure network which includes an access point, the ESSID (E for Extended) is used instead. Each of these different types may be referred to in general terms as SSID. A network's SSID is often referred to as the "network name" and is commonly set to the name of the network operator, such as a company name. An extremely weak form of wireless network security is to turn off the broadcast of the SSID: to the average user there does not appear to be a network in use; it is however still readily available to hackers using the appropriate tools.

SWGDE	Scientific Workgroup on Digital Evidence	An organisation composed of member agencies from all levels of government. It brings together organisations actively engaged in the field of digital and multimedia evidence to foster communication and cooperation as well as ensuring quality and consistency within the forensic community. www.swgde.org
TCP/IP	Transport Control Protocol/ Internet Protocol	TCP: Transport layer protocol defined for the Internet by Vint Cerf and Bob Kahn in 1974. A reliable octet streaming protocol used by the majority of applications on the Internet. It provides a connection-oriented, full-duplex, point-to-point service between hosts. IP: A protocol for communication between computers, used as a standard for transmitting data over networks and as the basis for standard Internet protocols. www.ietf.org
TID	Time stamps in Digital Forensic	Research project at NTNU
TTP	Trusted Third Party	In cryptography, an entity which facilitates interactions between two parties who both trust the third party; they use this trust to secure their own interactions. TTPs are common in cryptographic protocols, for example, a certificate authority (CA).
UUCP	Unix to Unix Copy Protocol	A computer program and protocol allowing remote execution of commands and transfer of files, email and netnews between Unix computers
VLSI	Very Large Scale Integrated (circuit)	
VPN	Virtual Private Network	A network that is constructed by using public wires to connect nodes. For example there are a number of systems that enable you to create networks using the Internet as the medium for transporting data. These systems use encryption and other security mechanisms to ensure that only authorized users can access the network and that the data cannot be intercepted.
WEP	Wired Equivalent Privacy	An implementation of RC4. It is part of the IEEE 802.11 standard (ratified in September 1999), and is a scheme used to secure wireless networks (WiFi). WEP was designed to provide comparable confidentiality to a traditional wired network, hence the name. www.ieee802.org
WIDS	Wireless Intrusion Detection System	
WLAN	Wireless Local Area Network	This is a generic term covering a multitude of technologies providing local area networking via a radio link. Examples of WLAN technologies include Wi-Fi (Wireless Fidelity), 802.11b and 802.11a, HiperLAN, Bluetooth and IrDA (Infrared Data Association). A WLAN access point (AP) usually has a range of 20 –300 m. A WLAN may consist of several APs and may or may not be connected to Internet.
WPA	WiFi Protected Access	An improved version of WEP (Wired Equivalent Privacy). It is a system to secure wireless (Wi-Fi) networks, created to patch the security of WEP. As a successor, WPA implements the majority of the IEEE 802.11i standard, and was intended as an intermediate measure to take the place of WEP while 802.11i was being prepared. www.ieee802.org
Y2K	Year 2000	

ISSN 0085-7130 © Telenor ASA 2005