telektronikk 1.08

Network Analysis

Telektronikk
Volume 104 No. 1 – 2008
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&I
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&I

Printing:
07 Gruppen as, Aurskog

Circulation:
4,000

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

Guest Editorial

GEOFFREY S CANRIGHT, KENTH ENGØ-MONSEN

Geoffrey S. Canright is senior researcher in Telenor R&I

Kenth Engø-Monsen is senior researcher in Telenor R&I

This issue of *Telektronikk* is devoted to "network analysis". You, the reader, may immediately think, "Of course, telecommunications companies have built and maintained and analyzed networks for many decades … and here comes more of the same." If so, please be assured that this issue is not an example of "more of the same". Here we offer studies in which the network in "network analysis" is not necessarily physical, and the important properties of the network are not primarily traffic and capacity. Instead we wish to present the general *concept* of a network (or graph), and a variety of interesting applications of this concept.

We motivate the idea with a visual example (Figure 1), taken from telecommunications data. Figure 1 shows a network that is not physical however – the links are not cables or fibers, and the nodes are not routers or switches. Instead, the nodes are *people*, and the links are *social relationships* (as measured by communication). Hence this is an empirically obtained *social network*.

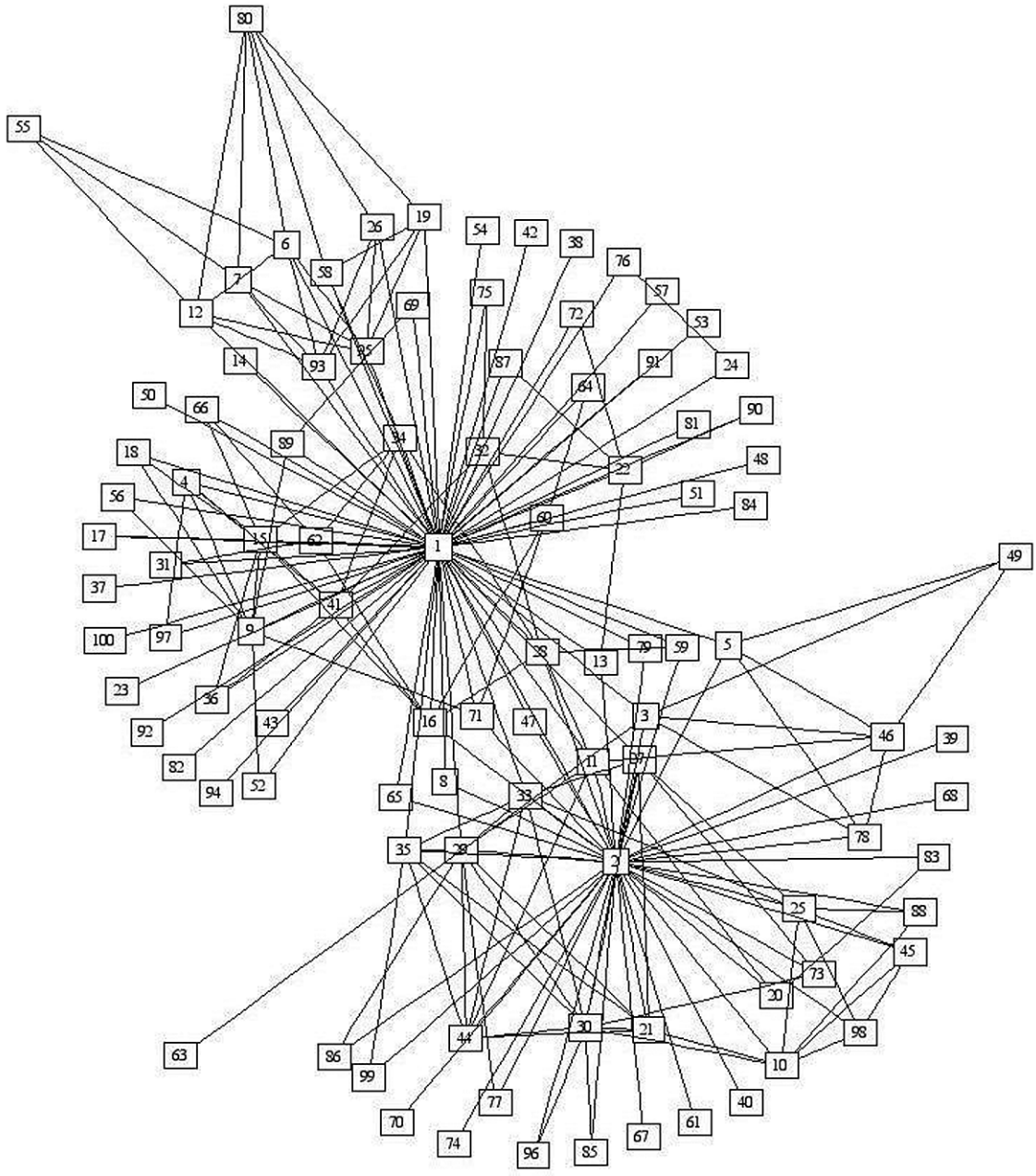


Figure 1 A (part of a) social network, as obtained from empirical call data

ISSN 0085-7130

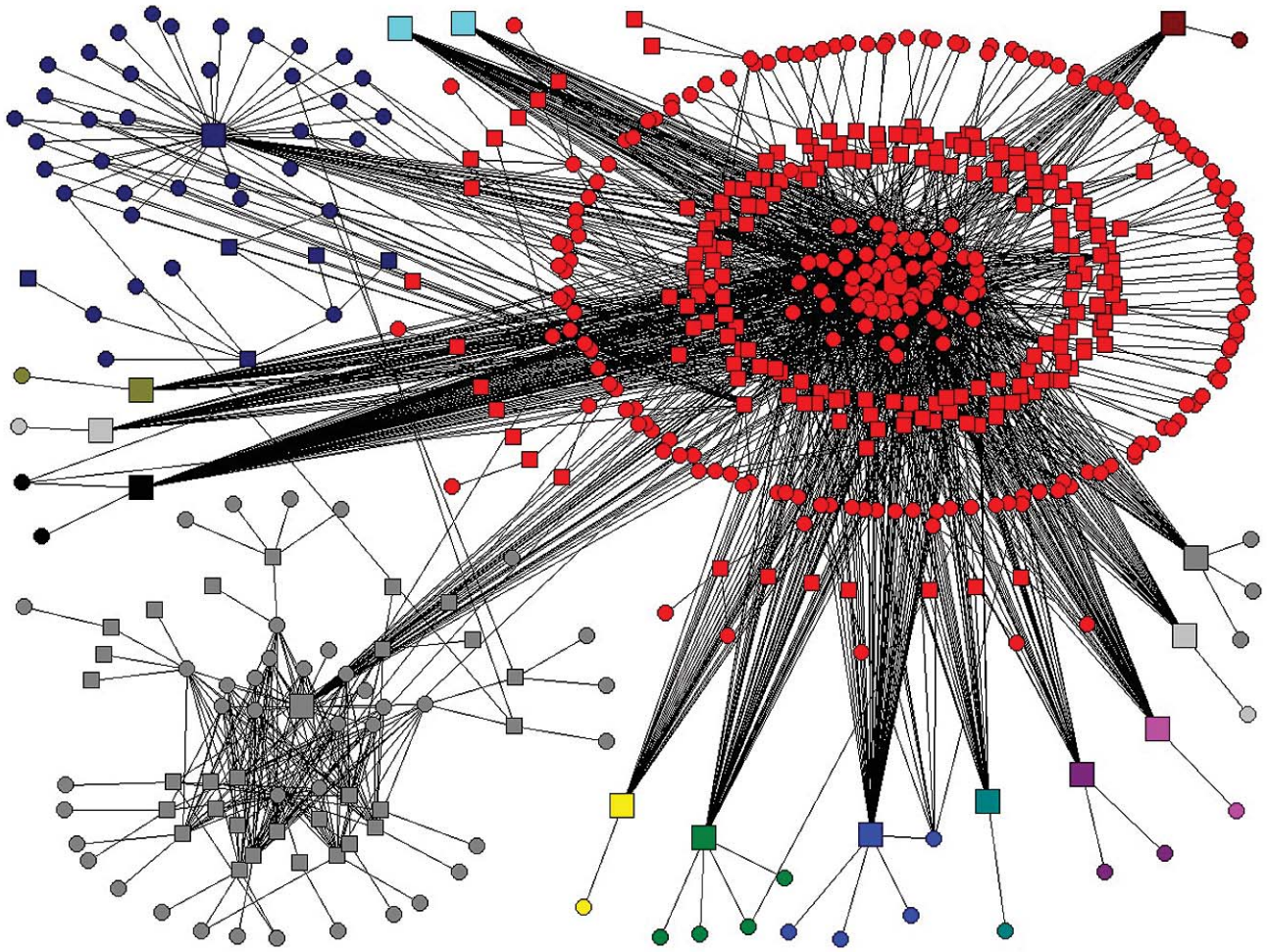


Figure 2 Another social network (the nodes are people); but here the links represent sexual contact

The network of Figure 1 is in a sense a fairly abstract thing – seemingly just points connected by lines – and yet it captures an important social reality, which is taking place in the physical world: humans communicating with one another. We see a lot of structure in Figure 1, even though it is highly abstract. Understanding this structure is an important part of the field of *social network analysis* or SNA – a field which is well represented in this issue, and one which is clearly highly relevant for the telecommunications industry.

Figure 2 shows another social network. It seems, as shown, much like that of Figure 1; but it represents something quite different. The difference is simply stated: while the nodes are again people, the links in Figure 2 now represent sexual contacts. This picture was generated as part of a study seeking ways to inhibit the spread of the HIV virus over such networks. Hence, understanding this network can be the difference between life and death for the people in this network – represented here as simple squares (men) and circles (women).

Figures 1 and 2 are each typical examples of the concept of a network: there are individual "units" (in each case here, people), connected by "links". The phenomena taking place over these networks, and the possible reasons for studying them, may be quite different – but the underlying network *concept* is the same. This concept underlies and unifies all of the studies presented in this edition of *Telektronikk*.

We want to justify the use of the word "unifies" in the previous sentence. Suppose researcher A wishes to study how a rumor is passed over the social network of Figure 1, while Researcher B wants to understand how the HIV virus is spread over the network of Figure 2. Clearly, researchers A and B can learn from one another's efforts! Not only can they share a common understanding of the *structure* of the networks of interest; they also study *dynamical processes* whose rules for propagation over the network are very similar.

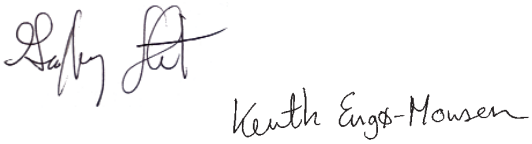
These brief examples are intended simply to whet the appetite of you, the reader. The points we wish to make may be stated thus:

- The network concept may be found and applied almost everywhere, in all disciplines of science and engineering.

- Because the field is so naturally interdisciplinary, understanding of networks gleaned from one discipline may often be usefully applied in one or several others.

In this issue, we present papers on telecommunications networks, but also on social networks, economic (value) networks, the 'Web graph' (a network composed of Web pages and their hyperlinks), and more. For a more thorough introduction, we encourage you to visit the article "Introducing network analysis"; this article offers a basic discussion of modern ideas in network analysis, and also serves as a guide to the other articles in this issue. We hope that you, the reader, will enjoy sampling the broad spectrum of problems and ideas which these articles represent, and that you will appreciate the basic utility and beauty of the network concept.

Welcome to today's world of networks!



Geoffrey S. Canright is a senior researcher with Telenor R&I. His background is in statistical physics. His current interests include network analysis and graph theory, social networks, Web search and mobile search, and self-organizing systems.

email: geoffrey.canright@telenor.com

Kenth Engø-Monsen is a senior researcher in Telenor R&I. 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 2008

Introducing Network Analysis

GEOFFREY S CANRIGHT, KENTH ENGØ-MONSEN

Geoffrey S. Canright is senior researcher in Telenor R&I

Kenth Engø-Monsen is senior researcher in Telenor R&I

This article serves as an introduction to this special issue of *Telektronikk* on network analysis. The introduction consists of three parts. In part 1 we give the motivation for studing networks as a general concept with wide applicability. We support this idea with a variety of example applications. Then in part 2 we offer an overview of many of the basic ideas and terms in modern network analysis, including where possible, clear quantitative definitions. The terms in part 2 are terms that turn up in a wide variety of fields where networks are important. Finally, in part 3 we give a "reader's guide" to the other articles in this issue.

1 Introduction

Networks are everywhere! Let us start with the familiar telecommunications industry. Telecommunications operators around the world have for more than a hundred years built infrastructure enabling people to talk to each other. The main building block of this physical network is copper wires and fibres, which connect people to each other in such a way that, in principle, anyone who is connected can call any other person connected to the same network. Another industry building a huge physical network infrastructure is the electric power industry. The power lines transport electricity from the point of production (i.e. a power plant) and all the way out to households that need electric power for lighting, hot water, and heating. It is not immediately apparent when you power up the DVD player or the laptop at home that you are tapping into a huge network to in fact get the electricity that you need. Still, a network is the main building block lurking in the background that enables you to draw off the electricity that you need for your household appliances.

Let us stop one moment, and think about the above two examples. Deliberately, the examples are two very familiar types of networks that everyone should know very well. The telecommunications network is a mediating type of network, where a physical net-

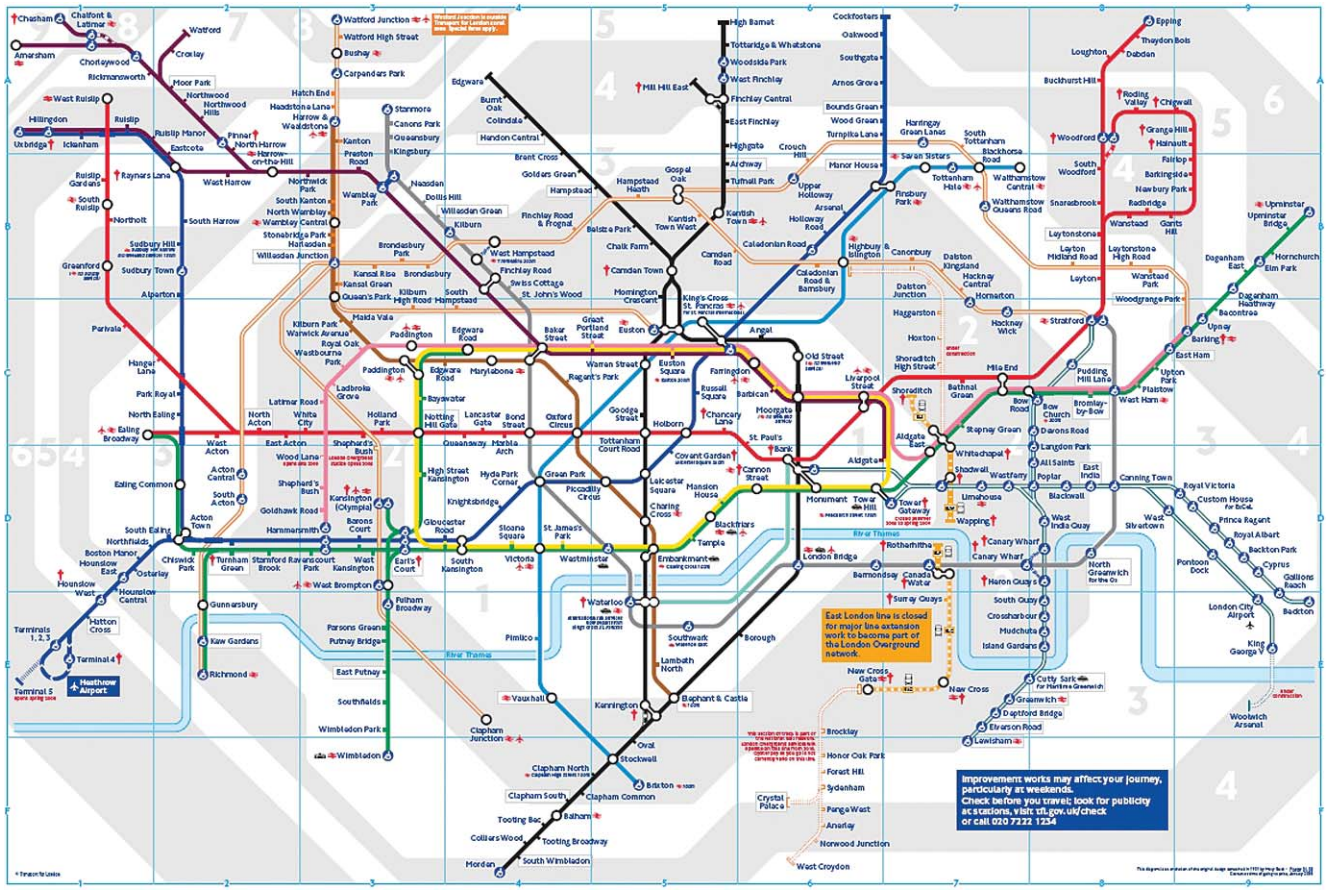


Figure 1 The London tube map!

ISSN 0085-7130 © Telenor ASA 2008

work has been built in order to support people in doing what they love the most; namely, to talk to each other. Before the era of telecommunications, people living far apart had to rely on a functioning postal service in order to communicate with each other, whereas with a telephone you just dial the number and when there is an answer at the other end you just start chatting. The electricity network, however, is an example of a transportation network. What is transported in this particular example is electric power, from the producer to the household. Maybe a more hands-on example of a transportation network is the network of public transportation in a large city. The unit being transported is people. These people are maybe going to work, visiting relatives or friends, or maybe they are just tourists travelling around sightseeing. What enables these people to travel around in the city is a network of bus lines, tram ways, and/or underground tube lines. To see the whole map of the London underground is nothing less than fascinating! And the map you see is a brilliant example of a transportation network (http://www.tfl.gov.uk/).

Often network systems are best understood as 'networks-on-top-of-networks'. This is also known as overlay networks. Again, let us consider a bus transportation network in a city. At the lowest level, there is a network of streets and roads that defines where cars and buses can drive. Similarly, the railroad tracks define where trains can go within a country. Then, on top of this physical infrastructure of streets and roads (railroad tracks) the public transportation company decides to operate bus lines on some of the streets. There does not need to be a bus line on every street; but we can view the set of all these bus lines as a bus transportation network that is running on top of – laid over – the network of streets and roads. Likewise, one can consider the network of fax machines to be a network that is running on top of the ordinary telecommunications network.

The examples so far have been mostly of a physical nature. Let us now look at networks that are of a more non-physical character. Even though owning a mobile phone enables you to call everybody else that also owns a mobile phone, you will most likely not do that. People are used to talking to their closest friends, family, and co-workers, etc., and it is not usual to call up a random person. Having a mobile phone enables you in principle to contact everyone else in the world having a phone, since the telecommunication infrastructure enables an all-to-all connection pattern among the end-users of a telecommunication service. However, people usually communicate with people within their social circles. The resulting logical, intangible network, with links defined by 'who communicates with whom', is an example of a social network. A telecommunications operator can in fact map out large parts of this social network by analyzing traffic records, and one will then most likely find that people are communicating with a few, on average tens, and in some rare extreme cases, hundreds of people. Nobody is talking to everybody else.

Other examples of networks are: airline networks; food webs (where the nodes are species, and a directed link S1 ==> S2 means "species 1 eats species 2"); film actor networks (where the nodes are actors, and actors are linked when they have played in one or more films together); scientific collaboration networks (like actor networks, but with links defined by joint papers rather than films); boardroom networks (here the nodes are companies, and a link means common board members); ownership networks of companies; trade networks; gas distribution networks; neural networks; gene regulatory networks (here the nodes are genes, and a directed link G1 ==> G2 means "gene 1 regulates the expression of gene 2"); value networks (see below); and there are many more. The examples continue into every field of science, and any conceivable application. The point is, whenever the objects of study are discrete, and the relationships between these objects can be defined and are of interest, a network model is appropriate and useful. By now we should have set the scene to be really thinking in terms of networks, and the reader should be able to spot networks everywhere!

The next fascinating thing about networks is how complex and counter-intuitive they can be when applied in modelling. For instance, model a system of city streets and roads as a network and analyze the throughput capacity of this network. What is important for planning purposes is to have a city traffic network that is able to handle varying demands in traffic. When adding a new tunnel or road, one is changing the capacity of the network, and one can in fact run the risk of making things worse traffic-wise, and causing more congestion for the cars in the streets after a new street or tunnel has been added to the traffic network. This phenomenon is usually called Braess' paradox, and can be stated as follows: adding extra capacity to a network, when the moving entities selfishly choose their route, can in some cases reduce overall performance [1]. This is because the equilibrium of such a system is not necessarily optimal. Hence, it can be of great importance to analyze systems that can be modeled as networks, because tampering with their topology can have the rather counter-intuitive effect of making things worse than what was planned for.

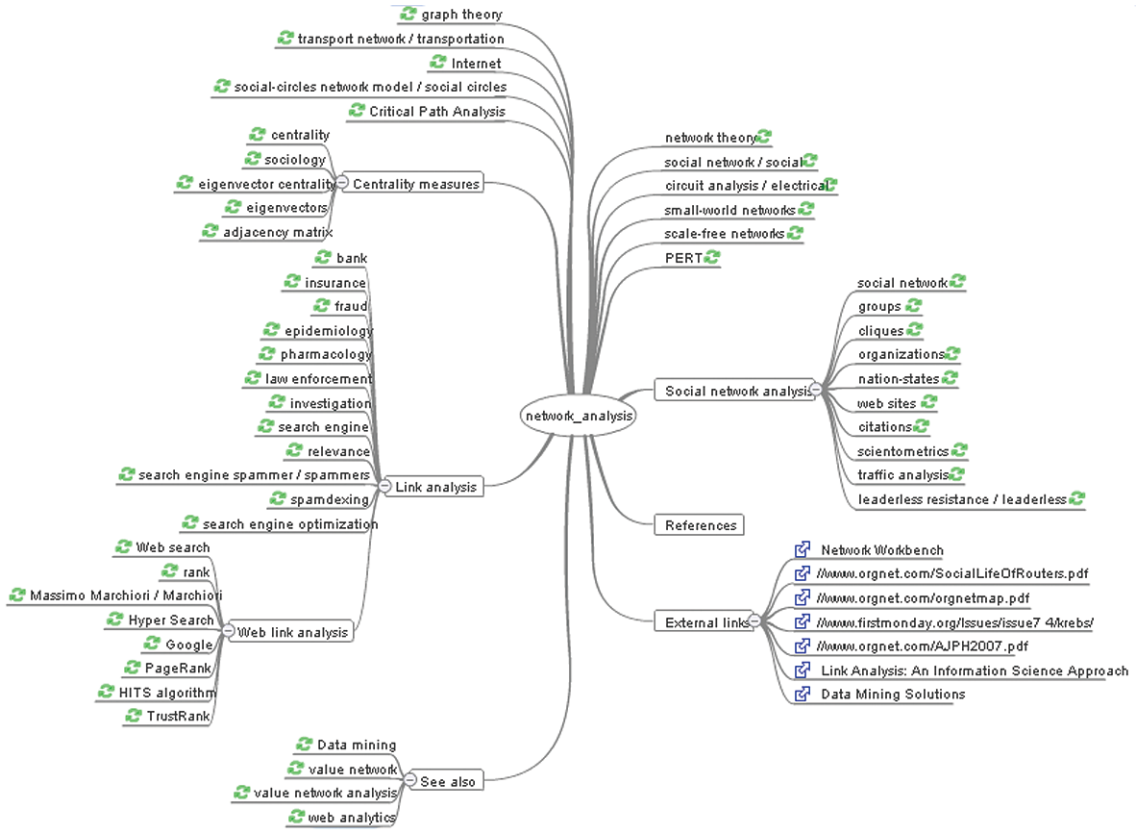


Figure 2 This shows the mind map obtained from the Network Analysis entry in Wikipedia, when using the WikiMindMap-tool at http://www.wikimindmap.com/

Combining economic and network modeling is a door into a world of problems that every telecommunications operator has to struggle with every day. Consider for example the following predicament: whom do you call, when you are the only one having a telephone? This is in fact the case for many people having a video-enabled 3G-phone today. They do not know whom to call (via video telephony) – hence the service of video telephony lacks utility for most users having the right type of mobile phone. Hence, the value for you – as an owner of a video-enabled mobile phone – depends on how many other users have the same video-enabled telephone for you to call. In economics this is called a network effect. A network effect is characterized by a good or service having a value to a potential customer which depends on the number (and/or identity) of other customers who own the good or are users of the service. Considering the whole life-cycle of a network service, network effects play the role of a barrier to adoption in the beginning, until the base of customers that have adopted the service reaches some critical mass. After that, network effects will make a positive contribution to the further adoption of the service, in that they will act as a catalytic force.

An example of everyday systems which display network effects is web marketplaces and exchanges. A well-known international example is *eBay*, while *finn.no* is a good Norwegian example. In each of these cases, the value of the marketplace to a new user is proportional to the number of other users in the market already. This is a good example of a rather pure "global network effect": because a new user is open to interacting with anyone who is already in the marketplace, the utility to new users depends only on the total number of users and not on who they are. The online, collaborative encyclopedia Wikipedia is another example illustrating global network effects. When the number of editors of the encyclopedia grows, the quality of the information on the website increases, which again encourages more users to turn to this website when looking for high-quality information. The process then continues when some of the new users are recruited as editors.

Social networking sites such as Facebook (http://www.facebook.com/), LinkedIn (http://www.linkedin.com/), and Nettby (http://www.nettby.no/), also display network effects, in that the more people that register for the social network service, the more useful the website is to its members. Here however we have a clear example of "local" network effects: the utility of Facebook to me is not determined only by the total number of existing Facebook members, but also by who they are: I want

my friends to be members, and am less interested in people I do not know.

We also see here that network effects are seldom purely global or purely local[1)]. After all, I as a new user will also want to use Facebook to meet new people – and the expected value of Facebook for meeting that need grows with the total number of Facebook members. Hence social networks like Facebook exhibit both local and global network effects.

The theme of this issue of *Telektronikk* is Network Analysis. In the above introductory remarks, we have tried to guide the reader into the world of networks through some examples. Contemporary network analysis is not just one, easily stated classical field of study – it is more the intersection of very different classical disciplines. And some of the ingredients might even come as a surprise to the reader.

What is network analysis? Let us try to answer this question by showing a visualization of the Network Analysis entry in Wikipedia [2]. We have used the Wiki mind-mapping tool at http://www.wikimind-map.org/, which is a very fascinating tool to use to learn about a particular topic (assuming that topic has an entry in Wikipedia).

Historically, network analysis has its origins in "graph theory", coming from mathematics and computer science. As can be seen from the mind-map in Figure 2, "graph theory" and "network theory" are present at the very top. Classically these fields involve the study of *graphs* – mathematical structures used to model pairwise relations between objects from a certain collection. Historically, graph theory has been considered as being born with the publication of a paper by Leonard Euler in 1736 on 'The seven bridges of Königsberg'. Euler considered the problem of finding a path crossing the seven bridges of Königsberg exactly once, and solved the problem by representing each connected piece of land as a vertex (node) of a graph, and each bridge as a link [3].

Mathematical graph theory was long occupied with sorting out the different structures that regular graphs have and witnessed a revitalization of the subject when the study of random graphs started in the middle of the 1900s. (For some recent overviews of graph theory, see [4] and [5].) Computer science addresses algorithmic aspects of graphs [6] – for example, how to find the shortest path between any two nodes in the graph, when knowing the global topology (the complete set of nodes and edges) of the graph. Another example is finding structural subgraphs within a larger graph (e.g. finding cliques – all-to-all connected subgraphs), finding link sets of maximum or minimum flow capacity, etc.

Returning for a moment to random graphs, the application of random graphs in describing real-world phenomena has had only limited success. Some recent developments that have advanced the understanding of naturally occurring network structures have been the understanding of small-world networks and scale-free networks.

Loosely speaking, a *small-world network* displays favorable properties according to both *local* and *global* criteria. First, a small-world network is locally well-connected – more quantitatively, one says that the network has high local *clustering* (see below). At the same time, the average shortest path (a global criterion) between two arbitrary nodes in a small-world network is small. This means that on average one can reach any other node in the network within a fairly small number of hops.

Let us set the small-world graph idea into context, by recalling a famous experiment done by Stanley Milgram in 1967 [7]. Milgram was speculating about how the network of acquaintances is structured: do we live in a 'small world', or not? The naïve assumption, leading to a negative answer, is this: people you know (acquaintances) usually live geographically close to you. By this assumption, it would seemingly require a very long chain of intermediate acquaintances for a letter/message to be delivered to some unknown person living geographically (and perhaps also socially) far away. A requirement for the experiment – which after all was seeking to probe properties of the acquaintanceship network – was that the letter/message can only be delivered to someone that you know personally. Milgram's experiment was constructed to find out how many intermediate acquaintances one would need on average to get a letter from the middle US (Kansas or Nebraska) to the east coast (Boston). What Milgram discovered was that those letters that reached their rightful destination used on average 5–6 hops. This small number of hops then contradicted the much larger number of hops that one expected based on the naïve assumption. The small-world network structure, in contrast, represents a good 'minimal' model to describe the network of acquaintances: one can model and measure local geographical acquaintances (and the fact that your friends are often friends of each other) by clustering, and the small number of hops needed to

1) *We thank Øystein D. Fjeldstad for introducing us to the ideas and thinking on local and global network effects in our joint research collaboration NABIRI.*

travel between arbitrary nodes in the network is well described by the small average shortest-path length.

The defining property of a *scale-free network* is that the distribution of the node degrees follows a power law. Roughly speaking, this means that, for a large scale-free network, there will be some (few) nodes with very large node degree, somewhat more nodes with slightly lower degree, and so on. This smooth variation of the node degree distribution (see the next section for details) – so that the fraction of nodes with very high node degree is not negligible – makes scale-free networks very well connected. Scale-free networks typically also satisfy the criteria for being small-world networks: they typically have high local clustering, and invariably have a small average shortest-path length. Some examples of scale-free networks are the World-Wide Web, Internet-scale router networks, social networks, and protein-protein interaction networks. The extremely well-connected nodes in a scale-free network are usually called 'hubs', and play a central role in the structure of a scale-free network. Simulation experiments have been done [8,9] in which random nodes have been removed from a scale-free network, resulting in little effect on the function (as measured by average path length) of the overall network. Hence, a scale-free network withstands random node (and link) attacks very well. However, if the targeting of the node removal is changed to include only hubs, that is nodes in the scale-free network with very many connections, the overall function of the network deteriorates very fast, and the network soon breaks down into many small, unconnected parts.

Social Network Analysis (SNA) can be considered as a large and important subfield of network analysis (see the right-hand side of Figure 2). SNA has reached maturity as a scientific discipline, after beginning in the late 1800s and picking up considerable momentum in the beginning of the 1900s [10]. The idea behind social network analysis can best be understood in terms of how it differs from classical psychological studies of the individual. Classically, one treats the individual as isolated from other individuals, and the behavior of the individual is viewed as the result of all the individual's attributes: gender, age, interests (football, technology savvy, etc.), personality, intelligence, etc. Social network analysis, on the other hand, realizes that the individual is not isolated and separated from the other individuals in the world. Social network analysis tries to uncover ties and relationships among the individuals, and tries further to understand the individual in the context of the other individuals and all the interdependences among them. This type of analysis is not only applicable to persons as individuals; thinking like this is also highly relevant for organizations, companies, and even countries.

Consider for example an organization. Every organization has implemented a hierarchical structure, which is the line-of-command. The implementation and utilization of the human resources in an organization can be efficient, and it can also be very inefficient. Inefficiencies are often due to overloaded bosses that are not able to respond to all information requests and decisions that need to be made. It is in the nature of a hierarchy that decisions tend not to be distributed, while the expert knowledge is. It is the job of leaders and management to make and implement wise decisions, but doing this without thoroughly aggregating and utilizing the wisdom and knowledge of the 'grass roots' of the organization is hard, if not impossible. Management and leaders can thus become bottlenecks for the whole organization, instead of being the driving force.

Performing a network analysis of the communication patterns of the workers in an organization can aid in identifying workers and leaders that are key nodes, and possibly bottlenecks, in the information flow. Having knowledge of for example who is discussing with whom, who seeks advice from whom, and who trusts whom, is very valuable. Utilizing this information in the correct manner can help the organization to implement a more efficient organization. It all boils down to analyzing internal communication within the organization, and, based on this analysis, pinpointing the right people that need to talk to each other. This is invaluable information for an organization – and social network analysis is a first step in the direction of finding this information.

One concept that plays a major role in social network analysis is the notion of *centrality*. There are many different centrality measures and metrics that are used in order to extract useful information from the network perspective that social network theory gives. Given a network, a natural question that can be asked is 'which node is the most central, relative to the other nodes in the network?'. Measures such as eigenvector centrality, degree centrality and betweenness centrality try to answer this question by mathematizing the following ideas: 'you are central if you know other central people', 'you are central if you know many different persons', and 'you are central if you lie on many short paths of information flow between individuals', respectively. Other useful descriptors of social networks include density, clustering, and cohesion. These measures seek to assess how complete the link structure of the network is (which can be assessed both on a local and a global level).

Another area where network analysis plays a major role is Link analysis (LA – see the left hand side of

Figure 2). In link analysis one seeks meaningful information from the regularities and irregularities in the patterns of linking which make up the graph's topology. As can be seen from Figure 2, link analysis can be used in a broad range of disciplines – for example, in the banking and insurance industries (in relation to fraud detection), in law enforcement (for example, analyzing communication patterns), and in the medical sector (epidemiology and spreading over networks). With link analysis one tries to find crucial relationships and associations between diverse objects of different types, so as to give a meaningful whole, in which all the pieces fit together. Each object treated by itself will just be an isolated piece of information that will not give any useful insight unless it is put into some type of context with the other objects and pieces of information. Link analysis aims at doing this, by putting the separate and isolated pieces of information into a useful *context*, from which a coherent and logical picture can be extracted. In a police investigation, it may be of crucial importance to examine the addresses of suspects and victims, the telephone numbers they have dialed, the financial transactions that they have partaken in during a given timeframe, and the family relationships between these subjects.

Web link analysis is that part of link analysis in which one uses the set of hyperlinks to place the web pages of the World Wide Web in context. Link analysis is used by web search engines to rank hit lists using (along with other measures such as text relevance) link-based centrality metrics – such as Google's PageRank algorithm [11], Kleinberg's HITS algorithm [12], and T-Rank [13]. The main motivation behind Web link analysis is the following: When you search for information on the Internet today, you will most likely find thousands (maybe millions!) of hits in response to the query you type into the web search engine. There is no way in which you can sift through all these hits to find the very best ones; hence the web search engine has to make the job easier for you by sorting (ranking!) the hits such that the most important hits are at the top. The link-based centrality algorithms contribute to the ranking by analyzing the network structure of all the web pages, and generating from this some kind of centrality or 'importance' score for each page. Remember that millions of people have carefully designed and uploaded their web pages to the Internet, and they have linked their web pages to other web pages that they think are important or relevant in relation to their own web page. Hence, one can think of the linking between web pages on the World Wide Web as a way of 'voting' for other web pages. There is thus a type of *recommendation* built into the network structure of the World Wide Web, and the link analysis algorithms try to harvest information from all these collective recommendations, with the aim of giving better ranking of long hit lists generated from search queries.

A field that does not lie centrally in our WikiMind-Map for 'network analysis', but that nevertheless can be considered to be very closely related, is value network analysis (bottom left, Figure 2). Value network analysis [14] is usually viewed as an economic discipline, and may be viewed as a generalization of the earlier, and simpler, notion of 'value chains'. Wikipedia (http://en.wikipedia.org/wiki/Value_network) defines a value network as "… complex sets of social and technical resources [that] work together via relationships to create social goods (public goods) or economic value". We profoundly believe that combining economic theory and network analysis will be an extremely fruitful line of endeavor, and that value network analysis as a separate discipline has much to gain from cross-fertilization with other areas of network analysis.

The last theme we would like to address, in this gentle introduction to network analysis, is something that Figures 1 and 2 are living examples of: visualization of networks. For a fairly small and hierarchical network, such as the mind-map visualized in Figure 2, the task of laying out the network on a 2D page is fairly straightforward. However, the complexity of the visualization task increases tremendously when the graph is not a hierarchical tree, when the nodes have different states, when the links can represent different types of relationships, and (not least!) when the number of nodes and links extend into the millions. For example: how can a telecommunication operator visualize the social network of all of its customers? What information is it possible to extract from a visual map of the underlying computer network on which the Internet is running?

Visualization of networks is about making accessible complex sets of data and interactions among them. In fact, the underlying data might yield zero insight or understanding when presented in tabular form; but a good visualization will literally make some underlying structure visible for any viewer, regardless of the complexity of the original data. In other words, visualization of networks is about communication, and how to communicate understanding of the data at hand.

In the next section you will find a 'beginner's guide' to some of the principal technical concepts of network analysis. We will introduce basic terminology which lies at the foundation of network analysis, and provide concise definitions wherever possible.

Understanding of the terms presented in this section should be useful for the reader who then moves on to the other articles in this issue.

Finally, in the last section of this Introduction, we will offer a reader's guide to the many and varied contributions in this special issue of *Telektronikk*. We will briefly introduce and describe each article, and also seek to put each contribution in context with respect to network analysis.

Welcome to the fascinating world of networks!

2 Basic Concepts – An Introduction to Network Analysis

Now we establish a basic foundation of ideas and terminology for network analysis. The terms 'network' and 'graph' will be used interchangeably throughout the presentation.

2.1 Networks Come in Two Flavors

2.1.1 Undirected Graphs

A graph/network is an abstract structure, composed of *nodes* connected by *links*. In fact, in the most simple picture of a network, there is no other information – only the set of nodes, and the specification of the links connecting the nodes. Such a simple network is typically termed an *undirected graph*. It is "undirected" because no information about the direction of the links has been given – just presence or absence. Such links (without any directional information) are also termed *symmetric* links. An undirected graph is usually represented in one of two ways; 1) the undirected network may be represented as a list of links (this link list implicitly also gives the nodes, assuming no isolated nodes), or by 2) the *adjacency matrix A*. The adjacency matrix has the same information, stored as follows: if there is a link from node i to node j, then, in the adjacency matrix, $A_{ij} = A_{ji} = 1$. All other elements of A are zero. Clearly, for symmetric links, the adjacency matrix is symmetric: $A = A^T$.
A simple extension of the simple undirected graph is to allow for *weights* (other than 1 or 0) on the links, where the weights now may represent, for example, capacities of the links, or probability of transmission of a virus. Such a graph is termed a *weighted undirected graph*. As real networks grow in the number of nodes N, the average number of links connecting to each node tends to be roughly constant, rather than being proportional to N. This makes the total number of links roughly *(constant)* $\times N$, resulting in the corresponding adjacency matrix being sparse. Sparse matrices are most efficiently represented by not storing the zero elements in memory.

2.1.2 Directed Graphs

The obvious extension of the above undirected graph is to allow for directionality of the links – that is, a link between i and j may be from i to j ($i \rightarrow j$), or from j to i ($j \rightarrow i$). Such links are termed 'directed' links, and the resulting graph is a *directed graph*. Directed links are used to represent asymmetric relationships between nodes. For example: Web pages are linked in the 'Web graph' [15,16,17] by hyperlinks. These hyperlinks are one-way; so the Web graph is a directed graph. Directed graphs can also be represented by an adjacency matrix A. The asymmetry of the links shows up in the asymmetry of the adjacency matrix: for directed graphs, $A \neq A^T$. One can also associate weights (other than 0 or 1) with the links of a directed graph. Thus directed graphs and undirected graphs have many features in common. Nevertheless (as we will see below), the structure of a directed graph differs from that of an undirected graph in a rather fundamental way.

2.2 Structural Properties of Graphs

How then can one describe the structure of a graph – both on a microscopic and macroscopic level? We refer the reader to [18] and [19] for some interesting ideas about what is meant by the 'structure' of a network, and begin the discussion here with a very basic notion: the network's size. How big is a network? A very crude measure of a network's size is the number N of nodes. Another measure of the network size is the distance between a typical pair of nodes. The distance between any two nodes can be counted as the length (in "hops", ie, number of edges) of the shortest path between them. This definition is quite precise – except for the misleading term "the" in "the shortest path" – since there may be many paths which are equally short. Hence a more precise wording for the definition of distance is the length (in hops) of *any* shortest path between the nodes.

The definition of distance may be modified for weighted graphs. Here the 'length' of any path from i to j may (as for undirected graphs) be taken to be simply the sum of weights taken from the links traversed by the path. The distance between i and j is then unchanged (in words): it is the length of any shortest path connecting i and j.

2.2.1 Undirected Graphs

We now restrict our discussion to the case of undirected graphs; we will treat directed graphs in a separate subsection, below. We return to the notion of the size of a network, based on distances between nodes. For N nodes, there are $N(N\text{-}1)/2$ pairs of nodes – and each pair has a shortest distance. So we want something like an "average distance" between nodes, that we denote by L. One can define L simply as the arith-

metic mean of the $N(N\text{-}1)/2$ path lengths (although other definitions of 'average' can be useful). Thus, we now have another notion of the size of a graph – the average distance L between pairs of nodes. Another common notion of the size of a graph is the graph *diameter D*. The diameter is simply the length of the 'longest shortest path'. That is, of all the $N(N\text{-}1)/2$ shortest path lengths, D is the greatest length found in this set. The diameter is clearly larger than the average path length; yet, for large N, it typically differs from L only by a constant (ie, N-independent) factor.

We make the following assumption (unless we explicitly state the contrary) throughout this presentation: we assume that any graph under discussion is *connected*. An undirected graph is connected if there is a path from any node to any other. Further, a connected graph does not 'fall to pieces': it is not in fact more than one graph. Our assumption requires then that any 'recipe' for 'growing' a network with increasing N gives a connected graph for every N. This requirement is readily satisfied for certain types of highly regular graphs; but it cannot always be guaranteed for graphs with randomness involved in their construction (see [20] and [5] for details). Finally, in this regard, we mention the *small-worlds* phenomenon. Recall that the idea of the small worlds came from the sociologist Stanley Milgram, in the 1960s [7]. Milgram performed letter-passing experiments which supported the idea that, in the United States with hundreds of millions of individuals, the typical path length between any two individuals (where a link represents acquaintance) is about 6 – an astonishingly small number. We note that a property of random graphs – $L \sim \ln(N)$ – offers an explanation for these small distances; but there is no reason to believe that the network of acquaintances is set up like a random graph. Watts and Strogatz [20] offered a new class of graphs which has much less randomness, but still gives (approximately) $L \sim \ln(N)$. These graphs are often referred to as Watts-Strogatz (WS-)graphs.

The Watts-Strogatz work has sparked a great deal of further research, which we will not attempt to review here. A principal lesson of the Watts-Strogatz result may be stated as follows: for networks for which the nodes' connections reflect some reasonable notion of geographic distance, the addition of a very *small* percentage of *long-range links* can drastically change the average path length. And, stated in this way, we believe that the Watts-Strogatz result does indeed explain the small-worlds effect.

2.2.2 Directed Graphs

Introducing a direction to the links of a network can have profound effects on the properties of that network.

The definition of a path length needs just a minor modification to still be valid; the updated definition of a path must take into account the direction. Hence, a path from i to j in a directed graph is a series of directed links, always traversed in the direction from 'tail' to 'head' of the arc. Then, with this definition of a path (a 'directed path'), the path length is the same as before (number of hops, or sum of weights).

The notion of connectedness for directed graphs is a bit more complicated. A directed graph is connected (also called *weakly connected*) if, when one ignores the directions of the links (thus making it undirected), it is connected. However, it is clear that there can be node pairs in a connected, directed graph, for which there is *no* (directed) path from one node of the pair to the other. Hence we need a further notion, namely that of being *strongly connected*. A directed graph is strongly connected if there is a directed path from every node to every other node.

We will as usual insist that any directed graph under consideration be connected. However, typically, connected directed graphs are not strongly connected; and we will lose generality if we only discuss strongly connected directed graphs. Instead, one can always find *subgraphs* of a directed graph that are strongly connected. By subgraph, we mean simply a subset of the nodes, plus all links (that are present in the whole graph) connecting these nodes. For example, if one can find, in a given directed graph, any two nodes that point to each other, then these nodes constitute a strongly connected subgraph of the given graph. *Strongly connected component* (SCC) is the term for a 'maximal' strongly connected subgraph – that is, a strongly connected subgraph that is 'as big as possible', in that it loses its property of being strongly connected if any other node is added to the subgraph. A typical directed graph is thus composed of multiple SCCs, with each node belonging to one, and only one, SCC. Also, the linking relationships between the SCCs are necessarily asymmetric, ie, one-way (since two hypothetical SCCs which are connected by links in both directions must in fact belong to a single SCC).

Using the SCCs, one can define a 'coarse structure' for a directed graph. This coarse structure is termed a *component graph* [21], where each SCC is represented as a node, and each pair of SCCs which is directly linked gets a one-way link connecting them in the component graph. The component graph is

loop-free, that is, it is a *directed acyclic graph* or DAG. If we think of any kind of flow or movement (say, of information, or of a computer virus [22,23]) which follows the direction of the arcs, then it is clear from the acyclic nature of the component graph that flow is essentially one-way. Thus we see that the notion of connectivity is considerably more involved for directed graphs than for undirected graphs. This makes a directed graph profoundly different from an undirected graph.

Now we proceed to the question of path lengths. For a typical directed graph with more than one SCC, there will exist pairs of nodes for which the path length is – as for some pairs in an unconnected graph – undefined (or infinite). This is because there will be node pairs (a,b) for which there is no directed path from a to b. Therefore, the definition of average path length and diameter for directed graphs is less straightforward than it is for undirected graphs.

One possibility is to ignore the direction of the links. However, if directionality of the links is meaningful, then it should not be ignored. Another is to only look at a strongly connected component – for which all nodes are reachable from all others – perhaps the largest SCC [24]. A third choice is to only average over those pairs for which a directed path exists [25]. This latter strategy actually somewhat resembles Milgram's [7]; after all, he did not count those letters which failed to reach their target.

2.3 Node Degree Distribution

Next we come to a concept which is heavily used in network analysis, namely the *node degree distribution* or NDD. We consider first the case of an undirected graph.

2.3.1 Undirected Graphs

When links are undirected, then for each node i we can define a *node degree* k_i which is simply the number of links to that node. Now we define p_k to be the fraction of nodes in a graph which have node degree k. The set of p_k then defines the node degree distribution. Clearly $0 \le p_k \le 1$; also, $\sum_k p_k = 1$. Thus the p_k have all the properties of a probability distribution, and can in fact be used to describe families of graphs, subject to the constraint that the probability (averaged over the family) for a node to have degree k is p_k.

In a *regular graph*, every node has the same degree K, so that $p_k = 1$, and all other $p_k = 0$. A complete graph is regular, with $K = N\text{-}1$. Small perturbations on regular graphs give graphs with a peak at $k = K$, and some weight around this peak at K. Such graphs are in many ways much like regular graphs; but at the same time, some properties of graphs (such as the average path length) can vary strongly upon only a slight deviation from a regular graph with an ordered structure of short-range links. Here we are thinking of the small-worlds effect [20].

In fact, the node degree distribution is far from offering a complete description of any graph. Two graphs with roughly equal NDD can have similar link structure, but very different average path lengths. Another two graphs, also with roughly the same NDD, can have very different link structures, but similar average path length. What's worth remembering is that graphs that are structurally very different can in fact have the same node degree distribution.

A very important class of graphs is the *scale-free graphs*. These are graphs whose NDD follows an inverse power law: $p_k \sim k^{-\alpha}$. They are termed 'scale-free' because the NDD itself is scale free: a power law remains a power law (with the same exponent) under a change of scale of the k axis. Scale-free networks are also called 'heavy-tailed', because they decay more slowly than exponentially at large k – so slowly, in fact, that they can have an unbounded variance. Scale-free graphs are also (typically) small-world graphs [26,27] – in the sense given above, ie, that their average path length grows no faster than $\ln(N)$. However, the reverse is not true: not all small-worlds graphs are scale free. An obvious counter-example is the class of Watts-Strogatz graphs, which are small-world graphs, but which have an NDD very different from power-law. A good discussion of the relationships between these two aspects of networks is given in [27].

2.3.2 Directed Graphs

For directed graphs we can define, for each node i, two types of node degree: the indegree k_i^{in}, and the outdegree k_i^{out}. Each of these types of node degree has its own node degree distribution over the entire directed network – denoted by p_k^{in} and p_k^{out}, respectively. Then, for a directed graph to be considered scale free, both of the node degree distributions should display a power-law. The two distributions typically have different exponents.

2.4 Clustering Coefficient

The clustering coefficient measures to what extent a network has dense local structure – ie, how close the network is to being a clique (a network with all-to-all connections). As an example, consider the acquaintanceship network. Suppose node j has five one-hop neighbors (ie, acquaintances). Do we expect these five people (at least, some pairs taken from them) to be acquainted with one another? The answer is surely yes: acquaintanceship tends to be clustered. Friends are not chosen at random; and the likelihood that

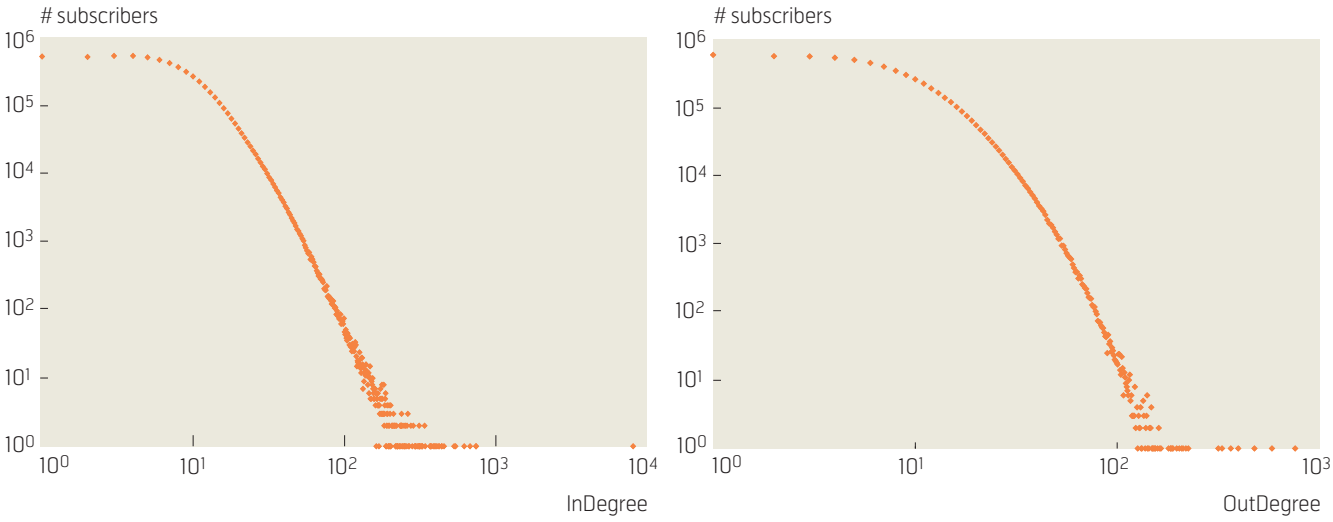


Figure 3 This is an example of the node degree distribution of a real communications network. The plot of the indegree is in the left, and on the right we see the outdegree. Both these plots display a typical scale-free structure observed in real systems

two of my friends are acquainted with one another is much higher than that for two randomly chosen people. Therefore, acquaintanceship networks will normally show higher clustering than a random network with the same number of nodes and links.

In order to quantify the extent of clustering, one can compare the ratio of the number of triangles in the network to the number of connected triples in the network. A triangle is a connected subgraph of three nodes which is a clique – that is; all the possible three links among the three nodes are present. A connected triple, in contrast, is a connected subgraph of three nodes which have only two links connecting the three nodes.

One approach to defining a clustering coefficient C for a graph is as follows. First one defines a *local clustering coefficient* around each node i:

$$C_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered on node } i};$$

then one takes the average of this local clustering coefficient over all nodes:

$$C = \frac{1}{N}\sum_i C_i.$$

This definition comes from the work of Watts and Strogatz [20]. Implicitly, in the above discussion, we have considered an undirected graph, but the same idea of defining the clustering coefficient as the ratio of triangles to triples may be generalized to the directed case. A definition of the clustering coefficient which is distinct from the above definition can be obtained by doing the averaging in a different manner. In the above definition, the C_i are local ratios, defined for each node i, and the average is taken over these local ratios. One can readily modify the definition to be the ratio of two whole-graph averages – that is, the ratio of the average number of triangles to the average number of triples [26].

Quite often one simply refers to "the" clustering coefficient – thus ignoring these differences. The point is that, for some types of graphs (ie, for some recipes for growing graphs), the clustering, by either of these definitions, tends to vanish as N grows; while for other types, neither of them vanishes, even for very large N. Intuitively, we expect "real" networks to have some degree of neighbourhood structure, and hence to fall into the second category. In contrast, random graphs fall into the first category – they have no correlations among the links, and hence no clustering. For this reason, even though random graphs have very small path lengths, they are not considered as realistic models for empirical networks.

Seeking a definition which is more in harmony with what we know and believe about empirical networks, Duncan Watts [20] has proposed a definition of a small-world graph which includes *two* criteria: (i) the average path length L should be like that of a random graph; but (ii) at the same time, the clustering coefficient C should be much larger than that for a random graph. That is, by this (widely accepted) definition, a small-worlds network should have a significant degree of neighborhood structure, and *also* enough 'long-range' links that the network diameter is very small. Scale-free networks fit this definition [27]. It seems in fact that an enormous variety of measured networks are small-world graphs by this definition.

As a final note of this subsection we mention that the above definition for clustering coefficients can be adapted to networks with weighted links.

2.5 Centrality on Networks

We now turn to the concept of centrality of the nodes in a network. The purpose of the centrality concept is to find out how "important" each node is, in the context defined by the other nodes in the network and the given link structure. For example, in a computer network, some computers are more important/central than other, more peripheral, computers, since the functioning of the computer network relies more on these central computers than on the others. Also, in the case of a social network, we humans (as social creatures) intuitively understand that there can be large differences in 'centrality' among the members of the social network. Hence, it is clear that ideas of centrality can play a significant role in the analysis of networks, and in the way that we are able to understand the functioning of large and complex networks – including those that usually defeat our intuition.

There exist many distinct definitions of centrality, and we will discuss some of them in the following subsections.

2.5.1 Degree Centrality

Degree centrality is the first and simplest example of a centrality measure for nodes in a network: *Degree centrality* of a node i is simply its degree k_i. This definition is consistent with the intuitive notion that having more contacts makes one more central. Also, it is very easily calculated, since for each node one just counts the number of nearest neighbors. Is there anything missing then from this definition? Well, it clearly ignores *every* aspect of the network's structure beyond the one-hop neighborhood view from each node, and is thus purely local in nature. Also, degree centrality takes no consideration of the "quality" of your neighbors – it is only their *number* that matters. Degree centrality can be used in assessing the immediate risk of a node for catching whatever is flowing through the network, such as a virus, or information. It is important then to understand that degree centrality measures this risk only locally, and applies when in fact the virus or information is within the one-hop neighborhood of the node.

In the directed case, the degree centrality is modified by taking into account the asymmetry of the links. Hence, for each node it is natural to speak about indegree and outdegree centrality.

2.5.2 Betweenness Centrality

Another type of centrality is captured with the *betweenness centrality* of a node. The idea of betweenness is simply this: a node has high betweenness centrality – also termed simply 'betweenness' – if it lies (in some sense) 'between' lots of other pairs of nodes. Thus, if I have high betweenness, then many node pairs within the network wishing to communicate with each other will likely do so through me. Clearly, the measure of betweenness centrality depends on the entire network topology.

Mathematically, one resorts to finding shortest paths among the nodes in order to find a quantitative measure of betweenness centrality. Look at each pair of nodes (i,j), and define g_{ij} as the number of shortest paths between i and j. The betweenness is then considered for some *other* node k; so we define g_{ikj} as the number of shortest paths connecting i and j which go through node k. The contribution to node k's betweenness coming from the pair (i,j) is then g_{ikj} / g_{ij}. Then, the betweenness of node k is simply the sum of this quantity over all pairs of nodes:

$$b_k = \sum_{\substack{\text{pairs } i,j \\ i \neq k; j \neq k}} (g_{ikj}/g_{ij})$$

This definition seems complicated; but note that, in the (rare, but simple) case that there is a unique shortest path between every pair of nodes, this definition for b_k simply reduces to counting the number of shortest paths that node k lies on (but is not an endpoint of). Thus the extra complications arise from the fact that there are often multiple shortest paths for a given pair of nodes.

Betweenness centrality thus measures the importance of a node in *transmissions* over the network – for example, of information, traffic, viruses, etc. A possible weakness is the choice to count only shortest paths, since all paths between two nodes can play some role in transmission – with the shortest paths having most weight, but not the only weight.

2.5.3 Eigenvector Centrality

Eigenvector centrality tries to capture the following idea: 'It's just not how many you know, but *who* you know that matters'. In other words: your centrality also depends on your neighbors' centralities. Actually, the first centrality measure we discussed above, degree centrality, is only counting the number of neighbors. In this sense, we can think of eigenvector centrality as a modification of degree centrality. The simplest way of making these words precise is to take the sum; so, letting e_i be the eigenvector centrality of node i, we write

$$e_i \propto \sum_{j=nn(i)} e_j .$$

Here "$j = nn(i)$" means only include nodes j which are nearest neighbors of i in the sum. We use a pro-

portionality sign rather than equality here, because we have made a circular definition: my centrality depends on my neighbors', but theirs depends on that of *their* neighbors – including mine. Hence we must seek a self-consistent solution; and this requires some factor in front of the summation:

$$e_i = \left(\frac{1}{\lambda}\right) \sum_{j=nn(i)} e_j .$$

Now our reasoning has given us this parameter λ: what is it? Can it be chosen freely? In fact the answer is simple. We rewrite the above as:

$$A\boldsymbol{e} = \lambda \boldsymbol{e},$$

where A is the adjacency matrix, and $\boldsymbol{e}$ is the vector of centralities. Here we have used the fact that

$$\sum_{j=nn(i)} e_j = (A\boldsymbol{e})_i .$$

Thus we have come to an eigenvector equation involving the adjacency matrix A. From this we know that there are N possible eigenvalues λ to choose from – each with its corresponding eigenvector. We can however quickly narrow down this choice: the adjacency matrix is non-negative and real, which (along with Perron-Frobenius theory [28]) tells us that one, and only one, eigenvector has all positive elements – namely, the eigenvector corresponding to the *principal* (largest) *eigenvalue*. We want a node's centrality to be a positive quantity. Hence we choose the *principal eigenvector* $\boldsymbol{e}$ as the solution to the above equation; and its i'th component e_i is then the *eigenvector centrality* (EVC) [29] of node i. Defined this way, a node's EVC e_i is sensitive to the EVC values of its neighbors – and their values in turn to those of *their* neighbors, etc.

Eigenvector centrality has been used in studies of epidemic spreading on undirected graphs. Simulations reveal that there is a close connection between the time at which the rate of new infections (or adoptions) takes off, and the time when the most central node, as given by the eigenvector centrality, is infected (or adopts) [30,31].

Carrying this seemingly orderly picture of the undirected case over to the directed case is not straightforward. The asymmetry of the adjacency matrix for a directed graph suggests that every node in the network plays two distinct roles: a node can *point to* other nodes (the set of out-neighbors), and at the same time a node can *be pointed to* (by the in-neighbors). These properties, 'pointing-to' and 'being-pointed-to', have found useful application in online web search engines for ranking hit-lists; in this field, these two properties correspond to, respectively, a node's *hub* and *authority* score. For a more thorough discussion of the directed case, see the paper by **Bjelland et al** in this issue of *Telektronikk*.

2.6 K-Cores of a Network

The k-core of a network is a subgraph, which may be defined by the procedure which generates it, for each k. The 1-core of a network is found by removing all the nodes of degree zero (which are isolated nodes). The 2-core is found by pruning off nodes of degree one or less in a recursive manner. That is: after removing all nodes of degree one or less, examine the resulting network, and remove all nodes which now have degree one or less; and repeat until nothing changes. Thus the 2-core of a network is the largest subgraph for which every node has degree at least two. We can illustrate these ideas by considering a network which has a tree structure (no loops). As the reader can verify (by recursively pruning leaves – end nodes – from the tree), a tree has a non-zero 1-core; but the k-cores for all higher k are empty (have no nodes). The general definition for any k should by now be clear: the k-core of a network consists of all remaining nodes and links, after the recursive pruning of all the nodes (and their respective links) having degree less than k. The concept of k-core is in fact a fairly recent construction [32,33], which has found applications in the structural study of networks, in biology, in information filtering, and in graph visualization. This last application brings us to the next subsection.

2.7 Visualizing Networks

One should never underestimate the power of communicating by visualization. Everyone has heard the saying: "a picture is worth a thousand words". This is more or less true also in the case of network visualization, or graph drawing [34]. Recall the fact that a network may be simply represented as two lists – a list of nodes, and a list of all the links between the connected nodes – and that such a representation in fact completely specifies the (abstract) network. However, humans have an easier time digesting visual information, and in order for humans to understand what their network data is telling them, the information contained in the lists may be better conveyed by a *drawing* of the network. We emphasize here that several visually different network diagrams can indeed depict the same underlying, abstract network, but the way the drawing has been done can reveal different aspects of the network. The point is then that different methods can be used to learn different things about the network.

There are several different graph layout strategies. (See [35,36,37] for sites with working visualization tools.) A very popular approach is the method of the

force-directed layout. Here the idea is to treat the network as a physical system, where the nodes are electrically charged and connected with springs (with the links weights as spring constants). So the electric charge acts to repel the nodes from one another, whereas the springs are working to keep linked pairs of nodes together. The force-directed layout then minimizes the energy function of this system, and produces very appealing network diagrams. Nodes that are tightly connected (large link weight) will be displayed close together, and weakly connected nodes will be displayed far from each other. For a human this is visually intuitive, and what one would expect.

3 What You Can Read About in this Issue on Network Analysis

In this *Telektronikk* special issue on network analysis you can dive deeper into network-analytic topics and closely related themes. The different contributions that are contained in this issue of *Telektronikk* are diverse and cover different aspects of network analysis.

Heegaard et al. in *The Cross Entropy Ant System for Network Path Management* address the problem of finding optimal routes in a network. "Isn't this an old and solved problem?" you might ask, but the difference lies in the information needed to find the routes. All the classical approaches and algorithms for finding shortest paths in a network assume that global information about the network is available. This means that the algorithm has information about the full network topology when running. The approach discussed in this paper, Cross-Entropy Ant System (CEAS), assumes no such knowledge. Hence, the approach is a fully distributed, and no global knowledge of the network topology is needed. The approach is based on borrowing ideas from biology, more specifically the foraging process used by ants. When ants are out exploring they tag their paths with a scent (that over time will decay). The more ants that walk down the same path, the stronger is the scent of that path. This tells newcomer ants that, if they go down paths of strong scent, they will reach an interesting target faster. Mimicking this on a network is what the CEAS approach does. The advantage of the approach is that it is fully distributed, and it is also very adaptable (unlike most global methods) in the case of node and link failure.

There are three contributions within social network analysis. The first, *The Social Networks of Teens and Young Adults*, by **Ling**, studies the social networks of teens and young adults. The question raised is how co-present (face-to-face) contact is an important factor in the development and maintenance of social groups. Mediating interaction will support the already existing social ties, but it is the copresence of the persons in the group that builds the social ties, which then later can be cultivated by mediating interactions, for example by means of telecommunication services.

Schnorf, in the second contribution *Forwarding Messages in Mobile Social Networks – an Exploratory Study*, discusses how social network analysis can be used as a tool for understanding diffusion of information in social networks. This exploratory study exemplifies how social network analysis can be used by for example a mobile telecommunication operator in gaining a new perspective on their customers. For example, customer segmentation should also be considered in the context of the customers' social network, and not only using the attribute-based approach of traditional marketing.

Thirdly, **Julsrud** in *Collaboration Patterns in Distributed Work Groups: A Cognitive Network Approach*, takes us into the realm of organizational network analysis, by investigating collaboration patterns in distributed work groups. The suggested method to use is the cognitive network approach, where *perceived* relations among the users should also be taken account of in the analysis. Interaction-based ties should be handled with care, and possibly supplemented with other relational network indicators in network studies of distributed groups.

Fjeldstad, in *Innovation in a Value Network Perspective*, guides us into that area of network analysis – value networks – that is closest in content to economic theory. In this article, an innovation is defined as an exploratory activity which results in new skills, practices, technologies, services or products of a firm. Fjeldstad discusses innovations in the context of value networks; the core innovations for a networked service are those that increase network connectivity (the number of people that can be reached by the service) and network conductivity (what can be transacted over the network). With this view in mind; the main task of a mobile telecommunications operator is to serve existing relations among its customers more efficiently, and to enable and serve potentially new relations.

Becker and Gaivoronski offer a different avenue into the world of networks in *Quantitative Network Analysis and Modelling of Networked Multiagent Environment*. By taking the reader on a trip into the world of stochastic optimization, coupled with contemporary investment science and game theory, they offer the reader a glimpse of a tool that can be used in the evaluation of business models and in the support

of strategic decision making in an uncertain, networked, multiagent telecommunication environment.

Bjelland et al in *Web Link Analysis: Estimating a Document's Importance from its Context*, take the reader on a journey into a contemporary part of network analysis called web link analysis. Every user of a web search engine knows that information overflow is becoming a larger and larger problem in the information age – search engines typically return far too many hits. Thus the hits must be ranked, such that the 'best' hits are placed at the top of the list. Web link analysis (as exemplified by Google's famous PageRank algorithm) is one tool which is used to help in this challenging ranking task. The article by Bjelland et al is an overview of the main ideas around, and mathematical approaches to, Web link analysis.

Lastly, **Aggarwal et al** in *Modelling Overlay-Underlay Correlations Using Visualization*, analyze overlay network structures in the context of the underlying network infrastructure. The approach is based on visualization. Using visualization analytically in this way is both novel and ingenious, and may prove to be a promising tool in the general process of engineering overlay network structures.

We wish the reader a stimulating journey in the world of networks when reading this issue of *Telektronikk*!

4 References

1 Braess, D, Nagurney, A, Wakolbinger, T. On a paradox of traffic planning. *Journal of Transportation Science*, 39, 446-450, 2005.

2 *Network analysis*. (2008, January 30). In Wikipedia, The Free Encyclopedia. Retrieved 09:29, March 3, 2008, from http://en.wikipedia.org/w/index.php?title=Network_analysis&oldid=188011760

3 See for example http://mathforum.org/isaac/problems/bridges2.html.

4 Bollobás, B. *Random Graphs*. New York, Academic Press, 1985, xvi+447pp.

5 Chung, F, Lu, L. Complex Graphs and Networks. *CBMS Lecture Series*, No.107, AMS Publications, 2006, vii + 264pp.

6 McHugh, J A. *Algorithmic Graph Theory*. New Jersey, Prentice Hall, 1990.

7 Milgram, S. The Small World Problem. *Psychology Today*, May 1967, 60-67.

8 Albert, R, Jeong, H, Barabasi, A-L. Error and attack tolerance of complex networks. *Nature*, 406, 378-382, 2000.

9 Holme, P, Kim, B J, Yoon, C N, Han, S K. Attack vulnerability of complex networks. *Phys. Rev. E*, 65, 056109, 2002.

10 Freeman, L. *The Development of Social Network Analysis*. Vancouver, Empirical Press, 2004.

11 Page, L, Brin, S, Motwani, R, Winograd, T. *The pagerank citation ranking: Bringing order to the web*. Stanford, CA, Stanford University, 1998. (Technical report)

12 Kleinberg, J M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46 (604), 1999.

13 Bjelland, J, Burgess, M, Canright, G, Engø-Monsen, K. *Eigenvectors of Directed Graphs and Importance Scores: Dominance, T-Rank, and Sink Remedies*. 2008. (DELIS technical report TR-0629) http://delis.upb.de/paper/DELIS-TR-0629.pdf

14 Stabell, C B, Fjeldstad, Ø D. Configuring Value For Competitive Advantage: On Chains, Shops, and Networks. *Strategic Management Journal*, 19, 413-437, 1998.

15 Kumar, R, Raghavan, P, Rajagopalan, S, Sivakumar, D, Tomkins, A S, Upfal, E. The Web as a graph. *Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS*, 2000.

16 Broder, A, Kumar, R, Maghoul, F, Raghavan, P, Stata, R, Tomkins, A, Wiener, J. Graph structure in the web. In: *Proceedings of the 9th International World Wide Web Conference*, 247-256, 2000.

17 Leonardi, S, Donato, D, Laura, L, Millozzi, S. Large scale properties of the web graph. *European Journal of Physics*, B38, 239–243, 2004.

18 Girvan, M, Newman, M. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99, 8271-8276, 2002.

19 Canright, G, Engø-Monsen, K. Roles in networks. *Science of Computer Programming*, 53 (195), 2004.

20 Watts, D. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, Princeton University Press, 1999.

21 Cormen, T H, Leiserson, C E, Rivest, R L. *Algorithms*. Cambridge, Massachusetts, MIT Press, 1990.

22 Kephart, J O, White, S R. Directed-Graph Epidemiological Models of Computer Viruses. *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*. Oakland, California, May 20-22, 1991, 343-359.

23 Newman, M E J, Forrest, S, Balthrop, J. Email networks and the spread of computer viruses. *Phys. Rev. E*, 66, 035101, 2002.

24 Adamic, L A. The Small World Web. *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, 1999.

25 Albert, R, Jeong, H, Barabási, A-L. Diameter of the World-Wide Web. *Nature*, 401, 130-131, Sept. 1999.

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

27 Amaral, L A N, Scala, A, Barthélémy, M, Stanley, H E. Classes of small-world networks. *Proc Natl Acad Sci USA*, October 10, 2000, 97 (21), 11149-11152.

28 Minc, H. *Nonnegative Matrices*. New York, Wiley Interscience, 1987.

29 Bonacich, P. Power and centrality: a family of measures. *American Journal of Sociology*, 92, 1170-1182, 1987.

30 Canright, G, Engø-Monsen, K. Spreading on networks: a topographic view. *Proceedings, European Conference on Complex Systems, (ECCS05)*, 2005.

31 Canright, G S, Engø-Monsen, K. Epidemic spreading over networks: a view from neighbourhoods. *Telektronikk*, 101 (1), 65-85, 2005.

32 Seidman, S B. Network structure and minimum degree. *Social Networks*, 5, 269-287, 1983.

33 Bollobas, B. The evolution of sparse graphs. In: *Graph Theory and Combinatorics, Proc. Cambridge Combinatorial Conf. in honor of Paul Erdos*. Academic Press, 1984, 35-57.

34 Di Battista, G, Eades, P, Tamassia, R, Tollis, I G. *Graph Drawing: Algorithms for the Visualization of Graphs*. Upper Saddle River, NJ, Prentice Hall PTR, 1998.

35 *Netdraw*. http://www.analytictech.com/Netdraw/netdraw.htm

36 de Nooy, W, Mrvar, A, Batagelj, V. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005. See also http://vlado.fmf.uni-lj.si/pub/networks/pajek/

37 *Tulip*. http://www.labri.fr/perso/auber/projects/tulip/

38 Barabási, A-L. *Linked: The new science of networks*. Cambridge, Massachusetss, Perseus Publishing, 2002.

39 Bornholdt, S, Schuster, H G (eds). *Handbook of Graphs and Networks: From the genome to the Internet*. Weinheim, Wiley-VCH, 2003.

40 Strogatz, S H. Exploring complex networks. *Nature*, 410, 268-276, 2001.

41 Golub, G H, Van Loan, C H. *Matrix Computations*. The Johns Hopkins University Press, Second Edition, 1989.

42 Motwani, R, Raghavan, P. *Randomized Algorithms*. Cambridge, UK, Cambridge University Press, 1995.

43 Rogers, E M. *Diffusion of Innovations*, 3rd ed. New York, Free Press, 1983.

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

The Cross Entropy Ant System for Network Path Management

POUL E. HEEGAARD, BJARNE E. HELVIK, OTTO J. WITTNER

Poul E. Heegaard is Associate Professor at NTNU and Senior Research Scientist in Telenor R&I

Bjarne E. Helvik is Professor at the Norwegian University of Science and Technology

Otto J. Wittner is a PostDoc at Q2S at the Norwegian University of Science and Technology

Finding paths between nodes is a basic enabling functionality in a communication network. At first glance, this may seem to be a trivial task. However, finding a path when no global information is available, is a challenge. Furthermore, paths should ensure an overall good utilisation of network resources, providing low delays and losses as well as the needed capacity between nodes. Paths should be altered as the network load and topology are changed, and paths should rapidly be recovered when network elements fail. The path management function has throughout the history of communication networks been designed to meet the prime requirement of the service provided by the network within what was technologically feasible. The future network will provide a multitude of services with, to some degree, conflicting requirements. At the same time inherent robustness and autonomy of network operation are of increasing importance. This invites new approaches relative to those used in the traditional communication network and the Internet. One such approach is to use swarm intelligence, where mobile agents explore, map and manage the network in a manner similar to the way insects, e.g. ants and bees, deal with their environment.

With the above in mind, we have developed a distributed, robust and adaptive swarm intelligence system for dealing with path management in communication networks. The system is called the Cross Entropy Ant System (CEAS), and is based on increasing the probability of finding a (near) optimal solution by an increasingly focused random search. As a background for the system, this paper gives a brief discussion on path finding challenges and trade-offs. Following up is a description of CEAS where its robustness and adaptivity are demonstrated on a variety of case studies using different management strategies, like: shared backup path protection (SBPP), p-cycles, resource search under QoS constraints and adaptive paths with stochastic routing. This paper also includes a description of a running implementation of CEAS based on small home routers. The implementation demonstrates and visualises the inner workings of the method.

1 Introduction

Being able to transfer addressed information between sources and destinations is the prime function of a communication network. Hence, how to find paths for the data flow between source and destinations through the network is one of the most salient issues and important functions in network architecture and operation. In this paper, the function is denoted *path finding*, irrespective of whether physically or virtually circuit switched paths (or circuits) are found, or stable routes for connectionless forwarding are obtained.

Throughout history, the path finding applied is a trade-off between requirements of the network service and available technology. The early POTS[1)] networks had a hierarchical routing scheme, which gradually evolved into a more non-hierarchical and adaptive schemes, see for instance [1]. In the early Internet, the prime objective was to have a routing scheme which was inherently robust and would find a path between source and destination irrespective of the current topology. This scheme has evolved to a range of routing protocols, see for instance [2]. The ability to deal with failures of network elements has always been an important issue, with schemes ranging from 1+1 protection of the physical circuits to elaborate end-to-end restoration schemes, as described in for instance [3,4].

A common characteristic of the state of art schemes, is that they apply some degree of preplanning, e.g. allocation of link weighs to links in OSPF[2)] and IS-IS[3)], introduction of operator policies in BGP[4)], planning of (G)MPLS[5)] shared protection paths. For a

1) *Plain Old Telephony Service*

2) *The Open Shortest Path First (OSPF) protocol is a hierarchical for routing in an Internet domain, using a link-state in the individual areas that make up the hierarchy. A computation based on Dijkstra's algorithm is used to calculate the shortest path tree inside each area. See IETF RFC 2328.*

3) *IS-IS is like OSPF a protocol for routing in an Internet domain, based on Dijkstra's algorithm, standardised as ISO10589. See IETF RFC 1195.*

4) *The Border Gateway Protocol (BGP) is the routing protocol between the domains (Autonomous Systems – AS) of the Internet. It is a path vector protocol and makes routing decisions based on path, network policies and/or rule-sets. See IETF RFC 4271.*

 ISSN 0085-7130

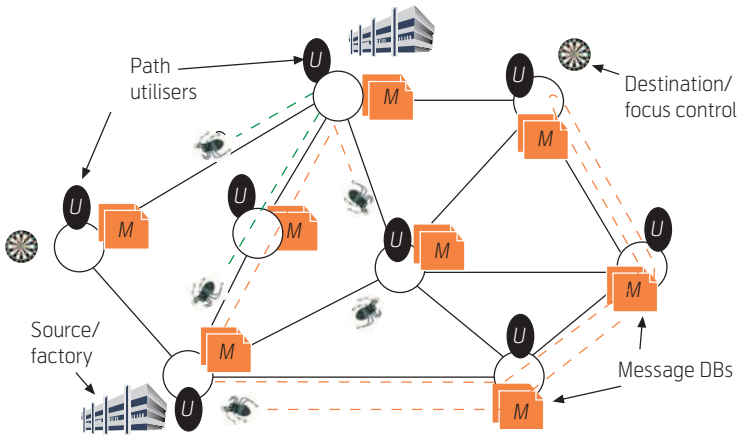


Figure 1 Virtual path management by ants

truly autonomous system, it is necessary that the path finding system itself discovers the network resources, changes in their operational status, their loading, etc. without reliance of a system external planning entity. This activity, as well as operation of the path selection should ideally be inherently robust and distributed.

Furthermore, vertically integrated networks, like for instance the PSTN[6)], were designed primarily to meet the requirements of one service. This is also reflected in the path management. The coming integrated multi-service network will transport services with highly differing QoS requirements with respect to *timeliness and performance* (delay, jitter, transfer rate, loss) as well as with respect to *trustworthiness* (availability, continuity, integrity, confidentiality, manageability). This poses a number of challenges on the path finding with respect to load sharing and QoS oriented routing [5], resilience differentiation [6], etc.

Current path finding schemes have problems in meeting these requirements, and there are ongoing research towards extending them or providing "add-on functionality". However, the self-management/ autonomy objective seems difficult to meet. For these reasons, emergent behaviour is investigated as a viable approach to routing and path finding in future networks. The approach adopted by us is inspired by the behaviour of ants. It is based on mobile agents[7)] representing ants swarming through the network. The idea is to let ants iteratively search for paths from a source to a destination node in a network. When a path is found the ant returns to the source on the reversed path and leaves markings, denoted pheromones (resembling the chemicals left by real ants during ant trail development), on every intermediate node. The strength of the pheromones depends on the quality of the path found. The subsequent searching ants stochastically select their next hops based on the current pheromone distribution. The overall process converges quickly toward a near optimal path. See Figure 1 for an illustration. The paper gives an outline of approach and summarises and discusses our findings.

The issue of path finding in communication networks and some background material on swarm based routing is elaborated further in Section 2. For dealing with path management in communication networks we have developed CEAS (cross-entropy ants system) which is based on Rubinstein's method for stochastic optimisation [7]. The theoretical background for CEAS and the principle of application are presented in Section 3. The CEAS basic technology is applied in a variety of studies using different management strategies. Section 4 presents four strategies; shared backup path protection (SBPP), p-cycles, resource search under QoS constraints, and adaptive paths with stochastic routing. For a proof of concept, and to gain experience with the implementation aspects of CEAS demonstrator systems have been made. This enables live, visualized demonstration of the inner workings of the CEAS. The current version consisting of interconnected small home routers, with a plug-and-play reconfigurable topology, are presented in Section 5. Some concluding remarks are given in Section 6.

2 Path Management

Paths between all source destination pairs in a communication network should be chosen such that an overall good utilisation of network resources is ensured, and hence high throughput, low loss and low latency achieved. At the same time the set of paths chosen must enable utilisation of the available spare capacity in the network in such a manner that a failure causes minimum disturbance of the directly affected traffic flows as well as other traffic flows in the network. The combinatorial optimisation aspects of this task are typically NP-hard, see for instance [8]. Nevertheless, considerable knowledge has been acquired for planning paths in networks. In addition

5) *(Generalized) Multi-Protocol Label Switching (G)MPLS are "Layer 2.5" protocols used to perform (virtual) circuit oriented switching in the Internet. See IETF RFC 3031 and 3945.*

6) *Public Switched Telephone Network*

7) *Mobile agents must be understood conceptually. Mobile agent technology may by used for implementation, but this has severe drawbacks with respect to security and performance. Hence, in our prototype realizations, the agents are realised by message passing between router kernels.*

to finding good paths, proper path management requires that: a) the set of operational paths should be continuously updated as the traffic load changes, b) new paths should become almost immediately available between communication nodes when established paths are affected by failures, and c) new or repaired network elements should be put into operation without unnecessary delays.

Insight and practical methods for obtaining paths *for connection oriented* networks by mathematical programming are available. For an overview, see the recently published book by Pióro and Medhi [9] and references therein. Several stochastic optimisation techniques which may be used to address these kinds of problems, have been proposed [10,11,12,7]. However, common to these are that they deal with path finding as an optimisation problem where the "solution engine" has a global overview of the problem and that the problem is unchanged until a solution is found. This differs from the requirement that path management should be truly distributed and adaptive. On the other hand, one should be aware that applying truly distributed decision-making typically yields solutions that are less fine tuned with respect to optimal resource utilisation.

Near immediate and robust fault handling advocates distributed local decision-making on how to deal with failures. This is reflected by the commonly applied protection switching schemes in today's telecommunication networks, e.g. in SDH and ATM [13,14]. Typically two (or more) disjoint paths are established, one serving as a backup for the other. Protection switching requires preplanning and is rather inflexible and not very efficient in utilising network resources.

If we turn to the *connectionless* domain; shortest path, distance vector and policy based routing as applied in the Internet, is distributed, has local decision-making and applies to some degree planning inherent in the network. See for instance [2]. However, routes (paths) are restored after a failure, which may incur a substantial delay before traffic flows along a route are fully reestablished. Furthermore, it is common that Internet operators use static link weights. This requires preplanning and lessens the adaptivity. In general, making plans that are able to cope efficiently with every combination of traffic load and network state is difficult, if at all possible.

There are two major "design axes" for management systems; a *spatial*, i.e. degree of centralisation-distribution, and a *temporal*, i.e. degree of preplanning. This is illustrated in Figure 2 where moments from the above discussion on planning for connection oriented and connectionless operation are indicated. In addition to spatial and temporal aspects, we have the degree of human involvement and control of the management. As illustrated in Figure 2, it is our research hypothesis that, relative to current approaches, self-managed path finding by emergent behaviour has the potential to provide combined advantages along both axes as well as minimise human involvement. Good resource utilisation currently obtained by centralisation and preplanning is potentially achievable, even combined with "real-time" adaptivity, inherent robustness of truly distributed schemes as well as continuity of service similar to what is today realised by preplanned "hardwired" protection. It is also an objective to overcome the trade-off, illustrated in Figure 3, between an efficient resource utilization with slow failure recoveries obtained by restoration tech-

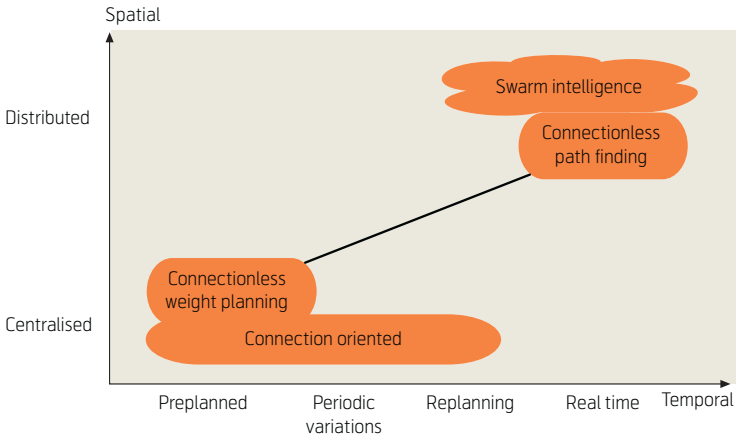


Figure 2 Illustration of the path planning activities connection oriented and connectionless networks, as well as networks managed by swarm intelligence

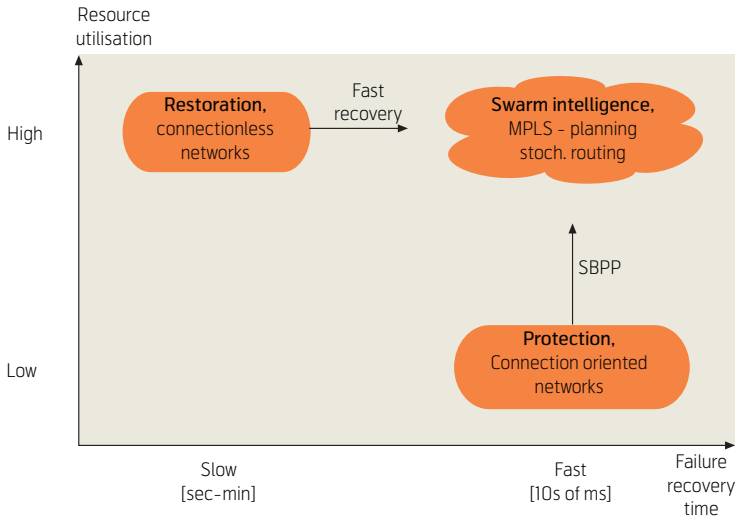


Figure 3 Approaches to achieving simultaneous fast failure recovery and high resource utilization in communication networks

ISSN 0085-7130 © Telenor ASA 2008

niques common in connectionless networks, and the short recovery times with better resource utilization obtained by protection schemes common in connection oriented networks. Overcoming this trade-off is an objective similar to what is the objective of the shared backup protection path (SBPP) techniques, e.g. [15,16,17,18], and fast recovery techniques, e.g. [19,20,21] currently heavily researched.

As mentioned above, self-management has the potential for autonomy, i.e. management without humans in the loop. A drawback is that in order to achieve autonomy, determinism is relaxed, both with respect to strict QoS guarantees and the overhead involved.

Research on swarm intelligence for path management has a ten year history. Schoonderwoerd & al. introduced the concept of using multiple agents with a behaviour inspired by ants to solve problems in telecommunication networks [22]. The concept has been pursued further by others, see for instance [23,24,25, 26,27,28] and references therein. Self-management by swarm intelligence is a candidate to meet the aforementioned requirements and to overcome some of the drawbacks of the current path and fault management strategies. In the next section we introduce the fundamentals of the technique investigated in this paper.

3 The Cross Entropy Ant System

In this chapter the virtual path management system by ants briefly introduced in Figure 1 is presented. Additional background is given in Section 3.1 before the Cross Entropy Ant System [CEAS] is introduced in Section 3.2. Performance and efficient use of network resources are important issues concerning the operation of CEAS itself. Hence, the basic CEAS is extended with mechanisms that improve its operation in these respects. These mechanisms are presented in Section 3.3. For readers interested in the inner workings of CEAS, three factboxes are added to provide more detailed insights.

Fact Box 1 – Property 1. Fundamentals

System. A network may be represented by a bidirectional graph $G = (V,E)$, where V is the set of nodes (vertexes) and E is the set of links (edges). The links, $(i, j) \in E$, are specified by their end nodes i and j. See Figure 4 for an illustration. A path sample is denoted $\omega_{[s,d]} = \{(d, i_1), (i_1, i_2), \cdots, (i_{k-1},d)\}$, where $(i, j) \in E$ denotes the link connecting node i and j, and k is the number of hops in $\omega_{[s,d]} \in \Omega_{[s,d]}$. Here, $\Omega_{[s,d]}$ is the set of all feasible paths between s and d. The objective of path management is to find a path, or a set of paths, from source node s to destination node d, with a minimum *cost*.

Cost. The link cost is denoted $L((i, j))$. It may vary with traffic load and time, and be a function of the link attributes as well as the message database of node i. The cost of a path is additive, $L(\omega) = \sum_{\forall(i,j)\in\omega} L((i, j))$.

3.1 Background

CEAS is based on two fundamentals; 1) the concept of emergent behaviour, and 2) the cross entropy method for stochastic optimization. Brief reviews of both fundamentals are given in the following.

Emergence is the behaviour of a system arising from a multiplicity of relatively simple interactions between system elements. This may be seen in nature, e.g. the behaviour of an insect colony emerging from the behaviour of swarms of insects. For an introduction, see for instance [29,30,31]. Emergence may be used to find solutions to optimization problems. An approach applying so-called 'swarm intelligence' is the Ant Colony Optimization (ACO) system [32]. An emergence analogy to path finding in networks is that mobile agents swarm randomly through the network from a source node to a destination node and communicate indirectly the quality of paths found by leaving messages, denoted *pheromones* (which also denotes chemical substances applied by insects to communicate), along their trail in a way similar to the foraging behaviour of ants. The use of swarm intelligence for path management was briefly reviewed at the end of Section 2.

The *Cross Entropy (CE) method for stochastic optimization* was first introduced by Rubinstein [7] and is applied for pheromone updates in CEAS. The basic notion of the method is that finding the best solution to a combinatorial optimization problem, e.g. a path in a network, by a random search has a very low probability when the search space is large, i.e. it is a *rare event*. Hence Rubinstein applies an *importance sampling* technique [33] where the random proportional rules are gradually and stepwise changed according to the importance (e.g. the cost) of the various paths found. The approach minimises the cross entropy between the random proportional rule matrices between two consecutive iterations considering the cost history. The CE-method is aimed at solving a wider range of discrete optimization problems (not only path finding). For a tutorial on the method, [34] is recommended.

With these fundamentals in mind, path management in arbitrary mesh networks is addressed. The structure of a network may formally be described as a bidirectional graph, see Fact Box 1. The overall objective is to, simultaneously and adaptively, find a set of paths between source – destination pairs in the network, minimizing the cost of the paths. Note that there may be additional requirements to the paths, for instance that paths should not contain loops/revisits

to nodes. Such requirements may also be more demanding. For instance, to require that a path is to form a Hamilton cycle visiting each node once[8], or require that corresponding working and protection paths must be disjoint.

The use of each link incurs a cost. The cost of a path should be the sum of the link costs along the path, cf. Fact Box 1. (Non-additive cost functions tend to create undesirable search spaces.) Hence the link cost will define the objective function of the swarm. For instance, if the link cost is the same for all links, paths with the minimum number of hops are sought; if the link cost is the measured short term average delay by using the link, the objective becomes finding the path with the shortest average end-to-end delay. Note that in the latter case, the link costs will change as the load of the network changes, hence the averaging interval applied will influence the reactivity and stability of the management network. The cost function may be compound and allow co-operation between swarms dealing with different tasks. An example: a penalty is introduced in the link cost for ants seeking a protection path if the link is likely to become a part of the working path for the same connection. Clearly the definition of the link cost plays a major role in the design of the emergent behaviour implemented by the management schemes presented in Section 4.

3.2 Cross Entropy Ant System (CEAS)

The Cross Entropy Ant System (CEAS) is illustrated in Figure 1. It was first introduced in [35] and is designed for robust and adaptive path management in communication networks. The source node of a path contains a factory that generates agents denoted (artificial) ants. Ants start their life cycles as *forward ants* searching for paths between a source and a destination node. At each step (intermediate node) along the path the next node is randomly chosen according to the random proportional rule in (1). The aggregated information in the random proportional rule is kept in *message databases* at each node as indicated in Figure 1. The paths and the behaviour of ants may also be governed by additional deterministic rules, e.g. paths without loops require that nodes are not revisited. At the destination node, the cost of a path is evaluated and a control variable, denoted *temperature*, is updated. The temperature indicates "the cooling level" of the search (cf. simulated annealing [10]), i.e. how close we are an optimal solution. From the destination *backward ants* return along the reversed path and update pheromone values in the message databases of each node visited. The better the path, the stronger the pheromone updates. The

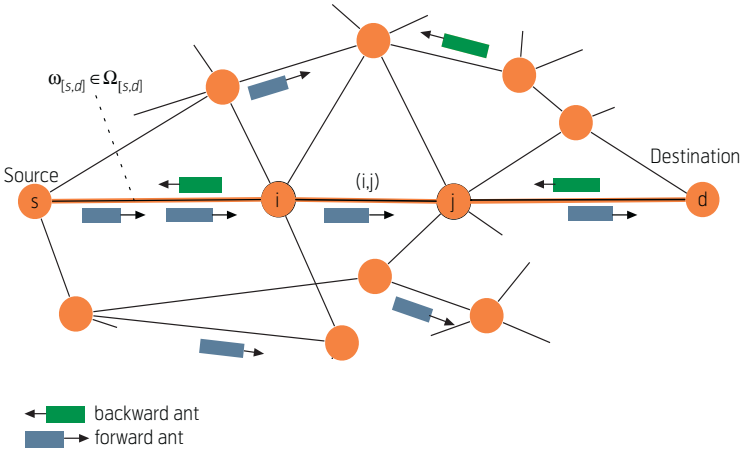


Figure 4 Graph representation of a network with forward and backward ants

corresponding change of the random proportional rule will guide future ants in their search for the same destination. A formal descriptions of the overall procedure is given in Fact Box 2.

The behaviour outlined above concerns *normal ants*. They follow the random proportional rule and maintain the paths with the best costs. In addition to normal ants, each factory generates a certain fraction of *explorer ants*. They perform random walks in the network to better detect new paths.

In Figure 1 it is shown that we may have more than one swarm searching for paths at the same time. In general, a CEAS may have an arbitrary number of species of ants as illustrated in Figure 5. Each of them deals with a single task, e.g. finding a good path

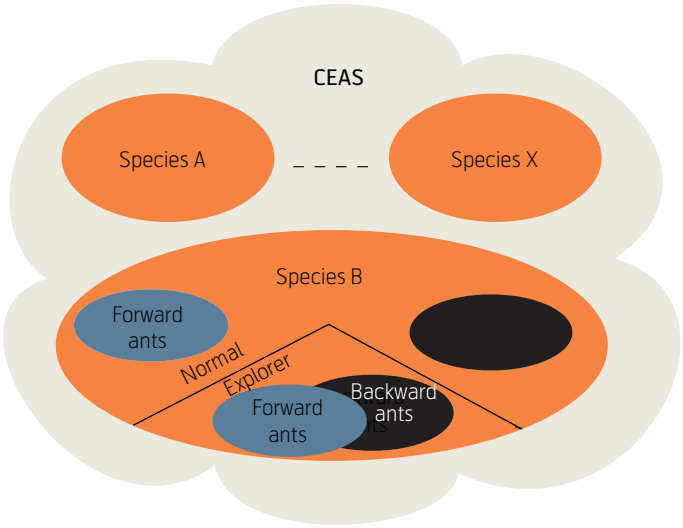


Figure 5 Classification of ants in CEAS

8) *Also known as a travelling salesman's tour. Finding a minimal such tour is an NP problem.*

ISSN 0085-7130 © Telenor ASA 2008

Fact Box 2 – Property 2. The Cross Entropy Ant System (CEAS) rules

Ant types. *Forward ants* are issued at the source s with rate λ_f, and are searching for the destination according to a random proportional rule. From the destination d, *backward ants* return on the reversed path and update the pheromone values and the random proportional rule in each node of the path. *Normal ants* follow the random proportional rule and maintain the paths with best cost, while *explorer ants* do random walks to detect new paths.

Random proportional rule. At visit t in node i, the normal forward ants sample their next hop according to a *random proportional rule* for link (i, j)[9)]

$$p_{ij,t} = \frac{\tau_{ij,t} \cdot I(j \notin U)}{\sum_{(i,l)\in E, l\notin U} \tau_{il,t}} \tag{1}$$

where $\tau_{i\,j,t}$ is the pheromone value of link $(i, j) \in E$ at update t, and U is a list of forbidden nodes according to deterministic rules associated with the ant. The random proportional rule matrix is denoted $p_t = [p_{i\,j,t}]_{\|V\|\times\|V\|}$.

Pheromone updating rule. The pheromone values are a function of the *entire history* of path cost values $L_t = \{L(\omega_1), \cdots, L(\omega_t)\}$ up to iteration update t. They are updated for *every path sample* according to (2), applying $H(L(\omega_t), \gamma_t)$, which is a *performance function* of the last cost value $L(\omega_t)$ and a control variable γ_t denoted the *temperature*.

$$\tau_{ij,t} = \sum_{k=1}^{t} I((i,j) \in \omega_k) \beta^{\sum_{x=k+1}^{t} I((i,\cdot)\in\omega_x)} H(L(\omega_k), \gamma_t) \tag{2}$$

The exponent of β is the number of ants that has updated node i at t since k, $\sum_{x=k+1}^{t} I((i,\cdot) \in \omega_x) \le t - k$ where $t - k$ is the total number of updates in the system at t since k.

Temperature update rule. The control variable γ_t, denoted the temperature, is determined by minimising γ_t subject to $H(L(\omega_t), \gamma_t) \ge \rho$ where ρ is a parameter (denoted *search focus*) close to 0 (typically 0.05 or less). In [7] a performance function recommended is $H(L_t, \gamma_t) = e^{-L_t/\gamma_t}$. To enable a continuous adjustment of γ at a small computational expense, an auto-regressive performance function $h_t(\gamma_t) = \beta h_{t-1}(\gamma_t) + (1 - \beta)H(L(\omega_t), \gamma_t)$ is applied. That gives

$$h_t(\gamma_t) \approx \frac{1-\beta}{1-\beta^t} \sum_{i=1}^{t} \beta^{t-i} e^{-\frac{L(\omega_t)}{\gamma_t}} \tag{3}$$

In [35] it is shown that the control variable γ_t is determined by minimising γ subject to $h(\gamma) \ge \rho$, which yields

$$\gamma_t = \left\{ \gamma \mid \frac{1-\beta}{1-\beta^t} \sum_{i=1}^{t} \beta^{t-i} e^{-\frac{L(\omega_t)}{\gamma_t}} = \rho \right\} \tag{4}$$

To avoid excessive storage and processing demands, it is assumed that the changes in γ_t are small from one iteration to the next. This enables a (first order) Taylor expansion of (4), providing

$$\gamma_t = \frac{b_{t-1} + L(\omega_t) e^{-L(\omega_t)/\gamma_{t-1}}}{\left(1 + \frac{L(\omega_t)}{\gamma_{t-1}}\right) e^{-L(\omega_t)/\gamma_{t-1}} + a_{t-1} - \rho\frac{1-\beta^t}{1-\beta}} \tag{5}$$

where $a_0 = b_0 = 0$ and $\gamma_0 = -L(\omega_0) / \ln \rho$, and

$$a_t = \beta\left(a_{t-1} + \left(1 + \frac{L(\omega_t)}{\gamma_t}\right) e^{-\frac{L(\omega_t)}{\gamma_t}}\right)$$
$$b_t = \beta\left(b_{t-1} + L(\omega_t) e^{-\frac{L(\omega_t)}{\gamma_t}}\right)$$

The pheromone values in (2) are updated by the result of (5). Again, to reduce processing and storage requirements, (2) is reformulated by a (second order) Taylor expansion

$$\tau_{ij,t} \approx I((i,j) \in \omega_t) e^{-\frac{L(\omega_t)}{\gamma_t}} + A_{ij} + \begin{cases} -\frac{B_{ij}}{\gamma_t} + \frac{C_{ij}}{\gamma_t^2} & \frac{1}{\gamma_t} < \frac{B_{ij}}{2C_{ij}} \\ -\frac{B_{ij}^2}{4C_{ij}} & \text{otherwise} \end{cases} \tag{6}$$

where

$$A_{ij} \leftarrow \beta\left(A_{ij} + I((i,j) \in \omega_t) e^{-\frac{L(\omega_t)}{\gamma_t}} \left(1 + \frac{L(\omega_t)}{\gamma_t}\left(1 + \frac{L(\omega_t)}{2\gamma_t}\right)\right)\right)$$
$$B_{ij} \leftarrow \beta\left(B_{ij} + I((i,j) \in \omega_t) e^{-\frac{L(\omega_t)}{\gamma_t}} \left(L(\omega_t) + \frac{L(\omega_t)^2}{\gamma_t}\right)\right) \tag{7}$$
$$C_{ij} \leftarrow \beta\left(C_{ij} + I((i,j) \in \omega_t) e^{-\frac{L(\omega_t)}{\gamma_t}} \left(\frac{L(\omega_t)^2}{2}\right)\right)$$

The initial values of (7) are $A_{ij} = B_{ij} = C_{ij} = 0$ for all $(i, j) \in E$.

In spite of the seeming complexity of the above equations, they yield a compact ant implementation with minimal storage and processing demands.

9) *In the initialisation phase the ants explore G with a uniform random proportional rule $p_{i\,j,t} = 1/(N_i - 1)$; $\forall i$. N_i is the number of neighbours to node i.*

between a source and destination pair. Co-operation between the species can be used to

- reach the overall objectives for the system through several parallel tasks, and
- increase the performance of the system (details in Section 4.5).

Different species may be independent. This is for instance the case when the network has fixed link costs and the objective is to find the shortest paths, e.g. the minimum number of hops. Note that this is similar to the common routing schemes applied in today's Internet, i.e. OSPF, IS-IS, BGP. There may be an implicit communication and co-operation between species through the performance of the network. This is for instance the case when we use measured (short term average) delay as path cost. Use of CEAS based path management enables load balancing and allows a gradual shift of load between paths. By this instabilities, which may occur with dynamic link measures in distance vector routing schemes, are avoided. Ant species may also co-operate through explicit indirect communication. In such a case, pheromones left by "alien" species, e.g. A-ants from Figure 5, will be included in the link cost function of B-ants. This enables a co-ordinated planning among the species. How the various co-operation schemes may be used to achieve management objectives are exemplified in the application discussion of Section 4.

3.3 Performance Improvements

Experiments show that CEAS is very robust. Paths are found and maintained even with a large number of lost messages (ants). CEAS can find new solutions and adapt to changes quickly. Increased ant rates speed up adaptation as long as they are well below the network capacity. If rates are very high, the processing and transmission capacity consumed by the ants run in danger of disturbing the network production capacity (forwarding of packets) and in the worst case cause routing instabilities. If ant rates are low, CEAS may react too slowly and the system will have a transient period after a change in the network conditions where routing is suboptimal or even not operational. To control and reduce overhead in terms of memory, processing and bandwidth consumption, we have made several extensions to the original CEAS. Elitism has been introduced in our *elite selection* approach improving convergence and reducing overhead in general. Furthermore, the overhead (in terms of number of ants) is reduced by two complementary extensions; self-tuning of ant rates in the source of the path, and self-tuning of rates in intermediate nodes. Finally, overhead in terms of memory consumed in each node is reduced by an extension introducing *cooperation between ant species* coming from different sources but with the same destination and cost objectives. Cooperation also improves convergence rates.

Elite selection. *Elite CEAS* [36] is introduced to speed up convergence and to reduce the overhead in terms of number of updates. For each ant that reaches its destination, the cost of the path found by the ant is compared with an elite selection level. If the cost is below the level, a backward ant is returned towards the source node updating all pheromone values along the reversed path. Otherwise, the ant is discarded and no updates of the pheromone values take place. The elite selection level is self-tuned and gradually tightened as better and better solutions are found. It finally converges to the cost of the best (optimal) path.[10)] In [38] the elite selection is applied only on normal ants and the results show a significant improvement in the performance compared to using elite selection on both normal and exploration ants.

Self-tuning ant rates. In [39,40] we propose two different, but complementary approaches to self-tune the ant generation rates in CEAS. The approaches are applicable to the adaptive path strategy variant of CEAS [41]. In short the approaches regulate the ant rate in the source of the path and in the intermediate nodes as a reaction to changes in network conditions. Results from simulation studies in [39,40] show that the overhead is significantly reduced without sacrificing performance in terms of convergence times.

Source rate. The source of a path generates ants at a given rate to search for a specified destination. If a (near) optimal path has been found and the network is stable, the source ant rate could be very low because a very limited number of ants is required to maintain and refresh the best path(s) in the network. If a change in the network occurs, the sending rate should be increased, and decreased again when the transient effect dissolves. To detect changes in the network conditions we propose to estimate and monitor the rates of sent (forward) and received (backward) ants in the source node. When a network is stable, nearly all ants follow a path with the same cost, and rates of forward and backward ants will be (almost) equal.

---

10) *An extension similar to Elite CEAS has also been introduced in ACO to realize the MAX - MIN Ant System (MMAS) [37] where convergence is sped up by increasing the exploitation of the* best sample*. In MMAS however, a batch oriented approach is taken. Pheromone values are updated after* m *path samples have been collected, and updates are based on only the path with the best (either among the last* m *paths or all paths up to now) cost value.*

This is because all ants that find the destination will meet the elite selection criterion and return towards the source, updating pheromone values on the reversed path. However, when not all forward ants reach the destination, or when some ants that reach the destination are discarded by elite selection, the backward ant rate will be lower than the forward ant rate. Such a rate difference indicates that it is not known what the best paths are under the current network conditions. Hence, ant rates should be tuned in proportion to the difference between the estimated forward and backward rates, see (9). Note that self-tuning of forward ant rates in source nodes will require some detection and reaction time from a change in network conditions occurs to a corresponding increase in forward ant rate takes place. Required detection time depends on the network topology, and the location and nature of the link state event. To avoid that ant rates increase infinitely, an upper limit of the forward ant sending rate is defined equal to the initial exploration rate.

Node rate. The normal behaviour of a forward ant visiting a node is to look up relevant pheromone values and apply the random proportional rule to select which node to visit next. To reduce detection and reaction times relative to what is achievable by self-tuning in the source nodes, we have proposed to enable self-tuning of ant rates also in intermediate nodes. Broadening the search for new paths by sending more forward ants when network conditions have changed is in general desirable. However, a broadened search is particularly desirable in the domain that is directly affected by the changed network conditions. Hence, local rate tuning has a potential. Two detection mechanisms that depend only on local node information are introduced; detection of link status changes (carrier/no carrier) of this node, and detection of changes in rates of forward and backward ants passing through the node. If no changes are detected this implies normal ant behaviour. If a change is detected the forwarding rate is increased by replicating the forward ants which continue with normal behaviour. This will broaden the search for the destination and produce more alternative paths in a shorter time after a change in network conditions than is possible with self-tuning in the source nodes only. An ant replica is a copy of the original forward ant (including the path history from the source to current node). An ant replica continues to search for a path applying normal ant behavior, however it will never be replicated again. Only original ants can replicate and only original ants are monitored to estimate forward ant rates in each node. The number of ant replica generated in a node is regulated by a replication forward ant rate that is proportional to the difference in rates of forward and backward ants passing through the node, to the number of outgoing links of the node, and to how important and critical a link is in the random proportional rule. The latter is quantified by means of Shannon's *entropy measure*, $E = -\Sigma_{\forall x}\, p_x \log\,(p_x)$ [42].

Ant species cooperation. The original CEAS was designed for finding paths for end-to-end connection and the ant species identifiers (pheromone IDs) are specified according to source-destination pairs, i.e. the end nodes of a path[11]. Such a design does not scale well. When a large number of paths need to be found, a large set of unique pheromone values need to be stored in most nodes in the network (minimum in all nodes that are part of the best paths for each of the source-destination pairs). The situation can be improved by letting ant species with partly common interests cooperate and share pheromone values. In this section we describe two such approaches.

Overlapping resource paths and profiles. In [44] a version of CEAS is designed where an ant's objective is to find a path to a resource type using a search profile that specifies a set of required resource attributes, e.g. capabilities, delay, content, and security requirements. This means that it differs from the path finding described above where only one (or at most two) attribute is considered. Each attribute in the search profile has a cost function related to it, with corresponding temperatures and pheromone values. A forward ant will apply a random proportional rule based on a combination of pheromone values, i.e. one pheromone value for each attribute in the search profile weighted by the corresponding temperature of each attribute. Backward ants will update pheromone values correspondingly. If the search profile of two ant species share attributes then they will both read and update the pheromone values for the shared attributes (as well as for attributes not shared). Hence, in each node a set of pheromone values will exist, which is the union of all attributes contained in the different search profiles of ants that have visited the node. A significant reduction in the number of unique pheromone values required to be managed in a node is now possible, compared to storing species specific sets of pheromone values. In addition, convergence times will be reduced since more ants (of several ant species) update the same pheromone value. See [44] for formal descriptions and performance studies.

Sub-path cooperation. Recall that the general objective, and overall problem, of path management is to find a path from a source to destination node with a

[11] *In [43] ant species identifiers are extended with primary or backup path indexes.*

Fact Box 3 – Property 3. Performance improvement

Elite selection. Let n be the number of forward ants from s to d, and $t \leq n$ the number of backward ants. ω_n^* is the path of forward ant n, and γ_n^* is the *temperature* determined by (2) using the cost values observed by all n forward ants, $L_t^* = \{L(\omega_1^*), \cdots, L(\omega_n^*)\}$. In [36] the pheromone values and the backward ant index t are updated only when the *elite selection criterion,* $L(\omega_n^*) \leq -\gamma_n^* \ln \rho$, is true. Hence, the *update path sample* ω_t is

$$\left\{ \begin{array}{l} \omega_t = \omega_n^* \\ t \leftarrow t+1 \end{array} \mid L(\omega_n^*) \leq -\gamma_n^* \ln \rho \right\}$$

The temperature in (2) and pheromone values in (6) are using the update path samples ω_t.[12)] The elite selection should be applied to *normal ants only*, and let all explorer ants lead to an update, see [38] for results.

Rate estimates. The forward and backward rates can be estimated by discretization of the time axis with granularity δ (and counting the number of ant arrivals in time intervals of size δ. A running average $\hat{\lambda}$ of these rate estimates is generated applying an auto-regressive formulation

$$\hat{\lambda}_{a,m} = \alpha \hat{\lambda}_{a,m-1} + (1-\alpha) \cdot \frac{N_{a,m}}{\delta}, \quad \hat{\lambda}_0 = 0, a \in \{f, b\} \tag{8}$$

where $N_{a,m} = \| \{T_a \mid (m-1)\,\delta \leq T_a < m\delta \vee T_a \in \mathbf{T}_a\} \|$ is the number of ant arrivals in time interval m, and $\mathbf{T}_a$ is the set of ant arrival events, where $a \in \{f,b\}$ (forward and backward ants). α is a memory factor that is tuned to capture the transient effects of the network.

Source rate self-tuning. The ant generation rate λ_f varies between $\lambda_0 \geq \lambda_f \geq \lambda_s$. The self-tuning rate introduced in [39,40] provides a low rate when $\lambda_f - \lambda_b$ is small (stable system), and a high rate when it is large (system not yet converged or network conditions changed):

$$\lambda_{f,m} \leftarrow \max\left(\lambda_s, \lambda_0\left(1 - \frac{\hat{\lambda}_{b,m}}{\lambda_{f,m-1}}\right)\right), \tag{9}$$

Node rate self-tuning. In [40] forward ants are replicated. The *replication rate* λ_r is self-tuned and proportional to the difference $\lambda_f^{(i)} - \lambda_b^{(i)}$ (where $\lambda_f^{(i)}$ and $\lambda_b^{(i)}$ are the forward and backward rates in node i), to the node out-degree ν_i, and to the knowledge about preferred links quantified by the Shannon's *entropy measure*, $E = -\Sigma_{\forall x} p_x \log(p_x)$ [42]. Hence,

$$\hat{\lambda}_{r,m}^{(i)} = \frac{E_m^{(i)}}{E_{\max}^{(i)}}\left(\hat{\lambda}_{f,m}^{(i)} - \hat{\lambda}_{b,m}^{(i)}\right)\nu_i = \nu_i \log(\nu_i)\left(\hat{\lambda}_{f,m}^{(i)} - \hat{\lambda}_{b,m}^{(i)}\right)\sum_{(i,j)\in E} p_{ij}\log(p_{ij}) \tag{10}$$

Maximum entropy $E_{\max}^{(i)} = \log(\nu_i)$ is when all links have the same (or no) pheromone values, i.e. uniformly distributed $p_{ij} = 1/\nu_i$.

Search profile cooperation. The objective of resource path management is to find a minimum cost path applying a chain of resources of a specific type, each with some given security r, capabilities f, content c and quality q. The random proportional rule, pheromone update rule, and temperature update rule in Fact Box 2 are no longer (s,d) specific, but instead applied individually for each attribute in the (r, f, c, q)-tuple. Note that each of the source profile attributes can be a set. If ant species i and j, as defined by their (r, f, c, q)-tuple, have (partly) overlapping tuples, $(r_i, f_i, c_i, q_i) \cap (r_j, f_j, c_j, q_j) \neq \emptyset$, they will (partly) use the same random proportional rules and update the same pheromone values of the attributes they have in common.

Sub-path cooperation. The objective of path management is to find a path from source node s to destination node d, with a minimum *cost*. In sub-path management the objective is to find a minimum cost path from a node i to the destination d. In node i the random proportional rule is (i,d) specific and can be shared and updated by all ant species that visit node i on their way to destination d.

minimum cost. Such problems may be split into subproblems [45], and hence the objective of subpath management can be to find a minimum cost path from an intermediate node to the destination. In such a case an intermediate node will apply a random proportional rule given for sub-paths from the node to a set of destinations. Hence, the pheromone values, and the corresponding random proportional rules, can now be shared and updated by all ant species visiting the same intermediate node and looking for the same destination. The sub-path concept was first introduced in [45], and improved and studied in detail in [38]. Simulation results from case studies applying a small and a medium sized network show significant savings in convergence time, memory storage and ant rates required.

12) *Note that if the elite selection criterion is not met the iteration index* t *is not updated and no path is assigned to the update sample set.*

4 Applications of CEAS

The basic CEAS technology has demonstrated its applicability through a variety of studies of different path management strategies, including; shared backup path protection (SBPP), p-cycles, resource search under QoS constraints, adaptive paths with stochastic routing, and traffic engineering of MPLS. This section provides highlights from these studies. All studies have been conducted on models of small and medium sized networks with different sets of network dynamics. Studies are conducted by simulation using ns-2[13] or Simula/DEMOS [46,47].

4.1 Disjoint Primary-Backup Paths with Performance Guarantees

In [48] a 1:1 protection scheme is adopted, i.e. every primary path is to have an independent and disjoint backup path ready for use if a link failure occurs in the primary path. Having dedicated capacity for backup paths implies 100 % redundancy in a network which is inefficient and expensive especially when failures are rare. Hence in [48] the capacity required by a backup path is sought shared with other (non-conflicting) backup paths, i.e. a shared backup path protection scheme (SBPP) is applied. Note that finding sets of such SBPP paths is complex and resembles proven NP-complete problems like "Path with Forbidden Pairs", "Disjoint Connecting Paths" and "Shortest Weight-Constrained Path" [49].

The SBPP version of CEAS in [48] lets each primary path and each back-up path be dealt with by a separate species of ant. The different species are made to detest each other in accordance with primary/backup optimisation criteria, i.e.

- Backup ants search for paths which are disjoint with their corresponding primary paths.
- Backup ants having overlapping corresponding primary paths search for disjoint paths.
- All ants detest other ants which represent a load that in addition to their own load may incur an overload of a link.

A novel cost function is devised to make the above behaviours emerge. The difference between the available and required capacity of all potential paths over a link is summarised for each link in a path, which results in an estimate of expected loss along the path:

$$L(\omega_m^r) = \sum_{(i,j)\in\omega_m^r} \mathcal{S}\left[a_m + \sum_{\forall ns:(,i,j)\in\omega_n^s} P_{ij}^{ns} V_i^{ns} Q_{mr}^{ns} a_n - c_{ij}\right]$$

where $S[...]$ is a shaping function applied to smooth the search space, c_{ij} is available capacity on link (i,j), a_m and a_n are required capacity of current (m) and competing (n) paths respectively, and $P_{ij}^{ns} V_i^{ns} Q_{mr}^{ns}$ is the total weight of the required capacity of competing traffic enforcing the detestation scheme. The index r is path rank, i.e. primary or backup. Results from case studies simulating SBPP CEAS show that near optimal sets of primary and backup paths can be found efficiently. See [48] for details.

4.2 Protection Cycles

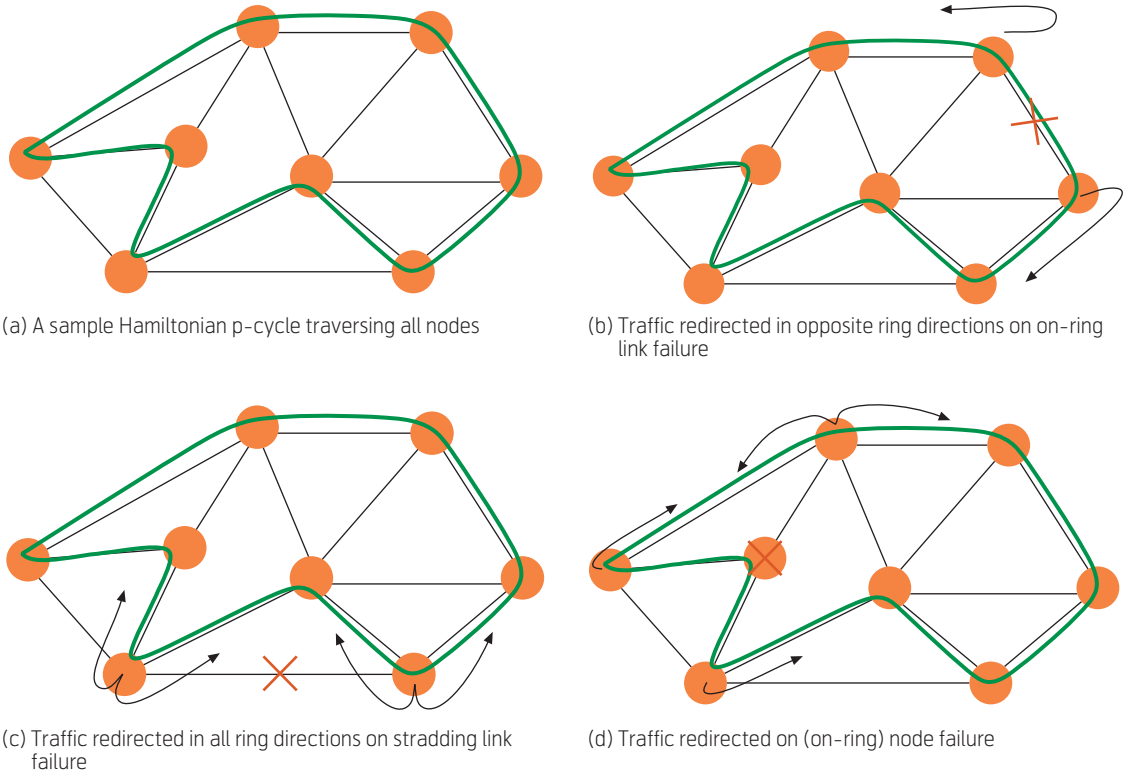
Protection cycles is a well know dependability measure and commonly applied in SDH networks with ring topologies. Applying protection cycles, known as *p-cycles* [50] in meshed networks has been shown to provide good protection against failures in both network links and nodes. Upcoming optical burst and packet switched networks will especially require node protection due to the expected longer down time of nodes compared to optical links.

A version of CEAS presented in [51] is capable of finding near optimal Hamiltonian cycles in meshed networks with respect to the amount of spare capacity on the links in the cycle. Hamiltonian cycles are good p-cycle candidates, enabling protection of all links and nodes in a network as may be seen in Figure 6. A new tabu memory as well as new cost functions are introduced in [51] to help CEAS find relevant cyclic paths. The capacity of the strongest "weakest" link of paths found is stored, shared and updated by all ants. A "weakest link" is the link with the lowest spare capacity in a path. Links with less spare capacity than the current strongest "weakest link" are tabued and hence avoided by ants during forward search. The cost functions take available capacity in both directions of links into account. See [51] for further details. Early results from simulations indicate that the new system has a promising ability to find good candidate p-cycles.

4.3 Path Management in Telenor's IP Network

The (former) backbone topology of Telenor's IP network has been applied in a series of simulation experiments with CEAS. The topology, as illustrated in Figure 7, consists of a core network with ten core routers (green) in a sparsely meshed topology, ring based edge networks with a total of 46 edge routers

[13] *http://nsnam.isi.edu/nsnam/index.php/Main_Page*



(a) A sample Hamiltonian p-cycle traversing all nodes

(b) Traffic redirected in opposite ring directions on on-ring link failure

(c) Traffic redirected in all ring directions on straddling link failure

(d) Traffic redirected on (on-ring) node failure

Figure 6 p-cycle *protection in a meshed network*

(black), and a dual homing access network with 160 access routers (orange). The topology consists of approximately 350 links where the relative transmission capacities are 1, 1/4 and 1/16 for core, edge and access links, respectively. The service studied in this network is IP connectivity and in particular establishing virtual connections (VCs) with performance guarantees. CEAS is used to establish, maintain and monitor VCs between one or several node pairs connected to access routers (green nodes). In Figure 7 the topology is illustrated with an example VC that is established between node 74 and 164. The best route is indicated by thicker lines. A distribution of the number of hops of the shortest paths[14] between any pair of access routes is included in a sub-figure. The average number of hops is 6.37 and the majority of paths have six or seven hops.

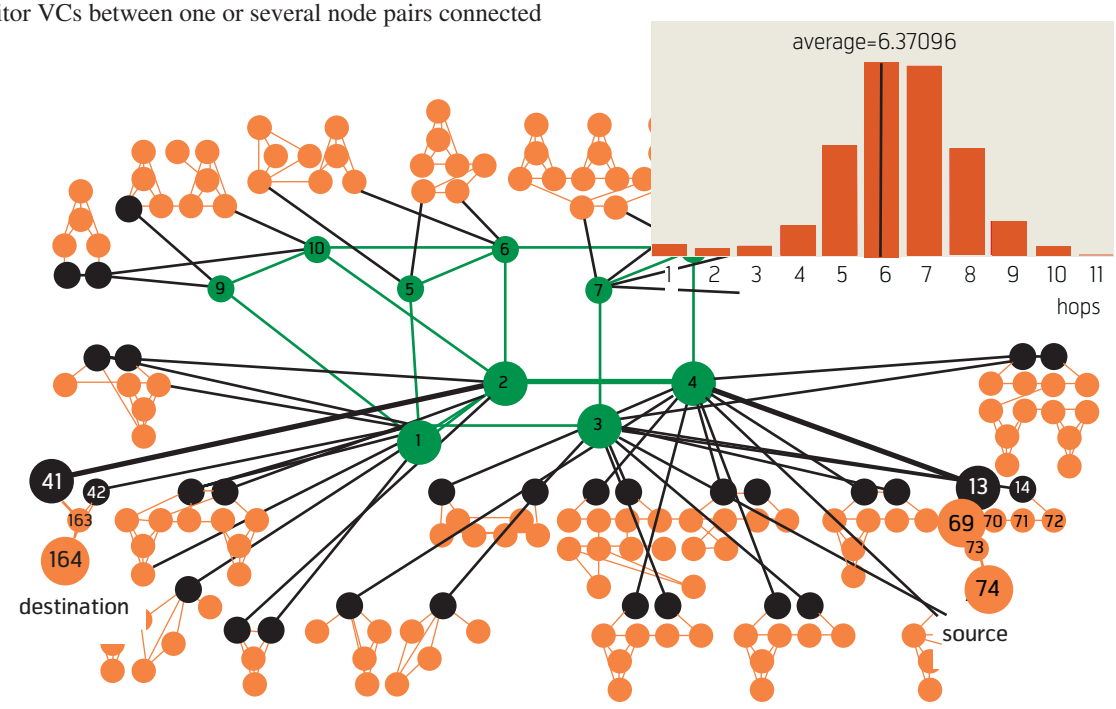


Figure 7 The simulated backbone network with 216 nodes and 350 links. The average number of hops between any pair of access nodes is 6.37. The shortest path between node 74 and 164 is given as an example

ISSN 0085-7130 © Telenor ASA 2008

The illustrated network has be applied to demonstrate overhead reduction techniques described in Section 3.3, for studies of the effect of pheromone sharing in Section 4.5, and as a large scale example to demonstrate the use of CEAS for performance monitoring, see Section 4.6.

4.4 Load Sharing and Protection

In [52] two path management strategies are investigated. The *primary backup* scheme has as its prime objective to establish disjoint primary and backup (MPLS) paths for (all) source destination pairs. As described in Section 4.1, primary and backup paths are to be established such that backup-paths reuse network resources without preventing (due to overload) the scheme to provide continuity of service when a network element fails. The scheme's main advantage is its explicit knowledge of the immediately restorable traffic. However, its scalability with respect to increasing network sizes and advanced priority policies is not yet investigated. The other strategy investigated is the *adaptive path* scheme, which applies stochastic routing of paths for all source destination pairs in all nodes of the network. Such a scheme pro-actively provides alternative paths in case of failure. The scheme's main advantages are simplicity and fast adaption to major chances in the network. It lacks, however, the ability to give explicit indication of the fraction of traffic that will experience continuity of service. Differentiation or priority is also difficult to provide.

The network presented in Section 4.3 has been applied for investigating the operation of the *primary backup* and the *adaptive path* strategies under dynamic network conditions. Aspects of establishing and managing 10 separate paths are studied in detail. Paths are exposed to network link failures, drops of management information, and changes in offered traffic loads, see Table 1 for a summary of the simulated changes. The objective is to study the transient behaviour, i.e. the adaptivity and robustness, of this distributed management approach. For a more comprehensive discussion, see [52].

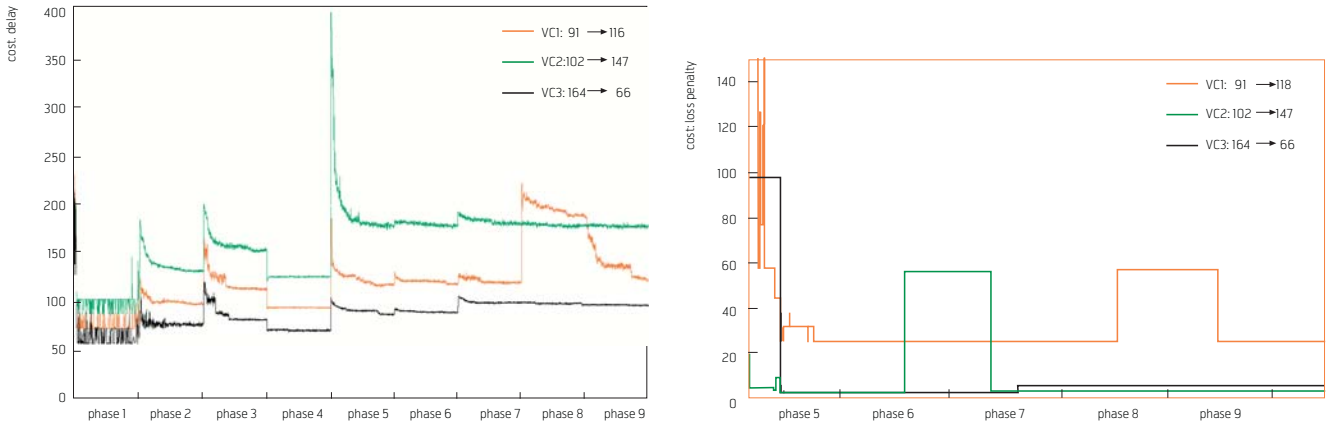
The results presented in Figure 8(a) shows the cost (e.g. delay) as a function of time. All results are based on ten simulation replications. Observe that the adaptive path strategy quickly switches to an alternative path on excessive load or link failure, and almost immediately back to the original path when load decreases or link is restored. In [52] it is also observed that the adaptive path strategy will distribute the load among paths with equal cost (e.g. delay) because a path is randomly selected according to the relative pheromone values that again are determined by the cost values.

Figure 8(b) shows the results from simulation of the primary-backup strategy for the same scenario. A switch-over from a disconnected operational path to an alternative path, either by protection switching (primary to backup) or by restoration (primary to a new primary), will cause an interruption of service. Observe for example the behaviour of VC2. After the core link failure at the beginning of phase 6, the primary path of VC2 is disconnected and VC2 is broken (regarded as down time). Explicit link failure notification will improve the path availability by making the protection switching mechanism more reactive.

Phase	Average load, ρ	Link events	Comments
-	0	-	Exploration phase
1	0	-	Initial topology
2	0.3	-	Increased load
3	0.6	-	Increased load
4	0.3	-	Decreased load
5	0.9	-	Sign. increase in load
6	0.9	Down [4,8], [6,8], [1,2]	Core links failed
7	0.9	Down [3,20], [1,42], [7,55], [3,22]	Edge links failed
8	0.9	Down [19,86]	Access link failed
9	0.9	Restored [19,86]	Access link restored

Table 1 Dynamic scenario for testing of adaptivity

[14] *Determined by Dijkstra's algorithm assuming static link costs.*



(a) Adaptive path strategy

(b) Primary-backup strategy

Figure 8 Comparison of management strategies in dynamic environment

4.5 Path Finding with Partly Overlapping Paths

In contrast to existing resource localization mechanisms for peer-to-peer systems a version of CEAS is presented in [44] which determines the path quality and cost by considering all resources involved, including peering and (client and server) middelware and network resource. Two types of profiles with QoS parameters are introduced. A *user request profile* specifies requirement to different resource-types relevant for a search. During a search the user request profiles are matched against *resource profiles* provided by resources visited. A resource profile indicates capabilities of a resource. Profile matching results in QoS loss vectors, and a set of such vectors finally generates a path cost vector as well as an overall path cost. The path cost vector (together with a temperature vector) is applied by backward ants such that a pheromone value for each QoS parameter is updated in all visited resources.

The CEAS version of [44] enhances the scalability by enabling cooperation between ants when they, fully or partly, have overlapping user request profiles. The total number of relevant unique QoS parameters $|\Xi|$ will be limited, hence a limited number of unique pheromones is required. However the total number of possible unique profiles $N_{\overline{\xi}}$, will still be large since

$$N_{\overline{\xi}} = 2^{|\Xi|} - 1$$

e.g. to enable a total of N, = 10^{100} different profiles only $|\Xi| \approx 333$ unique pheromones are required.[15)]

Cooperative behavior due to overlapping profiles will also increase general performance since more ants update the same pheromones (especially in popular resources). Results from simulations are promising and show that a set of near optimal resource paths conforming to a set of different but overlapping user request profiles may be found with improved performance.

Scalability as a result of cooperation between ants when paths overlap may also be achieved for virtual connections (VC) between specific source and destinations. In [45] extensions and changes to CEAS required to enable such cooperation are described and initial simulations conducted. The extended CEAS lets all VCs with the same destination, typically with different sources, update the same pheromones in every shared node along the route to the destination. Cost is recorded, and corresponding pheromones are updated, *from the shared node to the destination* independent of the original source of the ant that visits this node on its way to its destination. Hence, ant species are identified by their shared node (the new source) and destination node, and not by their source (origin) and destination nodes as in the original CEAS strategy described in Section 3. As an illustration consider Figure 9 where two different ant species search for the same destination resource. They share the sub-path from node 5 to the destination resource and have different sub-paths from the blue nest to node 5 and from the red nest to node 5. The figure is taken from [38]. While pheromones are better utilised, more temperature calculations per backward ant are required since each sub-path in a path must be considered separately when costs and temperatures are handled.

The effect of pheromone sharing has been tested on small network topologies with some network dynamics, e.g. with link failures as introduced in Figure 10.

15) *Deterministic requirements and binary matching of QoS parameters (loss or no loss) are assumed. By introducing non-deterministic requirements, i.e. having weighted loss output from a parameter matching, the profile space becomes even richer/larger.*

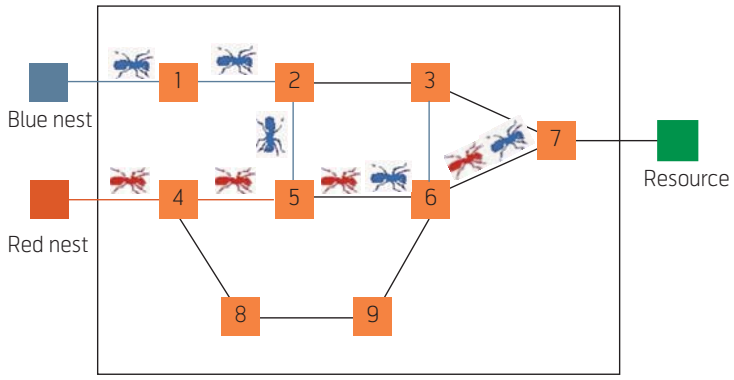


Figure 9 Two different ant species searching for the same destination (resource). In this example they share the sub-path from node 5 to the resource and have different sub-paths from the blue nest to node 5 and from the red nest to node 5. Figure obtained from [38]

The original source-destination and the new shared node-destination approaches are compared.

Results in Figure 10 show how the temperature for a VC changes as a function of iteration *t*. It can be observed that the convergence rate is improved without increasing the ant rate and with a significant reduction of the node storage demand with pheromone sharing compared to no sharing. It is also evident that the restoration time is significantly improved, and the failure detection is at least as good as in the shared strategy.

In [38] the shared pheromone strategy is studied in more detail and alternatives for cooperation are investigated. Several simulations have been conducted and demonstrate a significant effect of cooperation between ants with common interests. The simulation results using the Telenor example in Section 4.3 show that by introducing cooperation and reducing ant rates the number of temperature calculations may be kept at the same level as for original CEAS while overhead in terms of ant packets is reduced by 77 %.

4.6 Performance Monitoring of Path Quality

Monitoring of the quality of service is essential in the establishment and management of virtual connections. The ant system could be considered as a monitoring system. Several potential candidates for monitoring indices in an ant-based routing system are considered in [53] and the most promising with respect to detecting significant changes in the network conditions are (see summary in Table 2):

- Convergence index (temperature, or the elite limit that is a function of this);
- Cost value index (path delay, or loss ratio, available bandwidth);
- Pheromone values (in nodes).

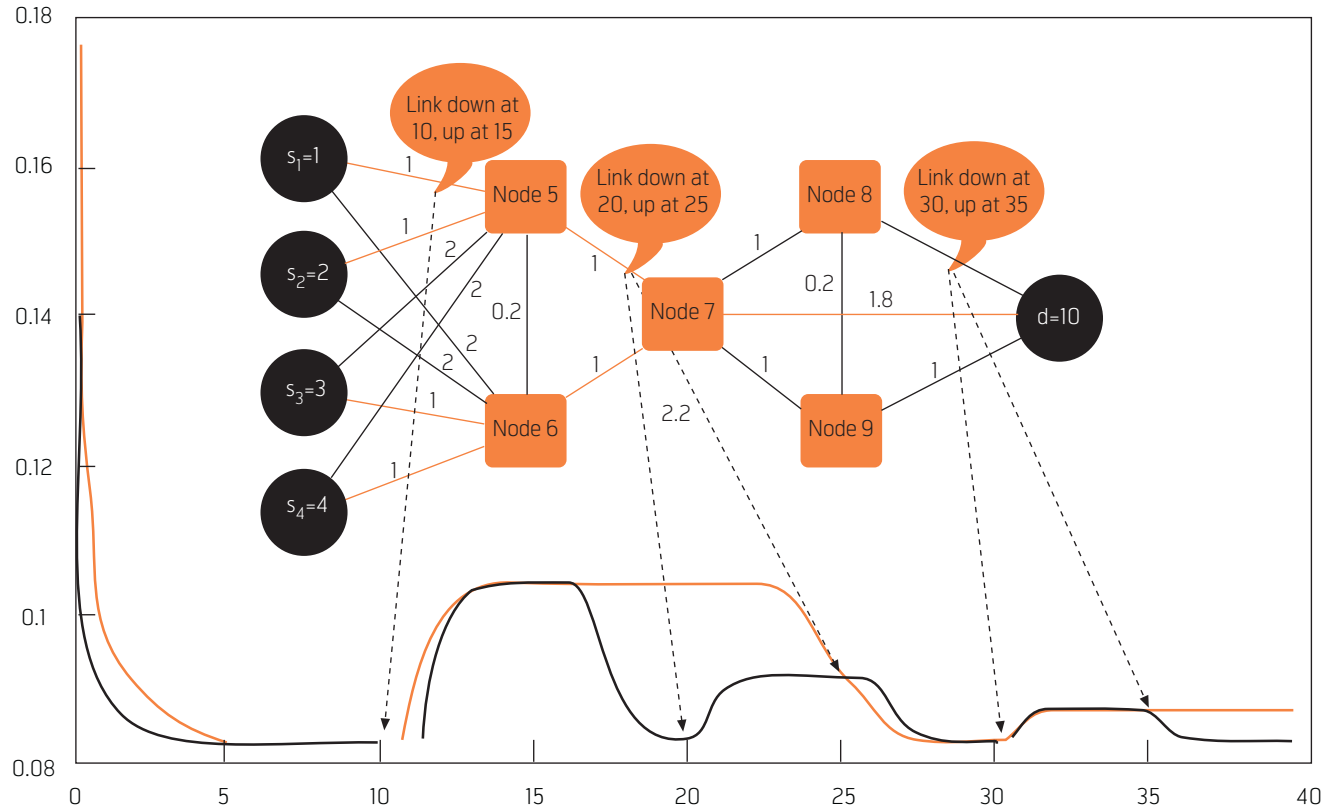


Figure 10 The temperature value for a VC as a function of iteration t *in a simulation case with topology and network dynamics as indicated*

ISSN 0085-7130 © Telenor ASA 2008

Metric	Observations	"Health"	Alarms
Ant route table	Deviation from data routing table	Misconfiguration in routing, interface overload	Significant deviation (in time or space)
Pheromone values	Increase by *x* % in *n* sec.	Node/link/path down	Check configuration
Convergence index	Decrease by *x* % in *n* sec.	New node/link/path discovered	A lower delay path for the VC exists
Cost value index	Average over *n* sec. decreased by x% last minute	After effect of change in network (still exploration)	None
Path probability	Close to max. for last minute	Stable networks	None

Table 2 Examples of use of CAS indices (from [53])

As an example, the cost (delay) of multi-VC connections is observed by simulation in *ns-2*. The simulator has implemented a Link state routing protocol that emulates the OSPF routing behaviour in an inter-domain and compares this with the CEAS behaviour. The network dynamics introduced in the simulations include node and link failure, and the variations in the cost and pheromone values in each node are observed.

In [54] it is observed that changes in the network topology are easily observed both by AntPing (probe packets routed by pheromones) and Ping (probe packets routed by link state tables) by use of time plot of cost and elite limit values. In Figure 11 is given an example. The plot includes observed cost values for the ants that return to the nest (AntPing: cost), the cost values for *all* ants reaching the destination (AntPing: costall), the elite limit that determines whether the VC should be updated or not (AntPing: elitelimit), and finally the one-way delay from the source to the destination observed by Ping packets. Observe that at time 50 a change occurs affecting the VC. At that point in time the change in delay of AntPing and Ping are different, hence they follow different paths. The reason is that the link state routing uses static cost values, while AntPing is sensitive to changes in link delays. If the cost metrics are not consistently set to reflect the (expected or observed) delays, the routing of AntPing and Ping might end up following different routes.

As an alternative to observations in end systems some information and indications of changes can be obtained by observing the pheromone values of the CEAS in the intermediate nodes. In Figure 12 the pheromone values for interface 2 and 3 in node 1 are

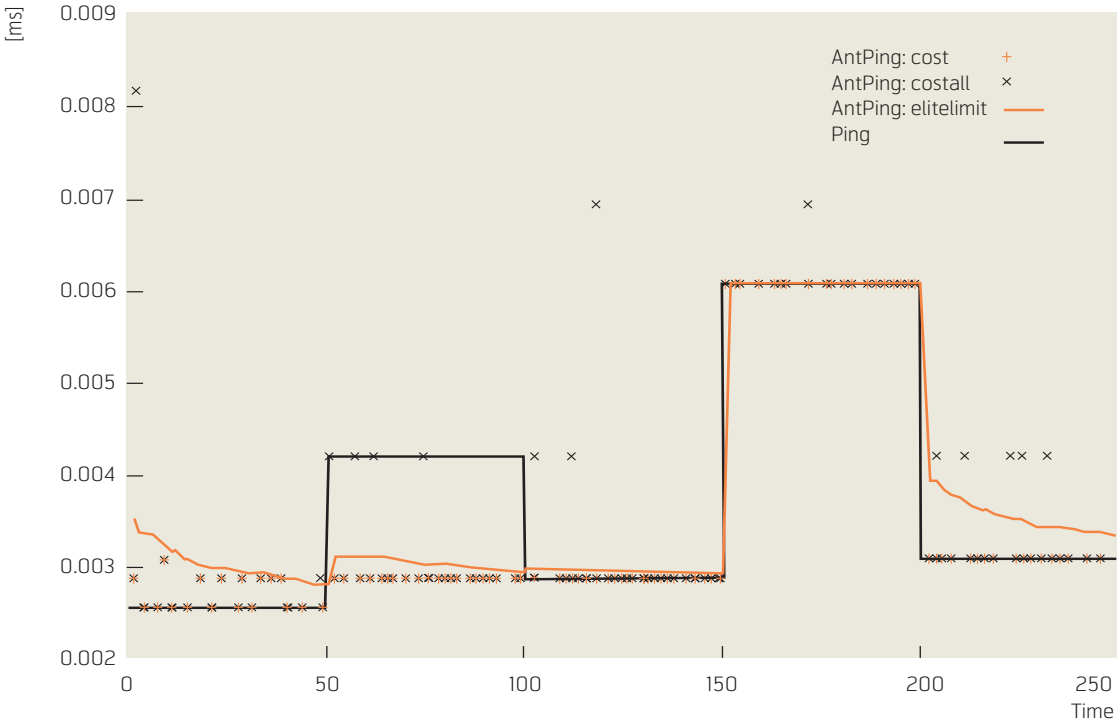


Figure 11 Cost and elite limit sample for typical time series

ISSN 0085-7130

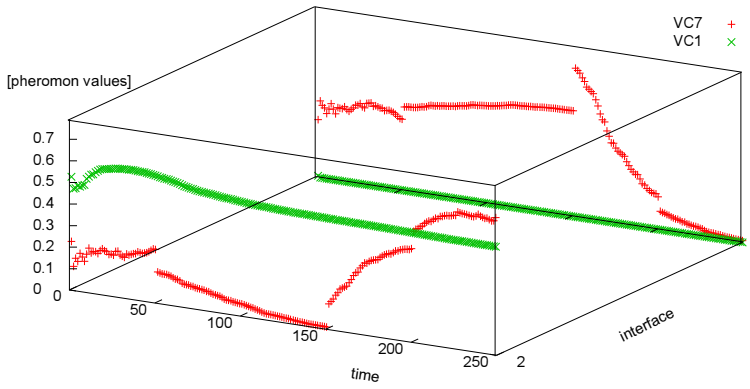


Figure 12 Pheromone values for VC3 and VC7 in node 1. It illustrates that VC7 changes preferred path at time 100 when link [2,4] fails. VC1 is not affected by the same dynamics, it has interface 2 as its preferred path for the entire period

plotted for two different VCs (denoted VC3 and VC7 in the figure). It can be observed that plotting all VCs in a node will visualize the *importance* (the number of VCs that have at least one preferred path through this node), *stability of node region* (the number of VCs that change the ant frequency and pheromone values) and *criticality* (the number of VCs with more than one preferred path through this node).

4.7 Establishing MPLS Label Switched Paths

In [55] the CEAS was applied for traffic engineering of Label Switched Paths setup by the Multi-protocol Label Switching (MPLS) protocol. CEAS is used to search, detect and monitor the best paths in the network and MPLS is applied to realize link disjoint primary and backup LSPs with specific QoS requirements. Control messages are specified to notify and change LSPs when something is changed in the network conditions.

Several strategies are studied for establishing virtual connections based on information from the underlying CEAS. The three most promising are discussed with respect to *speed of convergence* rather than globally optimal solution when network changes, and *stability* in solution ("route pinning") rather than switching to (temporary) slightly better solution when network is stable. The three strategies are

- **Check Periodically (the CEAS pheromone values)**. This strategy checks periodically the CEAS pheromone values to determine the current best path. The *best* path is here the path with the current highest selection probability, i.e. the highest pheromone value of the underlying CEAS. The speed of convergence depends on the periodic check interval and might slow down the reaction to changes, but will on the other hand not cause too rapid changes (maximum one per interval).

- **Check When (the temperature is) Crossing Limit**. By monitoring the temperature and its variation it is possible to determine if the CEAS has stabilised. This strategy triggers MPLS (re)establishment when the variance of the temperature falls below (or crosses) a certain limit. The stability is only checked once and therefor reduces the number of unnecessary changes of the LSPs. It is however important to choose a correct stability threshold.

- **Check When (the temperature is) Above Limit**. A variant of the previous strategy. The absolute level of the variance of the temperature can be used to specify if the system is considered to be in a stable period or not. If the variance is below a certain limit the system is said to be stable, otherwise it is unstable. This strategy reacts fast to system changes, but a correct variance stability threshold must be set.

The latter two strategies are sensitive to the estimation of the variance of the temperature and to the specified limits.

A series of simulation in *ns2* is conducted. The *Check When (the temperature is) Above Limit* strategy performs best with respect to the bandwidth of the connections, but trigger more path changes compared to the *Check When (the temperature is) Crossing Limit* strategy. The *Check Periodically (the CEAS pheromone values)* strategy is easy to implement and requires no calculation and storage of CEAS indices like temperatures and its variance. It reacts slower to changes than the other two since it is necessary to wait till the next CEAS check is conducted.

5 Demonstrator System

Implementing a working prototype provides useful insights into the complexity of swarmbased methods in real routers, and reveals potential implementation challenges and performance bottlenecks which are hard to predict through simulations and analysis alone. Hence, CEAS has been prototyped, both to gain implementation insights and to provide a running system able to demonstrate CEAS routing principles and illustrate the inner workings of the method. Both technical implementations of CEAS described in this section are based on the *Click* Modular software router system [56]. Section 5.1 describes the first pioneering prototype implementation of CEAS which uses Click for packet forwarding and a Java-based Mobile Agent System called *Kaariboga* [57] for the routing process. An upgraded prototype

implementation denoted *AntPing* is described in Section 5.2. AntPing provides improved performance by use of Click for both routing and forwarding. Section 5.2 explains how to use the AntPing to demonstrate CEAS and how a viewer can interact with the system.

5.1 Mobile Agent CEAS

The Mobile Agent CEAS was implemented as a "proof-of-concept" to gain experience with technical issues and effects that are hard to predict only through simulations. Maximising the performance of the working system was not a key consideration. The implementation was conducted as part of a Master assignment [58,59].

The system consists of two main components which interact to integrate the ants with the underlying network as shown in Figure 13. The software router is a customised version of the *Click modular router* system [56]. This implements the forwarding engine where data packets are forwarded according to the routing table that are updated by the ant system. The CEAS logic are implemented in a Java based Mobile Agent System framework called Kaariboga [57]. This ant-system receives ant packets detected by the kernel of the host machine. In case the ant packets are forward ants, they are routed stochastically according to the routing table, and otherwise the routing table is updated based on information in the ant packets and then forwarded to the next hop given in the header of the same packet.

The system was successfully implemented and tested in a small network. The test showed that the system is able to adapt to changes in traffic patterns and topology in the underlying network. Java and mobile agent systems are well suited for a rapid implementation of the system, however the implementation suffers from severe performance limitations even in a small-scale demo network.

5.2 AntPing

AntPing is also a prototype implementation of CEAS [54] developed as part of the final deliverable of the BISON project (IST-2001-38923). The main purpose was to learn more about the realization challenges of swarm intelligence on IP routers. The demonstrator visualises how ants are moving and dropped in the network. Animations are live and show how ants are searching and updating paths. Live plots of current and historical cost values of each virtual path are also provided as a function over time. The rest of this section includes a few details about the implementation and description about what it demonstrates.

Implementation. To achieve improved performance compared to the prototype system in Section 5.1, *AntPing* is implemented without use of the mobile agent system. Ants are no longer mobile agents but simple IP packets. AntPing extends the Click Modular software router system, and uses *hping3* [60] to generate and receive ant-packets from source to destination. The AntPing is running on home routers, with OpenWRT Linux [61], see [54] for more details. Figure 14 shows the functional blocks that extend Click (in the routers) and hping (in the end-systems). Figure 15 shows a picture from the lab. This implementation has moderate hardware and software requirements, which makes the demo inexpensive, flexible, and portable.

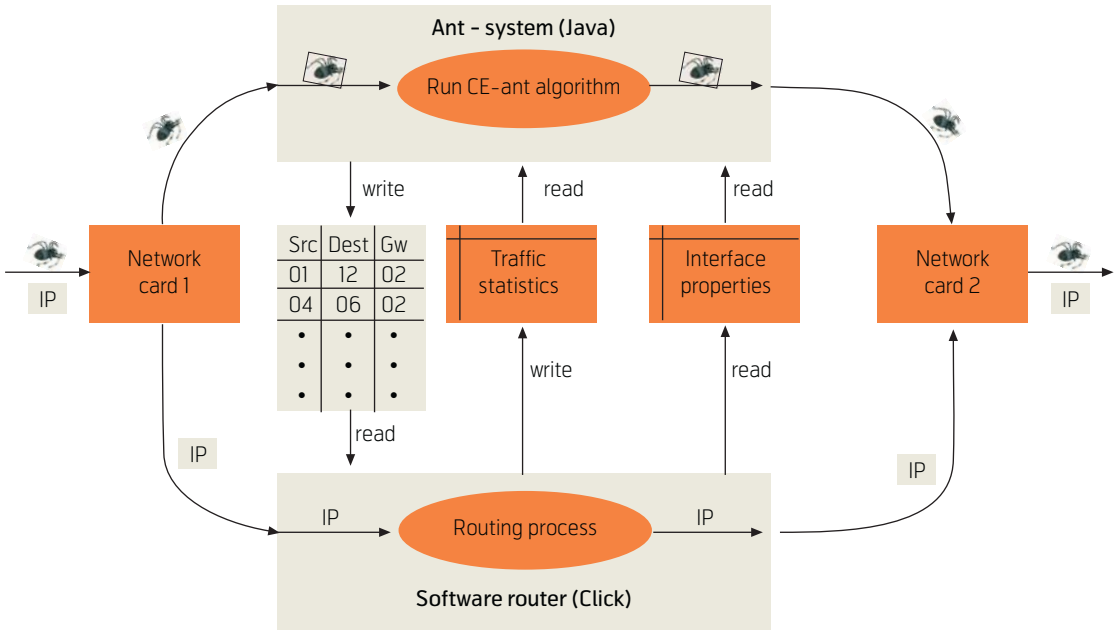


Figure 13 The main components of the hosts

ISSN 0085-7130 © Telenor ASA 2008

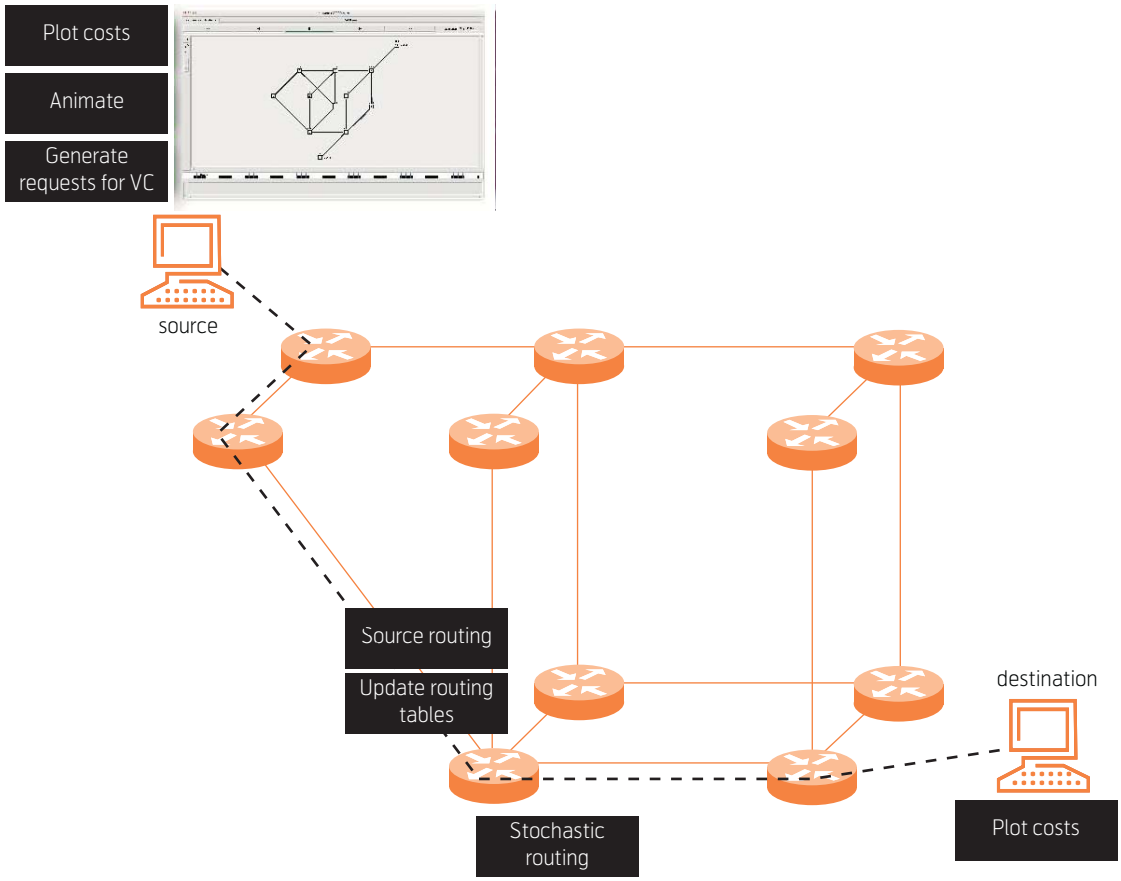


Figure 14 AntPing implementation

For the first version of this prototype the initial demo network topology is fixed and both sender and receiver processes are executed on external machines. In [62] the functionality of AntPing is extended to enable a self-configuring topology and to provide receiver processes in Click on every router. The latter implies that adding several destinations is possible without adding extra equipment.

Visualisation & interaction. The purpose of the demonstrator is to illustrate the inner workings of a swarm based method and to provide an interactive technical installation. The ant algorithm is animated live by use of Network Animator (*nam* [63]) showing how ants are moving and being dropped in the network, and how the topology is changing with link and node failures and restorations. The animation also shows ants that do not find the destination but are dropped because the Time-To-Live (TTL) is expired. Changes in cost values as a function over time of each virtual path are plotted live by use of *gnuplot* [64], both the cost of the current best path, and the cost of the last path found.

It is up to the users/audience to introduce network dynamics. They may unplug and re-plug cables between the nodes and/or the power supply to the nodes. Due to the extensions introduced in [62] new interfaces or links can also be added and several virtual connections can be established and monitored.

6 Closing Comments

Considering that finding paths between nodes is the basic and fundamental enabling functionality in a communication network, and that service handling in the future networks puts a wider range of requirements; an extension of the state-of-art in path management functionality is mandatory. Rather than pursuing “ad-hoc” improvements of current schemes or resorting to centralized management, we have addressed the problem by developing and applying the Cross Entropy Ant System (CEAS). This inherently robust, truly distributed and dynamically self-optimizing approach represents an important alternative paradigm for path management. The fact that CEAS has been applied to and is coping well with different relevant network management challenges, is a promising indication of its future.

There are, however, challenges ahead. We are confident that CEAS will operate in topologies typical for intradomain networks. Dealing with dynamic path management in the interdomain, as well as more comprehensive resource management tasks, requires additional functionality and improved insight into the

Figure 15 Setup from the lab (LinkSys WRT54GS (v4.0) routers)

rate of convergence and scalability issues. We intend to look further into these issues. With respect to the results presented in this paper, we will continue to look into how further improvements can be made to the system and merge our experience both with the simulated versions and the prototype implementations into an efficient design and implementation.

References

1 Ash, G R. *Dynamic Routing in Telecommunications Networks*. McGraw-Hill Professional, First ed., November 1, 1997. (ISBN 0-07-006414-8)

2 Huitema, C. *Routing in the Internet*. Prentice Hall PTR, 2 ed., November 1999.

3 Grover, W D. *Mesh-based Survivable Transport Networks : Options and Strategies for Optical, MPLS, SONET and ATM Networking*. Prentice Hall PTR, 2003.

4 Vasseur, J-P, Pickavet, M, Demeester, P. *Network Recovery : Protection and Restoration of Optical, SONET-SDH, IP, and MPLS*. The Morgan Kaufmann Series in Networking, Morgan Kaufmann, 2004.

5 Krunz, M, Matta, I (Eds.). Quality of service routing – Special issue. *IEEE Communications Magazine*, 40, Jun 2002.

6 Cholda, P, Mykkeltveit, A, Helvik, B, Wittner, O, Jajszczyk, A. A survey of resilience differentiation frameworks in communication networks. *Communications Surveys and Tutorials*, 9 (4), 2007.

7 Rubinstein, R Y. The Cross-Entropy Method for Combinatorial and Continuous Optimization. *Methodology and Computing in Applied Probability*, 127–190, 1999.

8 Ball, M O. *Handbooks in Operation Research and Management Science, Network Models*, 7. North Holland, 1995.

9 Pioro, M, Medhi, D. *Routing, Flow and Capacity Design in Communication and Computer Networks*. Morgan Kaufmann Publishers, March 2004. (ISBN 0125571895)

10 Kirkpatrick, S, Gelatt, C D, Vecchi, M P. Optimization by Simulated Annealing. *Science*, 220, 671–680, 1983.

11 Glover, F, Laguna, M. *Tabu Search*. Kluwer Academic, 1997.

12 Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, 1998.

13 ITU. *Types and characteristics of SDH network protection architectures*. Geneva, 1998. ITU-T G.841 (10/98).

14 ITU. *ATM protection switching*. Geneva, 1999. ITU-T I.630 (02/99).

15 Shen, G, Grover, W D. Survey and performance comparison of dynamic provisioning methods for optical shared backup path protection. *2nd International Conference on Broadband Networks*, 2005, 2, 1310–1319, 2005.

16 Jozsa, B G, Orincsay, D, Kern, A. Surviving multiple network failures using shared backup path protection. *Proceedings, Eighth IEEE International Symposium on Computers and Communication (ISCC 2003)*, 2, 1333–1340, 2003.

17 Zhou, L, Held, M, Sennhauser, U. Connection availability analysis of shared backup path-pro-

tected mesh networks. *Journal of Lightwave Technology*, 25, 1111–1119, 2007.

18 Mello, D A A, Pelegrini, J U, Ribeiro, R P, Schupke, D A, Waldman, H. Dynamic provisioning of shared-backup path protected connections with guaranteed availability requirements. *2nd International Conference on Broadband Networks 2005*, 2, 1320–1327, 2005.

19 Rai, S, Mukherjee, B, Deshpande, O. Ip resilience within an autonomous system: current approaches, challenges, and future directions. *IEEE Communications Magazine*, 43, 142–149, 2005.

20 Nelakuditi, S, Lee, S, Yu, Y, Zhang, Z-L, Chuah, C-N. Fast local rerouting for handling transient link failures. *IEEE/ACM Transactions on Networking*, 15, 359–372, 2007.

21 Menth, M, Martin, R, Hartmann, M, Spoerlein, U. *Efficiency of routing and resilience mechanisms*. Currently under submission, May 2007.

22 Schoonderwoerd, R, Holland, O, Bruten, J, Rothkrantz, L. Ant-based Load Balancing in Telecommunications Networks. *Adaptive Behavior*, 5 (2), 169–207, 1997.

23 Di Caro, G, Dorigo, M. AntNet: Distributed Stigmergetic Control for Communications Networks. *Journal of Artificial Intelligence Research*, 9, 317–365, 1998.

24 Wittner, O, Helvik, B E. Distributed soft policy enforcement by swarm intelligence; application to load sharing and protection. *Annals of Telecommunications*, 59, 10–24, 2004.

25 Wittner, O. *Emergent Behavior Based Implements for Distributed Network Management*. Trondheim, Norwegian University of Science and Technology (NTNU), Department of Telematics, November 2003. (PhD thesis)

26 Wedde, H F et al. Beeadhoc: an energy efficient routing algorithm for mobile ad hoc networks inspired by bee behavior. In: *Proceedings of the 2005 conference on Genetic and evolutionary computation (GECCO '05)*, New York, 2005, 153–160, ACM Press.

27 Di Caro, G, Gambardella, L M. AntHocNet: An Adaptive Nature-Inspired Algorithm for Routing in Mobile Ad Hoc Networks. *European Transactions on Telecommunications (ETT)* – Special Issue on Self Organization in Mobile Networking, 16 (5), 443–455, 2005.

28 Ducatelle, F, Gambardella, L M. Survivable routing in ip-over-wdm networks: An efficient and scalable local search algorithm. *Optical Switching and Networking*, 2 (2), 86–99, 2005.

29 Steels, L. Towards a theory of emergent functionality. In: Meyer, J-A, Wilson, S (eds.) *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior, SAB'90, Complex Adaptive Systems*, 451–461. Cambridge, MA, The MIT Press, 1990.

30 Damper, R I. Editorial for the special issue on 'emergent properties of complex systems': Emergence and levels of abstraction. *International Journal of Systems Science*, 31 (7), 811–818, 2000.

31 Engelbrecht, A P. *Fundamentals of computational swarm intelligence*. Wiley, 2005. (ISBN 13 978-0-470-09191-3)

32 Dorigo, M, Maniezzo, V, Colorni, A. The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, 26 (1), 29–41, 1996.

33 Heidelberger, P. Fast simulation of rare events in queueing and reliabili ty models. *ACM Transactions on modelling and Computer Simulation*, 5, 43–85, 1995.

34 De Boer, P-T, Kroese, D P, Mannor, S, Rubinstein, R Y. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134, 19–67, 2005.

35 Helvik, B E, Wittner, O. Using the Cross Entropy Method to Guide/Govern Mobile Agent's Path Finding in Networks. In: *Proceedings of 3rd International Workshop on Mobile Agents for Telecommunication Applications*, August 14–16, 2001, Springer Verlag.

36 Heegaard, P E, Wittner, O, Nicola, V F, Helvik, B E. Distributed asynchronous algorithm for cross-entropy-based combinatorial optimization. In: *Rare Event Simulation and Combinatorial Optimization (RESIM/COP 2004)*, Budapest, Hungary, September 7–8, 2004.

37 Stutzle, T, Hoos, H. MAX-MIN Ant System. *Journal of Future Generation Computer Systems*, 16, 889–914, 2000.

38 Kjeldsen, V. *Cooration through pheromone sharing in swarm routing*. Norwegian University of Science and Technology, 2007. (Master thesis)

39 Heegaard, P E, Wittner, O. Restoration performance vs. overhead in a swarm intelligence path management system. In: Dorigo, M, Gambardella, L M (eds.) *Proceedings of the Fifth International Workshop on Ant Colony Optimization and Swarm Intelligence (ANTS2006)*, LNCS, Brussels, Belgium, September 4–7, 2006, Springer.

40 Heegaard, P E, Wittner, O J. Self-tuned refresh rate in a swarm intelligence path management system. In: *Proceedings of the EuroNGI International Workshop on Self-Organizing Systems (IWSOS 2006)*, LNCS, University of Passau, Germany, September 18–20, 2006, Springer.

41 Heegaard, P E, Wittner, O, Helvik, B E. Self-managed virtual path management in dynamic networks. In: Babaoglu, O et al. (eds.) Self-* Properties in Complex Information Systems. *Lecture Notes in Computer Science*, LNCS 3460, 417–432, Springer, 2005. (ISSN 0302-9743)

42 Shannon, C E. A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656, July, October 1948.

43 Wittner, O, Helvik, B E. *Simulating mobile agent based network management using network simulator.* Poster in Fourth International Symposium on Mobile Agent System (ASA/MA 2000), September 2000.

44 Wittner, O, Heegaard, P E, Helvik, B E. Scalable distributed discovery of resource paths in telecommunication networks using cooperative ant-like agents. In: *Proceedings of Congress on Evolutionary Computation, CEC2003*, Canberra, Australia, IEEE, December 2003.

45 Kjeldsen, V, Wittner, O J, Heegaard, P E. Distributed and Scalable Path Management by a System of Cooperating Ants. In: *Proceedings of the International Conference on Communications in Computing (CIC 2008)*, Las Vegas, USA, July 14-17, 2008.

46 Kirkerud, B. *Object-oriented programming with SIMULA*. Addison Wesley, 1989.

47 Birtwistle, G. *Demos – a system for discrete event modelling on simula*. 1997.

48 Wittner, O, Helvik, B E. Cross Entropy Guided Ant-like Agents Finding Dependable Primary/ Backup Path Patterns in Networks. In: *Proceedings of Congress on Evolutionary Computation (CEC2002)*, Honolulu, Hawaii, IEEE, May 12–17, 2002.

49 Garey, M R, Johnson, D S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.

50 Stamatelakis, D, Grover, W. Rapid Span or Node Restoration in IP Networks using Virtual Protection Cycles. In: *Proceedings of 3rd Canadian Conferance on Broadband Research (CCBR'99)*, Ottawa, 7 November 1999.

51 Wittner, O, Helvik, B E, Nicola, V F. Internet failure protection using hamiltonian pcycles found by ant-like agents. *Journal of Network and System Management, Special issue on Self-Managing Systems and Networks*, 2005.

52 Heegaard, P E, Wittner, O, Helvik, B E. Self-managed virtual path management in dynamic networks. *Lecture Notes in Computer Science, Hot Topics*, 3460, 417–432, 2005.

53 Heegaard, P. *Performance monitoring of routing stability in dynamic networks*. Section 5 in Deliverable 05: Models for basic services in AHN, P2P and Grid networks in IST-FET Project BISON, December 2003. (IST-2001-38923)

54 Heegaard, P, Fuglem, I. *Demonstrator 1: Ant-based monitoring on software ip routers*. Tech. Rep., BISON, 2006. (IST-2001-38923)

55 Hesby, N, Heegaard, P E, Wittner, O. Robust connections in ip networks using primary and backup paths. In: *Proceedings of the 17th Nordic Teletraffic Seminar*, Fornebu, Norway, 25–27 August 2004.

56 Kohler, E et al. The click modular router. *ACM Transactions on Computer Systems*, 18, 263–297, 2000.

57 Struve, D. *Kaariboga mobile agents*. September 2003. http://www.projectory.de/kaariboga/, visited October 2003.

58 Mykkeltveit, A, Heegaard, P, Wittner, O. Realization of a distributed route management system on software routers. In: *Proceedings of Norsk Informatikkonferanse*, Stavanger, Norway, 29 Nov – 1 Dec 2004.

59 Mykkeltveit, A. *Realization of a distributed route management system by mobile agents*. Tech. Rep., Dept. of Telematics, NTNU, 2003.

60 Sanfilippo, S. *hping*. http://wiki.hping.org/ [online] – last checked 2007-09-05.

61 *Openwrt*. http://wiki.openwrt.org [online] – last checked 2007-09-05.

62 Kristiansen, T A. *Self-configured demonstrator of the Cross Entropy Ant System*. Norwegian University of Science and Technology, 2007. (Master thesis)

63 *Nam: Network animator*. http://www.isi.edu/nsnam/nam/ [online] – last checked 2007-09-05.

64 *Gnuplot*. http://www.gnuplot.info/ [online] – last checked 2007-09-05.

Poul E. Heegaard received his Siv.Ing. degree (M.S.E.E.) in 1989 from the Norwegian Institute of Technology (NTH), Trondheim, Norway and was awarded the degree Dr.Ing. (PhD) from NTH in 1998. Heegaard has since 2006 been Associate Professor at Department of Telematics, NTNU, where he is the coordinator of the Network Research area. In the academic year 2007/2008 he is a visiting professor at Duke University, Durham, NC, USA. Since 1999 he has been a Senior Research Scientist in Telenor R&I in Trondheim, where he currently holds a 20% position. He has previously been a Research Scientist and Senior Scientist at SINTEF Telecom and Informatics (1989–1999). His research interests are in the area of performance, dependability and survivability evaluation of communication systems. His special interests are rare event simulation techniques, and monitoring, routing and management in dynamic networks. His current research focus is on distributed, autonomous and adaptive management and routing in communication networks and services.

email: poul.heegaard@item.ntnu.no

Bjarne E. Helvik received his Siv.Ing. and Dr.Techn. degrees in 1975 and 1982, respectively. Since 1997 he has been Professor at the Norwegian University of Science and Technology (NTNU), Department of Telematics and is currently with the Norwegian Centre of Excellence (CoE) in Quantitative Quality of Service in Communication Systems (Q2S). He has previously held various positions with ELAB, SINTEF Telecom and Informatics, and as Adjunct Professor at NTH. Helvik's field of interests includes QoS, dependability modelling, measurements, analysis and simulation, fault-tolerant computing systems and survivable networks. His current research focus is on distributed, autonomous and adaptive fault-management in telecommunication systems, networks and services.

email: bjarne@q2s.ntnu.no

Otto J. Wittner received his Dr.Ing. degree from the Norwegian University of Science and Technology (NTNU) in 2003. He also joined the Norwegian Centre of Excellence (CoE) for Quantifiable Quality of Service in Communication systems (Q2S) in 2003 as a postdoc. His research interests focus on network control and management by emergent behavior principles, where he is looking into dependability issues, especially fault management.

email: wittner@q2s.ntnu.no

The Social Networks of Teens and Young Adults

RICH LING

Rich Ling is a sociologist in Telenor R&I

This article examines the way in which social networks operate within small groups. The study examines the social networks of teens and young adults. Groups of friends were recruited for the study and thus the unit of analysis is the group as opposed to the individual. The members of each group developed a contact diary (face to face, mobile voice, SMS, IM etc.) among the friendship group that also recorded their contacts with other "external" persons. The analysis shows the importance of copresence in the development of the groups and how mediated interaction can help to maintain and develop the group. The analysis also examines how different individuals played different roles within the groups.

Introduction

There is a link between the degree to which the members of a social network are in contact through various forms of media and the degree to which the group members are in physically copresent contact. It has been noted that there is indeed a type of stability in the interaction of social networks in spite of local disruptions associated with the interaction of particular dyads (Ebel, Mielsch and Bornholdt 2002; Kossinets and Watts 2006).

This suggests that the density of contact is a useful predictor of the propensity for adopting for example internet or mobile phone based social networking applications. The idea is that the more intense the web of interaction, the more the members have a common sense of the group. Given this common sense of one another, it is easy to suggest that social networking applications can be adopted, adjusted and used for the various purposes of the group.

The density of social interaction is not necessarily dependent on the density of only mediated interaction. Indeed, in many cases mediated interaction is only a small portion of the total interaction budget of a group. The interactions that are most essential for the development and elaboration of viable group interaction are co-located contact (Collins 2004; Ling forthcoming). Following this thought, teens often have a better milieu in which to develop these social networks since they are in daily contact with other same aged individuals both in the context of school and in their leisure time activities. The material here examines how the social networks of teen males and of young adult females function.

Method

The ambition of the work being reported here was to understand the internal network based interactions within teen peer groups. Thus, recruiting of randomly selected individuals was not an appropriate approach. Rather, we were interested in recruiting groups of peers and studying their interaction. To do this, a two-staged research process was developed. In broad strokes, we recruited several groups of friends, gathered information on their use of electronically mediated communication (as well as their face-to-face communication), and conducted a group interview with each group.

Recruitment

Often, social research is carried out in order to understand the way that individuals feel or think about various issues. The unit of research in this case is the individual. Indeed, when thinking of questionnaire based research or of focus groups (Kruegar and Casey 2000), it is the individual who is most often considered the unit of analysis.

Since this work examines the social networks of the teens and young adults – including their use of the so-called mediated communities (My Space, Facebook, etc) – it was not the individual who was the unit of analysis, but rather the group. This posed somewhat special problems both in terms of the recruitment and when conducting the focus groups (more on this second issue below). When recruiting persons for the study we first selected the general categories of groups to be included. These included younger teens (one group of males and another group of females) as well as young adults (again one of each gender).[1)]

1) *The young adult male group was actually somewhat younger than the female group, indeed they were still in high school. In many ways, they were more tightly tied into teen culture than into what might be considered the culture of young adulthood. For example, they were not regular wage earners nor did they live in their own apartments. Nonetheless, they approached the data collection and group interview in a serious manner and provided insightful information. In addition, two families were recruited. The results from this are reported on elsewhere.*

ISSN 0085-7130

Using these criteria, individuals in each of these categories were contacted and asked if they would be willing to participate.[2)] In addition to their personal participation, these "access point" individuals were also asked to help recruit four or five of those persons they considered to be their best friends. In each case, we were able to recruit a group of four to six individuals who were willing to be interviewed regarding their communication interactions with their friends and to participate in a focus group.

Both groups of younger teens were students in their respective junior high schools. Thus, they had daily contact in addition to several channels for mediated contact. The older males were high school students. They attended different high schools but they had all been neighbors during the earlier part of their education and they were all part of a skiing/snowboarding milieu. Indeed, this group was the very uniform in their choice of clothing style. Finally, the older female group consisted of four young adults. They had attended some school together in Kristiansand (about five hours drive from Oslo). They had then, at different times, moved to Oslo to either continue their studies or to work. Thus, beyond the social dimensions of the group, the main tie that bound them together was their common experience from Kristiansand.

Diary Data Collection

The focus of this study was to understand the way that teens and young adults use different forms of mediated interaction within the context of their peer groups. We were interested in gathering the stream of communications between the various group members over several days of interaction. It is this information that can be used to reconstitute the social networks of the groups and allow for the calculation of centrality in the context of the group.

The options for gathering the data included a retrospective questionnaire, a traditional paper based diary (Palen and Salzman 2002) or some type of automatic logging program (see for example, Raento et al. 2005). None of these approaches was practical. When thinking about retrospective questionnaires there is the problem of participant recall. It is possible to limit the period recalled to, for example, the previous day. However, when thinking about the ability to recall contacts that may have been somewhat minor and incidental, it can be difficult to summon up these details even when the time lag between the event and the questioning is moderately short (Freeman, Romney and Freeman 1989). A second approach is the use of paper diaries, i.e. a notebook or a sheet of paper carried by the participants to note their activities as they take place. Indeed, Hjorthol et al. (2007), Ling and Baron (2007) as well as Grinter and Eldridge have used the paper diary approach in their study of mobile communication (2001). These have the advantage of being more fine grained in their analysis. They allow the respondents to note their activities and behaviors as they happen. The disadvantage with the system is that it is intrusive and requires a large degree of commitment on the part of the respondents. Indeed the burden of having to fill out the diary can bias the number and type of respondents who choose to participate in a study. If the data entry is too cumbersome, the respondents might simply fill in their "best guess" at the end of the day or immediately before handing in their diary. Thus, there is the issue of recall as with the retrospective forms of data collection noted above.

A modification of paper based diaries is the Experience Sampling Model developed by Csikszentmihalyi and his colleagues (Csikszentmihalyi, Larson and Prescott 1977; Larson, Csikszentmihalyi and Graef 1980). In essence it employed mediated technology (in the case of the early work it was pagers) in order to elicit information of informants as they moved about their daily routines. In the early use of this method, a signal was sent to the informants that prompted them to fill out an entry in a diary at different points of the day. The approach was used to understand the daily flux of teens' activities and their mental state.

Palen and Salzman have employed an alternative form of electronic diary where they ask participants at various times to call into a voice mail system in order to gather the information. They found that this approach was less time intensive for the respondents since they did not have to stop their activity and manually record incidents. This was particularly important when the informant were mobile (Palen and Salzman 2002).

A final approach is to use some form of technology logging to capture the different communication situations where a particular type of mediation is used. For example, Diminescu et al. and his colleagues have used this in their investigation of geographic mobility and mobile communication (forthcoming). This provides the researcher with a rich and exact overview of the individual's use of a particular technology. In some cases, it can even provide details with regard who has been contacted and through which mode (voice telephony vs. SMS for example). A drawback with this approach is that it does not allow for multi-

2) *As an inducement, the individuals were paid for their participation with a gift certificate of NOK 500.*

modal forms of data collection. While it can provide information on, for example IM traffic, a completely separate apparatus would have to be used to gather co-present interaction or mobile telephony. Thus, a global form of data collection can become quite cumbersome.

A hybrid approach was used in this study. The individuals who were recruited were asked if they would be willing to receive three telephone calls every day from the data collection group, one at midday, a second in the late afternoon/early evening and a third later in the evening. The data collection period lasted for three days from Thursday thru Saturday to include interaction on the weekend.

In each call, the individual was asked to report on interaction with each of the other group members through various forms of mediation (face-to-face, mobile voice telephony, SMS, IM, etc.). In addition, they were asked about any other contacts they had had in the previous time span. This form of data collection when examined for the whole group resulted in a universal diary of interactions within the group. In addition, it provided an ego-based mapping of the individuals' contacts with those who were outside the group (Wasserman and Faust 1994, 53). The relatively short recall time combined with the mobile phone based collection of data eliminated some of the problems with other forms of data collection. Indeed, Hoppe et al. have found that telephone based "diary" studies result in better reporting and in cleaner data than did the more traditional paper diaries (2000). By way of critique, the three-day data collection period was somewhat short in terms of the total amount of data that was collected. In some cases, the interactions were rather sparse. However, issues of cost and issues of respondent willingness limited the time span of the data collection.

After the data collection period, the material was analyzed for broad trends, and the group was called in to a group interview. During the interview, they were asked about their interaction with one another and with other persons outside the immediate group. They were asked about the internal dynamics of their peer group and the types of activities and exchanges in which they participated. Given the form of recruiting, the focus groups had a particular dynamic. It is often the case in focus groups that the informants do not know one another. It is the job of the moderator to help the individuals feel comfortable speaking in front of others with whom they are not familiar. This is done by using a more extensive round of introductions and directing questions to individuals, etc. In the case examined here, there was a different situation. The informants were all familiar with one another and it became the job of the moderator to operate within the context of the groups' code of behavior. Where with normal groups of informants there are only the most basic forms of social contact between them, in this case there was a massive history shared by the informants. The moderator and those analyzing the material were left, however, to work out these internal group dynamics.

Analysis

The material from the series of "diary calls" was entered into a set of spreadsheets and summed. It was hoped that there would be enough material to allow examination of both the topography of copresent and mediated forms of interaction separately. The material indicated, however, that there was not a long enough time-period in order to accommodate a separate analysis and so the contact events were simply summed across all forms of interaction, both mediated and copresent.[3)] This material was then used to calculate the centrality of each individual within the group. In addition, the material on contacts with persons outside the group was tabulated and examined for overlap with the lists of persons mentioned by other group members. The material from the focus groups was transcribed and examined for themes that arose from the interaction.

Results

While there were four groups examined, it is perhaps most telling to contrast the situation of the younger teen males as opposed to the young adult females. In the former case, the group was in relatively tight contact. Their daily activities – school, sports, etc. – gave them the opportunity to interact on a regular basis. The young adult females were, on the other hand, a more diffuse group of individuals. The common thread that ran through the group was that they all had a history of spending at least a part of their youth in the town of Kristiansand and they found themselves collectively in Oslo.

Young Teen Males

The young teen male group was an active and relatively tight social network. The individuals in the group recognized that they were often together.

3) *It is clear that comparing a text message to a face-to-face chat is, to some degree, comparing apples to oranges. The text message is limited to 160 characters while the co-located interaction can take place over a longer period of time and can involve much more involved forms of interaction. However, when thinking about the calculation of centrality in a social network, it can be claimed, with only slight damage to the truth, that they are both expressions of the group's dynamics.*

Interviewer: I wonder about who is together with whom.

Andreas[4]*:* I am together with Håken, Andrew and a little with Erik sometimes.

Erik: We are all together, it is not like ...

Andrew: We play soccer on the same team together.

Andreas: Sometimes in the school band also.

Interviewer: [In the diaries] I saw that there were several common names, Oliver and ... But is it like you all know each other's friends?

Andreas: Yeah.

Interviewer: Can I ask in relation to groups. Often girls have a clique. Is it like that, that a well recognized gang hangs together.

Andreas: It is like, you know, but at school, then everybody hangs out with everybody. But in our free time it is more like that.

Håken: It is like that, that some are more together with each other.

Their comments indicate that while school was an important point in their daily interaction. It was in other milieus that they were able to work out their common identity. School was a venue for all the teens in their age group. It was however in the context of soccer and the school band that the small group had worked out their common identification.

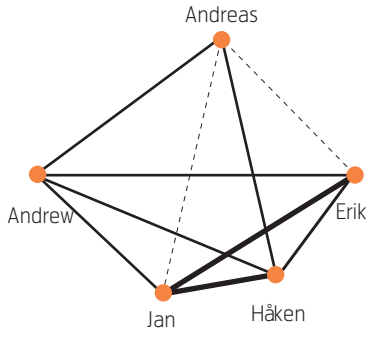


Figure 1 Social network of the teen males. The analysis shows that all the teen males reported interaction with all other members of the group. The most central members were Jan and Håken. Andreas was the least central (but the most verbal at the focus group)

The teens realized that there were some groupings that were more central than others.

The comments of the teens indicate that they have a somewhat imprecise idea of the contours of their social topography. There was the idea that some individuals were more directly linked to one another. However, the overall notions of the interactions were also mapped onto different situations (school time, soccer practice, free time, etc.). It is worth noting that the data collection period purposefully spanned these phases of their lives in order to capture the different aspects of their social interactions.

Looking at the material in the contact diaries, the five younger teen males had 167 contact events within the group and 218 contacts with persons outside the group during the three-day data collection period.

All the members of the group reported contact with all other members of the group during the data collection period. Some of the contact between the individuals was more tenuous than in other cases. In addition, some of the links were much more developed than other links. Nonetheless, all the members of the group reported being in contact with one another during the data collection period.

The most central person in the group was Jan (Figure 1). As can be seen in Figure 1, he had two very strong connections to Erik and to Håken in addition to a relatively strong connection to Andrew. Håken was the second most central.

The least central person in the group was Andreas. The material from the diary indicates that he had only limited contact with Jan (who was the most central) and with Erik. Andreas had the most contact with Andrew. Interestingly, Andreas was the "core" person with regard to the recruiting process; that is, he was the individual who helped to recruit the other group members. In addition, he was the most verbal person in the focus group. Of all the verbal turns taken by the five teens in the focus group, he took 37 %.

There were different modes of interaction, and the balance between them was also different. Among the younger teen males, copresent interaction dominated. The five individuals reported a mean of more than six face-to-face episodes with other members of the group during the data collection period (see Figure 2). This is as opposed to one voice based telephonic interaction and one using SMS. This was the approximate (percentage based) distribution pattern for the

4) *All names have been changed.*

teen female and the young adult male groups. Although the teen females had more contacts, their ratio of face-to-face vs. mediated contacts weighed heavily in the direction of the former.

Looking at both the internal interaction as well as the teen males' contact with persons who were outside the group the material reveals some interesting dimensions (Figure 3). First, one external individual was in contact with all the other group members during the data collection period but was not in the focus group nor were they a part of the diary study. Interestingly Andreas, who was the person who helped with recruiting the other group members, reported relatively heavy contact with this individual. Thus while in many respects this external individual could have been considered a member of the group, for one reason or another, he or she was not included.

Second, when considering Jan and Håken, the two most central individuals in the group, they also reported the fewest number of external contacts. This indicates that while they are core members of the group, they have, in effect, put all their eggs in one

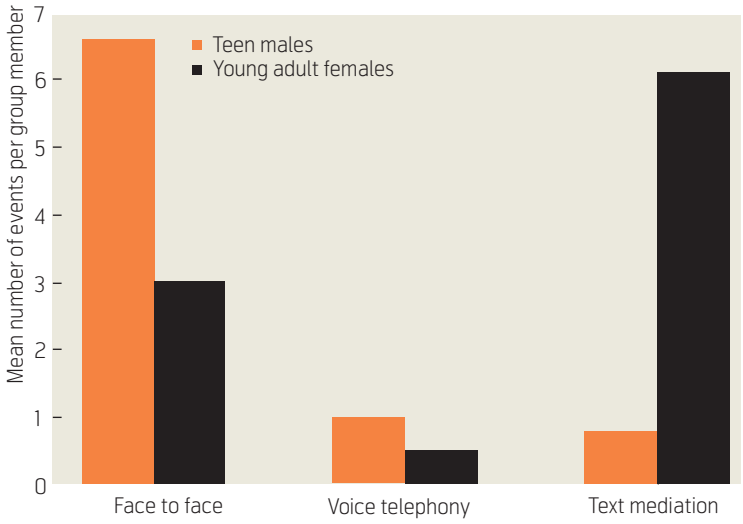


Figure 2 Mean number of contacts via different forms of interaction. The teen males have extensive co-present interaction at school and also during their free time when they play soccer together. This interaction is the main form of promulgation for the group. Texting and voice telephony is seen as a way to support the co-present interaction. With the young adult women the main form of interaction is texting, indeed this in spite of two of the women living together. Texting, it seems, is a convenient way for the group to maintain a type of lightweight contact

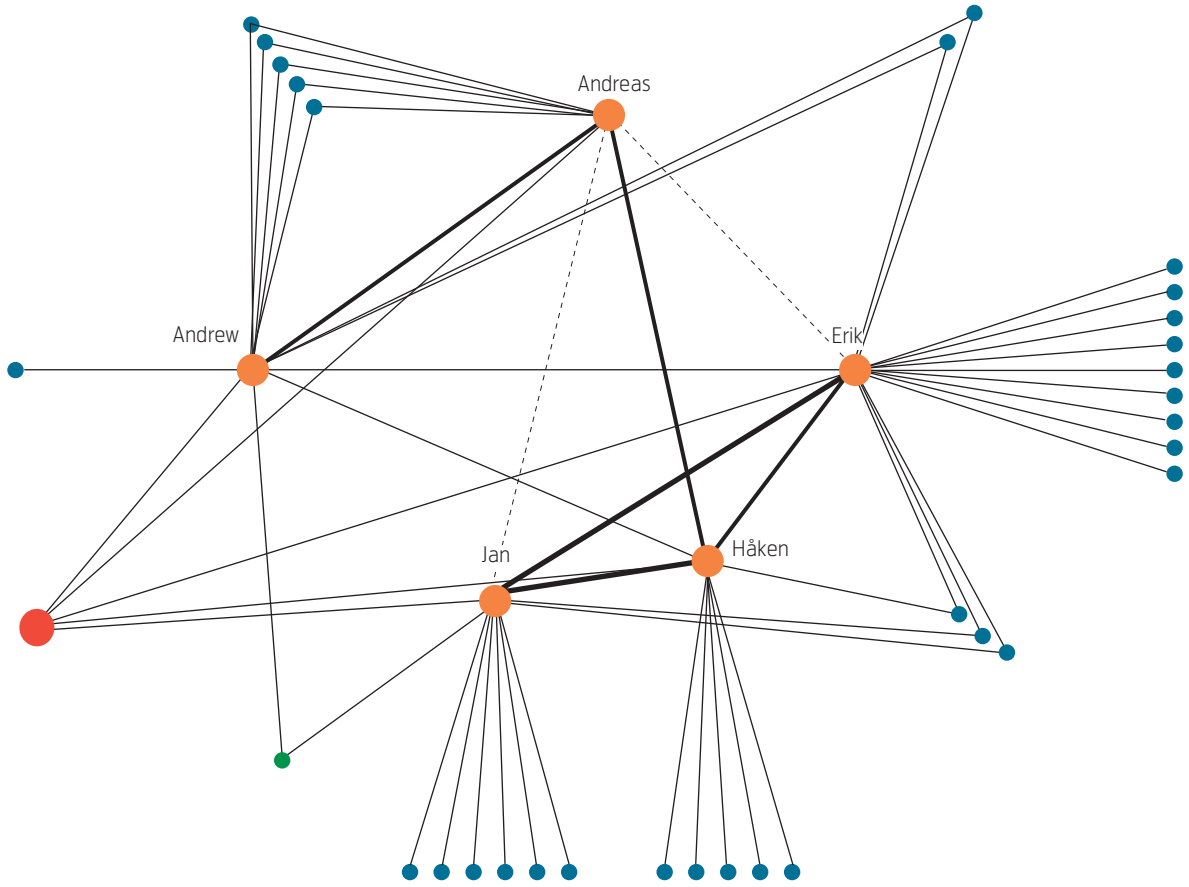


Figure 3 Social network of the teen males with their external ties. The material shows that there is a universally known "external" individual, some groups who circle around shared members (in particular Andrew and Andreas) and some individuals who have relatively few external ties (Jan). The situation of Jan and Håken is interesting since they are the most central members of the group and interestingly, they have a very small number of external ties. This indicates that the group is quite central to them. Alternatively Erik has a large number of external ties as well as a moderately central position in the group itself

ISSN 0085-7130 © Telenor ASA 2008

social basket. The group is central to them and they have committed themselves to it to a greater degree than several of the other individuals.

A third feature of this analysis shows that Andrew and Andreas reported a series of common friends who were not shared by the other members of the group. Andreas and Andrew had been classmates in elementary school and had continued to be classmates as they moved to junior high school. This group of common friends were also elementary classmates with whom they retained contact. The other group members had attended other elementary schools and so these individuals were not as closely tied to these individuals. Interestingly, neither Andrew nor Andreas had a significant number of their own exclusive external links (Andrew had one and Andreas had none). The material collected during the study seems to indicate that their social lives were bound either to the group being studied here, or to the group they had maintained from elementary school.

Finally, Erik reported the largest number of external ties. In addition to his participation in the group examined here, he also was active in another soccer group. Because of this he had a separate social sphere that was not particularly tied to the group of teens examined here.

The teen males had contacted 33 persons who were outside of the group during the data collection period. Of the external names, 36 % were "common" in the sense that more than one of the group had been in contact with the same person. In the case of one person, all five of the teen males had been in contact with them. The remaining 64 % of the names were unique in that only one of the teen males had been in contact with them.

While face-to-face interaction was the most common form of contact, the teens also used other mediation systems to help them keep in contact with others. So-called community services such as Facebook were popular at the time of the interviews.

The interview material indicated that photos were one way that the group interacted, but they were a part of the broader stream of interaction within the cohort of teens. Facebook had arisen as an application through which some of this interaction took place.

Interviewer: What do you use Facebook for?

Andreas: Write on each other's "walls".

Erik: Post pictures and wall-to-wall conversations, it is almost like MSN.

Interviewer: Do you send photos using your mobile?

Håken: Some take pictures and post them, but nobody sends them [with their phones using MMS].

Erik: I have only gotten 2 – 3 MMS since Christmas (note: the focus group was in June).

Interviewer: So you only send it to Facebook. Is it more practical?

Andreas: Yeah, it is free and that is why nobody sends [MMS] anymore.

Somewhat later in the interview, Erik noted, "I use Facebook mostly to see photos and to post photos." When compared with other forms of more direct interaction, this form of mediation seems quite indirect. However, it plays an important role in the way that the teens document their activities (be it parties, school activities, older photos or vacation memories). In addition, they used the functionality of the community application to point out particular individuals in the photos (to frame them). This generates a message to the individual being identified and indicates to them that they have been included in the photos of another person. Thus, the photo functionality, in addition to being an archive of activities, is also a type of gifting system.

The teen males were more likely to have contact with same sexed individuals. The material shows that on average each of the teen males had contact with about 4.5 males and slightly more than 1 female during the data collection period (see Figure 7).

The material provides insight into the functioning of the group and the different roles assumed by the teens. Jan and Håken were the two who were the core of the group. All the other members of the group referred to them and used them as a type of central clearinghouse. Andreas was the most verbal of the teens and was, in many ways the clown. He perhaps provided the levity when the group assembled itself. Finally Erik was what Burt might have called a network entrepreneur (2001). He had many exclusive contacts to another group. Through these contacts he could bring different types of influences and ideas back into the group where they could be acted upon or adopted.

The material here indicates that the intensity of the teens' interaction was, to a large degree, the result of the common parallels in their lives. The group was age sorted and indeed some of the members (for

example Andrew and Andreas) had been school chums since they were six years old. The age sorting function of the educational system in addition to their common interest in soccer and the propinquity of their homes meant that they were in almost daily contact and that they were able to cultivate a set of common interests and often had a common perspective on issues.

Interestingly, the material also shows that beyond these commonalities, there were also distinct roles within the group. There was the jovial verbosity of Andreas, the more sober anchoring of Jan and Håken and the linking of Erik. The members of the group seemingly each had their role and the interaction showed that they were more or less comfortable with their own positions and those of the other individuals.

Young Adult Women

The teen males had a daily location where they could meet (school) in addition to regular extracurricular activities. This regular and quasi-institutionalized form of interaction was visible in their contact diaries. In the case of the young adult females there was some common structure in their interactions, but it was not as strongly anchored in their daily lives.

Interviewer: Who is together with whom?

Astrid: It is a little mixed. Iris and I are together a lot, and I meet Vilde and Helene.

Iris: We, me and Astrid, live near one another and Vilde and Helene live together.

Interviewer: You all have separate networks? A lot of common friends?

Iris: Some.

Astrid: We had a lot of common friends from ... we were together a lot in Kristiansand. Now it is a little more like ...

Iris: We know a lot of the same people.

Interviewer: High school?

Iris: We got to know each other in high school, all of us.

Vilde: We knew about each other in junior high school.

Iris: Yeah, we went to the same junior high school, but we didn't know each other then.

Interviewer: Are there others?

Iris: But if there was something big that happened, it would be natural that we four would be a part of it.

The young adult females were in a different life situation. They had been school mates approximately ten years previous to the focus group. At the time of the focus group, each of them was pursuing somewhat different paths in life. Two of the group members (Helene and Vilde) shared an apartment while two others lived near one another. All of them use SMS and all, save one, studied at different locations in Oslo.

As with the teen boys, the establishment of the group was based in their experience as younger teens at junior high school.

Interviewer: Where do you all come from?

Astrid: From Kristiansand

Interviewer: Everyone?

Iris: Actually I am from here [Oslo], but I lived there. So we know each other from there. We didn't move [to Oslo from Kristiansand] at the same time, about a year in between, but almost at the same time.

The young adult females were not classmates or working colleagues. Three of the women were students (one at the University of Oslo, one at a junior college in Oslo and a third at a vocational school). The fourth woman worked. Because of this, there was not the same routinized structure nor was there a common milieu in which they could meet as with the teen males. Further, these women were in a life phase where they had left their family of orientation and had not yet entered into an eventual family of procreation that would demand its own time and attention. Thus, the group filled a certain social space in each of their lives. It had been more central when they were teens together in school and now it provided them with social contact and other logistical advantages (the common apartment and perhaps occasional transport back and forth between Oslo and Kristiansand). We can speculate however, that it was a social sphere that would – in anticipation of the completion of their education and the establishment of their own families – have to tolerate further adjustments.

Unlike the teen males, not all of the group members had been in contact with one another during the data collection period. Thus, the diary material shows that the social network for the young adult women was not completely linked. Of the six possible ties

between the four women, they only reported on five of them being used during the data collection period. Two of the group members did not interact during the data collection period and further, three of the six possible dyads had had only weak interaction. Perhaps most surprisingly, Helene and Vilde, the two who shared an apartment, reported only weak levels of interaction.

This group was unique among the four that we studied in that it was the only one that relied more on mediated than on face-to-face interaction (see Figure 2). Where the teen males reported a mean of more than six face to face interactions as opposed to about 2.5 mediated interactions,[5] the young adult women reported sending and receiving a median of six text messages to other group members during the data collection period. They reported a mean of three co-present interactions and somewhat less than one contact via voice telephony. The copresent interaction was clearly driven by the fact that Vilde and Helene lived together. Had they lived in separate apartments the co-present interaction would have been lower.

The network maps of the group give a very different picture when compared to the teen males. Where the teen males had a fully configured group (all individuals had interacted with all other members of the group), this was not the case with the young adult females.

The data also shows that it was Iris who was in many respects the key individual in the group. She was the most central when looking at the material from the interaction diaries. She was the "core" person who was used in recruiting the other individuals to the group and she is the one who made the most utterances during the focus group.

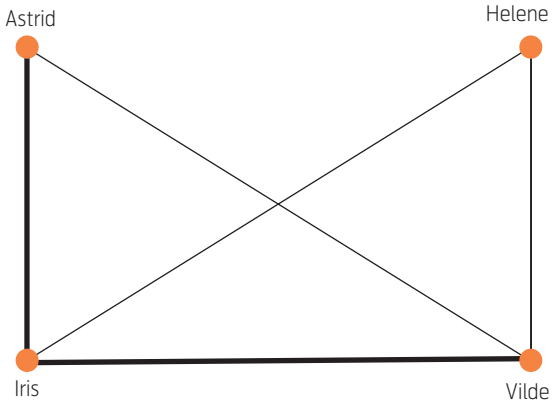


Figure 4 Social network of the young adult females. Iris is the most central member of the group with Astrid and Vilde being about equally central. Helene is the least central member and indeed has only moderate ties to Vilde (her room mate) and to Iris. There was no contact reported during the data collection period between her and Astrid

Her role as the leader was also recognized within the group:

> *Vilde:* We are maybe a core [group], with Iris as the leader. (Laughter)
>
> *Interviewer:* You will define Iris as the leader?
>
> *Vilde:* Because she says she is. (Laughter)
>
> *Interviewer:* Why do you define Iris as the leader?
>
> *Iris:* Me and Vilde have known each other the longest and then Astrid and Helene knew us from school and we were the ones that introduced them.
>
> *Interviewer:* Why is Iris the leader? Is she good at keeping appointments?
>
> *Iris:* Now there is a lot of silliness. However, I am good at keeping track of appointments.

The leadership of Iris is seen in the diaries, in the conversation of the focus groups and it is also recognized by other members of the group. With the teen boys, who were in a more integrated group, the different members had more explicit roles that they filled. The sense of this group however, was that while it is still vital, its role in the lives of the members is changing. Thus, rather than having a daily forum for of interaction, it is the individual members who must work to maintain the group. Iris is the most energetic in this context. The material from the interaction diaries shows that Astrid and Vilde were moderately central and that Helene was the least central.

The young adult females had contacted 46 persons outside the group. This is about twice as many as the teen males. There was a difference, however, in terms of the number of common and unique names. The data shows that only 7 % of the individuals were named by more than one of the four women in the group. None of the individuals had been contacted by all four of the women. The remaining 93 % of the external contacts had only been contacted by a single member of the group. The young adult females had a broader range of alternative friendship ties. It is perhaps an indication that since they did not have the

5) *This included a 1.5 mean of interactions via voice telephony and a mean of less than one text message from or to other group members.*

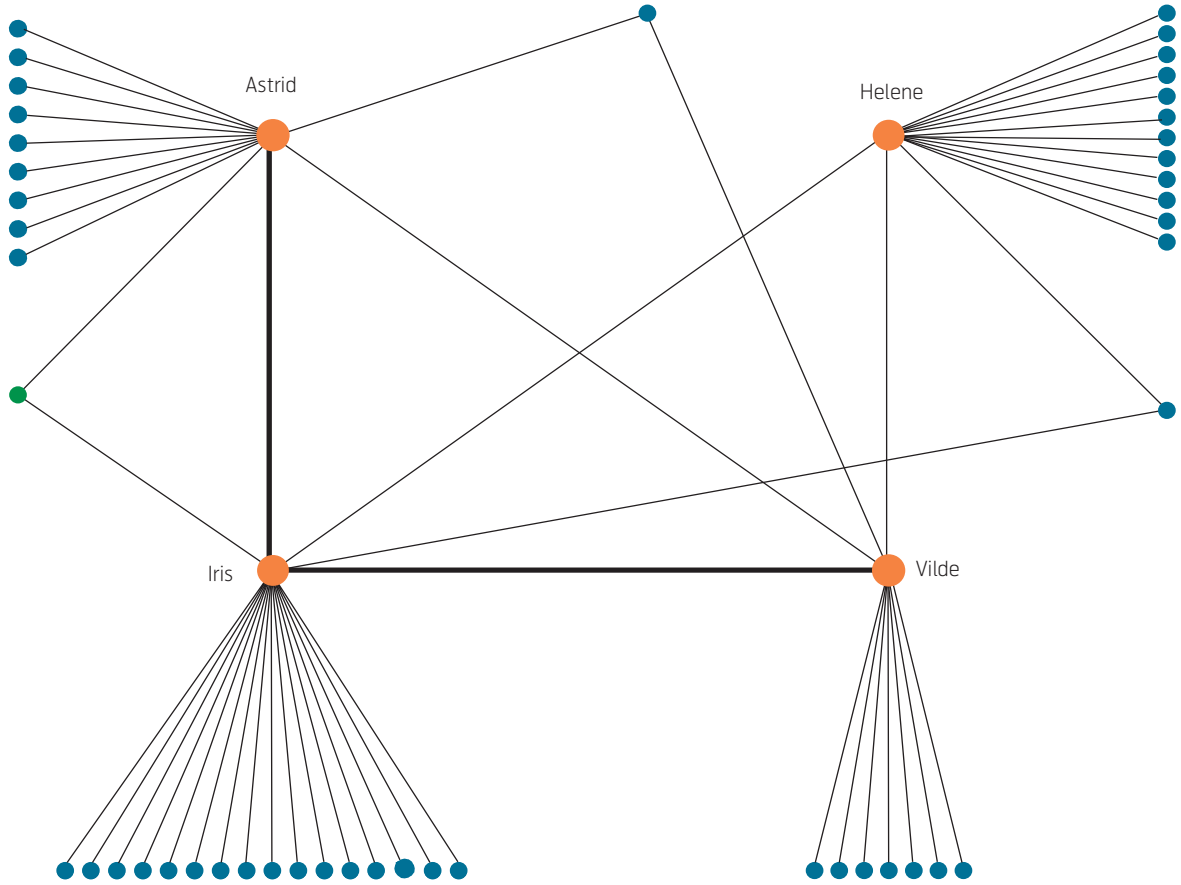


Figure 5 Social network of the young adult females with their external ties. The material shows that Iris reported the most extensive number of unique external ties. She was also quite central to the group. Helene was the least central group member but she reported the second largest number of external ties. This might indicate that she is "spreading her bets" in terms of social involvement. If her membership in the group falters, she will have other alternatives available

institutional support of a common school or a common leisure time milieu with which to support the group (as did the teen males) they needed to work harder at maintaining the group and that the existence of the group was perhaps more uncertain.

On the individual level, Iris also had the largest number of external ties (a total of 16). Thus, we see a situation where a single individual seemingly dominates the group in terms of centrality, links to other individuals and in the contributions to the conversation. A review of the focus group transcript does not show her to be overbearing. She did not, for example cut off other individuals when they were speaking or challenge the assertions that they made in any direct way. However, it is clear from this material that, as Vilde noted, Iris is the leader.

Interestingly it is also possible to see that Helene has a precarious position in the group. She was the least central person when examining the internal interaction within the group. She had only moderate ties to Iris and to Vilde and no reported interaction with Astrid during the period covered by the data collection. In addition, Helene reported the second largest

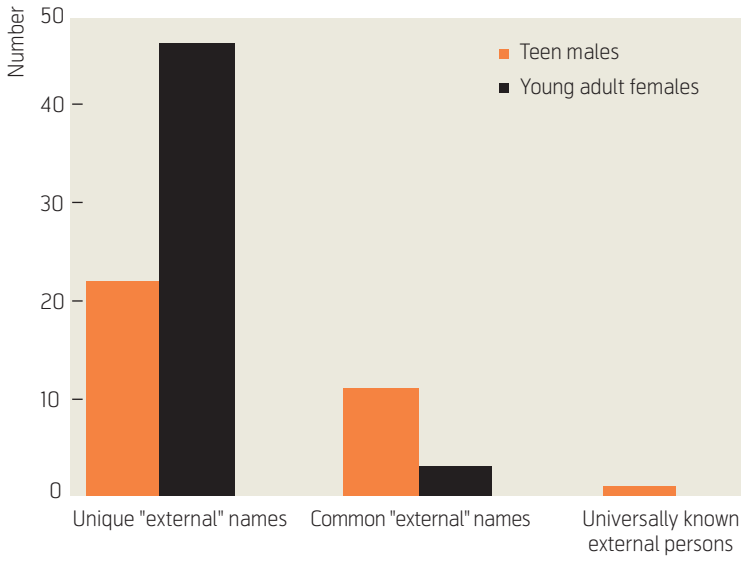


Figure 6 Total number of contact events with unique external individuals, external individuals who were in contact with more than one group member and external individuals who were in contact with all the group members. The material shows that the young adult females each had their own friends who were not in contact with other members of their group. By contrast, the teen males had more common friends who were outside the immediate group. (Note: the larger number of "unique" names generated by the women was in spite of their being only four group members as opposed to the five teen males)

ISSN 0085-7130 © Telenor ASA 2008

number of external persons with whom she had had contact during the data collection period (a total of 12). The material seems to indicate that the elements holding Helene in the group were the fact that she shared an apartment with Vilde and that Iris seemed interested in holding contact with her. Were these elements not in place, the centripetal elements of limited contact with the group and for example the pursuit of a more academic degree (a university degree as opposed to junior college or vocational school) along with the influences of her university peers could result in her focusing more on other groups.

It is perhaps not surprising that the young adult women had greater contact across the gender line than did the teen males. The material shows that on average each of the young adult females had been in contact with six females and slightly more than four males during the data collection period (see Figure 7). Where the teen males were only starting to explore romance, sexuality and relationships with women, the young adult females had had the opportunity to work through some of the main issues in this area. They were, for example, familiar with the dynamics of dating and the use of SMS and IM in the process of interacting with members of the opposite sex.

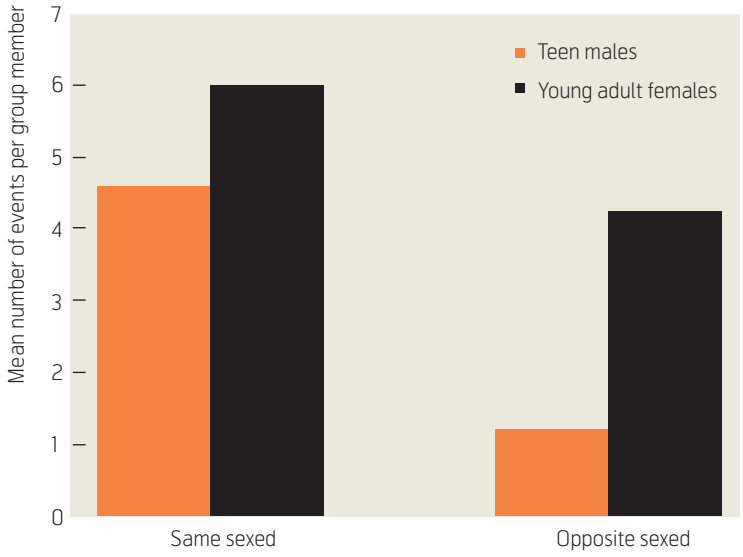


Figure 7 Mean number of contact events per group member with opposite sex individuals. The young adult females and the teen boys were each in contact with about the same number of same sexed individuals. The big difference between the two groups was in the case of contact with opposite sexed individuals. In this case the young adult females reported more contact with persons of the opposite sex. This is a reflection of their different position in the life cycle. While the teens were still unsure about the role of interaction with female peers, the young adult females were far more involved in that type of socializing

The image that one takes away from the analysis of the young adult females is that of a group that maintains itself based on its past associations (the experience of junior high and high school), on their common current situation (leaving Kristiansand for Oslo) and by having a dynamic person (Iris) who is willing to draw them into various social activities. This means that they maintain a certain sense of being a group. The teen males had an easier time of this in that they could devote much more time to common activities, both within the context of school and in their free time.

The common notion of the group developed by the young adult women can be seen, for example in their use of argot (or rather their ethics of which argot to use), which showed how the women used language in the definition of the group. The form of interaction was also important. The informants reported that they were not likely to use different formulations that might identify them as teens.

Interviewer: Do you use "koz"? [the respelling of the Norwegian word for "hug" with the substitution of z for the final s, a style that had been popular among teen users (see Ling forthcoming)]

All: No! Absolutely not!

Interviewer: Is that wrong?

Iris: We did it maybe before, but that was a long time ago.

Astrid: I have never written "koz" with z. I don't write "kos" so often. (Laughter)

Iris: No.

Astrid: "Klem" [another form of the word for hug] you can write that.

Iris: Sometimes. (Laugh)

This passage indicates that these women had adopted a carefully calculated style of writing and the range of endearments. The informants in this group were not as likely to rely on what they saw as immature and childish forms of interaction. They saw themselves as being more urbane and sophisticated. Thus, they eschewed formulations and argot that would potentially give the reader the wrong impression.

In the same way as argot can be used to help identify the group, so can the use of photos. They can be used to commemorate certain group events (Ling 2008) and they can also be seen as a type of gifting between

the individuals (Johnsen 2000; Taylor and Harper 2001).

Interviewer: Do you use your telephone to take pictures?

Astrid: Yeah, there is a little of that every once in a while.

Iris: It is usually when you are going out. I have sent some lately. I take more when I am out than normally, I do that.

Interviewer: Tell me. What do you do. What is typical?

Jeanne: Pictures, that is mostly like Iris says. If you are going to go out.

Iris: A fun picture, or an ugly picture of someone. Or a picture that someone wants.

Interviewer: What do you do with the pictures?

Iris: Send them to the person [who has been photographed]

Interviewer: A lot of jokes or funny ones? Or are they more serious?

Iris: Astrid and I have sent a lot.

As noted above, photos taken by the group members have both the function of being a type of informal archive of group activities and a way of giving an individual a type of special attention. Interestingly we also see again that Iris who is so central for the group in other contexts is also central here. She is recognized as the person who photographs different events and she is the one who sends them to others.

Conclusion

General Results

The results indicate that it is the strength of the co-present interaction that, in many ways determines the strength of the global social network. The younger teens, who met one another daily at school, during football practice and in their local neighborhoods were more strongly bound than the older women who had a relatively thin thread holding them together in the form of Iris. Where the younger teens were, in effect, thrown into a common environment, the older women had had some of the same background, but were in the process of establishing new identities and each developing their own sphere of friends.

The tightness of the teen male group, as opposed to the young adult females, can be seen in their contact with external individuals. Where the females had a larger number of contacts, there were far fewer that were common. In more than nine cases out of ten, the contacts were individual interactions and not contacts with common friends. This was only the case with about six in ten for the teen males. The network of external friends was more compact, but it was also more shared for the teen males.

The young adult females had more cross gender contact than did the teen males. Where about a third of the external contacts were males in the case of the young adult women, only 18 % were females in the case of the teen males. This is likely an indication of the unsure role that women had in the lives of the teen males as opposed to the more mature forms of interaction among the young adult women. Interestingly, the number of external persons who were family members was the same across the two groups.

Returning to the introductory question, the material here indicates that co-present contact is an essential element in the development and maintenance of social groups. This can be supported by mediated interaction, but it is through copresence that we are able to promulgate the social ties that are later cultivated via mediated interaction. In the process of developing social ties, the participation in common milieus is a common experience upon which the group members can develop their sense of their linked identity. This was obvious in the material here in that the teen males (school chums and members of the same soccer league) displayed a flourishing sense of group identity. In a similar way, the young adult women were still drawing on their sense of group identity that had its foundations in their common experience as teens. In this latter case, it is possible to see that the network was becoming somewhat frayed with time. In addition, as the women enter into a more established phase of life, the group ties will be further tested. Nonetheless, the fact that they have been able to maintain their common sense of a group bears witness to the strength of their earlier common experiences.

Bibliography

Burt, R S. 2001. The social capital of structural holes. In: Guillien, M F, Collins, R, England, P, Meyer, M (eds). *New directions in economic sociology*. New York, Russell Sage.

Collins, R. 2004. *Interaction ritual chains*. Princeton, Princeton University Press.

Csikszentmihalyi, M, Larson, R, Prescott, S. 1977. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, 6, 281-294.

Diminescu, D, Licoppe, C, Smoreda, Z et al. (Forthcoming) *Tailing untethered mobile users: A joint study of urban mobilities and communication practices by combining ethnography and cell-based mobile phone-supported localization journals*. (Mobile communications research series, 1)

Ebel, H, Mielsch, L-I, Bornholdt, S. 2002. Scale-free topology of e-mail networks. *Physical Review E*, 66, 035103.

Freeman, L C, Romney, A K, Freeman, S C. 1989. Cognative structure and information accuracy. *American anthropologist*, 89, 310-325.

Grinter, R, Eldridge, M. 2001. y do tngrs luv 2 txt msg? In: Prinz, W et al. (eds). *Proceedings of the seventh European conference on computer supported cooperative work ECSCW '01*, 219-238. Dordech, Netherlands, Kluwer.

Hjorthol, R, Jakobsen, M H, Ling, R et al. 2007. Det mobile hverdagsliv: Kommunikasjon og koordinering i moderne barnefamilier. In: Lüders, M, Prøitz, L, Rasmussen, T (eds). *Personlige medier. Livet mellom skjermene*. Oslo, Gyldendal Akademisk.

Hoppe, M J, Gillmore, M R, Valadez, D L et al. 2000. The Relative Costs and Benefits of Telephone Interviews Versus Self-Administered Diaries for Daily Data Collection. *Eval Rev*, 24, 102-116.

Johnsen, T E. 2000. Ring meg! En studie av ungdom og mobiltelefoni. In: *Department of ethnology*. Oslo, University of Oslo.

Kossinets, G, Watts, D J. 2006. Empirical Analysis of an Evolving Social Network. *Science*, 311, 88-90.

Kruegar, R A, Casey, M A. 2000. *Focus groups: A practical guide for applied research*. Thousand Oaks, Sage.

Larson, R, Csikszentmihalyi, M, Graef, R. 1980. Mood variability and psychosocial adjustment of adolescents. *Journal of Youth and Adolescence*, 9, 469-490.

Ling, R. 2008. *New Tech, New Ties: How mobile communication is reshaping social cohesion*. Cambridge, MIT Press.

Ling, R, Baron, N. 2007. The Mechanics of Text Messaging and Instant Messaging Among American College Students. *Journal of sociolinguistics*, 26.

Palen, L, Salzman, M. 2002. Voice-mail diary studies for naturalistic data capture under mobile conditions. In: *CSCW'02*, New Orleans, Louisiana.

Raento, M, Oulasvirta, A, Petit, R et al. 2005. ContextPhone: a prototyping platform for context-aware mobile applications. *Pervasive Computing, IEEE*, 4, 51-59.

Taylor, A, Harper, R. 2001. Talking 'Activity': Young people and mobile phones. In: Palen, L (ed). *CHI 2001 Workshop: Mobile communication: Understanding user, adoption and design*. Seattle, WA.

Wasserman, S, Faust, K. 1994. *Social network analysis: Methods and applications*. Cambridge, Cambridge University Press.

Rich Ling is a sociologist in Telenor R&I. He received his PhD in sociology from the University of Colorado, Boulder in his native US. Upon completion of his doctorate, he taught at the University of Wyoming in Laramie before coming to Norway on a Marshall Foundation grant. Since that time he has worked at the Resource Study Group and has been a partner in the consulting firm Ressurskonsult, which focused on studies of energy, technology and society. For the past 13 years he has worked at Telenor R&I and has been active in researching issues associated with new information communication technology and society with a particular focus on mobile telephony. He has led projects in Norway and participated in projects at the European level.
Ling has published numerous articles, held posts at and lectured at universities in Europe and the US and has participated in academic conferences in Europe, Asia, Australia and the US. He has been responsible for organizing scholarly meetings and editing both academic journals and proceedings from academic conferences. He has received recognition as an outstanding scholar from Rutgers University and Telenor, and his analysis has appeared in Norwegian newspapers.

email: richard-seyler.ling@telenor.com

Forwarding Messages in Mobile Social Networks: An Exploratory Study

SEBASTIAN SCHNORF

Sebastian Schnorf is a PhD student at University of Zurich, Switzerland

The telecom industry is investing a lot of resources into traditional advertising despite the fact that many users often turn to their social environment for advice. Social relationships are one of the oldest media which are nowadays "empowered" by communication services. This makes it possible to take a look at users' social context and gain insights into the referral behaviour of customers. In this exploratory study some results are presented which allow a new perspective on users and highlight the possibilities of social network analysis. Finally, it is stated that this knowledge is applicable for a new type of marketing and enhances the development of new services.

1 Introduction

Of course, it is not as simple as stated above but still there is less doubt that marketers are pushing the limits when it comes to shaping a shiny image, lowering churn, stimulating traffic or promoting new services. From the user perspective, it is very hard to decide whether a particular pricing plan is suitable or if a specific technical feature is really necessary, particularly if you are part of the majority of the customer segment which is not really into it. Some users are increasingly annoyed and react with denial despite having general interest in "useful" services.

User needs, attitudes and behaviour are not only shaped by individual characteristics, but also by their social environments. The referral behaviour is of particular interest because it can be a form of promoting products and services to others. On a more abstract and general level this process can be seen as the passing of information about innovations within a social network. Through social network analysis a clearer view on this referral behaviour is possible.

2 Social Networks and Passing Along Information

Social networks consist of relationships (ties) between individual actors. They build up a structure which defers from other networks with a more random tie distribution. For instance, people tend to reciprocate relations as a basic rule in human communication. Furthermore, actors with strong ties [1] – intense personal relationships – tend to build relationships with common actors as a result of their mutual exchange. In that way communication efforts are also being bundled, if you look for a formal explanation. Areas with clustered actors are effective for local transmission processes and, if seen as communities, they can provide social stability. Communities develop norms that are crucial for forming attitudes towards innovation [2].

2.1 Passing Along Messages

The spreading of information can be observed in daily online communication, as some people tend to forward e-mails with funny content to other users. Although the advertising industry tries to harness this by promoting "viral content", little is known about the actual underlying communication processes.

One study [3] examines the passing along of e-mails as an episodic process. These researchers examine the individual perception of such messages, as well as motivations to forward them. To a large degree, their findings implicitly point towards relational characteristics. In other words, users try to "stay in contact" with more distant persons or send specific content to persons with a similar mindset. The estimation of such aspects depends very much on previous social interactions. Therefore, besides looking at the message content itself, looking at communication ties is essential as this might give an explanation for the structural diffusion of information.

Forwarding content via the mobile phone has been observed within the realm of political mobilisation. An overview of descriptive case studies [4] shows that some mobile messages can reach a significant amount of people within a short period of time. This resulted in thousands of people gathering in public places to demonstrate against a political situation. In one case it is mentioned that an operator measured a significant increase in mobile communication traffic. Furthermore, the studies indicate that the forwarding behaviour not only occurs in times of "breaking news" but also happens with niche content within sociocultural milieus.

2.2 Qualitative Illustrations

The following example should illustrate that forwarding messages is also part of everyday mobile communication. It is taken from a 2005 qualitative study on MMS usage scenarios in Switzerland. In the very middle of the following chart (Figure 1) you can see

 ISSN 0085-7130

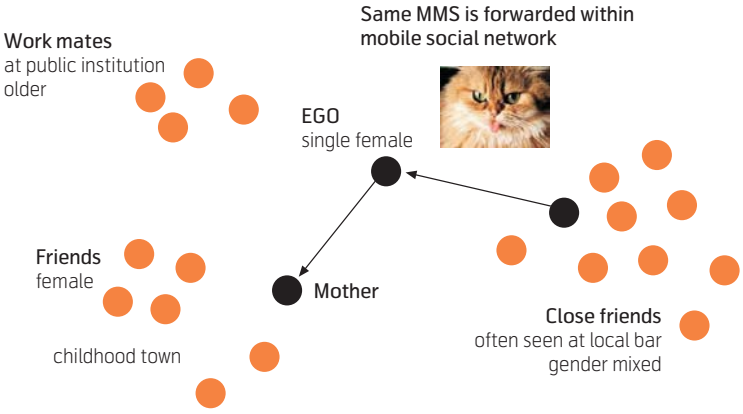


Figure 1 Sample sketch of a personal social network

what is called ego in social network analysis. It is a 30 year old single woman living on the outskirts of a Swiss city. During interviews persons wrote down names based on "name generating" questions such as who do you talk to about important issues. In the next step these name tags were positioned around "ego", so that closer persons indicate an intense exchange. Furthermore, names were grouped by the amount of contact these actors have with each other. You can recognise in our example that actors are grouped in different activities (see Figure 1). After this task the sketch was used as a reference during the interviews to reduce cognitive effort.

A starting point to the discussion was the pictures that participants had on their phone. A young woman described the following scenario: "I was hanging out with a friend at the local pub. For some reason we went through the pictures we had on our phones as is the case with picture books. Since I like animals my friend forwarded to me a neat picture of a cat, which I passed on to my mother as she also likes these kinds of pictures." This description provides rich contextual information about a particular forwarding scenario. The behaviour seems to rely on a very subjective salience of the content having the power to bridge a gap between local communities, as her mother lives in another town (see Figure 1).

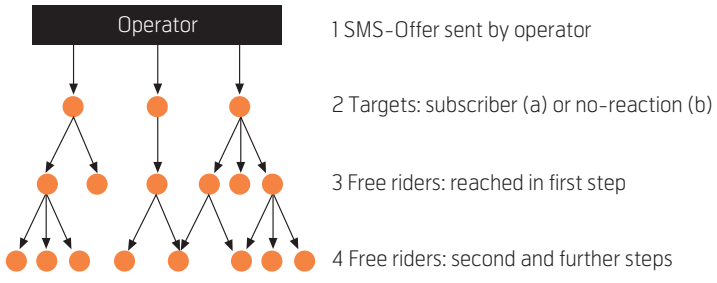


Figure 2 Assumed communication process

One could say that the previously described forwarding is only a marginal phenomenon. In another study, based on a four day private communication diary of about 60 people, there is evidence that this occurs more often than anticipated, as several participants reported such usage scenarios. According to a rough estimation about 1% of all private mobile communication can be categorized as mobile forwarding. At first sight, this might be a small figure, though from a revenue point of view it should not be underestimated. It is also worth mentioning that forwarding generates even more communication as people usually start follow-up communication mainly with peripheral members of their social network, as was mentioned earlier with the reciprocity mechanism in social networks.

3 The Free Rider Phenomenon in Campaigns

Forwarding behaviour has been assumed in campaign activities of a mobile operator. It has been noticed several times that a considerable amount of customers had not been directly targeted by an SMS-campaign but have participated in an exclusive promotional offer. In the following discussion these users outside the target group will be called "free riders".

3.1 Research Setting

Since the offer was very poorly communicated by other channels it was assumed that the target audience had forwarded the particular SMS-offer to others. Due to regulations it was impossible to exclude free riders from participating in the offer. These customers lowered campaign profitability. But from a research point of view the phenomenon provides an opportunity to study the tie structure of this diffusion network. The situation very much reminds us of a classical media influence model of a Two- or Multi-Step-Flow in communication [5], where certain members of society act as hubs or multipliers of certain messages (see Figure 2).

The aim of this exploratory study was to find a way to measure the diffusion process and provide more evidence for the forwarding assumption. The approach was to consider all direct communication that has occurred between involved actors. The direct forwarding itself could not be measured but was inferred from the communication ties.

The data was provided by a telecom operator and analysed with respect to privacy regulations. Due to regulations from the company absolute values can only be provided partially. The data consisted of the aggregated monthly mobile communication from a period after which a chosen campaign was conducted.

The data had to be carefully prepared so that connections represent as much social characteristics as possible and not artefacts such as those originating from reversed billing of content services. This connection data has been joined with demographic characteristics of actors, if available. The network consisted of several thousand nodes. This dimension of the network makes it difficult to apply regular methods of social network analysis.

For the case study a campaign was chosen aimed at all low/mid SMS users. The promotional offer was based on last month's billing amount and a small opt-in fee to send unlimited SMS for the coming month. The direct response rate was 2.1 %, which was relatively low compared to other campaigns. In contrast, the free rider population was significantly larger, consisting of mainly female actors, as it has already been noticed in other campaigns.

3.2 Indications for a Network Diffusion Process

In the following, single analyses are being described which by no means should be considered complete. Rather, it is the objective to show the possibilities and difficulties regarding social network analysis based on communication data. To start with, the results refer to the actors' function as message multipliers. Subsequently, the denser knit realm within the network is examined and finally the overall structure of the network is illustrated.

Direct ties between the target group (Figure 2: 2) and the free riders (Figure 2: 3) can be regarded as a first indicator for transmission of message multipliers. The calculations show that 18.1 % of the target group's participants (Figure 2: 2a) have direct ties to the free riders' group. Out of the large number of non-reacting people (Figure 2: 2b) in the target group, 7.3 % of the actors are in direct contact with a free rider. Considering absolute numbers non-reacting people still have more free rider contacts.

Due to these results the forwarding assumption is regarded as being supported in that there are direct ties between the target group and the free riders.

However, with regard to the interpretation of a possible spreading process via these ties certain aspects need to be considered, above all economies of scale. The more extensive a group of actors is within a network, the bigger is the likelihood of direct ties. Due to its size the large number of non-reacting people within the target group must therefore have more direct ties with free riders than the smaller group of participants. Due to the low reaction rate of the target group it can be assumed that the SMS offer is being ignored as unrequested advertising. On the other hand, the few participants in the campaign may perceive the offer positively and forward it for exactly that reason. The larger part of actors with direct ties to the free riders in the small group of participants could be an indicator of that. Of course, from a formal point of view, both one single and an indirect tie between the target group and the free rider can be sufficient for an overall spreading process.

Actors as Message Multipliers

Below the network function is addressed in which some actors forward their information to several other actors and therefore might promote the spreading as hubs (see Figure 2). Based on the data at hand this analysis clarifies if such behaviour is possible in the observed network.

The analysis shows that from the target group 87.1 % of the actors in touch with free riders have contact to a single follower and that almost 13 % of the actors have contact to more than one follower (Chart 1). If only the ties between the free riders are considered (Figure 2: 3+4) more than 44 % of them have more than one follower (not in the chart). Two participants have a maximum number of 13 followers. That reinforces the assumption that some actors can forward their message to several other actors and therefore act as multipliers.

Indication of Tie Preferences

According to a homophily mechanism, ties preferably exist between actors with similar characteristics, and it can be assumed that forwarding occurs the same way. As mentioned before, in most of the cases female actors are involved in the free rider phe-

Messages	Number of followers						
Multipliers	1	2	3	4	5	6<	Total
Target group	**87.1 %**	11.0 %	1.6 %	0.3 %	0.1 %	0.0 %	100.0 %
Free riders (1st to 2nd step)	79.7 %	16.3 %	3.3 %	0.6 %	0.1 %	0.0 %	100.0 %
Free riders (2nd to 3rd step)	81.7 %	15.4 %	2.6 %	0.0 %	0.0 %	0.4 %	100.0 %

Chart 1 Actors with ties to one or more followers

ISSN 0085-7130 © Telenor ASA 2008

Sender	Receiver		Total
	Female	Male	
Female	74.4 %	24.7 %	100.0 %
Male	56.6 %	41.9 %	100.0 %

Chart 2 Crosstabulation of communication ties according to gender

nomenon. This is particularly the case in the campaign examined. Whereas the majority of the participants are men (see Figure 2: 2a) the free riders consist of a majority of women (Figure 2: 3+4). This might be due to the fact that a large proportion of men were approached directly via SMS and that there is only a small number of male actors to be found outside of the target group. For a statistically significant statement, it is therefore vital to assess the relative under-, and over-representation. Calculations thereto show that women are over-represented by more than 20 % in the free rider population. For a spreading between genders explanations are sought in the formation of ties between the actors. According to the tendency towards homophily it can be assumed that there might be more ties between women than men. This analysis is applied to outbound ties of actors from the target group with direct contact to free riders (Figure 2: 2-3).

The results (see Chart 2) make clear that women as a group have significantly more communication ties with women than with men. On the other hand, men show a relatively balanced value with regard to communication between genders at the particular interface examined.

Analogue analyses with regard to age categories confirm this tendency to a "homogenous communication" (ref. Chart 3). Teens communicate mostly with teens, young adults with young adults, etc. An exception to this rule represented in the present analysis is older people; they have more contact to the younger age segments (see Chart 3).

Sender	Receiver				Total
	>19 years	20-30 years	31-54 years	<55 years	
>19 years	69.8 %	20.9 %	8.4 %	0.9 %	100.0 %
20-30 years	14.6 %	61.2 %	22.0 %	2.2 %	100.0 %
31-54 years	15.3 %	25.8 %	54.7 %	4.3 %	100.0 %
<55 years	12.2 %	30.9 %	47.7 %	**9.2 %**	100.0 %

Chart 3 Crosstabulation of communication ties according to age

Taken as a whole, the assumption is confirmed in that there is an increased contact between actors of similar socio-demographic characteristics. With regard to the spreading process, this serves as an explanation for the constitution of the free rider population. In a way, the analysis confirms some sort of a stereotype: women communicate more often with women than with men.

Tendencies for Local Clustering

Actors with a strong social tie tend to crosslink themselves. Such structures indicate local transmission processes in networks and explain, amongst other things, how this can lead to quick local spreading of information. The hypothesis postulates that so-called closed triads, three actors being completely interlinked, can be found in the network examined.

In the triad census all the possible tie formations between three actors are counted. In the network of the free riders a few closed triads can be found. For comparison the frequency expectation values in a random network with the same amount of actors and ties are being calculated. These estimations are at zero for the found triad formations. That means that there are social forces in the examined network which indicate a locally strong interlinked communication flow. This statement is confirmed by statistical comparison but does not offer additional insight. It is somehow apparent that structural characteristics in random networks differ from social communication networks particularly if they are comprehensive. Still, the present analyses show that there is a tendency for local clustering in certain areas of the examined network. The question arises whether further structural patterns can be identified.

Processing of extensive network data is feasible with the computer program Pajek [6]. It offers suitable algorithms for appropriate data reduction. The "Core" method reduces the number of ties in an iterative process in order to identify strongly networked areas within the network. These elements can be extracted. The procedure being done with free rider link data results in a small amount of cliques besides the

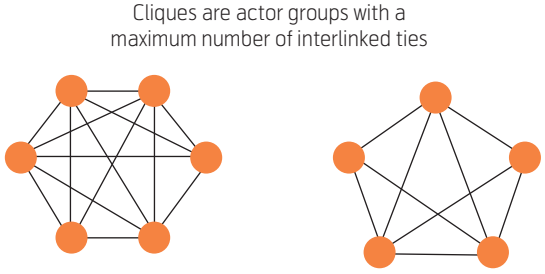


Figure 3 Identified cliques in the examined network

ISSN 0085-7130 © Telenor ASA 2008

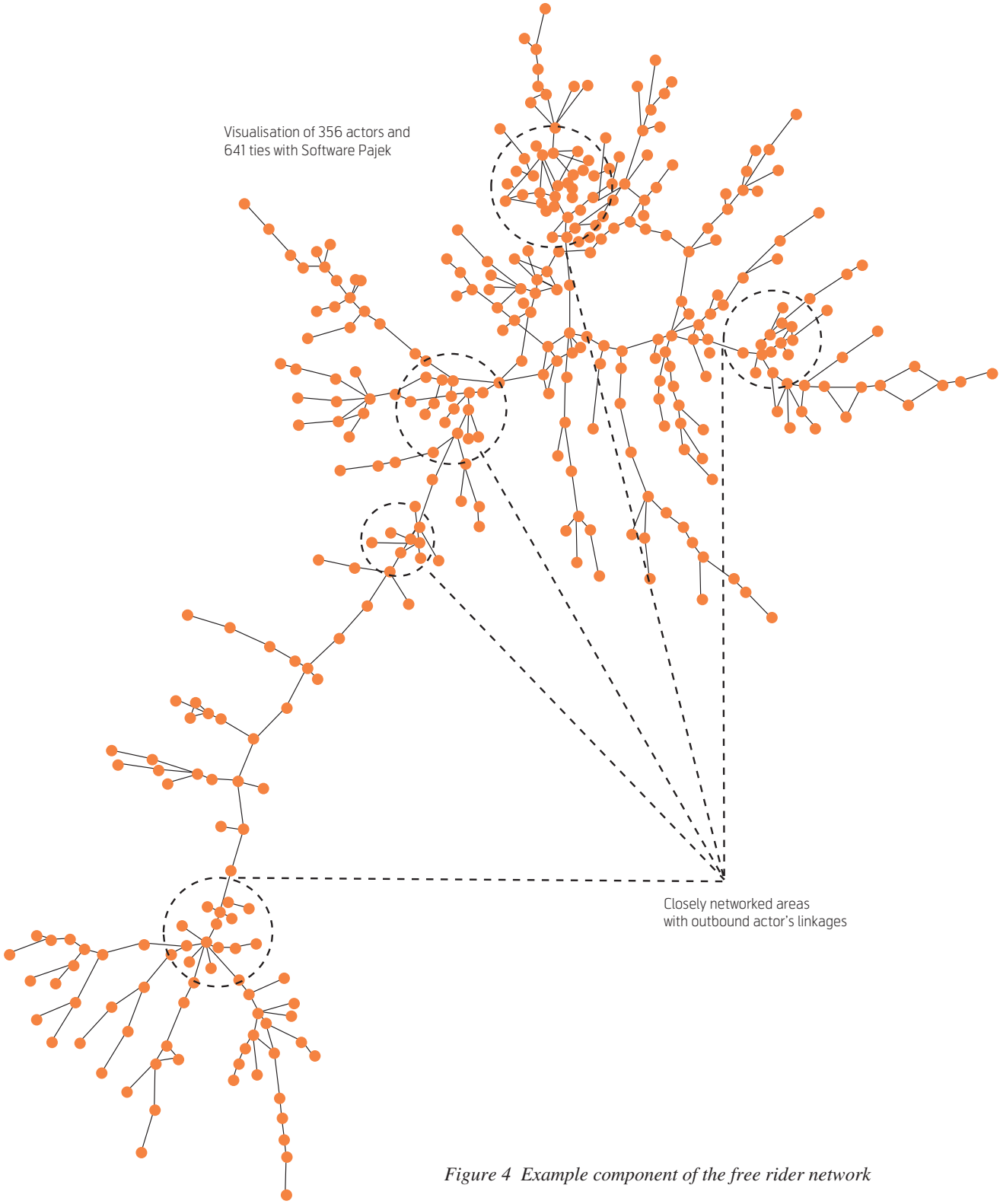


Figure 4 Example component of the free rider network

already mentioned triads. Cliques are actor groups with a maximum number of interlinked ties (Figure 3).

These cliques should not be overestimated as they are only rarely observed within the whole network. Nevertheless, the results testify closely networked areas which can be interpreted as communities. This particularly applies because the ties between the actors are based on the joint characteristics of the campaign participation and the assumed sharing of information. The sense of community is an important issue for a sustainable acceptance of innovation in general. This has been known for a long time but social analysis of communication data offers new opportunities to monitor structural properties of these processes.

Visualisation of Network Components

Components are stand-alone elements without ties between each other within a network. The forwarding

of messages may be reduced to those kinds of "islands". It can be assumed that the free rider network is not completely linked up but rather consists of several components with different sizes. The following analysis is limited to the ties between free riders.

The analysed network consists of components as was expected. But more than half of them consist only of isolated actors or one tie between two actors. 8.8 % of the free riders are in components with three to five members, 18.3 % in larger components of 6 to 356 actors. One component could be identified consisting of 1,518 free riders. The results clearly show that an overall snowballing effect, with a message being sent out by one actor to all the others, can be seen as a rather unlikely scenario.

The value of network visualisations is sometimes criticised in research for not offering added value compared to numerical figures. The advantage of visualisations lies essentially in the possibility to make complex structures more evident. Visualisations definitely provide an insight when dealing with comprehensive datasets. However, the illustrations must be reduced regarding the number of actors. An overview is hardly possible with visualisations of more than 1000 actors. Therefore, in the previous figure a single component of the entire network has been visualised. Some of the preceding triad constellations can be seen.

The visualisations of other components from the free rider network show similar structural patterns. In overlapping areas of actors' linkages paths there appear stronger linked actor communities (ref. Figure 4). This resembles less a typical social network than a social transport network similar to an underground system that provides transport from areas of high population density via single linkages to the periphery. According to other research [7] it is assumed that the spreading process has begun at the periphery of networked areas moving towards the centre and then causing a fit in the diffusion process.

4 Conclusion

This article is based on the proposition that the social environment of users must be given great importance. In order to be able to influence the needs and behaviour of users, not only individual psychological variables are important but also the social environment. This applies particularly if innovations are services that enable people to communicate. Therefore it is important to look at the use of communication services from the perspective of social network analysis. Before presenting some implications of this exploratory study about message diffusion the findings are briefly summarised. These results are not conclusive and are mainly intended to point out the possibilities of analysing the aspects of social relationships in communication.

The findings presented from qualitative case studies confirm that the forwarding of messages is also taking place via mobile phones and can be put down to tie characteristics. The starting point for the present case study is the assumption that actors from a campaign target group forward a promotional offer to actors outside of the target group by SMS and that they profit from the offer (free riders). From a communication research perspective this phenomenon offers the unique possibility to analyse "mobile word-of-mouth".

All in all, the findings support the assumption that messages are forwarded via the actors' social network. Most results can be interpreted as the network having an accelerating function for diffusion. This becomes apparent in the manner that some actors can act as multipliers in the network and that ties are predominantly set between actors with similar socio-demographic characteristics. The structuring of the network shows single triads and cliques which can be interpreted as communities. From these densely networked areas outbound actors' linkages can be identified in visualisations which point to a comprehensive spreading process.

The executed analysis illustrates that with SMS campaigns, a deliberate control of diffusion is likely to be difficult since the spreading can be caused by a rather large amount of actors within the target group. There is hardly a snowballing effect where only a few key actors infiltrate all free riders. It is also possible that free riders receive their information through one or more neighbouring actors. Furthermore, large numbers of independent groups can be found as single components within the network. The accelerating effect of networks should therefore rather be used in a positive manner and strenghthened. Taking these considerations as a starting point, one should concentrate less on nodal customer segmentation, as in traditional marketing, and more on network based interventions [8]. The basic idea is to consider actors' social environment and to take into account their natural informational and persuasive function as hubs within communities, instead of putting resources into conventional media campaigns.

Another implication of social network analysis concerns the development of new services. Web2.0-applications are elementarily based on connecting to other nodes. Lots of current online "networking platforms" are significantly determined by technical

possibilities so that some actors establish contact with as many actors as possible. There is of course a reason for this behaviour because every link acts as potential resources, but to enhance Web2.0-services crucial social mechanisms could be taken more into consideration and transferred into the online realm. For instance, this could be done on the tie level of networks by an increased support of reciprocity experience in the form of suitable presence information of remote actors. Another approach could be to make a community status of certain users more apparent based on their intense triadic communication between each other.

References

1 Granovetter, M. The Strength of Weak Ties. *American Journal of Sociology*, 78, 1360-1380, 1973.

2 Rogers, E. *Diffusion of Innovation*, Fifth Edition. New York, 2003, 26.

3 Phelps, J et al. Viral Marketing or Electronic Word-of-Mouth Advertising : Examining Consumer Responses and Motivations to Pass Along Email. *Journal of Advertising Research*, 4, 333-348, 2004.

4 Castells, M et al. *Mobile Communication and Society. A Global Perspective*. Massachusetts, MIT Press, 2007, 185.

5 Katz, E, Lazarsfeld, P. *Personal Influence*. New Jersey, Transaction Publishers, 2005 (first published 1955).

6 Nooy de, W, Mrvar, A, Batagelj, V. *Exploratory Social Network Analysis with Pajek*. Cambridge, Cambridge University Press, 2005.

7 Canright, G, Engø-Monsen, K. (2005): Epidemic spreading over networks : A view from neighbourhoods. *Telektronikk*, 101 (1), 65-85, 2005.

8 Valente, T. Network Models and Methods for Studying the Diffusion of Innovations. In: Carrington, J, Scott, J, Wasserman, S (eds.). *Models and Methods in Social Network Analysis*. Cambridge, Cambridge University Press, 2005, 111-113.

Sebastian Schnorf is a user researcher with a focus on communication services and social networks. He has conducted several projects in the telecommunication industry. Sebastian holds a Bachelor degree in the field of interaction design from the University of the Arts in Basel, Switzerland. He also has a Master degree in social sciences from the University of Zurich, where he is about to finish his PhD in Communication.

email: sebastian.schnorf@access.uzh.ch

Collaboration Patterns in Distributed Work Groups: A Cognitive Network Approach

TOM E. JULSRUD

Tom E. Julsrud is a research scientist in Telenor R&I

A promising and relatively new approach for studying distributed work groups is the application of social network techniques and theories (Ahuja and Carley 1999; Sparrowe, Liden et al. 2001; Cummings and Cross 2003; Hinds and McGrath 2006). A social network approach to distributed groups draws attention towards the social relations between the nodes in a network of distributed employees. Growth of new data coming out of on tele- and computer based traffic offers new opportunities to conduct such studies, and to analyse structures of collaboration patterns in distributed groups. Yet, the emerging field of network studies calls for a closer understanding of the relationship between patterns of mediated interaction and other more subjective relations, like trust and friendship. Based on a cognitive network approach (Krackhardt 1987; Corman and Scott 1994) this paper argues that observable interaction may be seen as indicative representations of a close relation, but not as confirmation. In particular in work settings, a high interaction level should not be seen as confirmative for close relationship, since much interaction is related to formal tasks. Interaction based ties should therefore be handled with care, and supplemented with other relational network indicators in network studies of distributed groups. The arguments are substantiated with results from a recent study of a group of distributed workers in an ICT company.

1 Introduction

While earlier studies of ICT-based work primarily focused on individual teleworkers (Jackson and der-Wilen 1998), recent studies tend to embrace the larger group of distributed and co-located employees. As such, work in the field of *distributed work* groups and virtual teams has accumulated rapidly the last decade (DeSanctis and Monge 1999; Lipnack and Stamps 2000; Hinds and Kiesler 2002; Duarte and Snyder 2006). This trend draws the study of telework closer to research fields concerned with teams and collaboration in computer-based environments[1)].

A promising and relatively new approach (in this field) is the application of *social network techniques and theories* (Ahuja and Carley 1999; Sparrowe, Liden, Waynes and Kraimer 2001; Cummings and Cross 2003; Hinds and McGrath 2006). Social network analysis is a theoretical and methodological paradigm that studies the patterning of relations among social actors, as well as the patterning of relationships among actors at different levels (Wellman 1988; Wasserman and Faust 1994; Scott 2000; Breiger 2004). A social network approach to distributed groups, then, draws attention to the social relations between nodes in a network of distributed employees or to the relations surrounding a single individual among distributed workers.

Growth of new user friendly software and access to new data coming out of on tele- and computer based traffic offers however new opportunities to conduct such studies and to analyze structures of collaboration patterns in distributed groups. This new wave of network mapping tools offers representations of distributed groups as dynamic communication networks. Yet, the question of how to interpret communication based networks is still often left unanswered. I will argue here that the field calls for a more elaborated understanding of the relationship between mediated interaction and other more affective relations, like trust, friendship, and identification.

This paper will illuminate some of these new opportunities and challenges represented by this development. It is suggested that a cognitive network approach (Krackhardt 1987; Corman and Scott 1994) may be a useful point of departure to entangle the diverse networks involved within a group. A Case study of distributed workers is used as illustrative evidence to the general argument proposed here; that different media may be related to different types of relations, and that intensive communication ties do not necessarily indicate affective closeness. The central objective of this paper is then 1) to suggest a theoretical framework useful to analyze patterns of mediated interaction as well as affective ties in distributed groups and; 2) present preliminary findings

1) *I will use the term* distributed work groups *in this paper, to denote groups that work together across time and space supported by the use of various ICT. Although several of the works cited in this paper use other terms like "virtual teams" and "computer supported collaborative work groups", I will stick to this term.*

ISSN 0085-7130 © Telenor ASA 2008

based on a case study of distributed workers utilizing this framework.

2 Distributed Group as Structural Networks

Social networks analysis is a strong approach to analysis of distributed collaboration. Based on a structural understanding of groups, it draws attention to the particular ties and bonds that hold a group of collaborators together. Groups are seen as a dynamic network of relations represented by constellations of direct and indirect ties (Scott 2000; Breiger 2004).[2] It can on the one hand be perceived as a methodological approach where the structural relations of a group is the basic area of interest, opening for a rich set of techniques for analysis of data (Wasserman and Faust 1994). However, in a wider perspective it may also be seen as a particular analytical approach founded on structural theories in social sciences (Wellman 1988).

Approaching distributed groups as networks of relations offers significant benefits: Firstly, it helps to locate, visualize and understand *roles* employees have within social networks. For instance, a network analysis of communication patterns can locate individuals that are disconnected from the information flow, or individuals that are particularly central in a dispersed structure. Secondly, the social network approach is useful to detect *cohesive sub units* and constellations within distributed groups. If there is, for instance, a clique[3] of strongly connected employees located at one geographical place, this might give a distributed group an imbalance. Thirdly, a network approach opens for comparative analysis of *different types* of structural networks. Comparing different distributed work groups' structural patterns might shed light on differences in performance and stability. And fourthly, if we have the opportunity to include also external ties in our analysis, it can help understand how well a group is connected to its surrounding environment. A group's performance is often strongly dependent on how well it is connected to its surrounding environments (Ancona and Caldwell 1992). In addition we should note the particular advantage of having access to the rich set of tools and techniques for analyzing network patterns developed within this field over the last three decades.

The social network approach is not all new to studies of telework and distributed work. Network studies of distributed workers and collaborative communities have occurred (although sparsely) during the last two decades.[4] These studies have on the one hand been focusing on structures of communication and coordination of tasks in distributed groups compared to collocated groups. The studies in this category have had a bias towards finding "the one best structure" for information flow in distributed environments. As such the emphasis has been on functional interaction and communication. So far, however, the evidence has been inconclusive: While some studies have found that a centralized structure works best for distributed groups (Ahuja and Carley 1999; Hinds and McGrath 2006), others have found that a decentralized structure works best for distributed groups with complicated tasks (Cummings and Cross 2003).

Another line of research has focused on the diverse use of ICT in distributed groups and communities (Haythornthwaite and Wellman 1998; Haythornthwaite 2001; Salaff 2002; Quan-Haase and Wellman 2006). These studies have had a descriptive orientation and have tended to include a broader set of relational variables. A central finding has been that affective ties are closely related to multiplex[5] use of media. Studies of students indicate that closer ties go together with more multiplex use of media. This line of research has included affective ties and bonds in their study. The focus here has had a bias toward looking at individual networks (ego networks) rather than complete networks of groups. As such the perspective of the group is often missing.

Yet, there is in the field of distributed work a growing interest for more affective relations and structures. It has been widely recognized that issues like trust and identity are highly important for the functioning of such groups (Jarvenpaa and Leidner 1999; Kanawattanachai and Yoo 2002; Zolin, Hinds, Fruchter and Levitt 2004). In this field the network approach is a largely untapped source that may help understand how and why affective relations develop within such groups[6]. Yet, as I will argue here, there is a strong need for a clearer understanding of the relationship between communication ties and other more affective

---

2) *Social network studies involve different types of approaches: Egocentrical data, focusing on individuals' network of relations; analyses of all relations within a restricted set of nodes (1-mode data); and studies of members' co-participation in groups or events (2-mode data). The focus for this paper is on 1-mode data.*

3) *A "clique" describes a maximal complete subgraph in a larger network structure. (Scott 2000, p. 117-118)*

4) *See for instance: Belanger 1999; Ahuja and Carley 1999; Salaff et al. 2002; Wellmann et al. 1996; Hinds et al. 2006; Yuan and Gay 2006.*

5) *"Media multiplexity" indicates the use of multiple media channels to support a dyadic social relation.*

6) *The potential use of a network approach for studying trust in distributed groups is outlined in more detail in Julsrud and Schiefloe (2007).*

Type	Sources	Example
Closer networks	Personal archives	Address book on mobile phones
	Direct interaction	Personal e-mail traffic or call-lists
Distant network	Public network archives	Belong to the same e-mail list, relations on Facebook
	Communication similarities	Use the same websites, read the same magazines

Table 1 Some new electronic sources for network analysis of organizations and groups

types of ties. I will return to this issue later in this paper. First, however, I will draw some attention to the new and emerging tools used to gather network data in distributed groups.

3 New Network Data Sources

The core of the network analysis is relational ties, and different social relations have been the objects of researchers' interest. Some of the most frequently studied are communication networks (for instance daily interaction pattern), formal ties (who is supposed to report to whom), affective ties (friends, trust), advice ties (who seeks advice from whom). In the early days of organizational networks studies, divergences between formal and emergent networks were much studied. A central finding coming out of this was that successful accomplishment of the work usually depended on regular use of informal networks (Tichy and Fombrun 1979).

The communication and interaction networks have always been at the centre of attention in network analysis, and in particular studies of media use in organizations.

A central facet of distributed work is the collaboration that takes part in a virtual space, and that communication through electronic media plays a more prominent role. Even though most such groups and teams have face-to-face communication, the balance is certainly pushed in the direction of mediated communication. Thus, the mediated ties are in particular relevant for studies of distributed groups.

Moreover, the application and use of new communication media have cleared the way for a number of new electronic sources, ready to use for network analysis. Records of communication through e-mail, mobile phones, SMS, and blogs can in many cases be used as inputs to network analysis. Traces of interaction can be concerted into network matrices and thus give researchers new data to conduction of network structures. A stream of new software that captures the mediated interaction within groups has emerged the last years. A random selection of some popular network-tracking tools includes TecFlow, Com-metrix, InFlow, Buddygraph and Enronic.[7)]

As a tool for analyzing communities and groups they are clearly on the rise (Tyler, Wilkinson and Huberman 2006). The importance of these electronic communication sources has been recognized by researchers almost since the dawn of the personal computer, and network studies of e-mail interaction and messaging systems have been conducted (sparsely) since the 1980s (ref: Rice et al.). The access to electronic sources has however now become much easier, due to software that is designed to track interaction based networks within predefined groups. For the most part these tools are directed towards analyzing e-mail interaction in a particular group of employees or collaborators, for instance Tyler and his colleagues who use e-mail files as a way to locate communities of practice in a research organization (Tyler, Wilkinson et al. 2006). Others have used mailing lists (Adamic and Adar 2002), mobile phone dialogues (Eagle, Pentland and Lazer 2007), and mobile phone address books (Lonkila 2004) as input sources for network studies.

In addition to the personal communication – accessible through e-mails and mobile phone logs – electronic sources include network sources like Facebook, Blogs or mailing lists. Such sites give the opportunity to capture more indirect relations based on the use of common network sources. As such, the new network sources seem to capture dimensions of both "weaker" and "stronger" ties (Granovetter 1973).

There are important differences embedded in networks based on personal interactions, public or private archives, direct interaction or similarities in activities visible in electronic sources. While interaction through e-mails is an intended and direct communication form, co-participating in the same e-mail

7) *For more information on these programs, see: http://www.buddygraph.com/ http://www.ickn.org/ickndemo/ http://www.orgnet.com/, http://www.orgnet.com/, http://jheer.org/enron/*

list is a much more indirect relationship. For network analysis of organizations and distributed groups the first two types are probably the most relevant. Therefore this paper will pay more attention to these types.

Interaction between individuals can of course also be captured without the use of software tracking the actual communication. This has been done by several earlier contributions in the field of network studies. Traditional survey instruments can be applied to ask individuals about their communication behavior. The software had, however, obvious advantages related to data reliability. Studies have indicated that memory based records in many cases have proved to be unreliable (Bernhardt, Killworth and Sailer 1982). As such the network tracking software has great advantages related to methodologies relying on individuals' recollection of communication activities.

The mediated interaction network diagram gives a picture of the mediated communication at a given time slot, and when various media are used in combination the media networks can be very useful as a point of departure for a study of distributed groups. The distributed group emerges out of the interaction data as a communicating electronic community. As a strategy for collecting reliable data of communication patterns within an organization or a group, the network-generating software is highly efficient. Yet, the strategy also has clear limitations: The first is that it is usually possible only to capture small pieces of the communication traffic that is going on within a group. Gathering all communication, including the informal talks and gestures, is an almost impossible task. Therefore the use of computer data rarely reflects the complete communication environment. Second, there are usually difficulties interpreting interaction-based ties and bonds. For instance, does an intense exchange of e-mails indicate the same kind of relations as an intense interchange of text messages on the mobile phone? Or does a high level of interaction between two collaborators indicate that they have a "close" relation? Although important, these issues have been the object of only limited discussions[8].

The work setting typical for distributed work groups appear as particularly challenging when it comes to analyzing interaction patterns. The reason is that the formal system or formal tasks heavily affect on the interaction. As described by Kadushin (2005) and others, the "pure" informal network interferes with a system of expected interaction ties described in work assignments and organizational charts. In a distributed group, the assigned tasks will be central for much of the interaction going on. Further, certain nodes in a network will be more "popular" than others due to formally assigned roles as leaders or specialists. Thus, work group ties are bound up with several factors that make interaction per se difficult to interpret directly.

4 A Cognitive Network Approach

To better understand and investigate networks of interaction in distributed groups it is necessary to ground this in a more elaborated theory of interaction and communication. As a point of departure I suggest drawing a clearer distinction between affective and interaction based relations. This perspective finds support in cognitive network theories, focusing on individuals' or groups' subjectively perceived relations in contrast to objective and interaction-based relations (Krackhardt 1987; Corman and Scott 1994). Corman and Scott have applied elements from Giddens' structuration theory to clarify the connections between observable communication networks and the latent networks of perceived relationships (Giddens 1984). They argue that social networks in general can be described as "cognitive" since they are based on individual perceptions of other individuals.[9] They argue further that different modalities explain the recursive relationships between cognitive social structure and interaction. Much in line with Giddens' description of "the duality of structures" they explain:

> *"… we define a communication system as a set of continually reproduced communicative interactions between individuals and collectives situated in time and space. The network is an abstract structure of rules and resources of communicative actors in a given social collective, instantiated in communication systems, but having only a 'virtual existence'"* (Corman and Scott, p 174).

The social network is here described as a cognitive resource embedded within a particular social community or culture, where spatial and temporal aspects are included in the analysis. Further, the authors propose that the cognitive network structures are activated through taking part in common activities (activity foci) or enacted through various triggering events.[10]

There is no room for further elaboration of the theoretical point made in the cognitive network theory

8) *For a discussion of possibilities and advantages related to the use of data based on computer mediated communication systems, see Rice (1990).*

9) *Note that Krackhardt uses the term cognitive network structures in a slightly different way than Cormann and Scott, denoting how individuals understand relations among other persons in their organization or community. (Krackhardt 1987)*

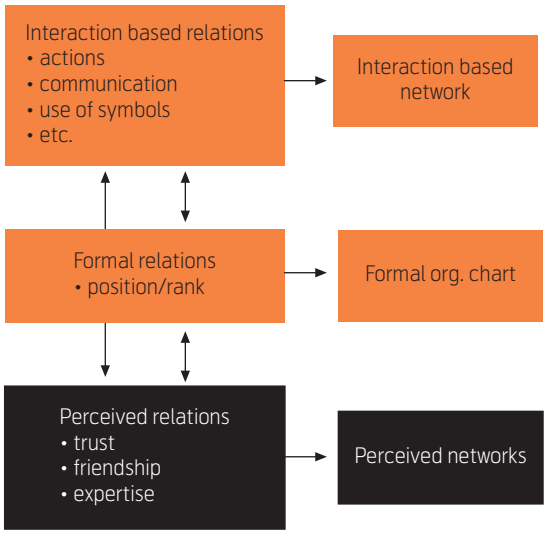


Figure 1 Interaction-based, formal and perceived relations

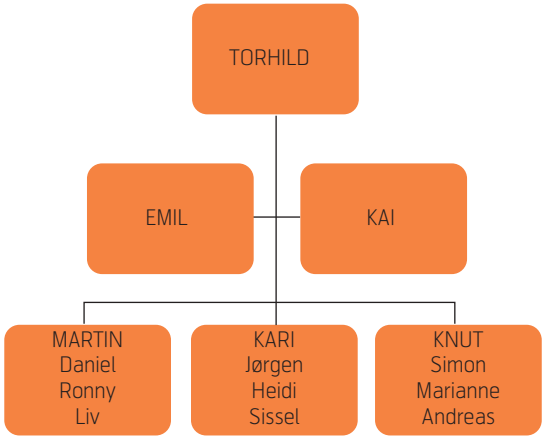


Figure 2 Omega's formal structure

here. The main point, however, is that social networks in general are seen as abstract cognitive constructs that are instantiated through participation in particular activities. The advantage of this perspective is that it helps to establish a clear distinction between a (cognitive) network structure, and a system of observable communicative actions. These structures are clearly related, but they are not isomorphic. Instead studies of the ways these structures are interrelated constitute an interesting and fertile area for empirical studies. I will in the next section of this paper use this approach as a point of departure for a case study of interaction patterns in a distributed work group.

5 The Case of Omega

The results presented here are based on a study of several work groups in a Nordic company, here called NOMO.[11] NOMO is a Norwegian ICT-provider with a fairly strong position in the Nordic markets. Approximately one year prior to our study, the company acquired and merged with a smaller Danish company to get an even stronger position in the Scandinavian market. This process was experienced as stressful for the employees in both companies. A major objective for the company after the acquisition was to integrate its operations across the national markets to create market synergies. This led to the setting up of a number of permanent work groups encompassing employees in different locations in Norway and Denmark. Since different functions now had to be coordinated across distances and national boundaries, distributed work was initiated and formalized in several different areas. The analysis in this paper will focus on one such group; Omega.

The core task of the Omega group was to manage and develop products for a particular segment of NOMO's customers. The group consisted of 16 product managers; 12 in Norway and four in Denmark, with the manager located at the headquarters in Norway.[12] Virtually all respondents had previous employment within the respective organizations, and most of them made deliberate efforts to maintain relations with previous colleagues.

The investigations followed the group from August 2005 to December 2006. When we first got in contact with members of Omega, they had operated as a distributed work group for about 15 months. The design of the study was based on a triangulation of different methodological strategies, including qualitative interviews with individuals as well as quantitative studies of group-based social networks.

5.1 Methodologies

The study started with an *explorative qualitative study* and was followed up with a *quantitative study* targeted at more specific issues evolving from the explorative phase. Yet, in the initial phase a general questionnaire was distributed to get baseline information about satisfaction, performance and interaction patterns. In this article we will mainly deploy the group-based network data supplemented with data from the qualitative interviews.

10) *The theory also draws on Feld's theory of activity foci, as well as Homans' theory of social groups (Homans 1950; Feld 1981).*

11) *Note that results from this case are to be published in Julsrud and Bakke (2008).*

12) *Danish and Norwegian were the working languages within the groups. The languages are fairly similar, but there are certain differences that can potentially lead to misunderstandings.*

Prior to the main quantitative network study, *semi-structured interviews* were conducted with employees and managers to get a better picture of their work situation. The interviews followed an interview guide focusing on the respondents' main work tasks, social relations, identity in group/organization and trust issues, and lasted 30-40 minutes. Fourteen of the sixteen employees in Omega were interviewed.

In the *social network part module*, interactions were registered through a web-based questionnaire and coded in a case-by-case social network matrix. We asked the persons to indicate interaction-based relations as well as perceived relations. A traditional "roster" design was used to the network study, where each group member received a list of the other members in the group (Wasserman and Faust 1994). The respondents were then asked to report the frequency of interaction with other members in the group as well as the type of media used in the interaction and the three perceived ties. We used a single question to map the *trust-based relationships*: "If you decided to search for another job similar to the one you have today, but in another company; whom on the list would you most likely talk to about this?" The idea behind this formulation is that this type of discussion would imply trustfulness, as disclosure of such plans would be negative for the reputation of the individual in question.[13)] Indirect questions are the most usual way to analyze trust-based relations in organizations. It should be noted, however, that such questions involve a risk of neglecting individuals that have a more introvert nature or simply prefer not to talk to anybody about such plans (even if they have trustful ties within the group).

The *expertise relations* were based on a question asking whom the informant preferred to speak to when facing problems in his/her work. Starting with the list of group members, we asked them to indicate whom on the list they would most likely turn to if they needed advice in their daily work. This expertise network does not address the affective aspect (like the trust ties), but the network with the most central professionals in the group.

The *friendship relations* were derived from our question whether there was someone on the list they considered as close friends in their group. Thus, we asked specifically for close friends, not regular work mates.

5.2 Description of Relations and Networks

As it turned out, the Omega group was handling the long-distance collaboration relatively well when measured along traditional network indicators for integration and coherence. For example, when looking at interaction via e-mail and mobile voice, none of the members were isolated from the others. All employees in the group were in contact with at least one other person during a regular week. In addition, the dialogues connected the employees through a network that crossed the geographical boundaries of the sub-units. It is easy to see, however, that e-mail interaction followed rather closely the formal interaction lines, in particular for the group managed by Martin (see Figure 3). It is also evident that much interaction seems to go through Martin, Kai and Emil. The manager Torhild was fairly central in the information flow. The mediated relations suggested that much of the information circulated between the sub-unit managers Martin, Kari, Knut, as well as Emil (see figures). It is also evident that most of the Danish employees were well integrated in the group, despite their geographical distance from the majority of employees in Omega.

Interaction based relations	Face-to-face meetings Mobile phone dialogues E-mail SMS Overall daily interaction
Prescribed relations	Formal work relations
Perceived relations	Trust (affective) Friendship Expertise

Table 2 Relations investigated in Omega

Table 3 provides more precise details for the networks based on interactions and on the perceived relations.

Among the *interaction-based relations*, the e-mail network was the most active, followed by mobile phone dialogues and SMS. The e-mail networks were denser and they also had higher reciprocity, indicating that they were not simply used to distribute information, but for two-way interaction. The lower level of *reciprocity* for SMS may suggest that this was a less formal channel, but also that the traffic here is less intense and task-related than in the e-mail network. The *average degree* score is a ratio of the number of incoming and outgoing ties for a network of relations (Freeman 1979). An average degree score reaching above six for e-mail relations then indicates that the average member had been in e-mail contact with approximately six other persons in the group during the last week. The corresponding numbers for mobile and SMS were 4.1 and 2.6. The score for e-mails was, interestingly, also higher than the face-to-face interaction (4), illustrating how e-mail connected

13) *This strategy is similar to the one used by earlier network studies on trust in organizations (Krackhardt and Hanson 1993; Krackhardt and Brass 1994; Burt and Knez 1996).*

	Relation	Scale	Links	Density	Average degree	Reciprocity	Core-Periphery	E/I index
Interaction based rel.	Interaction	Daily (weekly)	45	0.188 (0.546)	2.812	0.356	0.519	-0.301
	Mobile	1-4/5-10/ 11-20/>21	66	0.275	4.125	0.515	0.518	-0.208
	E-mail	1-4/5-10/ 11-20/>21	106	0.442	6.625	0.736	0.485	-0.083
	SMS	1-4/5-10/ 11-20/>21	43	0.179	2.688	0.512	0.370	-0.5
	Face-to-face	Daily (weekly)	64	0.267 (0.733)	4	0.688	0.829	-0.375
Perceived rel.	Trust	Yes/no	13	0.054	0.812	0.308	0.433	-0.818
	Friendship	Yes/no	10	0.042	0.625	0.4	0.466	-0.5
	Expertise	Yes/no	66	0.275	4.125	0.515	0.377	-0.250

Table 3 Selected network characteristics of interaction-based and perceived relations

far more people in the group than physical interaction. The *core/periphery* score indicates how well the registered values approximate to an ideal core/periphery structure (Borgatti and Everett 1999). This value was relatively high for the face-to-face networks due to the fact that there is a clear co-located core situated at theNorwegian headquarters, and that face-to-face follows close to this structure. This structure is softened in the mediated networks. However, while there was a relatively clear core/periphery structure reflected in the mobile network; this was less spelled out in the SMS network.

For distributed work groups it is of particular interest to see the extent to which the relations cross physical distance or not.[14)] To compare the number of ties within and across the two involved countries, we applied the *E-I Index*, as developed by Krackhardt and Stern (Krackhardt and Stern 1988). This indicator compares the external ties with the internal ties for groups within a network, ranging from -1 to +1. Given a partition of a network into a number of mutually exclusive groups, the E-I index is the number of ties external to the groups minus the number of ties that are internal to the group divided by the total number of ties. Maximum collaboration across the boundaries is then +1 (all links are external), while equally divided links will give an index equal to zero. We categorized the employees in Denmark as "external" and the Norwegian group as "internal". None of the interaction-based relations were equally divided, but e-mail messages was the form of interaction that was most boundary-crossing (considering the national boundaries) in this group. Interestingly, SMS was more frequently used within each of the national subunits, with mobile phone dialogues in a position in-between. This shows that – at least within this organization – the geography-bridging qualities of ICTs are selectively deployed; some are primarily used across larger distances, others are more commonly used within local regions. It is also interesting to observe that these technologies are important within collocated settings: While it is common to address the capacity of ICTs for bridging space and time, they are also used for communication with neighboring colleagues.

Turning to the *perceived relations* of trust, friendship and expertise, these were less cohesive in Omega than the interaction-based relations: only 10 friendship links (relations) and 13 trust links were reported. The expertise network, however, was about the same density level as for mobile communication (0.275). Reciprocity is often related to trust in organizations, as trust is often seen as stronger when relations are symmetrical (Kilduff and Tsai 2003). Interestingly, the trust relations have low levels of reciprocity, indicating that this is not a strongly interconnected network, but more open and "fluid". This indicates the "cognitive" nature of trust, since the existence of a trust-tie is not always perceived equally by two individuals in a network. The trust network was also strongly embedded in the national units within the group, indicated by the high negative E/I index, while the expertise relations had a much more boundary-crossing nature.

14) *In the case of mobile work, this can of course be difficult, as these boundaries are often blurred. Yet, in this group there was one important difference between individuals situated in Denmark and those in Norway.*

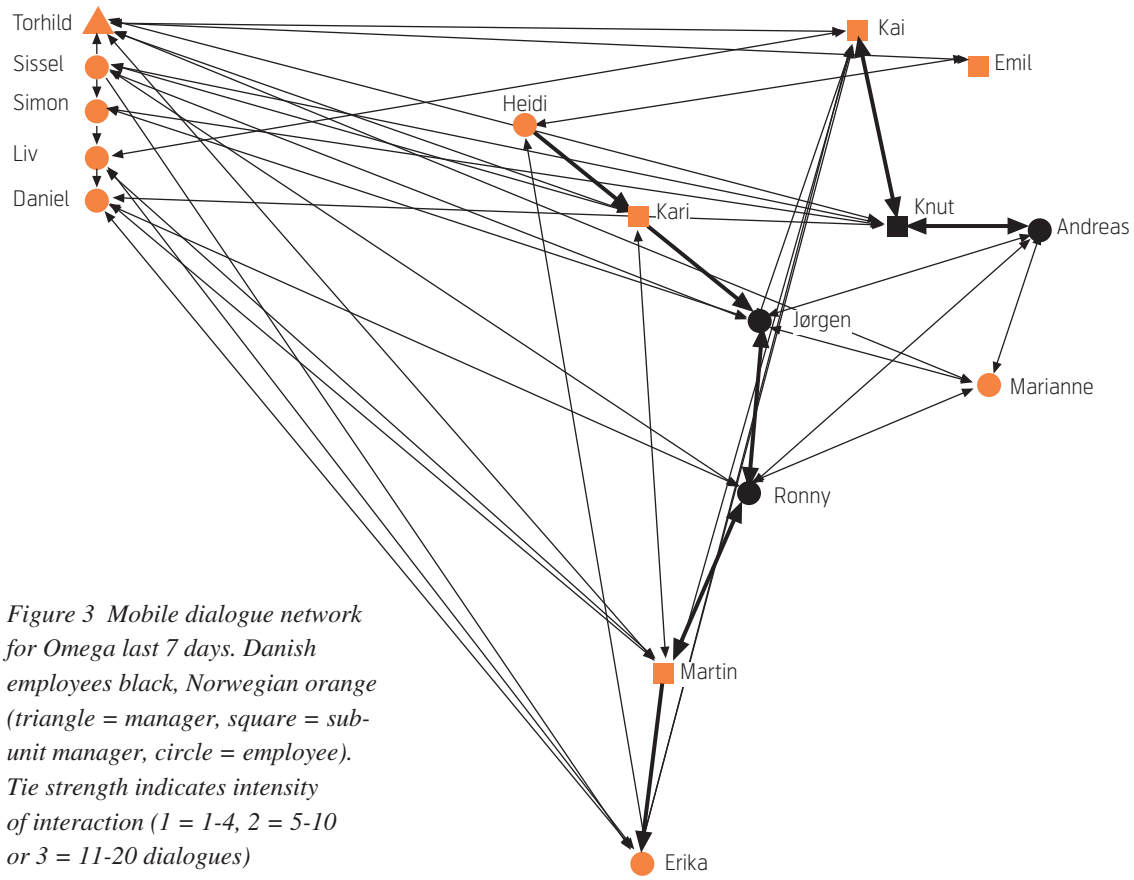


Figure 3 Mobile dialogue network for Omega last 7 days. Danish employees black, Norwegian orange (triangle = manager, square = sub-unit manager, circle = employee). Tie strength indicates intensity of interaction (1 = 1-4, 2 = 5-10 or 3 = 11-20 dialogues)

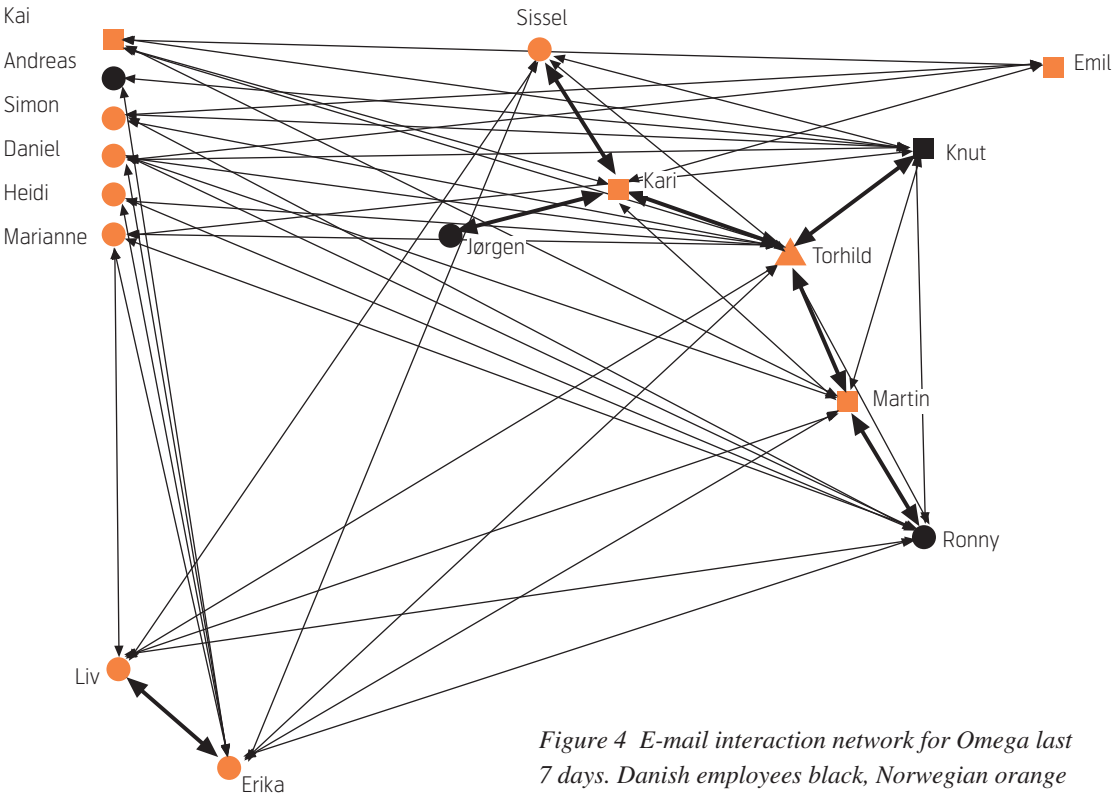


Figure 4 E-mail interaction network for Omega last 7 days. Danish employees black, Norwegian orange (triangle = manager, square = sub-unit manager, circle = employee). Tie strength indicates intensity of interaction (1 = 1-4, 2= 5-10 and 3 = 11-20 messages)

ISSN 0085-7130 © Telenor ASA 2008

	Formal	Expertise	Mobile	E-mail	SMS	Trust	Face-to-face
Expertise	0.316**						
Mobile	0.389**	0.435**					
E-mail	0.296**	0.522**	0.564**				
SMS	0.375**	0.38**	0.514**	0.393**			
Trust	0.068	0.113*	0.053	0.092*	0.12*		
Face-to-face	0.207*	0.236*	0.236*	0.295*	0.227**	0.054	
Friendship	0.1	0.086*	0.086*	0.074*	0.128**	0.278**	0.05

Note: * p < 0.05 ** p < 0.01

Table 4 QAP correlations for different networks (Jaccard coefficients)

5.3 Comparing Networks

Comparing networks through general indicators gives important information regarding the general use of interaction media and the general level of trust, friendship and expertise relations. Yet, to explore the similarities between the perceived networks and the four different interaction networks further, we conducted a QAP-correlation.[15] This procedure is often used to see to what extent there are similarities between two social networks containing the same actors (Hanneman 2001).

As indicated in Table 4 the expertise network, as well as the formal network, were closely related in all the media channels.[16] In particular, the relation between expertise, e-mail and mobile was strong (r = 0.522 and 0.435, respectively). The mediated networks of mobile phones, SMS and e-mail were all highly correlated, and in particular e-mail and mobile dialogues (r = 0.564) (all significant on a 0.01 level). This indicates that the media in Omega to a large extent followed the task related patterns of interaction, and that the media followed highly similar patterns, in particular in the case of mobile dialogues and SMS.

The trust network, however, had no significant relation to the formal network, the face-to-face network or the mobile communication network. It was however weakly (but significantly) related to the expertise network, the e-mail network and the SMS network. Trust relations were most strongly correlated to friendship relations (r = 0.278) but not at all with the formal relations (0.068). This indicates on the one hand that the perceived expertise relations were most closely related to the observable interaction that took place in Omega. This pattern also followed fairly close to the formal structure of the organization. On the other hand, the less intensive trust network diverged from the formal structure and was less similar to the mediated networks based on mobile dialogues. Yet it had high similarity to the friendship network, and also to the expertise network.

This then might suggest that trust relations are more strongly supported by text-based media like SMS and e-mail, while the more intensive work-related communication uses all media, and in particular e-mail and mobile dialogue. As such, it indicates that the instant problem-solving relations have other needs for communication than the more low-frequent trust and friendship ties. It is clear, however, that these relations do not operate as isolated structures, but have significant overlaps.

It should also be noted that physical closeness (i.e. face-to-face interaction) was positively correlated to the use of all media, and in particular the SMS network, indicating that mediated interaction is more intense among co-located workers.

6 Discussion and Conclusions

This paper has argued that interaction-based relations, represented by SMS, e-mail and so on, are important sources for analyses of distributed work. Still, the interaction-based relations and networks should not be taken as direct evidence of an affective tie or a group with high cohesion. I have proposed to draw a clear distinction between affective ties, formal ties and interaction based ties, relying in particular on cognitive network theories and structuration theory.

15) *UCINET's QAP correlation procedure is based on permutation of rows and columns together with one of the input matrices, and then correlating the permuted matrix with the other matrix. This is repeated hundreds of times to build up a distribution of correlations under the null hypothesis of no relationships between the matrices. A low p-value (< .05) suggests a strong relationship unlikely to have occurred by chance.*

16) *Table 4 presents Jaccard coefficients since some relations (trust, friendship, face-to-face) are binary .*

This framework is somewhat opposed to the much applied concepts of *weak and strong ties*, widely used in the social network field. The distinction was initially proposed by Granovetter who argued that the strength of a tie is a: "combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal services which characterize the tie" (1973, p. 1361).

The distinction between strong and weak ties is however often problematic. The operationalization of ties as a combination of four different qualities – as proposed by Granovetter – is not straightforward; and capturing weak ties is difficult due to the large number of potential ties, and the fact that weak ties are revealed in particular settings, whereas most of the time, they are 'latent' (Krackhardt 1992). Further, studies of vocational networks have often found that relations between colleagues rarely fall in one of the two (Wegener 1991; Nardi, Whittaker and Schwarz 2000).

The advantage of the cognitive network approach is that it offers a clearer distinction between observable and "cognitive networks". For studies of mediated networks this opens for a more systematic investigation of how different types of ties are supported by the use of various media.

In this paper the cognitive network approach has been used as a theoretical framework for an analysis of various relations within a distributed work group. Clearly, the case presented has several shortcomings. Firstly, it is based on self-reported ties and not traffic generated data. Secondly, it represents just a snapshot of interactions over one week. Thirdly, only one simple case of distributed workers has been investigated. The findings should thus be seen as preliminary findings.

Still, the results suggest that the perceived networks of trust, expertise and friendship are supported in different ways by the media. Interaction through e-mail and mobile phones was following closely the expertise based ties, but not always the friendship and trust relations.

As already mentioned, some earlier studies have found that stronger relations tend to communicate more intensively and also use more numerous media (Haythornthwaite 2002). There may be different reasons why this study paints a slightly different picture: Firstly, our case involved a group of technical professionals working in permanent work groups – and although the group was recently established, the members had a history within the organization. In contrast, former studies of relations and media-use in distributed environments have used empirical data from ad hoc teams of students collaborating in temporal, virtual teams (Haythornthwaite 2001; Haythornthwaite 2005) or in a community of scholars at a university (Koku and Wellman 2002). In Omega, the task-related ties were very much in focus, whereas the trust and friendship relations were less explicit. Also, this group was relatively recently established, connecting experts located in different geographical units due to a company merger. This might have made the friendship relations within the group less dense and more weakly supported by media, as compared to networks of students or university scholars. Another reason for the differences may be that we applied the term trust in addition to friendship; a term that is rarely studied in relation to mediated ties in organizations. We believe, however, that this actually unveils a relational dimension that is different from close friendship in organizations, but still important. In modern organizations it might be that it is more important – or more achievable – to have someone that you trust to discuss difficult personal matters with, than someone you consider as close friends.

The conclusion we may draw is that we should be cautious and not jump to conclusions when analyzing interaction based structures in distributed groups. There is no such thing as "the social network" of a group, but multiple different interconnected relationships. Patterns of mediated interaction – together with face-to-face interaction – are crucial as indicators of the coordination going on in a distributed group, and can be important as indicators of affective ties like trust and friendship. Yet, different media may play different roles in such networks. As such, future studies should not only rely on communication networks, but as far as possible try to capture also more affective relations as well as the formal structures. The cognitive network approach, as applied here, might be a promising point of departure.

References

Ahuja, M, Carley, K M. 1999. Network Structure in Virtual Organizations. *Organization Science*, 10 (6), 741-57.

Ancona, D G, Caldwell, D F. 1992. Bridging the Boundary:External activities and Performance in Organizational Teams. *Administrative Science Quarterly*, (37), 634-665.

Bélanger, F. 1999. Communication Patterns In Distributed Work Groups: A Network Analysis. *IEEE Transactions on Professional Communications*, 42 (4), 261-275.

Bernhardt, H R, Killworth, P, Sailer, L. 1982. Informant accuracy in social network data V. *Social Science Research*, (11), 30-66.

Borgatti, S P, Everett, M G. 1999. Models of core/periphery structures. *Social Networks*, (21), 375-395.

Breiger, R. 2004. The Analysis of Social Networks. In: Hardy, M, Bryman, A (eds). *Handbook of Data analysis*. London, Sage.

Burt, R, Knez, M. 1996. Trust and Third-Party Gossip. In: Kramer, R M, Tyler, T R (eds). *Trust in Organizations. Frontiers of Theory and research*. Thousand Oaks, Sage, 68-89.

Corman, S R, Scott, C R. 1994. Perceived networks, activity, foci and observable communication in social collectives. *Communication theory*, (4), 171-190.

Cummings, J N, Cross, R. 2003. Structural properties of work groups and their consequences for performance. *Social Networks*, (25), 197-210.

DeSanctis, G, Monge, P. 1999. Introduction: Communication processes for virtual organizations. *Organization Science* (Special Issue), (10), 693-703.

Duarte, D L, Snyder, N T. 2006. *Mastering Virtual Teams*. Josey-Bass.

Eagle, N, Pentland, A, Lazer, D. 2007. Inferring Social Network Structure using Mobile Phone Data. In: Liu, H, Salerno, J J, Young, M J (Eds.). *Social Computing, Behavioral Modeling, and Prediction*. New York, Springer.

Feld, S L. 1981. The Focused Organization of Social Ties. *The American Journal of Sociology*, 86 (5), 1015-1035.

Freeman, L C. 1979. Centrality in social networks: Conceptual clarification. *Social Networks*, (1), 215-39.

Giddens, A. 1984. *The Constitution of Society*. Berkley/Los Angeles, University of California Press.

Granovetter, M S. 1973. The Strength of Weak Ties. *American Journal of Sociology*, (81), 1287-1303.

Hanneman, R. 2001. *Introduction to Social Network Methods*. Dep. Of Sociology, University of California, Riverside.

Haythornthwaite, C. 2001. Exploring Multiplexity: Social Network Structures in a Computer-Supported Distance learning Class. *The Information Society*, (17), 211-226.

Haythornthwaite, C. 2002. Strong, Weak, and Latent Ties and the Impact of New Media. *The Information Society*, 18 (5), 385-401.

Haythornthwaite, C. 2005. Social Networks and Internet Connectivity Effects. *Information Communication and Society*, 8 (2), 125-147.

Haythornthwaite, C, Wellman, B. 1998. Work, Friendship, and Media Use for Information Exchange in a Networked Organization. *Journal of the American Society for Information Science*, 12 (49), 1101-1114.

Hinds, P, Kiesler, S. 2002. *Distributed Work*. Cambridge, Massachusets, MIT press.

Hinds, P, McGrath, C. 2006. Structures that Work: Social Structure, Work Structure and Coordination Ease in Geographically Distributed Teams. *CSCW'06*, Banff, Alberta, Canada.

Homans, G C. 1950. *The Human Group*. New York, Harcourt Brace.

Jackson, P, derWilen, J M V. 1998. *Teleworking: International Perspectives. From Telecommuting to the Virtual Organisation*. London, Routledge.

Jarvenpaa, S L, Leidner, D E. 1999. Communication and Trust in Global Virtual Teams. *Organization Science*, 10 (6), 791-815.

Julsrud, T, Bakke, J W. 2008. Trust, friendship and expertise: The use of email, mobile dialogues and emails to develop and sustain social relations in a distributed work group. In: Ling, R, Campbell S (eds). *The mobile communications research annual: The reconstruction of space and time through mobile communication practices*. New Brunswick, NJ, Transaction, 1. (Forthcoming)

Julsrud, T E, Schiefloe, P M. 2007. The development, distribution and maintenance of trust in distributed work groups. A social network approach. *International Journal of Networking and Virtual Organisations*, November.

Kadushin, C. 2005. Networks and Small Groups. *Structure and Dynamics:eJournal of Anthropology and Related Science*, 1 (1).

Kanawattanachai, P, Yoo, Y. 2002. Dynamic nature of trust in virtual teams. *Strategic Information Systems*, (11), 187-213.

Kilduff, M, Tsai, W. 2003. *Social Networks and Organizations*. London, Sage.

Koku, E F, Wellman, B. 2002. Scholarly Networks as Learning Communities: The Case of TechNet. In: Barab, C, Kling, R (eds). *Designing Virtual Commu-*

nities in the Service of Learning. Cambridge, Cambridge University Press.

Krackhardt, D. 1987. Cognitive Social Structures. *Social Networks*, (9), 109-134.

Krackhardt, D. 1992. The strength of strong ties: The importance of philos in organizations. In: Nohria, N, Eccles, R (eds). *Network and Organizations; Structure, Form and Action*. Boston, Harvard University Press, 216-239.

Krackhardt, D, Brass, D. 1994. Intraorganizational Networks. The Micro Side. In: Wassermann, S, Galaskiewicz, J (eds). *Advances in Social Network Analysis*. Thousand Oaks, Sage, 207-229.

Krackhardt, D, Hanson, J R. 1993. Informal Networks: The Company behind the Chart. *Harvard Business Review*, 71 (4), 104-111.

Krackhardt, D, Stern, R N. 1988. Informal Networks and Organizational Crises: An Experimental Simulation. *Social Psychology Quarterly*, 51 (2), 123-140.

Lipnack, J, Stamps, J. 2000. *Virtual Teams. People Working Across Boundaries with Technology*. New York, John Wiley.

Lonkila, M. 2004. Phone notebooks as data on personal networks. *Connections*, 25 (1), 53-61.

Nardi, B, Whittaker, S, Schwarz, H. 2000. "It's Not What You Know, It's Who You Know: Work in the Information Age." *First Monday*, 5 (5), May.

Quan-Haase, A, Wellman, B. 2006. Hyperconnected Net Work: Computer-Mediated Community in a Hich-Tech Organization. In: Hekscher, C, Adler, P (eds). *The Firm as a Collaborative Community. Reconstructing Trust in the Knowledge Economy*. New York, Oxford University Press, 281-333.

Rice, R E. 1990. Computer-mediated communication system network data: theoretical concerns and empirical examples. *International Journal of Man-Machine Studies*, (32), 627-647.

Salaff, J W. 2002. Where home is the Office. The New Form of Flexible Work. In: Wellmann, B, Haythornthwaite, C (eds). *The Internet in Everyday Life*. Oxford, Blackwell, 464-495.

Scott, J. 2000. *Social Network Analysis: A handbook*. Thousand Oaks, CA, Sage.

Sparrowe, R, Liden, R, Waynes, S, Kraimer, M. 2001. Social networks and the performance of individuals and groups. *Academy of Management Journal*, 44, 316-325.

Tichy, N, Fombrun, C. 1979. Network Analysis in Organizational Settings. *Human Relations*, 32 (11), 923-965.

Tyler, J R, Wilkinson, D M, Huberman, B A. 2005. Email as spectroscopy: automated discovery of community structure within organizations. *The Information Society*, 21, 143-153.

Wasserman, S, Faust, K. 1994. *Social Network Analysis. Methods and Applications*. Cambridge University Press.

Wegener, B. 1991. Job mobility and social ties: social resources, prior job, and status attainment. *Annual Sociological review*, 56, 60-71.

Wellman, B. 1988. Structural analysis: from method and metaphor to theory and substance. In: Wellmann, B, Berkowitz, S D (eds). *Social Structures: a network approach*. New York, Cambridge University Press, 19-61.

Wellman, B, Salaff, J, Dimitrova, D, Garton, L, Gulia, M, Haythornthwaite, C. 1996. Computer Networks as Social Networks: Virtual Community, Computer Supported Cooperative Work and Telework. *Annual Review of Sociology*, 213-38.

Yuan, Y C, Gay, G. 2006. Homophily of Network Ties and Bonding and Bridging Social Capital in Computer-mediated Distributed Teams. *Journal of Computer-Mediated Communication*, 11 (4).

Zolin, R, Hinds, P J, Fruchter, R, Levitt, R E. 2004. Interpersonal Trust in Cross-Functional Global Teams. A longitudinal Study. *Information and Organization,* 1(14), 1-26.

Tom E. Julsrud is a research scientist in Telenor Research and Innovation and is also associated to the Norwegian University of Science and Technology (NTNU). His research areas of interest include social networks, distributed work, mobile work, workplace changes, trust in organizations and social capital. He has co-authored books on telework and distributed work, and has published several articles and papers on collaboration in distributed and virtual teams. His latest work focuses in particular on the development of trust in virtual environments and distributed groups.
email: tom-erik.julsrud@telenor.com

Innovation in a Value Network Perspective[1)]

ØYSTEIN D. FJELDSTAD

Øystein D. Fjeldstad is Professor at BI/Norwegian School of Management

A growing number of firms create value by networking their customers. Classic examples include telecommunication operators, logistic services and banks. The primary innovations of such firms increase the connectivity of a network, that is who or what can be reached and the conductivity of a network, that is what can be exchanged. The article discusses exploration and exploitation in networking services. It has implications for how innovations are conceived, managed and measured.

Exploration and Exploitation

An innovation is a significant change in knowledge or practice. Innovating is exploratory activity that results in new skills, practices, technologies, services or products of a firm (Greve & Taylor, 2000, p. 55). Organizations innovate in order to gain competitive advantage from new products, processes and organizational resources. Two important misunderstandings about innovation are, first, that innovations are primarily technical in nature and second, that the primary output of product development is innovations. Innovations are more than just technical devices, and much – if not most – product development does not result in innovations.

Because the word innovation is often interpreted narrowly, the term exploration is used to describe a broader idea of organizational generation of variation. Exploration takes many forms and includes, in addition to the more narrow term innovation, things such as search, variation, risk taking, experimentation, play, flexibility, and discovery (March, 1991). Furthermore, it is well established that the sources of innovation are not found exclusively inside firms. They are commonly found in the space between firms, universities, research laboratories, suppliers and customers (Powell, 1990; Powell, Koput, & Smith-Doerr, 1996). In short, firms innovate when they conduct and participate in exploratory activity inside and outside of their organizational boundaries or absorb novel practices from outside their organizational boundaries (Cohen & Levinthal, 1990). In contrast, exploitation is activities that seek to capture the gains from innovations. Organizations are structured to exploit their competencies (Greve & Taylor, 2000) and exploitation includes refinement, choice, production, efficiency, selection, implementation and execution (March, 1991). The goal of much product development is exploitation through modifying the results of exploration activities to fit a given market or through making incremental improvements in order to beat the competition.

Over time organizations need to engage in both exploration and exploitation because exploration is the seed of competitive advantage, but exploitation is required to translate competitive advantage into commercial success. The form that exploration takes and the balance between exploration and exploitation depend on the type of business the firm is in, on properties of its environment and on timing. In other words where, how and when exploratory activity takes places will vary from organization to organization. This is a challenge with respect to the measurement and evaluation of innovative activity. It is almost trivial to measure classic internally organized research and development activity. However, this accounting category does not correspond well to exploration because it omits exploration done by importing novel practices from elsewhere, it blends exploration (research) with exploitation (development), and it misses exploration in non-technical areas such as development of new business models, and entry into new markets. Hence it accounts for only a portion of organizational innovation, or exploration, particularly since knowledge in rapidly developing fields is both sophisticated and widely dispersed across a number of organizations. In industries in which know-how is critical, companies must be expert at both in-house research and cooperative exploration with universities, customers, competitors and firms that complement their value creation (Powell et al., 1996). However, even more refined accounting measures of exploration would face the conceptual problem that inputs into the innovation process are used as substitutes for the output of innovation.

Exploration in Network Service Industries such as Telecommunication and Inter-related Transaction Services

Telecommunication operators, banks and parcel services are examples of organizations that create value by linking customers who are, or wish to be interdependent (Stabell & Fjeldstad, 1998; Thompson,

1) *The discussion has benefited greatly from comments by Professor Henrich R. Greve, BI, and from Dr. Knut B. Haanæs, BCG.*

ISSN 0085-7130 © Telenor ASA 2008

1967). Their activities are modeled by the Value Network configuration consisting of three simultaneous activities: network promotion and contract management; service provisioning and infrastructure operations. Figure 1 depicts the value network configuration for a mobile network operator. Organizations with a Value Network configuration co-produce (Ramirez, 1999) value in a horizontally interconnected and vertically layered value system (Stabell & Fjeldstad, 1998) exemplified in Figure 2. The actors between whom exchange is facilitated, such as end user customers and content creators surround the system as do suppliers of hardware and software to both operators/service providers and end users (see Andersen and Fjeldstad, 2003).

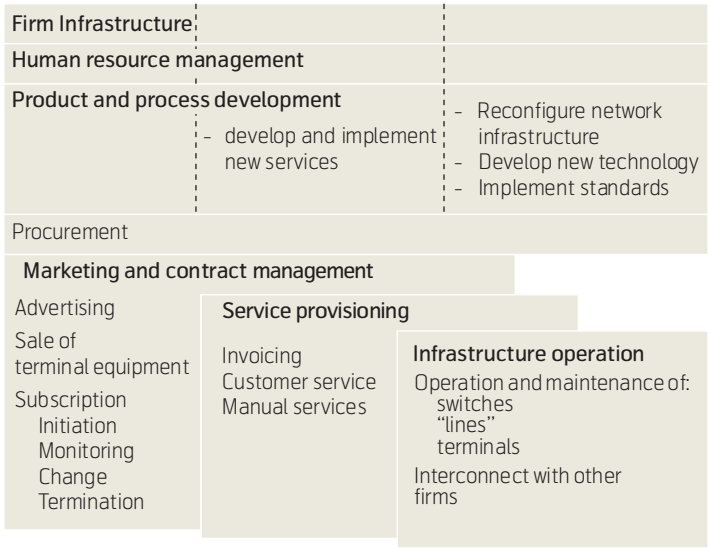


Figure 1 Value Network diagram for POTS provider

The below discussion of innovation references both the value network configuration as it applies to an individual firm and the overall value system.

The value of their offerings is thus strongly influenced by network effects (Katz & Shapiro, 1994) associated with existing and/or potential inter-customer relations for which they can provide service. There is a network effect when the value of a product depends on the size of the network or on its composition, i.e., who the other customers are (Rohlfs, 1974; Stabell & Fjeldstad, 1998). Simplified we can distinguish among innovations that increase the connectivity of a network, that is who can be reached, the conductivity of a network, that is what can be transacted, and other innovations that simply improve on either the willingness to pay for service or on the cost of providing the service, for example a better user interface or a more efficient process.

Network effects are important to the creation and capture of value for a wide range of transaction services (Economides, 1993, 1996; Encaoua, Moreaux, & Perrot, 1996; Rohlfs, 1974), and because of their strong importance, the issues associated with investment in network effects and the ability to earn profits from such investments are analogous to the issues associated with investment in technological development and patenting (Katz & Shapiro, 1985). It follows that exploration of more effective ways of diffusing new networks, that is improved practices for recruitment of customers and the roll-out of physical, human and software infrastructures, or upgrades to existing networks, is at the heart of exploratory activity in telecommunication and other transaction service firms. Such innovations contribute to increasing the connectivity of the network serviced. Firms use a variety of strategies in order to establish critical mass within the targeted user groups. Some of these include "give-aways", and "Trojan Horses", a product whose standalone features get it diffused, but which has a future potential as a network access terminal, and features that make the service diffusible, for example self installation. The diffusion of the PC was independent of its future potential as an Internet terminal, but it greatly aided the diffusion of the Internet (Fjeldstad & Haanæs, 2001).

Network service firms obviously also engage in other forms of exploratory activity. They conduct classic technologically oriented R&D activities. It is common to distinguish between product and process innovations (Wheelwright & Clark, 1992). Product innovations improve the characteristics of the product that the customer buys whereas process technologies improve on processes by which the product is created and delivered. In service industries such as telecommunication, the distinction between product and pro-

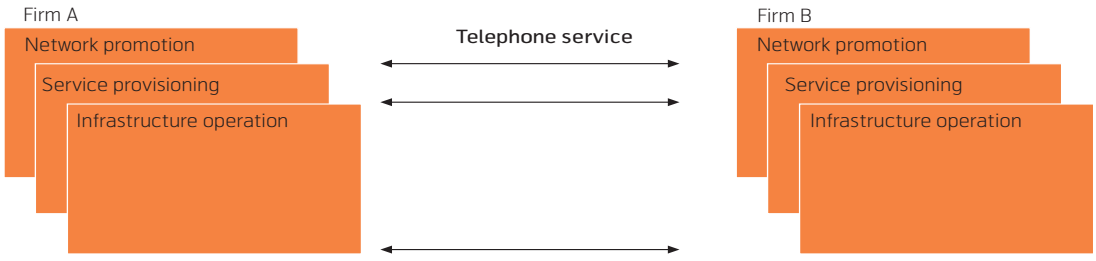


Figure 2 Horizontal co-production with cooperating and competing networks

ISSN 0085-7130 © Telenor ASA 2008

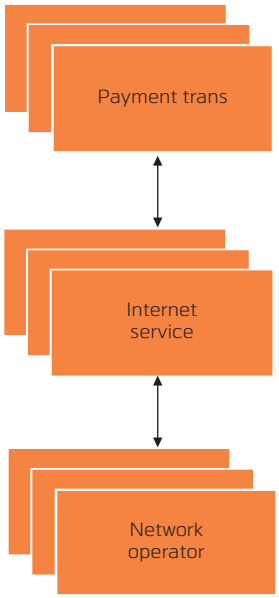


Figure 3 Vertical co-production

cess is blurred and one should expect a greater portion of process innovations. These are not typically the domain of traditional R&D activity. In fact a number of process innovations related to IT system development, change of business processes etc., are created with assistance from external consultants.

Some technological innovations directly increase connectivity. These are frequently related to the interconnection of one network to another, for example creating a bridge between e-mail software and mobile phone SMS. Technological innovations that increase conductivity, that is what can be transacted and how it can be transacted, for example speed and security become valuable to customers only by the diffusion to a sufficient number of relevant other customers. The technological innovation process must therefore also closely consider the diffusion process, including to the extent that particular features of the product makes it diffusible. Much of the success of the Internet telephony system Skype is due to its ease of spread; it is easy to download. For each user it suffices that only one additional user with whom he or she is interested in communicating with also uses Skype, and by the SkypeOut interconnection the company has increased the product's connectivity greatly beyond the network of Skype users.

A better service interface and a new type of content are examples of innovations that don't necessarily improve on network connectivity or conductivity. They improve the basic willingness to pay for a service regardless of adaptation by other users. However, exploitation of such innovations can either come directly from the customers' increased willingness to pay for the service or perhaps more important from gains to the effectiveness of the diffusion of a service as described above. Therefore strategic choices about the aim of such developments, as discussed above, and the pricing of the results in the market are very important.

Network services are basic societal infrastructures. They provide virtual spaces for flows in the modern global economy (Castells, 1996). The value is increased each time they can be used to carry another type of transaction. There are therefore two additional important forms of innovations related to network services. First, there are innovations that improve the ability of other network services to use a more basic service, for example mobile data communication, as a medium for their own type of exchange. Examples are innovations that enable banks, real-estate brokers, ticket agencies and others to facilitate their type of transactions via the mobile phone network. Such vertically aimed innovations promote the development of multi-layered network structures (Fjeldstad, 1999) important for the overall value of the service. Second, there are user-located innovations that improve users' ability to use the network to transact. Innovations made in "Tele-medicine" at hospitals and clinics to adapt their equipment and processes to network use are examples of this category. Both forms of innovations take place at the boundary of the organization. They obviously have to be carried out in some form of cooperation with the complementing providers, often competitors, when the aim is to create a sufficiently large combined network and customers.

Telecommunications innovations directly illustrate the features of exploration discussed above. First, innovations take place in the technology, in the business model for pricing the use of the technology, and in the services running on the technology and their business models. Technological exploration is sometimes integrated with exploration in the market domain, but the two domains can also be explored independently. Second, many innovations in the market domain are done once, somewhere in the world, and then imported and adapted elsewhere. This is possible because services that increase the general user's ability to pay cannot be protected as readily as technological innovations, and it can be profitable for all firms taking part in the exploration because innovations targeting willingness-to-pay are progress for both the firm and the customer even when they spread to all competitors, and thus do not alter the competitive balance. Third, some innovations in the market domain are developed jointly with specific users. The costs of developing these innovations are shared across participating organizations, but they still represent significant exploration from the viewpoint of the telecommunications provider. Multi-lay-

ered network structures are examples of such innovations.

In summary, exploratory activity is business type dependent, it takes many forms beyond classic R&D activity and it takes place in the space between a large number of different forms of actors. In network services, innovations are in particular related to technologies, processes, practices and competencies that increase network connectivity and network conductivity. Such innovations are at the heart of exploration in telecommunication service. They are exploited by willingness to pay for network membership and through transaction volumes and transaction prices. In the case of Telenor, Telenor's substantial competence in how to roll-out networks and diffuse services that lead to high network use is globally explored and globally exploited.

References

Andersen, E, Fjeldstad, Ø D. 2003. Understanding inter-firm relations in mediation industries with special reference to the Nordic Mobile Communication Industry. *Industrial Marketing Management*, 32, 397–408.

Castells, M. 1996. *The Rise of the Network Society*. Oxford, Blackwell.

Cohen, W M, Levinthal, D A. 1990. Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35 (1), 128–152.

Economides, N. 1993. Network Economics with Application to Finance. *Financial Markets, Institutions & Instruments*, 2 (5), 89–97.

Economides, N. 1996. The Economics of Networks. *International Journal of Industrial Organization*, 14 (6), 673–699.

Encaoua, D, Moreaux, M, Perrot, A. 1996. Compatibility and competition in airlines Demand side network effects. *International Journal of Industrial Organization*, 14 (6), 701–726.

Fjeldstad, Ø D. 1999. The Value System in Telecommunication. In: Eliassen, K, Sjøvaag, M (eds). *Liberalising Eurpoean Telecommunication*, 238–256. London, Routledge.

Fjeldstad, Ø D, Haanæs, K B. 2001. Strategy Tradeoffs in the Knowledge and Network Economy. *Business Strategy Review*, 12 (1), 1–10.

Greve, H R, Taylor, A. 2000. Innovations as Catalysts for Organizational Change: Shifts in Organizational Cognition and Search. *Administrative Science Quarterly*, 45 (1), 54–80.

Katz, M L, Shapiro, C. 1985. Network Externalities, Competition and Compatibility. *American Economic Review*, 75 (3), 424–441.

Katz, M L, Shapiro, C. 1994. Systems competition and network effects. *Journal of Economic Perspectives*, 8 (2), 93–115.

March, J G. 1991. Exploration and Exploitation in Organizational Learning. *Organization Science: A Journal of the Institute of Management Sciences*, 2, 71–87. INFORMS: Institute for Operations Research.

Powell, W W. 1990. Neither Market nor hierarchy: Network forms of organization. In: Staw, B M, Cummings, L L (eds). *Research in Organizational Behavior*, 12, 295–336.

Powell, W W, Koput, K W, Smith-Doerr, L. 1996. Interorganizational collaboration and the locus of innovation: Networks of learning in Biotechnology. *Administrative Science Quarterly*, 41 (1), 116–145.

Ramirez, R. 1999. Value co-production: Intellectual origins and implications for practice and research. *Strategic Management Journal*, 20, 49–65.

Rohlfs, J. 1974. A Theory of Interdependent Demand for a Communication Service. *Bell Journal of Economics and Management Science*, 5 (Spring), 16–37.

Stabell, C B, Fjeldstad, Ø D. 1998. Configuring value for competitive advantage: On chains, shops, and networks. *Strategic Management Journal*, 19 (5), 413–437.

Thompson, J D. 1967. *Organizations in Action*. New York, McGraw-Hill.

Wheelwright, S C, Clark, K. 1992. *Revolutionizing Product Development*. New York, The Free Press.

Professor Øystein D. Fjeldstad holds the Telenor Chair of International Strategy and Management at BI/Norwegian School of Management. He researches the Value Creation and Strategic Management of network facilitating and problem solving organizations. He holds a PhD in Business Administration from the University of Arizona as well as an MSc in Management Information Systems. His undergraduate degree is from the Norwegian School of Economics and Business Administration.

email: oystein.fjeldstad@bi.no

Quantitative Networks Analysis and Modeling of Networked Multiagent Environment

DENIS BECKER, ALEXEI GAIVORONSKI

Denis Becker is a PhD candidate at the Norwegian University of Science and Technology (NTNU)

Alexei Gaivoronski is Professor at the Norwegian University of Science and Technology

This paper gives an overview of the current research program and some recent results in modelling of modern telecommunication environment obtained in the Department of Industrial Economics and Technology Management, NTNU. We concentrate on different strategic decision problems when it is necessary to take into account uncertainties in demand, technology and other important variables which characterize this rapidly changing environment. Besides, we look at cases that are characterized by the interaction of different agents engaged in relations of competition and collaboration. One such example deals with a quantitative evaluation of business models for collaborative provision of advanced mobile data services while another looks at the relations between network operators and service providers.

Besides, we put our research focus in a wider perspective by presenting a survey of several promising modelling approaches of quantitative network analysis and formation. These include stochastic optimization, statistical mechanics of networks, network formation games and agent based computational economics. All of these research domains study complex interactive systems that are either explicit networks or can be described as such. The paper explores similarities, dissimilarities and linkages of these concepts including approaches on the border between these methodologies. Finally, implications for the research in telecommunications will be given.

1 Introduction

Modern telecommunications environment represents a challenge for industrial planners and academics alike. Only a few decades ago it was a rigidly controlled and static monopolistic environment with a handful of mature services with long life cycles, simple business models and predictable decision outcomes. Now we have a totally different picture of a dynamic industrial reality with a multitude of actors assuming a wide variety of roles, rapidly changing technological solutions, innovative business models, increased uncertainty and risk.

In our research program we focus on the development of methodological tools for an adequate treatment of risk and uncertainty in order to support strategic decisions and the evaluation of business models in telecommunications and the information industry. In this paper we present some of our recent results in this direction. Two examples are described in some detail. The first deals with the modeling of relations between network operators and service providers or virtual network operators with the aim to produce advice for a network operator on his policies towards these actors. The second example deals with the evaluation of service platforms and business models for collaborative service provision of advanced mobile data services. The aim here is to design business models which would induce independent agents to contribute their expertise towards the creation of a successful service or successful service platform.

Both cases deal with situations with considerable uncertainty and risk where several independent agents possess incomplete information about environment and each other's aims. For their analysis we employ the modern modeling methodologies for risk management and optimal decision support under uncertainty developed in operations research and investment science. More specifically, we develop stochastic programming models with bi-level structure enhanced with certain notions of game theory and modern quantitative finance.

In order to put our methodological choices in a perspective we also present a survey of different relevant quantitative methodologies for modeling networks and explore their similarities, dissimilarities and linkages. The first of them is *stochastic optimization* which is specifically developed for the support of optimal decisions under uncertainty. We present an example of an application of this methodology to the planning of service provision. The next two domains are *social network analysis* and *statistical mechanics of networks*. The first evolved in social sciences while the latter has its origins in natural sciences. Key issues are the mathematical description of properties of networks and the exploration of principles behind the network generation and evolution that lead to specific network properties. The reader is referred to "introducing network analysis" by Canright/Engø-Monsen in this issue. Another research domain which is referred to as *network formation or link formation games* applies game theoretic concepts to the analysis

ISSN 0085-7130 © Telenor ASA 2008

and formation of networks, where self-interested economic entities build interconnections. Contrary to social network analysis and statistical network mechanics the network generating process focuses on the concepts like Nash-equilibrium, efficiency, and stability. As far as methods of simulation are applied in network formation games this research area intersects with *agent based computational economics*, where the focus lies on the economic network simulation. This approach allows departures from the traditional game theory like imperfections of markets and bounded rationality of the interacting agents. Yet another area of network analysis contains *network design* problems. Here the focus is set on managing physical or material networks in telecommunications, energy transmission, transportation, and others. The key concepts for analysing or evaluating networks are profitability and efficiency. When other concepts like the reliability of data transfer in case of breakdown of links, or the security of data against malicious attacks are introduced then the network design problems find common ground with the other network analysis methodologies mentioned above.

The remainder of this paper will be organized as follows. Section 2 contains a survey and comparative analysis of the network modeling methodologies mentioned above. An overview of our research results with two modeling examples of relations between network operators and service providers and of collaborative service provision is contained in Section 3. The paper concludes with a summary, acknowledgement and an extensive list of the literature.

2 Methodologies for Quantitative Network Analysis

In this section we survey different modern methodologies for network analysis and prepare the ground for a discussion of our current research in Section 3.

2.1 Stochastic Optimization

2.1.1 Introductory Comments on Stochastic Optimization in Telecommunications

Telecommunications has a long tradition concerning the application of advanced mathematical modeling methods. Besides being a consumer of mathematical modeling, telecommunications provided a motivation for the development of areas of applied mathematics. Important chapters of the theory of random processes have their roots in the work of telecommunication engineers. So far this mutual influence was mainly limited to the queuing theory and the theory of Markov processes, but now new decision problems arise which require the application of optimization methods. The recent trends in telecommunications have led to considerable increase in the level of uncertainty which became persistent and multi-faceted. The decision support methodologies which provide adequate treatment of uncertainty are becoming particularly relevant for telecommunications. Here stochastic optimization is the methodology of choice for optimal decision support under uncertainty; see [6],[14],[25]. We start by defining a classification which will serve as a roadmap for the exposition. This classification is made according to the scale of the decision, its relevance within the telecommunications value chain, and the types of uncertainty to be controlled. Besides, different types of uncertainty come into play at different levels. We distinguish three scale levels: *technological*, *network*, and *enterprise* shown in Figure 1. The technological level corresponds to the smallest scale and the enterprise level to the largest and the most aggregated scale.

The *technological level* deals with the design of different elements of telecommunication networks, including switches, routers, multiplexers. Uncertainty on this level is a salient feature of communication requests and flows in the network. Besides, it can arise due to equipment failures. The key decisions are the engineering decisions which define the design for blueprints of these elements. Such blueprints depend on a number of parameters which should be chosen from the point of view of performance and quality of service. Traditionally, performance evaluation of the elements of telecommunication networks was the domain of queuing theory [34]. To be successful the methods of this theory require a specific probabilistic description of the stochastic processes which govern the behavior of communication flows. Usually such a description is not available for new data services, and when it exists, it does not satisfy the requirements of the queuing theory. Stochastic optimization may help to obtain the performance estimates in the cases when

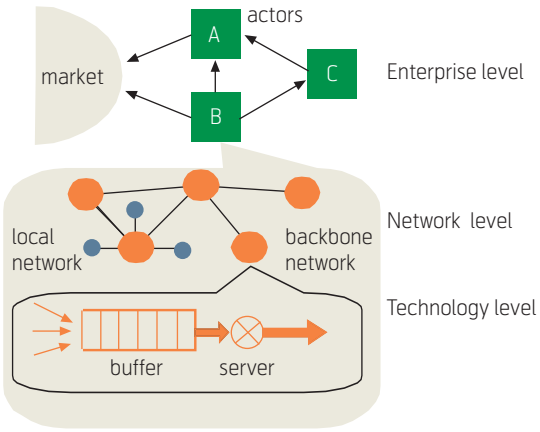


Figure 1 Three modeling levels of telecommunications environment

ISSN 0085-7130 © Telenor ASA 2008

more traditional methods are difficult to apply. See Gaivoronski [18] for one such example.

Network level problems deal with the design and planning of different kinds of networks. The application of stochastic programming on the network level will be discussed in more detail in the next section. Section 2.1.3 also contains an example of a stochastic optimization model for design problems at the network level. For related examples, see [7][40][44][15][13][50][2].

Finally, the *enterprise level* is the highest level of aggregation and looks at the telecommunication enterprise as a member of a larger industrial environment which includes other industrial actors and different consumer types. Decisions involve the selection of the range of services which the enterprise will provide to the market, strategic investment decisions, and pricing policy. Market acceptance of services, innovation process and actions of competition constitute the sources of uncertainty which are not present at the lower levels. Telecommunications and, more generally, the information industry differs in important ways from traditional industries due to the rapid pace of innovation. This leads to the absence of perfect markets and to fundamental non-stationarity which makes it difficult to apply traditional microeconomic approaches based on equilibrium. Stochastic programming models enriched with selected notions of game theory can provide more adequate decision recommendations here. We outline one such model in section 3.1. There is no rigid boundary between various levels since decisions made at each level influence decisions on other levels.

2.1.2 Stochastic Programming for Physical or Material Network Design Problems

Network design issues arise in a variety of industries like for aviation [9], shipping [12], water distribution [39], energy distribution [11], and telecommunications [2][18][29][40], and similar problems. In a simplified manner a network design problem can be described as follows. In different geographic locations demand for or supply of commodities or services can be observed. The network has to be designed such that the supply of the service and demand for it are aligned to each other, i.e. a path must exist from the supplying nodes to the nodes where the demand for the services and commodities occurs. The demand depends on the price of the service. Besides, the demand may not be completely specified for a given price but is subject to uncertainty. Furthermore uncertain events may cause components of the network to break down, such that the transportation or distribution needs to take another path through the network if such a path is available. The network designer is confronted with the decisions of pricing, installing links between nodes and routing the traffic through the network. However, the installation, expansion and maintenance of links and capacity as well as processing and transport of data or commodities are costly. Furthermore if the service delivery must be rejected, penalty costs may be incurred. This is especially the case if network failures are observed or if the demand has been wrongly anticipated. The designer seeks for the optimal network design with respect to the expected return on investment, the expected total costs or another equivalent objective.

In telecommunications, networks differ by scale, purpose, and technology involved. We find access networks, local area networks, fixed or mobile networks, and voice or data networks. The decisions involve the placement of processing and link capacities provided by a given technology in a given geographic area with the aim to satisfy aggregated demand for telecommunication services from different user groups. Decisions are often of a dynamic nature and include several time periods. The main uncertainty here is related to the demand for telecommunication services. Due to quantitative and qualitative explosion of such services, this kind of uncertainty increased considerably during the last decade. There are important additional sources of uncertainty connected with possible network failures and future technology development. Stochastic programming methods provide an added value of identifying the robust network design which within reasonable bounds will accommodate the future demand variations. This is particularly true for stochastic programming problems with recourse and multi-period stochastic programming problems which provide intelligent means for mediation between different and often conflicting scenarios of the future. While traditional design approaches are centered on the minimization of the network costs under technological and quality of service constraints, a systematic application of stochastic programming techniques includes the incorporation of modern tools from corporate finance like the evaluation of real options. Comprehensive models which include pricing decisions and binary variables provide a motivation for further development of this methodology.
In the following section we illustrate the general considerations outlined above by one simple yet typical example of the application of the stochastic programming methodology to network planning under uncertainty.

2.1.3 Example: Planning of Internet-based Information Service

We consider here the deployment of an Internet based information service on some territory like a country

or a region. The service provider on behalf of which the problem is solved can be the network owner, but can also be a virtual service provider which does not possess its own network and leases network from some network owner. We assume that the network itself exists already and that the decision consists in the deployment of servers at the nodes of this network and the assignment of demand generated in different geographical locations to these servers. More particularly we consider a phased introduction of a service where the deployment in phase 1 with unknown future demand is followed by further deployment in phase 2 that is contingent to the trends in the market. The decisions in the latter phase depend on the project profitability which in turn depends on various options embedded in it, e.g. the option to expand, to abandon and to upgrade the technology. Among various aspects of the problem one can also consider the geographical dimension, the uncertainty of demand and costs, the cost structure which includes fixed and variable costs, the competition and substitution between services as well as relations between different market actors, e.g. network providers and service providers.

In the following we present two steps of the model development. Step 1 represents the simplest possible deterministic planning model which assumes the total knowledge of the market and its future development. Step 2 shows how this model with the help of stochastic programming can be transformed into a more adequate model which takes into account the possibilities to adapt to market reactions and to newly available information.

Step 1: Single Period Deterministic Cost Minimization Model

We start by considering only one decision period and full knowledge about demand and other important parameters. Although these assumptions are highly unrealistic, the resulting model sets the stage for more realistic models. In this setting we assume that the deployment program has to satisfy the known demand fully. The service price is assumed to be given such that the revenues become fixed. For this reason the only way of influencing the profit is by minimizing the costs. Let us introduce some notations.

Notations

$i = 1, ..., n$ – index for regions which constitute a territory where a user population generates demand,

$j = 1, ..., m$ – index for possible server locations,

y_j – binary variable which takes the value 1 if the decision is made to place a server at location j, and 0 otherwise,

x_{ij} – amount of demand from region i served by server placed in location j,

f_j – fixed costs for setting up a server in location j,

c_{ij} – costs for serving one unit of demand from region i by server at location j,

d_i – demand generated at region i,

g_j – capacity of server placed at location j.

Model 1. Find the server deployment program $y = (y_1, ..., y_m)$ and assignment of user groups to servers $x = \{x_{ij}\}$, $i = 1, ..., n$, $j = 1, ..., m$ as solution of the problem

$$\min_{x,y} \sum_{j=1}^{m} f_j y_j + \sum_{j=1}^{m} \sum_{i=1}^{n} c_{ij} x_{ij},$$

$$\sum_{j=1}^{m} x_{ij} \geq d_i \text{ for } i = 1, ..., n,$$

$$\sum_{i=1}^{n} x_{ij} \leq g_j y_j \text{for } j = 1, ..., m,$$

where y_j takes values from $\{0, 1\}$ and $x_{ij} \geq 0$. Here the first term in the objective function from the first line represents the fixed costs of the deployment of servers while the second term represents the variable costs for serving demand. The objective function is followed by two groups of constraints. The first group is imposed in order to obtain full demand satisfaction, while the second group shown on the last line contains the capacity constraints. This is a well known facility location model and it will serve as a starting point for developing a stochastic programming model with different scenarios of the future demand and a larger number of deployment phases.

Step 2: Two Period Stochastic Cost Minimization Model

We use the previous model as a building block for creating a more adequate stochastic optimization model which takes into account the key uncertainties of the problem. There are several such uncertainties, and most important here is the uncertain user demand. A natural way to describe this uncertainty is the formulation of several *scenarios* about the future demand development. These scenarios can be obtained from market analysis of similar services and expert estimates. In the simplest case we may think about average, optimistic and pessimistic demand scenarios. Each such scenario is described by the

value of the demand in different regions and by the probability of this scenario.

Two deployment phases are considered: present Phase 1 with known demand, and future Phase 2 with uncertain demand which is described by a finite number of scenarios. The Phase 2 decisions include additional deployment of servers and reassignment of demand to servers in response to the demand development. The decision made during Phase 1 strikes a tradeoff between the minimization of immediate deployment costs and the minimization of average anticipated costs on Phase 2 for additional deployment when demand becomes known. The model follows the framework of stochastic programming with recourse [6]. The formal description of the model is as follows.

Additional notations

$r = 1, ..., R$ – index for demand scenarios,

d_i^r – demand generated by region i under scenario r,

p^r – probability of scenario r,

z_j^r – binary variable which takes the value 1 if under scenario r the decision is made to place a server at location j, and 0 otherwise,

x_{ij}^r – amount of demand from region i served by a server placed in location j under scenario r,

α – coefficient for discounting of the Phase 2 costs to the present.

Each scenario is characterized by a pair (d^r, p^r) where $d^r = (d_1^r, ..., d_n^r)$.

Model 2. Find the Phase 1 server deployment program $y = (y_1, ..., y_m)$, and assignment of user groups to servers $x = \{x_{ij}\}$, $i = 1, ..., n,\ j = 1, ..., m$, as the solution of

$$\min_{x,y} \sum_{j=1}^{m} f_j y_j + \sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij} x_{ij} + \alpha \sum_{r=1}^{R} p^r Q(r, y)$$

subject to the constraints of Model 1. The third term in the expression above represents discounted costs of the Phase 2 deployment averaged over scenarios. The costs associated with scenario r is $Q(r, y)$ and it depends on the Phase 1 deployment decision y. These costs are obtained from the solution of the *recourse* problem for each scenario r:

$$Q(r, y) = \min_{x^r, z^r} \sum_{j=1}^{m} f_j z_j^r + \sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij} x_{ij}^r,$$

$$\sum_{j=1}^{m} x_{ij}^r \geq d_i^r \text{ for } i = 1, ..., n,$$

$$\sum_{i=1}^{n} x_{ij}^r \leq g_j(y_j + z_j^r) \text{ for } j = 1, ..., m,$$

which is similar to Model 1 and chooses the Phase 2 deployment $z^r = (z_1^r, ..., r_m^r)$ and a new assignment of user groups to servers $x^r = \{x_{ij}^r\}$, $i = 1, ..., n,\ j = 1, ..., m$, according to the minimization of fixed deployment costs and variable service costs for a given scenario r. The modern optimization technology permits to solve it for practically important cases, using a combination of commercial solvers like CPLEX or XPRESS with decomposition techniques.

It is important here to note that the deployment decision obtained from the solution of this problem does not aim at the best deployment for any given scenario. This is because the optimal solution for a fixed scenario can be grossly non-optimal if this given scenario does not materialize. Instead, stochastic programming solution aims at obtaining the *robust* decision which will make adaptation to changing demand patterns less painful. More details of stochastic programming approach for network planning are given in Gaivoronski [18].

Evaluation of investment opportunities, real options

The stochastic programming approach allows embedding the modern notions of financial theory and investment science into the process of evaluation of industrial projects. One such important notion is *real options* which represent flexibilities inherent in telecommunication projects [45]. An example where the real option approach can be utilized is the gradual development of a mobile network where new cells are added contingent to an increase of traffic, as opposed to full scale deployment from the start. While for more traditional industries the evaluation techniques can be similar to the evaluation of financial options, for innovative industries with unique projects such approaches are difficult to apply. Stochastic programming models represent an important tool for real option evaluation in such cases. Let us consider some of the options inherent in the example of the service development from above. Here we deal with options to expand, to upgrade technology, to abandon or to convert a part of the infrastructure.

Option to expand (wait and see option). This option is already imbedded in the model outlined above which contains the possibility to add additional servers during Phase 2 contingent to the market trends. The value of this option is obtained by comparing the solution of this model with the solution of the restricted model where there is no additional deployment during Phase 2.

Option to upgrade technology. This is a valuable option because it can dramatically change the project evaluation, especially in an innovative industry like telecommunications. In order to evaluate this option it is necessary to have a closer look at the ways the technology development can affect various components of the model of our example. Namely, the technology development can lead to a decrease in the fixed costs for server installation and/or an increase in the possible server capacities during phase 2. In this case it is necessary to introduce these features into the definition of the scenarios.

Option to abandon. This is a valuable option when the market reaction is uncertain. If demand does not catch up it is reasonable to cut maintenance costs in the regions where demand is weak and possibly recover part of the fixed costs by selling or leasing the server infrastructure.

Results of one such evaluation are represented in Figure 2. This figure shows the dependence of the project value on the service price charged to customers. Three alternatives are shown in this figure. The first alternative is depicted by the green curve and describes the dependence of the project value on the price in the case when no option to expand and no option to upgrade the technology are considered during Phase 2. The second alternative allows an option to expand, but not an option to upgrade technology and is depicted by the black curve. The third alternative shown with the orange line allows both options during Phase 2.

First of all, one notices the jumps on the curves which are due to the discrete character of the decisions. The objective in all three cases is full demand satisfaction. A small increase in price leads to a small decrease in demand which can make a given server redundant with a corresponding stepwise decrease in fixed costs. Another observation confirms the added value of flexibility provided by the options. The value of the project without options is barely positive even for the best choice of the service price. The project becomes decidedly profitable when the option to expand is allowed. There are two regions of profitability with respect to the service price. The first corresponds to an aggressively low service price designed to stimulate large demand and the second corresponds to a less aggressive behavior with high prices and smaller demand. These profitability regions expand when an additional option to upgrade technology is considered. In the absence of options the model recommends defensive behavior with high pricing, while flexibility imbedded in options allows stimulating demand more aggressively with lower prices.

For further details and additional examples of using of stochastic programming models for finding optimal planning decisions under uncertainty in telecom see Gaivoronski [18].

Interaction of market participants

In the problems presented so far the decisions are made by single decision makers who do not have to take into account the strategic behaviour of other market participants. Price choices, traffic routing decisions, and the network deployment are independent from reactions of customers, suppliers and competitors. In reality, however, we have a variety of interacting and mutually reacting players on different decision layers. One player's decisions affect the other players' strategies, and vice versa. This constellation is considered in the approaches presented below. In particular, these are network equilibrium problems (Section 2.2), network interdiction (Section 2.3) network formation games (Section 2.4) and constellations as described in chapter 3 where interdependencies of several network operators and service providers are modelled.

2.2 Network Equilibrium Problems

Network equilibrium models are commonly used for the analysis and prediction of traffic patterns in transportation, distribution or telecommunication networks where congestion occurs. The reader is referred to Nagurney [36] who gives an outline of the historical beginnings of network equilibrium models. Assuming a given network, the users or applications compete for the given resources. They analyse the state of the network and individually optimize flow and routing from a supplying node to a demanding node. Each application's decision changes the state

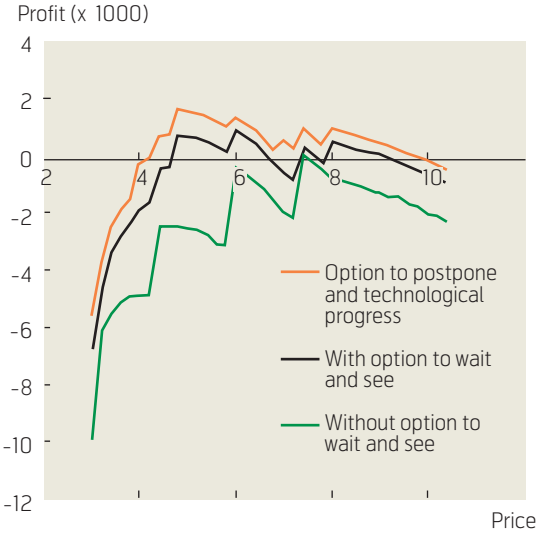


Figure 2 Evaluation of real options in the case of service introduction

ISSN 0085-7130 © Telenor ASA 2008

of the network and so do the optimization problems of the other users or applications. In this section we focus on the network design which is finalized before the network usage and the occurrence of equilibrium. Mathematically we have a bi-level planning problem à la von Stackelberg [46]. In the first level the network designer makes his choices to install, upgrade, or abandon parts of the network such that the operator maximizes revenues or social wealth that stems from the usage of the network. In the second stage multiple network users maximize their wealth under the given network topology and under anticipation of the other users' behaviour. The formulation of network design problems as bi-level mathematical programs goes back to LeBlanc [28] (see also [42]) who studies a public highway network which is planned and implemented by the public sector and used by private individuals. While the government attempts to maximize the social welfare, each individual selfishly pursues own interests like minimizing the travel time when moving from point A to point B in the network.

Marcotte [31] represents a network design model where the equilibrium flow problem is formulated as a variational inequality. The objective of the network designer is to minimize the total traversal time and investment costs of the network, while the users optimize individually their flow traversal time.

2.3 Studies of Survivable Network Design or Network Interdiction

Most of the studies in network design, may they be of an optimization approach or with a game theoretic background, are designed from a cost minimizing or profit maximizing perspective. The field of *survivable network design* adds objectives and measures for maintaining a reliable network in case of failures of network components [35]. However, these failures are assumed to be of an accidental and random nature. Targeted attacks of rational agents who aim at a substantial loss of network performance are studied in the so-called approaches of *network interdiction*. For example Smith/Lim/Sudargho [41] consider a three-level, two-player framework, where the first level network designer constructs a network and sends multi-commodity flows through the network. In the second level an attacker attempts to destroy the network performance by destroying links. Three strategies are considered for the attacker: (a) destruction of the links with largest capacities, (b) destruction of the links with highest initial flow, (c) destruction such that the maximum post-interdiction flow is minimized. While (a) and (b) are heuristics of a bounded rational attacker, (c) is the strategy of a rational player.

2.4 Network or Link Formation Games and Games on Networks

A *network formation game* (also referred to as *link formation game*) is given by a set of players where each player decides individually with what other players he/she wants to create links (connections). The formation of a link causes costs that are either carried by the node that initiated the link or will otherwise be shared by both the nodes. In the link creating process each player pursues individual interests, i.e. he weighs the benefits from being directly and indirectly linked against the costs from initiating, installing and maintaining links. The utility that a player can receive depends on his/her own actions as well as the actions of other directly or indirectly connected players. Links may represent friendships, co-authorships, common research projects, trade agreements, political or economic alliances, and others. Models of network formation can be classified as either static (Jackson/Wolinsky [24]) or dynamic (Jackson/Watts [23]). In the first kind of models the issues involved are the following:

a) Which network topology is efficient?
Different concepts of efficiency can be applied. In the case of strong efficiency the total value of the obtained network is higher than the total value of any alternative network structure. In the concept of Pareto efficiency the value of each single player is considered rather than the total value of the network: for a given value function and allocation rule, a network structure is Pareto efficient if no other network structure exists that gives a higher pay-off to at least one agent, without reducing the pay-off of at least one other agent.

b) Which network topology is stable, i.e. does an equilibrium exist for the network game?
An equilibrium is reached if no player has an incentive to unilaterally change its own prevailing linkages to other individuals. Hence, the network structure will come to a resting point. For the treatment of these issues in static settings see Jackson/Wolinsky [24].

c) How is the value allocated to the individuals in equilibrium?

d) What pay-off structures or allocation rules are necessary for the network to become efficient or stable?

The section on network formation in the survey of Kosfeld [27] of network experiments also gives an overview of some interesting applications. Two of the often cited network formation games are the connections and the co-author model (Jackson/Wolinsky [24]). In the connections model social relations

between individuals are represented as links. Having both direct and indirect relationships to other individuals incurs benefits which may be in the form of friendship, social integrity, access to information, and others. Direct relationships offer the highest benefit, and the longer the path to another individual, the less this benefit becomes. In the specific case of the symmetric connections model, depending on the parameters of the pay-off function, the unique strongly efficient network is either a star, a complete graph or an empty graph. Here the complete graph is a unique pairwise stable network while the star does not necessarily reflect this property. However, an empty graph is not stable. In the co-author model each node represents a researcher who works on different projects. Links represent the fact of two researchers being involved in the same project. The time that two researchers spend within the project determines their synergy. The more projects one researcher has the less time he can spend within the project; hence the less synergy will occur. In this model strong efficiency occurs if there are separate pairs of authors that are connected, and the pairwise stable network can have fully intra-connected components that vary in size.

Slightly different from network formation games are theoretical and empirical studies with respect to games that are played between individuals in populations (see Kosfeld [27] for a survey). In many of these studies the network structure is given and individuals play games on this particular topology. The purpose here is to evaluate the affect of network structure on how the individuals play games with each other. Then different network topologies can be compared with respect to stability and efficiency of the individual decisions. Phan [37] for example studies the prisoner's dilemma played among individuals on different network constellations. In particular the dominance and transition of strategies are compared for a regular network on the one side and a small world network on the other. The players do not have complete information on the whole network. Each player only observes the pay-offs and strategies of his/her neighbours. The decision rule of an agent is to apply the strategy within his clique (consisting of him and his restricted number of neighbours) that gives the maximum payoff. Furthermore, accidental defection by a certain number of players is introduced symmetrically into the network. The results of Phan show that in regular networks the whole population will tend to defect instead of cooperate, i.e. the welfare of the population is reduced to its minimum. When the small world property is introduced the defection does not necessarily spread over the whole population. Hence, the small world network allows obtaining the higher welfare of the population.

In another study of Goyal/Vega-Redondo (2005) costs for establishing links are introduced. These links are then preconditioned for playing the stage game. In this study conditions for connectedness or emptiness of the network are derived. In this case individual decisions take affect on the composition of the network.

However, many network formation games lead to topologies that are not in alignment with the properties of real world networks found in social network analysis and statistical network mechanics (see "Introducing network analysis" by Canright/Engø-Monsen in this issue). Hence, it remains interesting to investigate what economically driven decision rules and processes result in network topologies observed in practical cases.

2.5 Agent-based Computational Economics and Multiagent Networks

Within agent-based computational economics (ACE) complex agent-based systems are studied by means of computerized simulation. The objective is to analyse the dynamics, global properties and patterns of complex systems (like networks or societies) at the macro-level and analyse their emergence from the autonomous, heterogeneous, individualistic, idiosyncratic, self-interested and interacting behaviour of individuals on the micro-level. ACE follows the traditional studies on self-organizing economies originated by Smith, Hayek and Schumpeter. However, only the recent developments in computational power made ACE possible. The advantage of ACE compared to conventional quantitative modeling of agent-systems is that the agents can have a richer heterogeneous internal cognitive structure. However, departing from traditional game theory the individuals are characterized by bounded or procedural rationality.

Normative recommendations are derived on how the individual actions are successful in complex environments or how mechanisms can be imposed by regulators to take a desired effect on the complex system. Adding to Tesfatsion's extensive internet presentation on this topic (http://www.econ.iastate.edu/tesfatsi /ace.htm, 2006) the survey article by the same author [43] gives an introduction and an overview of several applications of ACE. Tesfatsion [43] addresses various fields of ACE studies, among which the following are of particular interest for the discussion of network formation and analysis: (a) ACE research on learning, (b) bottom up modeling of market processes, and (c) formation of economic networks.

In (a) researchers are motivated to find out how different learning schemes affect the outcome of the

simulated system with respect to improved efficiency, global optimality, selection from multiple equilibria, etc. often in contrast to traditional models that presume rational choice as individual behaviour. Learning might be simply imposed or empirically substantiated, and learning schemes may contain self-reflection as well the reflection of other players' strategies (see for example Vriend [47]). One research direction that benefited from the studies of learning within ACE is that of the application of evolutionary algorithms to economic problems. In this area Arifovic [3] gives a survey of research that addresses the following issues: "(1) the convergence and stability of equilibria in the models with unique rational expectations equilibria, (2) the use of the algorithms as equilibrium selection devices in the models with multiple equlibiria, (3) the examination of transitional dynamics that accompanies the equilibrium selection process, (4) examination of learning dynamics that are intrinsically different from the dynamics of the rational-expectations versions of the models." (See Arifovic [3], p. 374.)

Issue (b) addresses the question of how markets organize themselves or how transitions from and to market equilibria take place. This issue is strongly connected to the issue (a) mentioned above, since the market outcome strongly depends on the learning schemes applied by the modeller. This issue is considered in the studies by Balmann et al. [5], who look at the application of a parallel genetic algorithm to an agricultural market problem.

For issue (c), the formation of economic networks, Tesfatsion [43] narrows the research focus with the following questions: "What drives the formation of interaction networks among buyers and sellers? How do these networks evolve over time? What are the social welfare implications of these networks?"

Some studies can be placed at the intersection of statistical network mechanics and ACE. For example Wilhite [48] compares four types of trade networks: (1) completely connected networks, (b) a network of disconnected trade groups, (c) a network of trade groups that are aligned around a ring where one trader of a group is connected to one trader of the neighbour trade group, and (d) small-world networks. The consequences of these network structures for a bilateral trade are studied with respect to the trade-off between market efficiency and transaction costs. Wilhite finds that the small-world trade network provides market-efficiency close to completely connected networks and a reduction of transaction costs as in locally connected networks. He also hypothesizes the existence of micro-level incentives for the evolution of such a network structure, i.e. due to the advantages of a network with small world property, the agents self-organize to such a network type. Other researchers focus even more explicitly on the formation of such networks (see Vriend [47]).

Another ACE approach that uses the results from social network analysis and statistical mechanics of networks is provided by Phan/Pajot/Nadal [38]. Basically, they study regular, random and small world networks of individuals. They study the case of a monopoly that sells a single product to their customers. Customers interact with each other and influence each other's surplus function that each customer maximizes. These network externalities depend on the topology of the network. The surplus function is defined as the idiosyncratic preference for the product plus the social influence through neighbours who also use the product minus the price that needs to be paid for the product. The monopoly's objective is to maximize the profit considering the individual choices of the customers who are affected by their interaction that depends on a certain network structure. The paper shows that the monopoly's price depends on the structure of interaction between customers. Hence, it is recommended for a monopolist to analyse the network structure for deriving optimal decisions. The optimal price and the profit increase with the degree of connectivity and with the range of interaction.

The concept of *agent nets* developed in Gaivoronski [17][19] and Bonatti, Ermoliev and Gaivoronski [8] also belongs to the class of ACE models. In these papers the formal definition of agent nets was developed particularly suited for modeling of industrial relations in the information economy. Based on these ideas the modeling system MODAGENT was developed and used for the analysis of typical constellations of industrial agents in the telecommunications sector.

3 Some Current Research Issues: Competition and Collaboration in the Networked Telecommunication Environment

The previous section gave a broad overview of quantitative models for the analysis of different kinds of networks. Here we give two examples from our current research which utilize some of the methodologies described above for modeling strategic decisions in the telecom market. Both the examples are united by the common objective: provide quantitative models for support of strategic decisions in the situations which are characterized by the following two features:

- Uncertainty about important parameters which influence decisions, like demand, technology, user behaviour, market conditions, etc.;

- Presence of several independent actors who assume different roles and engage in complex relations of competition and collaboration.

Our focus on these two features is due to the observation that they play a more and more important role in advanced industries like telecom or more generally, the information industry compared to more traditional industries. Consequently, from the methodologies presented above we select stochastic programming as an adequate methodology for dealing with complex decisions under uncertainty. It is enhanced by certain concepts borrowed from game theory and network games, a natural choice to represent actors who take independent decisions. In our future research we are planning to expand this analysis by integrating concepts from the agent based computational economics similar to how it was done in Gaivoronski [17][19], and incorporating insights from the statistical mechanics of networks or social network analysis. A promising direction to go is the representation of the market as an interaction system that can be described as a network which shows properties like high clustering, small worlds, and power laws in node degrees. Instead of using traditional aggregated demand functions the market is modelled as a social network and as such builds into the hierarchical decision models of different actors in telecommunications.

For now we look at the following two situations:

Virtual network operators. There are two or more telecom operators who provide a similar service to a population of users. One of these operators, called *network operator* (NO), possesses the entire network infrastructure to provide this service, while others, called *virtual network operators* (VNO), do not operate the network themselves. They need to lease the network capacity from the network operator to provide their service. There is a lot of uncertainty in this environment, including market projections, user response, and mutual knowledge of the operators about parameters of their respective business models.

We develop a model that allows answering the following questions: What are the market conditions under which this relationship will be mutually beneficial? When will all operators continue to offer a service, and when will some of them have to exit from the service provision? What are the responsible bounds that a regulator can impose on the leasing prices? What is the pricing scheme for virtual operators to bear a fair share of the costs for maintaining and developing the network infrastructure?

Provision of advanced mobile data services. Provision of such services involves concerted effort of many actors which assume different roles in the service provision. Some of them will contribute with network capabilities, others with content, still others with organizational effort like brokering or billing. They are all independent actors pursuing their business objectives, and yet they should decide to unite their efforts if a service is to come into being. Services are united in bundles or platforms and they compete between themselves and with traditional services for users' attention. The following issues are addressed: What will distinguish successful services or service platforms from unsuccessful ones in such a dynamic and uncertain environment? Which traditional and new business models should be adopted for service provision? What roles can the actors combine and which combinations are detrimental for the business? We will try to answer these questions by drawing upon developments in stochastic programming and ideas from modern finance and investment science.

3.1 Virtual Network Operators

We use this example to describe a modeling approach for the provision of decision support and strategy evaluation of an industrial agent in complex relations of competition and collaboration with other agents in the telecommunication environment. This is the situation of many telecom service providers nowadays, with a deregulation process and convergence between telecommunications, computer industry and content provision being well under way. The objective of the approach is to provide a set of quantitative decision support tools which would enhance the quality of strategic and tactical decisions.

Microeconomic theory [33] provides important theoretical insights into these issues, especially when the studied system is under conditions of equilibrium. However, classical theory often treats uncertainty inadequately. Unfortunately, central features of today's telecommunication environment are the presence of uncertainty and, usually, the absence of equilibria. This makes many established approaches inapplicable. Therefore we employ techniques that are specially designed to incorporate uncertainty and dynamics in decision models, namely approaches and methods related to stochastic programming [6], [14]. On the theoretical level, such techniques have been under development for a few decades, but only relatively recently has the state of software and hardware allowed large scale applications. We supplement this by selected ideas from game theory because a part

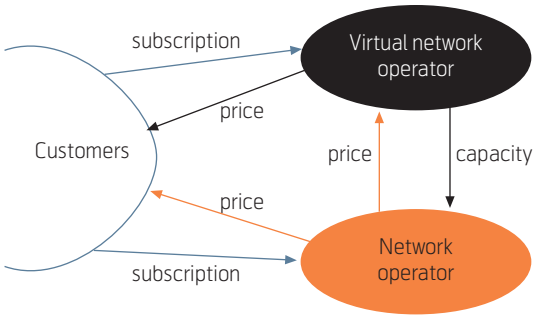


Figure 3 Relations between customers and network operators

of the uncertainty that a given decision maker faces results from actions of other decision makers.

Figure 3 shows relations between service providers and a customer population which we are going to study. The considered time horizon consists of several time periods. We assume that the two operators provide a common market with the same or similar type of service based on the telecommunication network. For delivery of this service they utilize network capacity. Whereas one of the providers owns the network, the other one is a virtual operator without network facilities. The latter needs to lease capacity from the network owner in order to provide the service

Since the aim is to provide decision support tools for a given actor we do not follow the usual economic view on a market "from above", i.e. the maximization of a general welfare [33]. Instead, the point of view of one of the providers is adopted here. His main focus lies on maximizing his own profit or another business performance measure. We take the point of view of the network operator, but the virtual operator could be considered similarly. In order to achieve his goal the network provider formulates predictions of the customer behavior and his rival's responses to his policy. The prediction models depend on a number of parameters with uncertain values, which makes an adequate treatment of uncertainty particularly important.

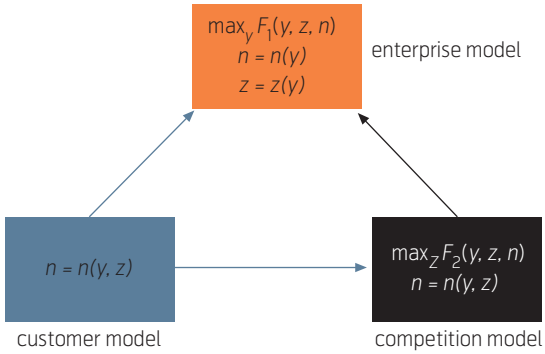


Figure 4 Structure of the model bundle

Following this approach, the decision support model of the network owner consists of a coordinated bundle of submodels: *enterprise model*, *competition model* and *customer model* that are connected as illustrated in Figure 4.

At the beginning of each time period the network operator performs the following steps to determine his optimal decision under the current circumstances:

- Predict the customer response for a given decision and a given competition response using the customer model. This yields the estimate of the customer numbers for both the network operator and the competition.
- Predict the competition response for a given decision using the competition model.
- Select an optimal policy from the enterprise model by using the predictions of the customer and the competition response obtained in the previous two steps.

The following notations are utilized in Figure 4:

y – decisions of the network operator (NO): price y_1 for service provision to customers and price y_2 for capacity leased by his competitors. Besides, the decisions for upgrading and expanding the network capacity can be included here.

z – decisions of the virtual network operator (VNO): price z_1 for service provision and amount z_2 of capacity leased from the NO.

$n = (n_1, n_2)$ – total number of customers of the NO and the VNO respectively. These numbers depend on the respective decisions y and z.

$F_2(y, z, n)$ – performance measure of the VNO like profit, revenue or market share. It depends on both provider's decisions y and z and on the number of his customers $n = n(y, z)$ obtained from the customer model. It comprises the network operator's knowledge about his rival's aims, namely the NO thinks that the VNO chooses his decisions from maximization of this performance measure. More formally, the network operator uses the predicted decision $z(y)$ of the virtual operator which is the solution to the following problem:

$$\max_{z \in Z} F_2(y, z, n(y, z))$$

where Z is the set of admissible decisions of the VNO.

$F_1(y, z, n)$ – performance measure of the NO, which depends on decisions of both providers, y and z, and on the number of his customers $n = n(y, z)$ obtained from the customer model. For a fixed decision y the value of this function is computed using the prediction $z(y)$ of the virtual operator's response and the prediction $n(y) = n(y, z(y))$ of the network owner's customer number. Consequently, the decision y is found by solving the problem:

$$\max_{y \in Y} F_1(y, z(y), n(y))$$

where Y is the set of admissible decisions of the network operator. Both functions are average performance measures where the averages are taken with respect to the values of random parameters which enter the description of the problem, like the customer response to the price change, reciprocal knowledge about the production costs, etc. Besides, the gradual acquisition of information by the actors in a dynamic setting and their response to changing market conditions are also included in the model.

A typical example of this modeling advice is given in Figure 5. It shows how the expected profit of the network operator depends on his pricing decisions y_1 and y_2. The decision space in this example can be divided into four regions:

- *Normal competition.* This regime happens when both service prices and leasing prices are moderate. Both providers are present on the service market and the revenue of the network operator is composed from two parts: service provision and network provision.

- *Network operator service monopoly.* This regime is the result of high leasing prices and moderate service prices. The price of entry to the service market becomes prohibitive and only the network provider develops the service provision capabilities while VNOs stay away.

- *Core business solution.* This is the regime with moderate leasing prices and high service prices. All operators concentrate on their core business, i.e. the network provider maintains and develops the network and leases capacity to VNOs who concentrate on the service provision to customers.

- *Market collapse.* It happens with high leasing prices and high service prices. High leasing prices prohibit the entry of the VNO to the service market while the high service prices scare off the customers. As a result, there is no service provision by any of the operators. Obviously, this regime is to be avoided.

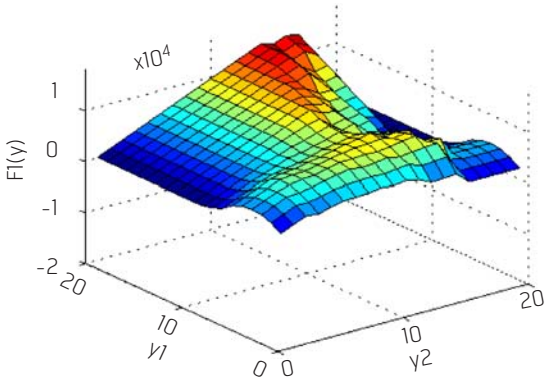


Figure 5 Dependence of profit of the network provider on his decisions

Having this decision support tool, the network operator can decide which regime is more profitable to him or corresponds better to his aims. The network operator also obtains insights into how other business decisions like production cost reductions or technology upgrades will affect his performance, and how his knowledge about competition can affect his strategy.

More details of this example can be found in Audestad/Gaivoronski/Werner [4].

3.2 Provision of Advanced Mobile Data Services

In this section we draw upon the modeling experience of multi-agent environments obtained during the studies of relations between network operators and virtual network operators and enrich it with some modern notions of financial theory and investment science.

General Setting

The design of advanced mobile data services to be carried on 3G networks and beyond is a hot topic in the telecommunication industry and academy. This is because the business success of the provision of such services will define the business success of the mobile operators and other relevant industrial actors in the near to medium future. In this respect considerable attention is given to the design and development of service provision platforms which support a set of tools and basic services that facilitate the development, deployment and customization of specialized services by service providers and even nonprofessional end users. Such platforms are yet to appear in commercial use in the mobile environment, but they already exist on the Internet.

Deployment and operation of service provision platforms and provision of individual services require collaboration of different industrial actors who contribute to the common goal with their individual

capabilities and expertise. One can think about fixed network operators, mobile operators, providers of different information content, internet providers, software developers and other actors who will join their forces to provide a successful service. This gives a rich picture of a service provision environment where a multitude of actors cooperate and compete in order to deliver a wide range of services to customers in a profitable manner.

Understandably, the main efforts in research and development so far have been concentrated on technological and engineering aspects which enable the provisioning of advanced mobile data services. The history of information technology testifies, however, that the possession of the best technological solution is not necessarily enough to assure the business success of an enterprise. A very important and sometimes neglected aspect is the design and evaluation of an appropriate business model which would support the service provision. Business models for service provision by a single actor are pretty well understood, both organizationally and economically. This is the case, for example, for the provision of the traditional voice service over a fixed network. When an actor evaluates the economic feasibility of entering the provision of such service, he can employ quantitative tools developed by investment science, like the estimation of the Net Present Value of such a project [30]. Usually an actor should choose between several service provisioning projects, each characterized by return on investment and the risk involved. Then the portfolio theory [32] suggests a way to balance between return and risk and to select the best portfolio of projects taking into account the actor's risk attitudes. An adequate risk management is especially important in a highly volatile telecommunication environment. Industrial standards in this respect are starting to emerge, originating from the financial industry [1]. Industrial projects in high-tech industries are often characterized by considerable uncertainty and at the same time carry different flexibilities.

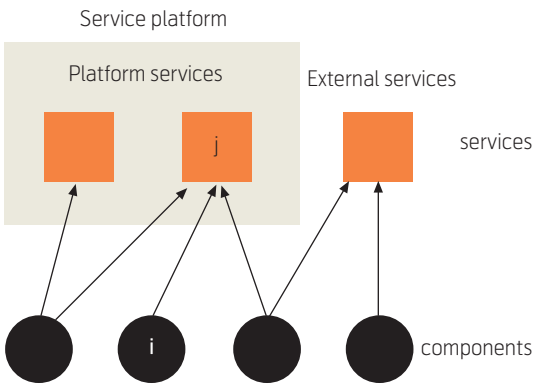


Figure 6 Two level service composition

Stochastic programming provides the optimization models for adequate treatment of uncertainty and flexibilities in the planning of service provision.

Business models for cooperative service provision that involve different constellations of actors are studied to much lesser extent and the quantitative analysis similar to what exists for the single actor case remains a challenge. The methods mentioned above are all developed to be used by a single actor engaged in the selection and risk management of his portfolio of industrial projects. The influence of other actors is present only implicitly on the stage of the estimation of the future cash flows. This is not enough for an adequate analysis of collaborative service provision. Suppose, for example, that a service provider delivers a service to a population of users and receives revenues for this delivery. If a service is composed from modules and if the enablers are provided by different actors then the service provider has to decide about the revenue division between these actors such that it becomes attractive for them to participate in the service composition and provision. This revenue sharing decision together with a concept of what is attractive to other actors should be explicitly incorporated into the evaluation of the profitability of this project.

Our aim here is to contribute to the adaptation and further development of the methods of evaluation and risk management of business models and industrial projects for the case of the collaborative service provision. We look at the actors engaging in a service provision as making a decision about the composition of their portfolio of services to which they are going to contribute. They do this independently following the risk management framework of portfolio theory. The pricing and revenue sharing schemes induce the actors to contribute the right amount of provision capacity to participation in the service provision. We develop a two tier modeling framework which results in the optimal selection of pricing and revenue sharing. This is done by utilizing the approach of stochastic optimization with bi-level structure [4].

Let us outline how this methodology coupled with notions of investment science can be used for decision support and evaluation of business models for collaborative service provision.

Model of Service Structure and Provision

The composition of a service can be quite complex. For the purposes of clarity we use here a simplified description which still possesses the main features of the provision environment important for business modeling. Namely, two levels of the service composition will be considered here as shown in Figure 6.

In this case the service environment is composed of two types of services. The first type comprises services whose structure and provision we are interested in and which we are going to consider in some detail. They can be provided in the context of a service platform and therefore they will be referred to as *platform services*. There will also be *external services* whose structure is of no concern to our modeling purposes. They are present in the model in order to adequately model the environment in which the provisioning of the platform services happens.

The main building blocks of the platform services are service components and/or enablers indexed by $i = 1 : N$ and services indexed by $j = 1 : M$. Here and in the rest of the section we shall use the term *components* as a generic term for software components, enablers and enabler services which compose a service. Components are measured in units relevant for their description, like bandwidth, content volume, etc. The relation between components and services is described by coefficients λ_{ij} which measure the amount of component i necessary for provision of the unit amount of service j. Thus, a service j can be described by vector

$$\lambda_j = \lambda_{1j}, ..., \lambda_{Nj}.$$

A service j generates a revenue v_j per unit of service. This quantity depends on the service pricing which in turn depends on the user behavior and market structure. For the moment let us assume that v_j is a random variable with known distribution which can be recovered from expert estimates and from simulation models that explore the structure of user preferences and market features. The random variables v_j can be correlated due to the service substitution, macroeconomic phenomena and other causes.

Services can be provided by different constellations of actors. Here we consider one such constellation where the actors are the enterprises which have the capability to provide service components assuming different roles. Different constellations can be considered in a similar manner. In this section we shall focus on the two tier structure of the service provision.

- *Component provision layer.* For the matter of simplicity we consider generic actors who provide just one specific component for different services. Such *component providers* can correspond to real actors or to business units of real actors if the real actors fulfill several roles and provide several components. The objective of a component provider is to select a portfolio of services to which this actor will make a contribution. This decision is made on the grounds of balance between projected profit from component provision balanced against the risk of variations in demand and service acceptance among the prospective users of services. In order to quantify this decision process it is necessary to use a simplified profit model for an actor.

- *Service provision and platform provision layer.* There is one actor who provides the service aggregation and organizes the overall service delivery to the end users, this actor will be referred to as a *service provider*. This actor can provide the whole bundle of platform services and will decide which services to include in this bundle, and is then called a *platform provider*. He will collect the revenue from the end users and distribute it among the component providers according to some revenue sharing scheme. This scheme is defined by a vector of revenue shares decided by the service provider

  $$\gamma_j = (\gamma_{1j}, ..., \gamma_{Nj}), \gamma = (\gamma_{11}, ..., \gamma_{N1}, ..., \gamma_{1M}, ..., \gamma_{NM})$$

  such that an actor that contributes with the component i receives the revenue $\gamma_{ij} v_i$. Determination of these revenue sharing coefficients is one of the objectives of the design of the business model for service provision.

Besides platform services the actors can supply components also to external services. The structure of these services is not specified and it is assumed that they are fully described by the revenue v_{ij} generated by provision of the unit of component i to external service $j, j = M + 1, ..., K$.

Component Provision Layer

Let us describe how the component providers decide to join the provision of a particular service. We assume them to be rational economic agents that pursue the aim of maximizing their profit. They select the services to which their efforts contribute similar to how an enterprise will select its portfolio of industrial projects or how a bank would select the portfolio of financial assets for investment. Therefore the set of services to which a provider of a given component contributes will be called his *service portfolio* and we shall utilize portfolio theory [32] in order to model the composition of his portfolio. Portfolio x_i is defined by shares x_{ij} of the component provision capacity that the component provider i allocates to service $j, j = 1, ..., K$:

$$x_i = (x_{i1}, \ldots, x_{iK}).$$

The next step is to define the revenues, costs, profit and return on costs of the component provider.

Revenues:

$$V_i = W_i \left(\sum_{j=1}^{M} \nu_j x_{ij} \frac{\gamma_{ij}}{\lambda_{ij}} + \sum_{j=M+1}^{K} \nu_{ij} x_{ij} \right)$$

where W_i is the provision capacity of the provider of component i.

Costs:

$$C_i = c_i W_i$$

where c_i is unit component provision cost. Here we assume that all component provision capacity is utilized and that the fixed provision costs are included in the variable costs.

Profit:

$$\pi_i = W_i c_i \left(\sum_{j=1}^{M} x_{ij} \left(\frac{\nu_j \gamma_{ij}}{c_i \lambda_{ij}} - 1 \right) + \sum_{j=M+1}^{K} x_{ij} \left(\frac{\nu_{ij}}{c_i} - 1 \right) \right)$$

Return on costs:

$$r_i(x_i) = \sum_{j=1}^{M} x_{ij} \left(\frac{\nu_j \gamma_{ij}}{c_i \lambda_{ij}} - 1 \right) + \sum_{j=M+1}^{K} x_{ij} \left(\frac{\nu_{ij}}{c_i} - 1 \right)$$

Expected return of service portfolio x_i of component provider i:

$$\bar{r}_i(x_i) = \sum_{j=1}^{M} \mu_{ij} x_{ij}$$

This is the performance measure of the service portfolio. Here μ_{ij} is the expected return associated with provision of component i to service j:

$$\mu_{ij} = E(r_{ij}) \text{ where } r_{ij} = \frac{\nu_j \gamma_{ij}}{c_i \lambda_{ij}} - 1 \text{ for } j = 1, ..., M$$

$$\text{and } r_{ij} = \frac{\nu_{ij}}{c_i} - 1 \text{ for } j = M+1, ..., K.$$

and r_{ij} is the random return associated with provision of component i to service j. Its randomness is connected with the uncertainties of revenues, costs and even service composition. It brings risk that the realized return will differ from the expected one. This risk should be measured and controlled.

Risk associated with service portfolio x_i of component provider i:

$$R(x_i) = \text{StDev}(r_i(x_i)) = \text{StDev} \left(\sum_{j=1}^{K} r_{ij}(x_{ij}) \right)$$

We take here the traditional way of financial theory to measure risk with standard deviation of portfolio return [32]. It is also possible to include modern risk measures like Value at Risk or Cash Flow at Risk [1] into the analysis. After having defined the notions of performance and risk we can now follow the approach of portfolio theory [32] in order to obtain the composition of the component provider's service portfolio. This theory looks at the portfolio selection as a trade-off between risk and performance and proceeds as follows.

1. *Construction of the efficient frontier.* Some target average return η is fixed. The risk of the service portfolio is minimized with subject to this target return, i.e. the following problem needs to be solved:

$$\min_x R(x_i)$$

$$\bar{r}_i(x_i) = \eta$$

$$\sum_{j=1}^{K} x_{ij} = 1, \quad \sum_{j=1}^{K} x_{ij} \geq 1$$

The solution of this problem for all admissible values of the target return η will provide the set of service portfolios which are the reasonable candidates to be selected by the component provider i. They constitute the efficient frontier of the set of all possible service portfolios. This concept is illustrated in Figure 7.

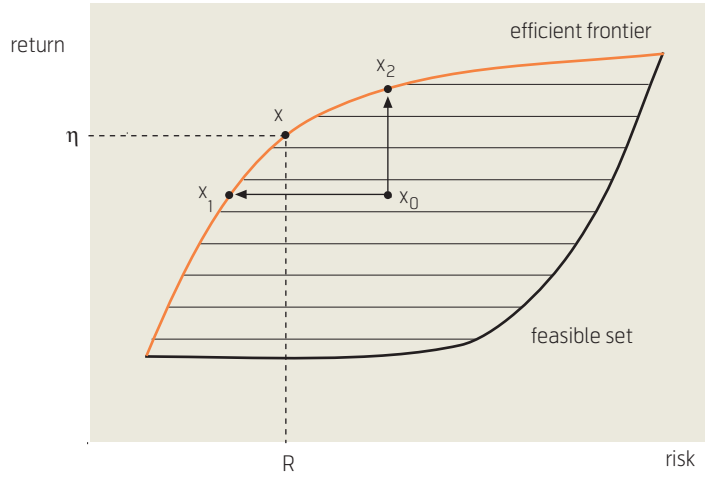


Figure 7 Selection of an efficient service portfolio

Each service portfolio x can be characterized by the risk-return pair defined above. Therefore it can be represented as a point in the risk-return space depicted in Figure 7. The set of such points for all possible portfolios describes all existing relations between risk and return and is called the feasible set. However, an actor will seek the highest possible return among equally risky alternatives or she will seek the lowest possible risk among equally profitable alternatives. Considering Figure 7 it becomes

clear that some of the service portfolios should be preferred to others. For example, let us consider the feasible portfolio x_0. It is clear that portfolio x_2 should be preferred to x_0 by an agent who makes his decision on the basis of return and risk. This is because portfolio x_2 has the same risk as portfolio x_0 and a larger return. Similarly, portfolio x_1 should be preferred to x_0 as well because it provides the same return with a lower risk. Thus, portfolio x_0 is dominated by both portfolios x_1 and x_2 and should not be taken into consideration. The actor whose decisions are guided by risk and return should only consider non-dominated portfolios which constitute the efficient frontier which is depicted by the orange curve in Figure 7.

2. *Selection of the target service portfolio*. The previous step resulted in the selection of a much smaller set of efficient service portfolios from the set of all possible service portfolios. An actor selects his target service portfolio from this efficient set by choosing the trade-off between risk and return. One way to achieve this trade-off is to consider the largest risk an actor is willing to take. Suppose that the value of such risk is R (see Figure 7). Then the actor should choose the portfolio x on the efficient frontier with this value of risk. Suppose that this service portfolio yields a return η. No other portfolio yields a better return without increasing the risk. If an actor is not satisfied with the return η then she should increase her risk tolerance or look for opportunities to participate in the service provision not yet described in this model.

Service or Platform Provision Layer

Here the service or platform provider decides about revenue sharing, pricing, and the composition of the bundle of platform services. Different component providers select their service portfolios observing these decisions as exogenous inputs and having their targets described in terms of return on investment and risk tolerance. However, a service or a platform can become a reality only if the participation in its provision will be consistent with these individual targets. This means that all actors which cover the roles indispensable for provision of a particular service should have this service in their efficient service portfolio. In other words, the service portfolios of the relevant actors should be coordinated and compatible.

Thus, the service or platform provider should make his decisions in such a way as to assure this coordination and compatibility. He does this by choosing his own trade-off between return and risk similar to how it is done on the component layer. The resulting decision structure is similar to what is described in Section 3.1 and is obtained by solving the stochastic optimization problem with bilevel structure.

Architecture of the Decision Support System

We now develop a prototype of a decision support system for the assistance of strategic decisions and the evaluation of business models in multi-agent environment under uncertainty typical in telecommunications. It combines a customized implementation and model development with the use of general purpose mathematical modeling systems and commercial software. The architecture of this system is shown in Figure 8.

The system is composed of four components: data and user interface, a library of service models, a library of mathematical models and a library of solvers.

The *Data and user interface* is implemented in Excel due to its familiarity to potential users. Its purpose is to provide an easy tool for storing and changing the data that describe the service and customer properties, for the presentation of results of business modeling and for providing the capability to the system user to ask what-if questions pertaining to different scenarios. For example, the efficient frontier in Figure 7 is presented to the user through this component.

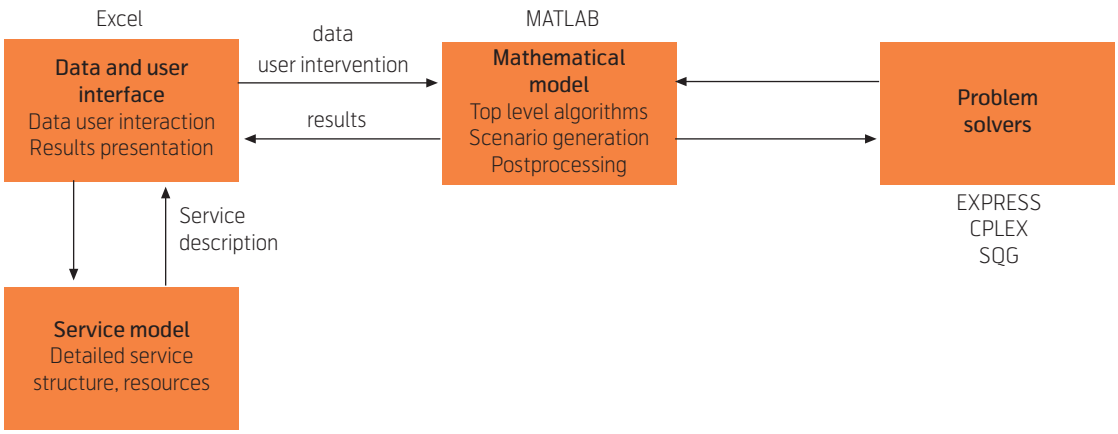


Figure 8 Architecture of decision support system for evaluation of business models of service provision

ISSN 0085-7130 © Telenor ASA 2008

Service models provide the capability to perform modeling of advanced data services with the aim to obtain the aggregated description of the services' composition λ_j.

The *Library of mathematical models* implements the quantitative description of the business decision process of the collaborative service provision from the previous sections. It imports data from the data interface and implements the top level structures and algorithms necessary for the representation and solution of the models described above. The custom algorithms for an analysis and solution of these models are implemented in MATLAB. This component is also responsible for calling external commercial software for solving sub-problems with standard approaches.

The *Library of solvers* contains solvers for linear and nonlinear programming problems and some specialized solvers for stochastic programming problems like SQG in Gaivoronski [20].

The system depicted in Figure 4 is now in an advanced stage of development. In particular the service model component and some mathematical models of service provisioning were implemented in MATLAB.

4 Conclusions

Stochastic optimization coupled with the notions of modern investment science and game theory constitute a powerful tool for evaluation of business models and support for strategic decisions under risk and uncertainty in the multi-agent networked telecommunications environment.

Many relevant issues remain beyond the scope of this paper and will be treated in our future research. These include different actor constellations, combinations of roles by an actor, evaluation of the whole service provision platform, modeling of flexibilities and uncertainties inherent in the service provision, the life cycle of a service, and others.

Another important objective to pursue on the methodological level is to integrate approaches and findings of computational multi-agent economics and statistical mechanics of networks. Particularly relevant is the description of market trends and behavior by means of these approaches and to consider them appropriately when evaluating decisions and strategies of telecommunication companies.

5 Acknowledgement

The authors are grateful to Dr. Kenth Engø-Monsen and Dr. Josip Zoric for stimulating discussions.

6 References

1 *Amendment to the capital accord to incorporate market risks*. Bank for International Settlements, 1996.

2 Andrade, R, Lisser, A, Maculan, N, Plateau, G. B&B frameworks for the capacity expansion of high speed telecommunication networks under uncertainty. *Annals of Operations Research*, 140 (1), 49–65, 2005.

3 Arifovic, J. Evolutionary Algorithms in Macroeconomic Models. In: *Macroeconomic Dynamics*, 373-414, 2000.

4 Audestad, J-A, Gaivoronski, A, Werner, A. Extending stochastic programming framework for modelling of several decision makers: pricing and competition in telecommunication sector. *Annals of Operations Research*, 142, 19-39, 2006.

5 Balmann, A. Applying Parallel Genetic Algorithms to Economic Problems: The Case of Agricultural Land Markets. In: *Proceedings of the Tenth Biennial Conference of the International Institute of Fisheries Economics and Trade*, Corvallis, Oregon, USA, July 10-14, 2000.

6 Birge, J R, Louveaux, F. *Introduction to Stochastic Programming*. New York, Springer, 1997.

7 Bonatti, M, Gaivoronski, A, Lemonche, P, Polese, P. Summary of some traffic engineering studies carried out within RACE project R1044. *European Transactions on Telecommunications*, 5, 207-218, 1994.

8 Bonatti, M, Ermoliev, Y, Gaivoronski, A. Modeling of multi-agent market systems in the presence of uncertainty: The case of information economy. *Journal of Robotics and Autonomous Systems*, 24 93-113, 1998.

9 Büdenbender, K, Grünert, T, Sebastian, H-J. A Hybrid Tabu Search/Branch-and-Bound Algorithm for the Direct Flight Network Design Problem. In: *Transportation Science*, 364-380, 2000.

10 Canright, G S, Engø-Monsen, K. Roles in Networks. In: *Science of Computer Programming*, 195-214, 2004.

11 Carrano, E G, Takahashi, R H C, Cardoso, E P, Saldanha, R R, Neto, O M. Optimal substation location and energy distribution network design using a hybrid GA-BFGS algorithm. In: *IEE Proc.-Gener. Transm. Distrib.*, 919-926, 2005.

12 Christiansen, M, Fagerholt, K, Ronen, D. Ship Routing and Scheduling: Status and Perspectives. In: *Transportation Science*, 1-18, 2004.

13 Dempster, M, Medova, E. Evolving system architectures for multimedia network design. *Annals of Operations Research*, 104, 163-180, 2001.

14 Ermoliev, Y, Wets, R J-B (Eds.). *Numerical Techniques for Stochastic Optimization*. Berlin, Springer Verlag, 1988.

15 Fantauzzi, F, Gaivoronski, A A, Messina, E. Decomposition methods for network optimization problems in the presence of uncertainty. In: Pardalos, P, Hearn, D, Hager, W (Eds.). *Network Optimization, Lecture Notes in Economics and Mathematical Systems*, 450, 234-248, 1997. Berlin, Springer.

16 Gaivoronski, A, Zoric, J. Evaluation and design of business models for collaborative provision of advanced mobile data services: portfolio theory approach. *Proceedings of 2008 INFORMS Telecom Conference*.

17 Gaivoronski, A. Modeling of Complex Economic Systems with Agent Nets. In: Banzhaf, W et al. (Eds.). *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference*, Orlando, FL, USA, July 13-17, 1999. San Francisco, CA, Morgan Kaufmann.

18 Gaivoronski, A. Stochastic optimization in telecommunications. In: Pardalos, Resende (Eds.). *Handbook of Optimization in Telecommunications*. Springer Science+Business Media, 2006, 761-800.

19 Gaivoronski, A. Agent Nets: a methodology for evaluation of enterprise strategies in information economy. *Telektronikk*, 94 (3/4), 33-48, 1998.

20 Gaivoronski, A. SQG: Stochastic programming software environment. In: Wallace, S W, Ziemba, W T (Eds.). *Applications of Stochastic Programming*. SIAM & MPS, 2005.

21 Gaivoronski, A. Stochastic Programming Approach to the Network Planning under Uncertainty. In: Sciomachen, A. *Optimization in Industry 3: Mathematical Programming and Modelling Techniques in Practice*. Wiley, 1995, 145-163.

22 Goyal, S, Vega-Redondo, F. Network Formation and Social Coordination. In: *Games and Economic Behavior*, 178-207, 2005.

23 Jackson, M O, Watts, A. The Evolution of Social and Economic Networks. In: *Journal of Economic Theory*, 265-295, 2002.

24 Jackson, M O, Wolinsky, A. A Strategic Model of Social and Economic Networks. In: *Journal of Economic Theory*, 44-74, 1996.

25 Kall, P, Wallace, S. *Stochastic Programming*. New York, Wiley, 1994.

26 Klos, T B, Nooteboom, B. Agent-based computational transaction cost economics. In: *Journal of Economic Dynamics & Control*, 503-526, 2001.

27 Kosfeld, M. Economic Networks in the Laboratory: A Survey. In: *Review of Network Economics, CRA International*, 19-41, 2004.

28 LeBlanc, L J. An Algorithm for the Discrete Network Design Problem. In: *Transportation Science*, 183-199, 1975.

29 Lisser, A, Ouorou, A, Vial, J-P, Gondzio, J. *Capacity Planning Under Uncertain Demand in Telecommunication Networks*. Research Funded by CNET-France Telecom. 1999.

30 Luenberger, D. *Investment Science*. Oxford University Press, 1998.

31 Marcotte, P. Network design problem with congestion effects: A case of bilevel programming. In: *Mathematical Programming*, 142-162, 1986.

32 Markowitz, H. *Portfolio Selection*. Blackwell, second edition, 1991.

33 Mas-Colell, A, Winston, M D. *Microeconomic Theory*. Oxford University Press, 1995.

34 Medhi, J. *Stochastic Models in Queueing Theory*. Boston, Academic Press, 1991.

35 Myung, Y S, Kim, H J, Tcha, D W. Design of communication networks with survivability constraints. In: *Management Science*, 238-252, 1999.

36 Narguney, A. *Influence of Beckmann, McGuire, and Winsten's Studies in the Economics of Transportation on Innovations in Modeling, Methodological Developments, and Applications*. Prepared for the Panel: Studies in the Economics of Transportation: A Retrospective, at the 50th North American Regional Science Association Meeting in Philadelphia, Pennsylvania, November 2003.

37 Phan, D. Small Worlds and Phase Transition in Agent Based Models with Binary Choices. *4° workshop on Agent-Based Simulation*, Montpellier, April 28-30, 2003.

38 Phan, D, Pajot, S, Nadal, J P. The Monopolist's Market with Discrete Choices and Network Externality Revisited: Small-Worlds, Phase Transition and Avalanches in an ACE Framework. *Ninth annual meeting of the Society of Computational Economics University of Washington*, Seattle, USA, July 11-13, 2003.

39 Savic, D A, Walters, G A. Genetic Algorithms for Least-Cost Design of Water Distribution Networks. In: *Journal of Water Resources Planning and Management*, 67-77, 1997.

40 Sen, S, Doverspike, R D, Cosares, S. Network planning with random demand. *Journal of Telecommunications Systems*, 3, 11-30, 1994.

41 Smith, J C, Lim, C, Sudargho, F. Survivable network design under optimal and heuristic interdiction scenarios. In: *Journal of Global Optimization*, 181-199, 2007.

42 Suh, S, Kim, T J. Solving a nonlinear bilevel programming model of the equilibrium network design problem for Korea. In: *Papers in Regional Science*, 47-59, 2005.

43 Tesfatsion, L. *Agent-Based Computational Economics*. ISU Working Paper No.1, Iowa State University, 2003.

44 Tomasgard, A, Audestad, J, Dye, S, Stougie, L, der Vlerk, M V, Wallace, S. Modelling aspects of distributed processing in telecommunications networks. *Annals of Operations Research,* 82, 161-184, 1998.

45 Trigeorgis, L. *Real Options: Managerial Flexibility and Strategy in Resource Allocation*. Cambridge, MA, MIT Press, 1996.

46 von Stackelberg, H. *The Theory of the Market Economy*. Oxford University Press, 1952.

47 Vriend, N J. An illustration of the essential difference between individual and social learning, and its consequences for computational analyses. In: *Journal of Economic Dynamics and Control*, 1-19, 2000.

48 Wilhite, A. Bilateral trade and 'small-world' networks. In: *Computational Economics*, 49-64, 2001.

49 Wood, R K. Deterministic network interdiction. In: *Mathematical and Computer Modeling*, 1-18, 1993.

50 Yen, J, Schaefer, A, Smith, C. A stochastic SONET network design problem. In: *9th International Conference on Stochastic Programming*, Berlin, August 25-31, 2001.

Denis Becker graduated from the Chemnitz University of Technology, Germany in 2001 with a Master in Business Management with emphasis on Finance, Investment Analysis, and Management Accounting. He then joined the Chemnitz University of Technology as an assistant lecturer for a period of five years. Since 2006 he has been a PhD candidate at the Department for Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU), Trondheim. In his research he pursues the application of financial and economic models to decision problems in telecommunications under the presence of multiple decision makers and under uncertainty. Besides, his interests lie in network analysis and formation.

email: denis.becker@iot.ntnu.no

Alexei Gaivoronski obtained his PhD from Moscow Institute for Physics and Technology in applied mathematics specializing in operations research and optimization. He has since worked in academy (Ukraine, Austria, Italy, Norway) and telecommunications industry (Italy). He is Professor of Industrial Economics and Operations Research at the Department of Industrial Economics and Technology Management of Norwegian University of Science and Technology (NTNU). His research interests are optimization and decision support systems under uncertainty ranging from theory to industrial applications to software, and he has published numerous papers on these subjects. His applied intersts are focused on telecommunications planning, tele- and information economics and finance.

email: alexei.gaivoronski@iot.ntnu.no

Web Link Analysis: Estimating a Document's Importance from its Context

JOHANNES BJELLAND, GEOFFREY S CANRIGHT, KENTH ENGØ-MONSEN

Johannes Bjelland is researcher in Telenor R&I

Geoffrey S. Canright is senior researcher in Telenor R&I

Kenth Engø-Monsen is senior researcher in Telenor R&I

The World Wide Web is a fascinating, useful, and important phenomenon of our time. It is also a graph – that is, a network, in which the nodes are Web pages and the links are one-way hyperlinks. This network view of the Web is basic to the challenging but practical endeavour of Web link analysis. Web link analysis takes the network of hyperlinks as a given input, and uses this information to give a quantitative score which estimates the importance of each page in the graph. In this article we give a pedagogical overview of the main ideas around, and mathematical approaches to Web link analysis.

1 Introduction

Consider a reader who has just gotten a machine to help her to find a document of interest. More precisely: the reader hopes that the machine has delivered one or more documents of interest. This is an extremely common situation in the Information Age; and it presents interesting challenges – some of which we will address in this article.

The reader normally has two types of criteria in mind for a 'good' document. First, it should be *relevant* to the user's current focus or need – for example, the purchase of new cars in Malaysia. Second, the document should be of high *quality*. Clearly, each of these criteria is important for the reader. Furthemore, each is difficult for a machine to evaluate; but the second one – quality – is perhaps the most difficult. After all, human readers can debate endlessly what constitutes (or exemplifies) high quality, and what does not; how then can a machine say something useful about this question?

Herein lies the beauty of the approach termed 'Web link analysis' (or simply 'link analysis' or LA): in essence, link analysis involves using a machine to 'digest' the collective recommendations of millions of human readers. The output of this process is a numerical score, which estimates a document's 'quality'. Since a high LA score for a document may be interpreted to mean that the document has a high degree of collective recommendation, it is perhaps more precise to replace the word 'quality' with the word 'importance' or 'authority'. We will use all of these terms fairly loosely in this article, as they all involve the *interpretation of a score* – which is itself normally defined quite precisely, as the output of the mathematical process termed link analysis.

Let us now flesh out and justify the claims in the preceding paragraph. We focus on documents which are Web pages. These documents together make up the World Wide Web or WWW. A common, and highly useful mechanism which is operative in the WWW is the *hyperlink*. A hyperlink is essentially a pointer from one Web page (document) to another. Many (but not all) hyperlinks are laid down by human writers – typically, the author of the 'pointing' page. This author is then, with this pointer, sending a reader of her own page from that page to another page – which she may or may not have written, but has (typically) read. It is implicit in this offer to send the reader that the author of the pointing page is (again, typically) *recommending* the pointed-to page, as being worth a visit.

Web link analysis then involves gathering information on all hyperlinks existing among the document set of interest (the WWW, a domain of the same, an Intranet, etc), building (in some form, in a machine) the Web graph from these hyperlinks, and extracting, for each document in the set, a numerical score from this Web graph, which estimates the importance of the document. Thus we see that Web link analysis is justified, to the extent that the following statements are true:

1 The documents are well linked by hyperlinks.

2 The hyperlinks are laid down by humans, who have read the pointed-to pages.

3 The hyperlinks may be interpreted as a kind of recommendation.

These three statements sum up the 'why' of Web link analysis. None of them is entirely true; but, in the short history of Web link analysis, they have proven to be true enough to give useful results.

Besides the 'why', we have also in the above summarized the 'what' of Web link analysis (hyperlink graph $\rightarrow$ importance score). There remains of course the 'how', as symbolized by the arrow '$\rightarrow$' in the preceding sentence. There are many approaches to this problem. We will devote the bulk of this article (Section 3) to a discussion of several approaches to the

 ISSN 0085-7130

problem of how to get a useful importance score from a hyperlink graph. Prior to that, in the next section, we will give a brief discussion of the nature and structure of the Web graph. Clearly, this graph is a kind of network, in the sense of this issue of *Telektronikk*; but it is special in that the links are one-way. A useful and compact name for such a network is *'directed graph'*. Section 2 gives an overview of the properties of directed graphs in general, and of the WWW more specifically. Then Section 3 presents and discusses a number of approaches to link analysis. Section 4 places Web link analysis in context, by sketching how link analysis, along with the resulting scores, is integrated into a 'generic' search engine. Finally, in Section 5, we look to the future, presenting some challenges to the current approaches to Web link analysis, and speculating on future developments. In particular, we will in the end tie the entire discussion (at least loosely) to telecommunications, by focusing on the likely migration of search to mobile devices, and on the challenges and trends that we see in that arena.

2 The Web Graph

We recall that the Web graph is – like other networks/graphs – a set of nodes, connected by links. Here the nodes are documents, and the links are one-way hyperlinks. The resulting graph is both abstract and concrete. The concreteness comes from the fact that a hyperlink is a measurable, physical object, and a node is a Web page from which we as human readers may derive useful information, pleasure, pain, annoyance, etc. Once one resorts to the language of nodes and links, however, the Web graph quickly becomes an abstract object as well. We will seek to keep both of these sides of the Web graph alive in the discussion to follow. We start with an abstract discussion, focused on this question: how does a directed graph differ from an undirected graph?

2.1 Directed Graphs

We start with the simpler case – an undirected (or symmetric) graph. Undirected links are termed 'symmetric' because an undirected link **A–B** implies a path from A to B and from B to A, and hence is equivalent to **A↔B**. We typically study *connected* graphs – i.e. graphs which do not decompose into several pieces, without removing one or more links. Connected symmetric graphs are very simple: for *any* two nodes *i* and *j* in the graph, there is a path from *i* to *j*.

Directed graphs are built up of one-way links: **A→B**. This statement does not exclude the possibility of reciprocal one-way links, i.e. that both **A→B** and **A←B** are present; however, neither is this guaranteed. The above definition of a connected graph may also be applied to directed graphs. This is equivalent to the more conventional definition – that a directed graph is connected if, when all links are made symmetric, there is a path from *i* to *j* for any two nodes *i* and *j* in the graph. We will (unless stated otherwise) focus entirely on connected graphs in this article.

Now we can say something new about directed graphs: in general, when we do *not* ignore the direction of the links, there is *not* necessarily a (directed) path from *i* to *j*, for any two nodes *i* and *j* in a connected directed graph. This is clear from the 'simplest possible' directed graph, the two-node graph **A→B:** there is a path from A to B, but not from B to A.

This single, negative statement distinguishes directed graphs from undirected graphs. Now we build from this property, in a positive way. One can find a set C of nodes in a directed graph, such that, for any two nodes *i* and *j* in the set C, there is a path from *i* to *j*. This set is called *strongly connected*; and if we make this set as large as possible – without killing its property of being strongly connected – then we get a set (subgraph) which is called a *strongly connected component* or SCC. SCCs are very useful in thinking about directed graphs. Any directed graph may be uniquely decomposed into its strongly connected components – that is, every node belongs uniquely to one and only one SCC. Furthermore, given two SCCs C1 and C2, any links connecting the two are either (i) all in the same direction, e.g. **C1→C2**, or (ii) non-existent. A two-way connection such as **C1↔C2** is impossible, since then the two SCCs collapse to one. For the same reason, there are no cyclic connections of the form **C1→C2→C3→ … →C1**, since the entire cycle comprises in fact a single SCC.

Thus, movement along directed paths in a directed graph takes two forms: 'circulation' from any node to any other *within* an SCC; and one-way or 'downhill' flow *between* SCCs. This overall one-way flow in directed graphs, which is found at the coarse (inter-SCC) scale (see Figure 1), is quite distinct from the purely circulatory nature of connected undirected graphs; and it has interesting consequences. For example: unless the directed graph is composed of a single SCC, there must always be one or more SCCs which have no inlinks from other SCCs; these are called *source* SCCs. Similarly, there must be one or more 'ends' to the downhill flow – SCCs without outlinks, termed *sink* SCCs.

If we now recall the real, concrete Web graph, our intuition might lead us to say that 'most' Web pages participate in some kind of large-scale community of pages, with circulation from any page to any other.

We will see that this intuitive picture is not particularly misleading. Yet, at the same time, we reiterate that – because the Web graph is a directed graph, which is almost certainly not a single SCC – it *must* have one or more sets of pages for which there are no inlinks to the set. These are the source SCCs of the Web graph. Similarly, the Web graph is guaranteed to have one or more sink SCCs – sets of pages such that a reader who arrives to the set via a hyperlink will not find any path leading out again. Furthermore, we cannot expect just one single, almost-all-encompassing intermediate SCC lying between these sources and sinks. The Web graph is built up by billions of uncoordinated individual acts; and its structure (as we will see in the next subsection) reflects its anarchic and self-organized nature.

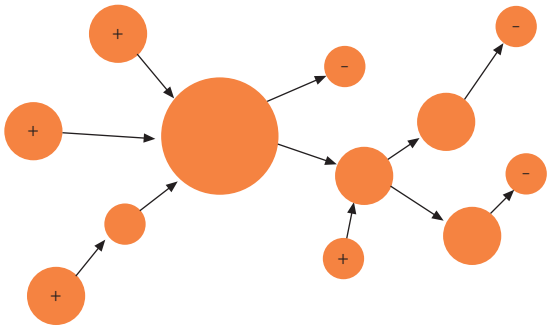


Figure 1 Generic structure of a directed graph. Each strongly connected component or SCC (see text) is indicated with a circle here. The (one-way) direction of links between SCCs is indicated with arrows. We see that the arrows take one from 'source' SCCs (marked with '+'), via (typically) intermediate SCCs, to 'sink' SCCs (marked with '-')

2.2 Structure of the Web Graph

Figure 2 shows schematically the structure of the Web graph, as measured by Broder et al. in 2000 [1]. Subsequent measurements have confirmed this structure – while of course the number of pages in each component has grown. (As of the time of writing, the number of 'indexable' Web pages is estimated to be over 20 billion [2].)

The structure in Figure 2 has been called a 'bow tie' – for obvious reasons. The 'knot' of the bow-tie is a giant SCC (GSCC), which lies *centrally* (in an important sense) in the Web graph. The IN component represents another huge set of nodes. A Web surfer who starts in the IN component may reach the giant SCC from a node in IN; but no node in IN is reachable from the GSCC. IN should be thought of as many SCCs, many of which (but not all) are source SCCs. Similarly, nodes in OUT may be reached from the GSCC – but there is no path from any node in OUT into the GSCC. Also, we see that some paths which leave IN do not reach the GSCC; instead they terminate in 'out-Tendrils' to which IN is linked. OUT has its corresponding 'in-Tendrils' (which must include some source SCCs).

Still we have not exhausted the possible categories of Web page positions in the Web graph. We see from Figure 1 that there are paths from IN to OUT which never reach the GSCC. Nodes in such paths lie in the 'Tubes'. Finally, there are pages which are not even connected to the large, connected structure which is composed of IN, GSCC, OUT, Tendrils, and Tubes. These pages are noted as 'Disconnected components' in Figure 2.

We illustrate the schematic structure of Figure 2 in more explicit form in Figure 3. Here we have taken data from a small subgraph (1222 nodes) of the WWW. These nodes are Web pages of political bloggers in the US [3]. We note that even this small piece mimics remarkably well the overall structure shown in Figure 2 (disconnected components have been omitted, so that only the largest connected component is shown in Figure 3). In particular, there is a large and well-linked GSCC, with 793 nodes. Also, all the other sets illustrated in Figure 2 (IN, OUT, Tendrils, and Tubes) are present in our 1200-node 'slice' in Figure 3 – with the one exception that the 'Tubes' have no nodes – they are only represented by direct links from IN to OUT.

We see from the numbers in Figure 2 that, even though the GSCC of the WWW is in an important sense both central and dominant in the Web graph, it does not contain a majority of the pages. This struc-

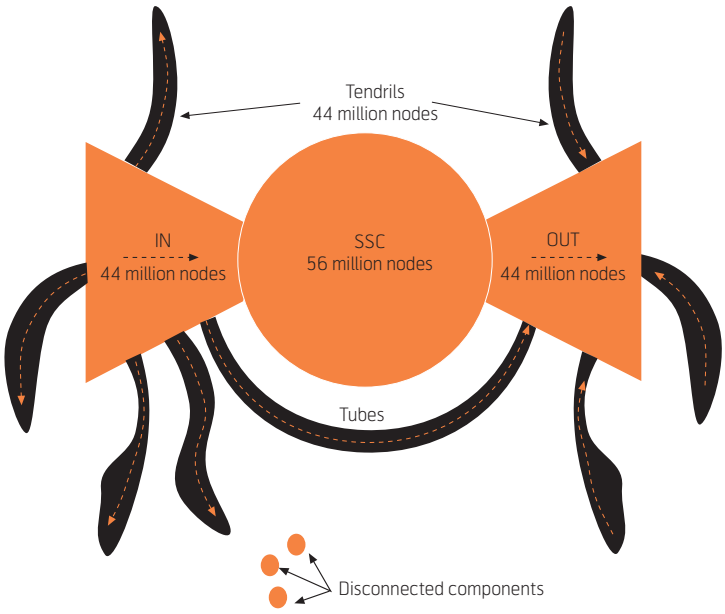


Figure 2 Schematic structure of the World Wide Web. Taken from [1], 2000

ISSN 0085-7130 © Telenor ASA 2008

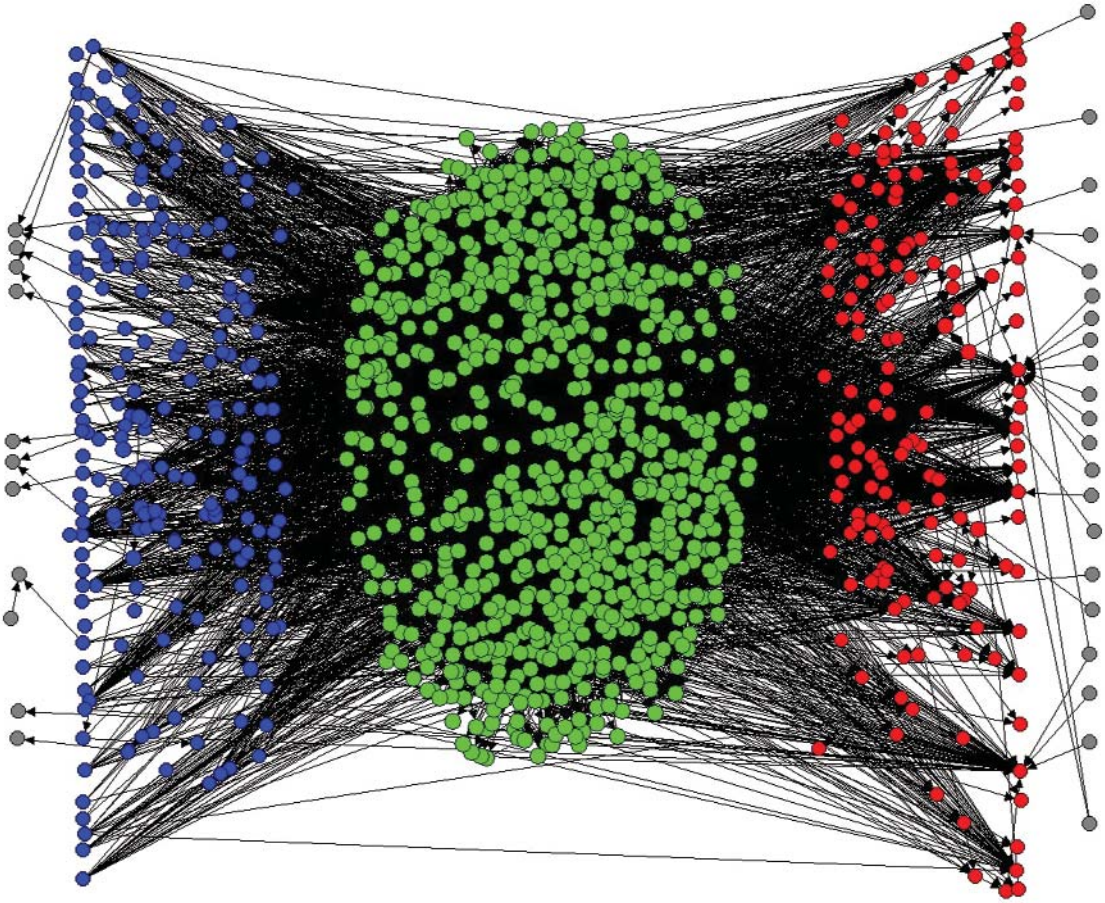


Figure 3 A visualization of a small subgraph of the Web. The nodes are political bloggers. Blue color denotes the IN set, green the GSCC, red is OUT, and grey represents Tendril nodes

ture thus poses some clear challenges to Web search engines. For one thing, how does one find all the pages? Clearly – because of the one-way nature of movement on the coarse scale – it is not enough to simply start somewhere and follow all outlinks. Furthermore, it is impossible to move 'upstream' from any given Web page, since all inlinks pointing to any page currently in view are themselves invisible, as they lie in another page. The problem of finding all the Web pages – given the structure shown in Figure 2 – is the problem of *crawling* the Web. This problem must be solved in order to build an adequate Web graph in a database – a prerequisite to good link analysis.

We see clearly from Figure 2 a second problem, which is even more pertinent to link analysis: if we view hyperlinks as recommendations, then there are clearly pages which receive little or no collective recommendation from the Web graph. For example, pages lying in source SCCs may recommend one another; but no page from 'outside' this set recommends any of these. Similarly, IN is in some sense more weakly recommended than OUT. The point is that the asymmetric nature of hyperlinks seems to favour 'selfish' pages (and sets of pages) which manage to accumulate recommendations, without giving out recommendations. (Perhaps the reader can call to mind individuals which fit this description.) Put differently: the arrows themselves in Figure 2 show that there is a real tendency for 'recommending weight' to accumulate on the 'downstream' side. In some cases (as we will see below) this tendency takes the extreme form that 'upstream' SCCs may receive *no* weight (and hence a score of zero), even though they are far from being isolated, and may be, by most criteria, important. We will call this generic tendency the 'sink problem'. We use the word 'problem' because some approaches to link analysis give poor or useless results unless some corrective measure is applied – a 'sink remedy'.

We mention briefly another important practical problem with Web link analysis. The fact is that statement 3 (of the three statements proposed above for justifying Web link analysis) is not true for most links. In fact, most hyperlinks are "internal" links, i.e. links laid down so as to connect pages which each lie internal to a single site. Such links may also be called navigational links, since their typical purpose is skewed much more towards navigation ('Return to Home Page') than towards recommendation. A search

engine using LA must find practical methods for classifying hyperlinks (external/internal, i.e. roughly, recommendation/navigation), and then for removing – or at least reducing the effects of – the internal links.

In summary: Figure 2 reveals the Web graph as a somewhat wild and woolly community of Web pages, which is built up by the combined, but largely uncoordinated actions of humans and machines. It is the task of Web link analysis to extract meaningful and useful information from this structure – one which, furthermore, is constantly growing and changing. This is a formidable task; and yet, remarkably enough, it is not impossible.

3 Approaches to Web Link Analysis

Now we address some of the most basic approaches to Web link analysis. We recall that the problem is to take the Web graph (suitably corrected for internal links, sinks, etc) as input, and to produce a set of scores – one for each Web page – as output. A good LA algorithm will of course give scores which are reasonably interpreted as importance scores.

We discuss the main algorithms roughly in order of historical appearance.

3.1 Link Popularity

Here the idea is very simple: a Web page is a good page if it has many pages pointing to it – and hence recommending it. Therefore, one can simply count the inlinks to a given Web page to get a quantitative measure of its quality or importance. The resulting score k^{in} is commonly termed the 'link popularity' of the page – since it is entirely determined by the number of 'votes' given to a page.

This idea is intuitively reasonable. Also, it is very easy to calculate: once the full Web graph is obtained, the 'invisible' inlinks become fully visible, and the counting of these is a computational task which can be done very easily. The obtaining of k_i^{in} for every page i is equivalent to multiplying a matrix times a vector once. To see this, let A be the *adjacency matrix* of the Web graph, such that $A_{ij} = A_{i\to j}$ = 1 if i points to j, and 0 otherwise. Then

$$k_i^{in} = \sum_j A_{j\to i} = (A^T \mathbf{I})_i. \qquad (1)$$

Here A^T is the transpose of A, and $\mathbf{I}$ is a vector of ones.

Now we come to the 'dark side' of Web link analysis, which is commonly known as 'spam'. More specifically, the application of link popularity as an importance measure is vulnerable to corruption from *'link spam'*. That is, many Web page authors have a strong interest – financial, and/or otherwise – in enhancing the visibility of their page(s), by enhancing their importance scores. An unscrupulous person UP can then simply make a large number of 'dummy' pages, which are created for the sole purpose of pointing to UP's real Web pages, and so giving them a very high link popularity score. This kind of artificial boosting will not be detected by the simple application of Eq. (1) above; instead, one must examine the pointing pages, their contents, the patterns of their links, etc. for suspicious signs. And of course once UP knows what kinds of suspicious signs are used as tests, he or she can often find ways to camouflage the dummy pages so as to pass the current tests.

In short, link spam is a problem that (i) can never be totally and perfectly solved, and so (ii) will exist as long as there are real incentives to boost one's own visibility. The link popularity approach to LA is particularly vulnerable to link spam – and it is this vulnerability that has prompted the investigation of other methods for Web link analysis.

3.2 HITS

The fundamental idea that may be expected to correct for the effects of link spam – largely, if not completely – is to insist that the *quality* of the pointing pages should be taken into account in assigning the importance of the pointed-to page. This means, among other things, that worthless dummy pointing pages should contribute nothing to the score of the pointed-to page.

This fundamental idea forms the foundation for all the other methods for LA that we will discuss. The idea clearly can be expressed in many ways; but it is important to note that, since the quality of the pointing pages is used to find the quality of the pointed-to pages, one finds oneself in fact in possession of a circular definition. We may express this circularity most succinctly as follows:

Good pages are pointed to by good pages. (2)

The HITS approach of Jon Kleinberg [4] is an implementation of the idea embodied in (2), but in a form in which every document gets *two* types of importance scores – one for the extent to which it is a good *'hub'*, and one expressing the extent to which it is a good *'authority'*. These two roles may be expressed (again circularly) in terms of each other, as follows:

Good hubs point to good authorities. (3)

Good authorities are pointed to by good hubs. (4)

Note that (3) and (4) are redundant – that is, (4) is simply a restatement of (3). Together, these statements amount to a circular definition of two kinds of quality: 'authoritativeness' and 'hubness'. Kleinberg's idea was that – just as with people – some documents may be viewed as good sources to consult, because of the special knowledge that they have about some theme. These documents are then the authorities. Similarly, other documents may be the best places to go if one wants to find out where are the best authorities. These 'other' documents are hubs – they are like the friend that does not know so much about repairing cars, but does know where to find a good mechanic.

Kleinberg proposed that we view all Web pages as having – to some degree – both of these functions. We can quantify their two roles – hub and authority – by translating (3) and (4) into equations. We let $h(i)$ be the hub score of node i, and $a(i)$ its authority score. Then we relate the two as follows:

$$h(i) = (const) \times \sum_j A_{i\to j} a(j) = (const) \times (Aa)_i \quad (5)$$

$$a(i) = (const) \times \sum_j A_{j\to i} h(j) = (const) \times (A^T h)_i \quad (6)$$

Here we have used the symbol a to represent the vector of authority scores, while h is the vector of hub scores. We insert an undetermined constant (*const*) in each case because, without it, a consistent solution is in general impossible – as we shall soon see. Eqs. (5) and (6) are simply the expression of (3) and (4), respectively, in mathematical form. For example, in (5), we estimate the hub score of page i by simply using the authority scores of all pages j that i points to – consistent with the verbal expression (3).

Eqs. (5) and (6) may be rewritten as

$$h = \left(\frac{1}{c_h}\right) Aa = \left(\frac{1}{|Aa|}\right) Aa \quad (7)$$

$$a = \left(\frac{1}{c_a}\right) A^T h = \left(\frac{1}{|A^T h|}\right) A^T h \quad (8)$$

Here (following Kleinberg) we choose the constant, in each case, to give a vector of unit length as a result. Kleinberg in fact presents the process of assigning authority and hub scores as an iterative application of (7) and (8), and shows that this iterative process converges. Calling the converged score vectors a^* and h^*, we find from (7) and (8) that these must be solutions to

$$h^* = \left(\frac{1}{c_h^*}\right) Aa^* = \left(\frac{1}{c_h^*}\right) A \left(\frac{1}{c_a^*}\right) A^T h^* = \left(\frac{1}{c_h^* c_a^*}\right) AA^T h^* \quad (9)$$

$$a^* = \left(\frac{1}{c_a^*}\right) A^T h^* = \left(\frac{1}{c_a^*}\right) A^T \left(\frac{1}{c_h^*}\right) Aa^* = \left(\frac{1}{c_a^* c_h^*}\right) A^T Aa^* \quad (10)$$

Thus we see that the hub and authority score vectors h^* and a^* must be *eigenvectors* of, respectively, the matrices

$$\mathcal{H} = AA^T \quad (11)$$

and

$$\mathcal{A} = A^T A \quad (12)$$

We note that both the Hub matrix $\mathcal{H}$ and the Authority matrix $\mathcal{A}$ are symmetric matrices. Also, it is clear from our discussion of Eqs. (5) and (6) that right multiplication by the matrix A sends weights "backwards" – i.e. against the arrows, hence towards the pointing node – while right multiplication with A^T sends the scores forwards – with the arrows, towards the pointed-to node. We can see this by examining a typical matrix element of (for example) $\mathcal{H}$:

$$\mathcal{H}_{ij} = \left(AA^T\right)_{ij} = \sum_k A_{ik} A^T_{kj} = \sum_k A_{i\to k} A_{k\leftarrow j} \quad (13)$$

Thus, when $\mathcal{H}$ acts (by right multiplication) on a vector h of candidate hub scores, it will send scores from a node j, first forward to j's outlinks k (which are thus candidate authority nodes), then backwards to these nodes' inlinks – for example, i – which is thus rated for its role as a hub. In short, $\mathcal{H}$ allows nodes to send one another (via two hops) their hub scores; and a node j with a high hub score (which thus "thinks" it points to good authorities) will send (via action of $\mathcal{H}$) larger hub weight to other nodes i which point to the same nodes (good authority nodes) as j does. This sending is iterated until convergence – at which point all nodes agree on their respective hub scores. The same logic applies to authority scores; one simply exchanges, in the above discussion, authority ↔ hub, forward ↔ backward, pointing ↔ pointed-to, etc.

As noted after Eqs. (9) and (10), the converged scores h^* and a^* are eigenvectors of (respectively) $\mathcal{H}$ and $\mathcal{A}$. However, we know that these matrices, for a Web graph with N nodes, will be $N \times N$ matrices, and thus have (in general) N distinct eigenvectors. Thus we must ask: which eigenvector do we want?

The answer is simple. As noted by Kleinberg, the iterative process of sending scores will converge (almost always) to a unique vector, called the *principal eigenvector* of the matrix in question ($\mathcal{H}$ or $\mathcal{A}$). The principal eigenvector is that corresponding to the

largest eigenvalue. For nonnegative matrices such as $\mathcal{H}$ or $\mathcal{A}$, the largest eigenvalue is positive, and the principal eigenvector is nonnegative (positive or zero) at every node. In fact, as long as the graph represented by $\mathcal{H}$ or $\mathcal{A}$ is connected, the principal eigenvector will be strictly positive at each node – because these matrices are symmetric. Hence we can expect a positive, nonzero hub and authority score for each node as a result of this process.

Kleinberg designed the HITS method to work on subgraphs of the WWW. That is, he assumed that one is given (from some search engine) a hit list that is large, because the topic of the search is not very narrow. The hit list likely includes many good authorities, but also many uninteresting pages. To find the good authorities (and hubs) for the given topic, he proposed to take the top-ranked t pages from the hit list (thus forming the *root set* of pages), and then to augment this root set by including (i) all pages pointed to by the root set, (ii) some or all pages pointing to the root set, and (iii) all links internal to the resulting, augmented, *start set*. Kleinberg argued that the resulting start set should contain most of the good authorities for the topic in question. Because of this property of using hyperlinks to 'zero in' on a topic, the method acquired the name Hypertext-Induced Topic Selection or HITS.

The basic HITS method described here, and modifications which build from it, have been granted numerous patents [5]. It is however not known to us whether any commercial search engine is currently using this method.

The key ideas of HITS are by now (we hope) clear:

(i) Define the quality of Web pages in a circular way – that is, in terms of the quality of other Web pages.

(ii) Use the link structure of the Web graph to express these definitions as equations.

(iii) The resulting equations call for the principal eigenvector of a matrix; this eigenvector gives the desired quality scores.

These three basic steps underlie all of the methods we will discuss, beyond the most basic method of link popularity.

3.3 PageRank

Now we discuss a method which – in contrast to HITS – is meant to be used on the entire Web graph. The PageRank method [6] gives a single score for each Web page – its PageRank. This score is best regarded as a kind of authority score, in that it is generated (like link popularity) from inlinks.

We begin with the basic concept as expressed in (2). However, we will only consider one type of "goodness" or quality score; hence we need not expand (2) to (3)+(4). Instead we express (2) mathematically as:

$$p_i = \sum_{j \to i} \left(\frac{1}{k_j^{out}} \right) p_j = \sum_j \left(\frac{1}{k_j^{out}} \right) A_{ij}^T p_j$$
$$= \sum_j \left(A^T \right)_{ij}^{(col)} p_j \qquad (14)$$

The first equality in (14) tells us that – as prescribed by (2) – node i gets its PageRank score p_i from all nodes j pointing to i. However, it also says that nodes j do not send their full quality score (weight) to i; instead, sending nodes j must divide their weight by the number of outlinks k_j^{out} which they have. Thus, the PageRank approach, as expressed in (14), is rather "democratic": every page can only "vote" once (using its current score). That is, a good page cannot give its full weight to each of its outlinks, because the *total* weight it sends must be conserved – hence, divided up among its outlinks. This is expressed in another way by saying that the forward-sending matrix A^T is *normalized*. More precisely, $(A^T)^{(col)}$ is a *column-normalized* version of A^T, since the entries in each (nonzero) column are adjusted so as to sum to one.

Note that there is no constant prefactor in (14). The reason is that we know the constant before we do any calculation; and it is one. We can (as before) reformulate (14) into an eigenvector equation:

$$p = (A^T)^{(col)} p; \qquad (15)$$

but in this case we know that the (principal) eigenvalue is 1 – because of the normalization property of $(A^T)^{(norm)}$ [7].

On the face of it the PageRank approach (14) – (15) appears to be simpler than the HITS approach: just one equation, giving one set of scores p. This simplicity has a price however.

We recall that the HITS approach gives a positive score (actually, two positive scores) to every page, as long as the effective graph ($\mathcal{H}$ or $\mathcal{A}$) is connected. This nice positive definite property comes from the fact that the effective links (in $\mathcal{H}$ or $\mathcal{A}$) are symmetric. PageRank uses the simpler operator A^T, which is *not* in general symmetric. This means that, even given a connected Web graph, one can expect in general exactly zero weight for a large fraction of all the pages. (For a detailed justification of this fact, see

[8]). This is an undesirable outcome: it means that, for the set of nodes with exactly zero weight, (i) one is forced to regard them as having zero quality, and (ii) LA provides one with no basis whatsoever for comparing them.

One can in fact show [8] that all pages in the Web graph – except those lying in sink SCCs – will get zero weight in the principal eigenvector of $(A^T)^{(col)}$. This fact agrees with our intuition: all SCCs lose weight to downstream SCCs, except sinks. Thus we have come up against the 'sink problem' mentioned earlier – and in a rather severe form that clearly justifies its name.

Brin and Page [6] offer their own solution to this sink problem. The normalization of their forward operator allows them to think of the score p_i as a *probability* – more precisely, the probability that a random walker, following outlinks from one page to another, will (in the limit of many hops) visit page *i*. They then allow this "random Web surfer" to hop (again at random) from any page to any other, with probability $\in/N$. To avoid violating weight (ie, probability) conservation, they correspondingly adjust the probability of following an outlink from *j* to be $(1 - \varepsilon)(1/k_j^{out})$. In terms of matrices, this amounts to adding an all-to-all 'random surfer matrix' $\hat{R}$ to the given matrix $(A^T)^{(col)}$, and weighting the sum to retain probability conservation. The PageRank eigenvalue equation then becomes

$$p = \left((1-\varepsilon)\left(A^T\right)^{(col)} + \varepsilon\hat{R}\right)p. \qquad (16)$$

The all-to-all matrix $\hat{R}$ (with $\hat{R}_{ij} = 1/N$) makes the effective network which is represented by the sum $\left((1-\varepsilon)\left(A^T\right)^{(col)} + \varepsilon\hat{R}\right)$ a single SCC. Hence there are no sinks in this effective network, and thus no sink problem: all entries in *p* are positive definite.

Eq. (16) presents the PageRank approach in terms of a simple eigenvalue equation. PageRank is (like HITS) patented [9]; and its application in the search engine Google [10] is well known. We will briefly describe, in Section 4, how LA scores are integrated into a working search engine. For now we will simply remind the reader that an LA score such as PageRank is only one part of a highly complex evaluation that must be performed in order to give good, relevant search results. In particular, a whole-graph score such as PageRank can say something about the quality or importance of a page, but nothing directly about the relevance of a given hit to a given search query.

3.4 T-Rank

We, in collaboration with Mark Burgess, have explored [8] the idea of abandoning the normalization constraint that is present in PageRank. As such an idea has not (to our knowledge) been studied by others, we have chosen to give it the name 'T-Rank'. The non-normalized (T-Rank) modification of (14) is simple:

$$t_i = (const)\sum_{j\to i} t_j = (const)\sum_j \left(A^T\right)_{ij} t_j. \quad (17)$$

Here we have used the symbol *t* for the vector of T-Rank scores. Note that the constant prefactor has reappeared – because the loss of normalization means that the eigenvalue will be, not one, but greater than one.

The abandonment of the normalization constraint means that a good page can send its *full* weight to all of its outlinks. This is less 'democratic' than PageRank in that a very strong page can substantially raise the score of all of its outlinks – no matter how many they may be. (In practice, of course, it is likely that, when there is a page with very many outlinks, most of them will be internal or navigational links, and thus eliminated or given little weight.)

Another consequence of dropping normalization is that the sink problem is less severe. That is [8], the principal eigenvector of A^T for the Web graph (Figure 2) will (with very high probability) have nonzero weight in the entire GSCC, and in all of OUT. Nevertheless, there remain a large number of pages (IN, Tendrils, Tubes) with zero weight; hence we expect that a sink remedy is needed for this approach also. The 'random surfer operator' $\hat{R}$ may again be used here; but since the original T-Rank operation is not weight conserving, there is no need to have a weight-conserving sum. Instead we simply weight the $\hat{R}$ operator with the tuning parameter $\in$. The resulting eigenvector equation is

$$t = \left(\frac{1}{\lambda}\right)\left(A^T + \varepsilon\hat{R}\right)t. \qquad (18)$$

Here λ is the dominant eigenvalue of the operator $\left(A^T + \varepsilon\hat{R}\right)$.

We offer a brief summary of results from [8] which are relevant to the T-Rank approach. Perron-Frobenius theory [7] tells us that the set of eigenvalues of the full graph is simply the union of the eigenvalue sets of the SCCs. Hence, the largest eigenvalue of a graph lies in the spectrum of one or more SCCs. In [8] we define a 'dominant SCC' as one whose eigenvalue spectrum includes the largest eigenvalue for the entire graph (in addition to satisfying some other conditions when there is more than one such SCC). We have shown that the principal eigenvector for the T-Rank matrix A^T for the full graph is positive in the dominant SCC (DSCC), positive in all SCCs lying 'downstream' (a precise term) of the DSCC, and zero

in all other SCCs. For the Web graph, we expect the GSCC to be the dominant SCC; hence we expect positive weight in the GSCC, and in the downstream part OUT, for the dominant eigenvector of A^T for the Web graph. We have also shown that a new type of sink remedy may be used with T-Rank – one that is not possible with the normalized approach of PageRank.

We see that the simple action of dropping normalization can have large consequences:

- Less 'democratic';
- Significantly different weight distribution in the unmodified graph;
- New type of sink remedy is possible.

We have to date performed only limited tests of the T-Rank approach, with each type of sink remedy mentioned here. Hence it is too early to say conclusively whether the T-Rank approach can be a viable alternative to PageRank. However it is clear that it is a distinct alternative, and one that merits further study. The T-Rank approach has also been patented [11].

3.5 Other Approaches

We note some basic choices that have appeared in our discussion of link analysis:

- We can normalize the matrix (PageRank), or not (HITS and T-Rank).
- We can use a compound, symmetric form (à la HITS), or not (PageRank and T-Rank).
- We can use a forward (A^T) (PageRank and T-Rank) or backward (A) matrix, or both (HITS).

Most approaches to LA involve some set of choices from this list, with some creative modifications. It is already clear from our short list that there remain some 'basic' options that we have not discussed. For example, one might choose to use the non-compound 'backward' operator A – with or without normalization. Also, one can try using normalized versions of the compound operators $\mathcal{H}$ and $\mathcal{A}$.

The use of the backward operator A – in a non-compound form – has received almost no attention to date, and so we will not discuss it further here. We summarize the remaining 'basic' options in Table 1.

Note that we have introduced (following Kleinberg [4]) the terminology 'one-level' and 'two-level' in Table 1 – corresponding to, respectively, the use of simple or compound matrices. We think this terminology is useful: as pointed out by Kleinberg, in the one-level approach, authority weight comes directly from other authorities, while in the two-level approach, authority weight comes from hubs – and so only indirectly (via two hops) from other authorities (which after all define what is a good hub).

	Normalized	Non-normalized
1-level Forward	PageRank	T-Rank
2-level (compound) $\mathcal{H}$ and $\mathcal{A}$	SALSA	HITS

Table 1 A simple categorization of approaches to Web link analysis

We note also the new entry in the 'compound, normalized' cell in Table 1: the 'SALSA' algorithm (Stochastic Approach for Link-Structure Analysis). This option was investigated by Lempel and Moran in 2001 [12]. In the SALSA approach, the normalized Hub operator is defined as

$$\tilde{\mathcal{H}} = A^{(row)} \left(A^T\right)^{(col)} \tag{19}$$

and the normalized Authority operator is correspondingly

$$\tilde{\mathcal{A}} = \left(A^T\right)^{(col)} A^{(row)}. \tag{20}$$

Here $A^{(row)}$ is the row-normalized (i.e. by indegree) adjacency matrix, and $(A^T)^{(col)}$ is the column-normalized transpose which we have seen before in the PageRank algorithm. The logic is thus (we hope) clear: SALSA seeks to combine the two-level approach of HITS with the normalization of PageRank. For example, (20) says that authority weights are sent via two hops: first backwards to hub candidates, using $A^{(row)}$ – with the row (indegree) normalization ensuring that each candidate authority can only use its full authority weight once in 'voting' for candidate hubs. Then $(A^T)^{(col)}$ allows these candidate hubs to 'vote' once for new candidate authorities, by dividing the weight they have gotten among their outlinks.

The SALSA approach clearly offers a new set of basic choices for link analysis. However – remarkably – the eigenvectors of the SALSA operators $\tilde{\mathcal{H}}$ and $\tilde{\mathcal{A}}$ can be proven [12] to give the same scores as simple link popularity. That is, to within vector normalization, the score of a node in the dominant eigenvector of $\tilde{\mathcal{A}}$ is its indegree, and the score of a node in the dominant eigenvector of $\tilde{\mathcal{H}}$ is its outdegree.

This result is very similar to a familiar result for undirected graphs. That is, it is known [13] that, when A is symmetric (as is the case for undirected graphs),

the dominant eigenvector of the column-normalized $A^{(col)}$ is (except again for vector normalization) simply the node degree. Since column normalization gives weight conservation, one can also interpret the (vector normalized) scores as probabilities. Hence, this known result for undirected graphs says that, in the limit of a very long random walk, the probability of visiting each node is proportional to the node degree. Similarly, the SALSA results say that a random walk which always follows a forward hop with a backward hop (and vice versa) will (in the limit of very long time) reach a node after a forward hop with probability proportional to that node's indegree, and will reach a node after a backward hop with a probability proportional to that node's outdegree.

This result is even more remarkable in light of the fact that the naïve generalization of these ideas to the one-level, normalized case (PageRank) does *not* work. That is, it is not the case that, in the dominant eigenvector of $(A^T)^{(col)}$, each node's score is proportional to that node's indegree.

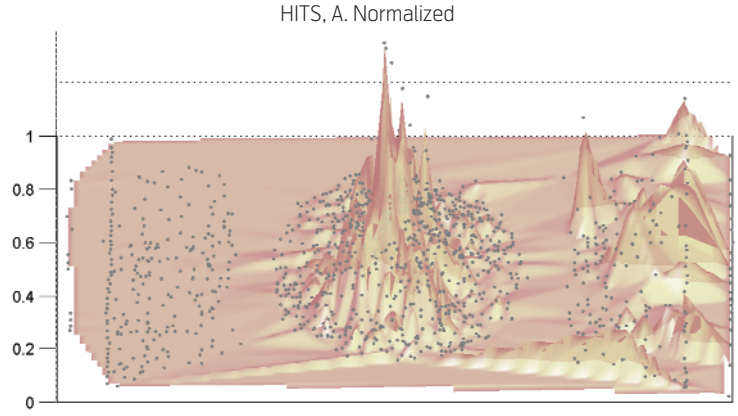


Figure 4 A 3D visualization of the Authority eigenvector obtained using the HITS algorithm. The underlying graph is that of Figure 3, and the nodes' layout in the 2D plane is the same as in that figure. The height of a node is the node's score in the normalized eigenvector

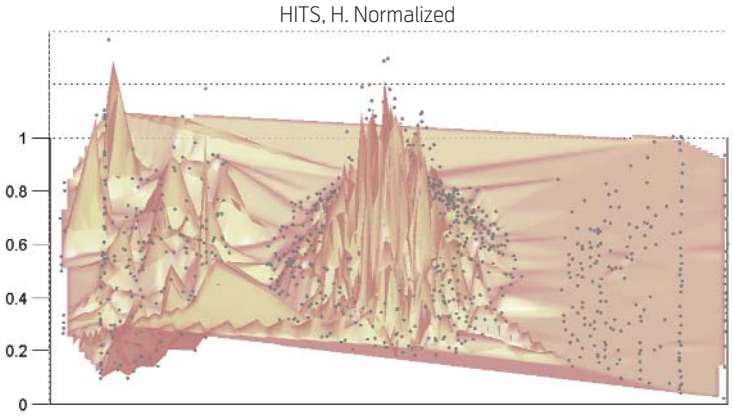


Figure 5 A 3D visualization of the Hub eigenvector obtained using the HITS algorithm

3.6 Visualization

In this section we present visualizations of the eigenvectors produced by the various approaches to link analysis discussed above. We use the graph of Figure 3 as our illustrative case. We use a three-dimensional visualization, in which each node's position in the xy plane is the same as in Figure 3, while the node's 'height' is determined by its score in the visualized eigenvector.

We begin with the two HITS eigenvectors. Figure 4 shows the authority eigenvector obtained from the HITS operator $\mathcal{A}$. We note that our visualization tool (taken from MATLAB) allows us to see the eigenvector as a smoothed surface in 3D, rather than a set of isolated points.

The picture we see in Figure 4 has no great surprises to offer. We expect large weight in the GSCC – and that expectation is confirmed. Also, since we are measuring authority of the nodes (which they get by being linked to from good hubs), the IN set has very low score, while there is considerable weight distributed among the nodes of the OUT set.

In Figure 5 we show the Hub scores (obtained from $\mathcal{H}$) which complete the HITS picture. The contrast with Figure 4 is striking. Of course, we are not surprised to see almost no Hub weight in OUT, and considerable HUB weight in IN. However, it is also obvious from comparing Figures 4 and 5 that the two distributions (Hub and Authority weight) are quite different even within the GSCC. Thus these visualizations confirm the underlying notion of the HITS approach – namely, that the two roles (Hub and Authority) are quite distinct, and that each node has in general some 'Hubness' and some 'Authoritativeness'.

We comment in this regard that the approaches found in the top row of Table 1 are also intended to measure 'authoritativeness' – but (as noted above) via a direct process, where good authorities point to good authorities. Mathematically this rule amounts to taking an eigenvector of a forward operator (normalized or not). The essential feature of authoritativeness is that a node be pointed to by good nodes. And, in this language, 'hubness' is distinct: a good hub must point to good nodes. Hence, one can in principle also measure 'hubness' in a one-level approach, by using the (normalized or not) 'backward' operator A. Having noted this point, we will here (as in Section 3.5) concentrate primarily on the more well-studied authority approaches discussed above.

Figure 6 shows the eigenvector for the PageRank approach (16) – but with an extremely small value for

ϵ (10^{-3}). Figure 6 thus shows the 'sink problem' in its most extreme form; it is essentially what one gets with $\epsilon = 0$. As we noted in Section 3.4, all the weight in the eigenvector for $(A^T)^{(col)}$ goes to the sink SCCs – in particular, to those which are not single nodes and so can be normalized. We find that there is only one such sink SCC in the OUT set of the blog graph in Figure 3; it is a two-node SCC (two blogs that cite one another), and it gets all the weight in the eigenvector in the absence of the random surfer remedy.

We obtain a more realistic view of the PageRank approach by setting $\epsilon = 0.15$ (Figure 7). (This value is believed to be close to that used today by Google.) We see from Figure 7 that one gets (at least superficially) sensible authority results: the PageRank distribution shown here is (very roughly) like that for HITS authority scores (Figure 4). However this likeness only holds in a very coarse sense: it is clear from inspection that the two distributions differ in many ways, not least in the GSCC.

We remind the reader that the HITS approach was developed for use on a small topic-focused subgraph of the entire Web, while PageRank is meant for use with (and is used with) the entire Web graph. The graph we use here is clearly only a small subgraph. However, we believe there is value in visualizing all these different approaches on the same subgraph; the aim is to convey some impression of how these various approaches differ from one another, and the best way to do that – purely mathematically, i.e. without resorting to evaluating ranked hit lists – is by applying all methods to the same graph. (More thorough comparisons for the well studied methods (PageRank, HITS, and SALSA) may be found in the published literature [], while a more thorough comparison including T-Rank may be found in [8].)

One known feature of PageRank [8] is clear already from comparing Figures 6 and 7. We see that the weight distribution in the eigenvector is extremely sensitive to the chosen value of ϵ – at least, around $\epsilon = 0$. This sensitivity decreases for larger ϵ. As ϵ approaches 1, the eigenvector approaches a completely 'flat' (uniform) distribution, in a rather smooth fashion: the lowest values, representing the 'floor' of the distribution, simply increase smoothly, while otherwise the overall shape of the distribution as seen in Figure 7 changes only slowly.

Next we show the T-Rank eigenvector in Figure 8. We note several things with regard to this figure. First, we see that, even without help from the random surfer operator, we get a reasonable authority distribution from the T-Rank approach. The reason for this was noted in Section 3.4: without normalization, the

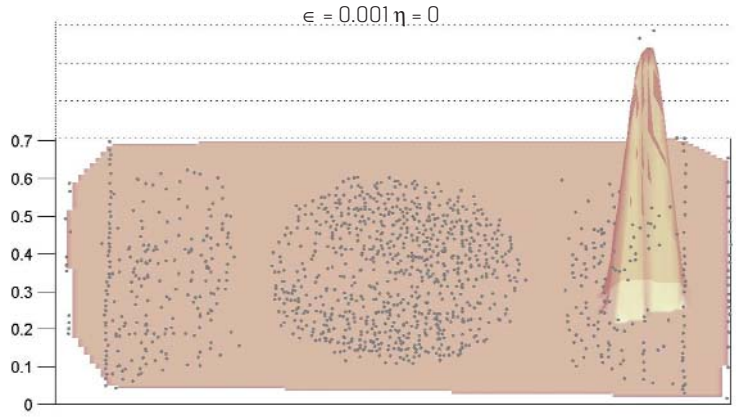


Figure 6 The eigenvector of PageRank's normalized forward operator $(A^T)^{(col)}$, without random surfer

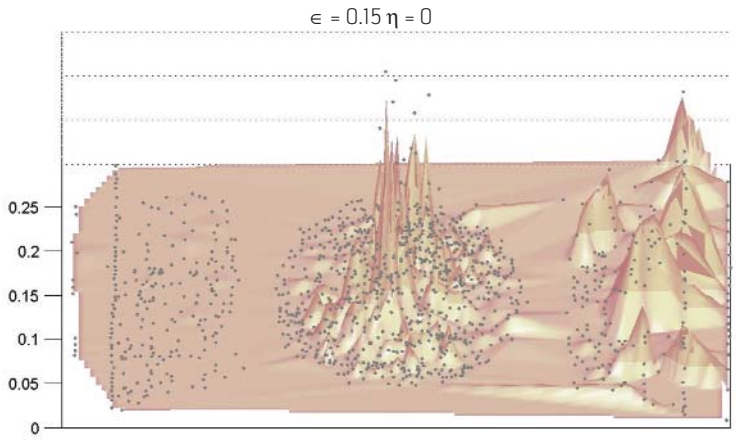


Figure 7 Same as Figure 6, but with $\epsilon = 0.15$

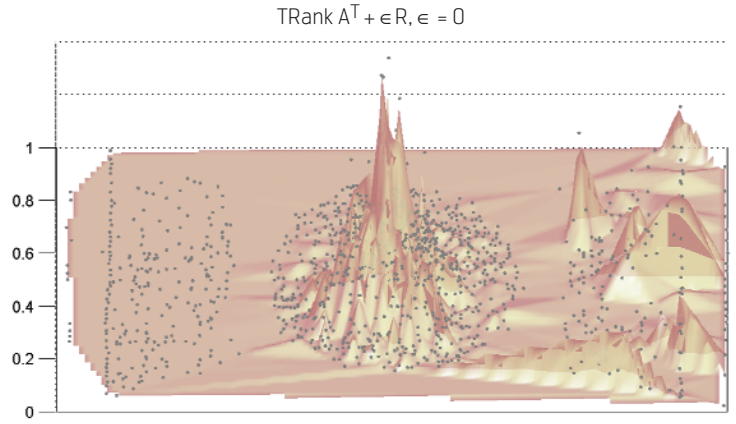


Figure 8 The eigenvector of the T-Rank forward operator AT, without random surfer

GSCC has the largest eigenvalue of any SCC, and so is able to generate positive weight for itself and all downstream SCCs [8]. The IN set gets of course exactly zero weight here. However, setting $\epsilon = 0.15$ gives a picture which is visually indistinguishable from Figure 8 – while at the same time giving some small weight to the IN set.

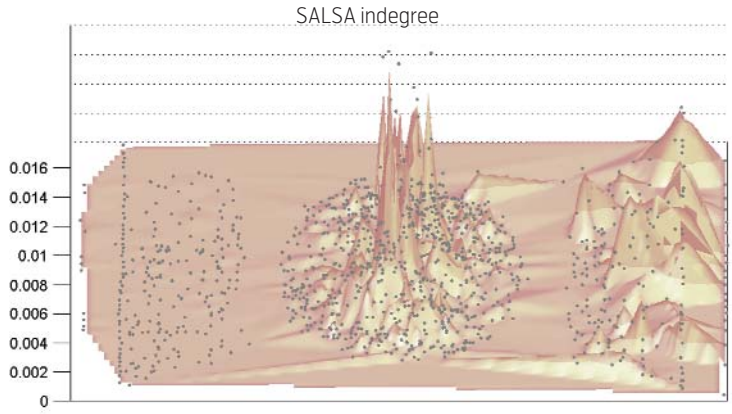


Figure 9 The SALSA authority vector, which is the indegree for each node, divided by the total number of links to give vector normalization

This brings us to our second observation: the T-Rank eigenvector is much less sensitive to the value of ϵ than is the PageRank approach – even at $\epsilon = 0$. Of course, $\epsilon = 0.15$ is in a sense a 15 % perturbation on the normalized PageRank operator (which has $\lambda_{max} = 1$), but only a much smaller perturbation on the non-normalized operator (with $\lambda_{max} \approx 34$). We find however that even setting $\epsilon = 5$ (about 15 % of λ_{max}) has very little effect on the eigenvector. In fact, we see little significant change before ϵ is increased to values $\epsilon \geq 100$. Reference [8] offers a more quantitative study of this clear lack of sensitivity to ϵ.

We also see, by comparing Figures 4 (HITS), 7 (PageRank), and 8 (T-Rank), that the HITS and T-Rank eigenvectors appear to be much more similar to one another than either is to the PageRank eigenvector. This suggests that the choice of matrix normalization (or not) has more significant effects on the resulting eigenvector than the choice of one-level or two-level authority scheme. We reserve also this question for future work.

We can provide another non-conclusive test of this conjecture by examining one last authority vector, namely, the SALSA authority vector – i.e. the node indegree (its link popularity). We show this vector in Figure 9.

We find from this figure some support for our conjecture. That is, the SALSA authority vector appears to be closer in structure to the normalized one-level PageRank vector (Figure 7) – in particular, in the GSCC. Since the SALSA vector is simply the indegree, this suggests in turn that the PageRank score is more like link popularity than is the T-Rank or HITS authority score.

Of course, we remind the reader once again that we can offer no definite conclusions from these visualizations. The trends and correlations that we see and discuss here are only suggestive – but such suggestions are a valuable source of questions and ideas for future work.

4 Application of Web Link Analysis to a Search Engine

Now we place the whole idea of link analysis (LA), as applied to Web search, in context, by presenting a very simplified picture of a working search engine. Reference [15] offers a more thorough discussion.

The WWW is – as noted earlier – unmanaged, dynamic, and complex. The tasks of a working Web search engine are then: (i) obtaining an overview of what there is to be found 'out there' on the Web; (ii) storing that information in a form which allows rapid retrieval when a query is presented; (iii) evaluating the relevance of each page on a hit list to a given query, by looking at the *content* of the page; (iv) evaluating the quality or authority of a page by looking at its *context*; (v) ranking the results (the pages in the hit list) according to some weighted combination of scores from (iii) and (iv), and presenting the ranked results to the user; (vi) possibly, clustering the results according to how well they 'belong together' thematically; and (vii) fighting spam.

Below we describe briefly several of these components. Using the numbering scheme of the previous paragraph, the tasks which we will further discuss are: (i) crawling, (ii) indexing, (iii) text relevance analysis, (iv) link analysis, and (vii) spam fighting. Task (v) (ranking) amounts to the problem of combining, in a useful way, all the score information – both search-dependent and search-independent – that is available for a given page. We are not aware of any theory beyond 'black art' for this task. Task (vi) is an interesting problem which we touch on briefly in Section 5.

4.1 Crawling

We begin with crawling. A search engine provider must maintain an up-to-date picture of the contents of the enormous WWW (now over 10 billion pages, and always growing). This requires that the Web be 'crawled'. That is, a 'robot' or crawler must find, visit, and fetch every page of interest. This is a huge task: known pages are constantly updated by their authors, and so must be re-crawled, while at the same time new pages turn up, and must be found. These latter may be found if they are linked to by existing pages. Otherwise, they are very difficult to find. Furthermore, there are *dynamic* Web pages, which are only generated when triggered by a visit from a user – in the form of entering some information which a

crawling robot cannot generate. These pages cannot be crawled – by definition, that is, barring changes in what a crawler can do.

The result is that, for any given approach to crawling, there will be a 'visible Web' (that set of pages which is crawled) and a 'deep Web' (the invisible part). It goes without saying that only the visible Web can give useful results for a search. The deep Web – consisting of those pages which are not crawled due to limited resources, those pages which have been changed since the last crawler visit, those pages which the crawler cannot find, and dynamic Web pages – represents an enormous store of (so far) untappable information.

4.2 Indexing

The information present in crawled pages must be stored in a form allowing fast retrieval – that is, in an *index*. The index is a database which is so constructed that, given one or more input terms of text, it allows rapid retrieval of all documents containing the terms, plus information about the frequency, location on the page, etc of these terms. The schematic form of the entry for (term) in an index is thus

(term) (Doc1) (freq1) (title1) (url1) (anchor1) …
(Doc2) (freq2) (title2) (url2) (anchor2) …

Besides the term's frequency (freqx) in Docx, there may be information about the term's placement in the title, or in the url of the page. Furthermore, it is often useful to include the fact that other pages (say, Docy) than Docx have pointed to Docx, using "anchor text" associated with the outlink of Docy, such that this anchor text includes the term. For example, Docx's relevance to the term 'bread' may be judged as enhanced if one fetches, from the index, the knowledge that 10 other pages have linked to x, with each of these 10 pages having the word 'bread' in the anchor text associated with the link to Docx. These 10 pages clearly are saying that Docx has something to do with bread.

4.3 Text Relevance Analysis

A hit list is generated by fetching, from the index, IDs for all documents containing the term or terms in the query. If the query is very special, the hit list may be very short – in which case, one may simply present the entire, unranked, list. Most commonly, the hit list is long (recall the size of the WWW). Thus one needs to *rank* the pages in the hit list, so as to only present the best pages; and one obvious criterion for ranking them is to look at their *content*, and to evaluate (by machine) the relevance of that content to the query. We call this process 'text relevance analysis'.

As may be deduced from our generic index entry above, there are a number of simple, standard ways to help a machine perform the task of text relevance analysis – specifically, to give a numerical score which quantifies the relevance of Document x to term t. One can of course look at the *frequency* of the term in the document – typically, corrected for the frequency of the term in the entire document set, so as to avoid giving too much weight to common terms. Often, this correction includes the extreme measure of excluding 'stop words' such as (in English) 'and', 'the' etc. These stop words are viewed as carrying zero information about the content of the document; while the information content of other words may be first quantified in terms of the statistics for the entire document set, and then used to weight the term frequencies fetched for each document.

It is also common to give extra weight for a term's appearance in the *title* and/or *url* of the document in question. Another source of text relevance score is the 'external' text associated with a document – typically, in the form of *anchor text* in its inlinks. Inlinks (and their anchor text) are of course not readable from a visit to a document; hence this information is only available if the inlinking pages themselves are crawled. However (as pointed out by Kleinberg [4], and many others) a page with high relevance for a given term may in fact have low or even zero frequency for that term – for example, the Toyota home page may be expected to have a low incidence of the term "automobile manufacturer". In contrast, anchor text is often just a few words which summarizes the linking author's perception of the linked-to page – so that these few words are likely to be relevant to the page pointed to (and vice versa).

The use of anchor text may be thought of as a kind of hybrid of text relevance and link analysis. In its simplest version, application of anchor text information does not take into consideration the quality of the inlinking page. This version thus corresponds to a query-dependent form of link popularity. Other forms are possible, in which some quality score (eg. from link analysis) is used to weight those inlinks having the appropriate anchor text.

Finally, we mention that *query analysis* is also of significant interest. Here, the aim is to "read the mind of the user", and the tools needed are essentially statistical. For example, if a query is recognized as having a strong news slant, then freshness of a page is given extra weight in ranking the hits retrieved.

4.4 Link Analysis

Link analysis has already been discussed in this article at a theoretical level. Now we mention a few practical considerations.

First we mention the obvious one: it is necessary for the crawling process to fetch and store information about the hyperlinks found on each crawled page, in order to build a representation of the Web graph for the crawled pages in a hyperlink database. As noted more than once above, inlinks to a crawled page are invisible to the crawler – it can only find, store, and follow outlinks. Of course, it builds up the Web graph by crawling both ends of a hyperlink; but it must do so in spite of the fact that it cannot directly read inlinks.

The query-dependent application of LA, as exemplified in the HITS approach, requires that the analysis be run in 'real time'; ie. in response to each query. This presents a severe challenge, since it is of utmost importance to get results to the user in a very short time. This 'response time problem' is already difficult enough, because (i) the document set is large (making retrieval of the hit list time consuming), and (ii) the number of searches per unit time (query traffic rate) will also be very large for a successful search engine. Hence any post-processing (after the hit list is generated) must be fast. Since the hit list gives the root set, this postprocessing includes building the start set (requiring fetching links from the hyperlink database), and then iteratively finding eigenvectors for the Hub and Authority matrices. Since these two matrices are so tightly coupled, however, the time to find both eigenvectors is essentially the same as the time to find one. The HITS approach has the further advantage that its start set may be adjusted to be quite small (a few thousand pages is typical). Furthermore, for ranking purposes, one needs not an exact eigenvector, but one which is roughly correct. Hence, iterative approaches (which are normally needed to find the eigenvector) may be practically terminated after a fixed, and not too large, number of iterations.

Search-independent (whole-graph) link analysis approaches (such as PageRank and T-Rank) have the advantage that they may be run 'offline'. The corresponding disadvantage is clear: the graph to be analyzed is huge. Thus one must meet a simply stated technical challenge: the time needed to update the LA score vector should be much less than the time needed for the crawler to update the full index (and the hyperlink database); otherwise, the LA is essentially always out of date. Whole-graph LA (as with the query-dependent version) may be speeded up by demanding only a fixed (and not too large) number of iterations. Also, this fixed number may be further reduced due to the fact that the whole-graph LA may use the previous eigenvector as a starting vector for the current update.

Finally, of course, LA scores must be combined with text relevance scores, before a ranking can be performed. These two scores are of rather different natures: quality – especially a whole-graph measure for quality – is quite distinct from relevance (which can only be evaluated with relation to a query). We choose to omit even a brief discussion of this 'combination problem' because we know of no solid theoretical background for this problem.

4.5 Spam

Most methods for scoring (and subsequently ranking) documents can be spammed. In addition, since high ranking gives visibility – while low ranking gives invisibility – there are many reasons why an unscrupulous person UP would wish to spam a search engine. We briefly mention three types of spam here.

4.5.1 Text Spam

Text spam is in principle easy to perform, and (sometimes) difficult to counter. We suppose UP is author to a page whose visibility he wants to boost. Suppose further that the page is about bread. Then UP may boost its visibility by embedding in the page a large number of occurrences of the word 'bread' – in such a way that they are visible to the crawler, but not to any user visiting the page.

A more malicious form of text spam is to set in large numbers of invisible 'spam' words which are unrelated to the content of the page. For example, UP has a pornographic page, and wishes to catch the attention not only of those interested in sex, but also of other unsuspecting users – such as those interested in bread.

4.5.2 Anchor Text Spam

Anchor text spam is more subtle than simple text spam. In order to have an effect, anchor text spam requires many pages to cooperate, in linking – all with consistent anchor text – into a given page that one or more UPs wish to boost. Also, if the inlinking pages are actually 'dummy' pages, they may be detected by the same techniques as those used against link spam.

There is an 'anti-boosting' form for anchor text spam which is called 'Google bombing' [16]. Here, a set of Web page authors conspire to link to the same page with the same consistent anchor text – but the anchor text is negative, so that the linked-to page is boosted when the chosen negative text is in the query. The famous example is the negative anchor text 'miser-

able failure', with the linked-to page being the official biography page for George W Bush. The effect of the campaign was sufficient (for a time) to make the George Bush biography page the top hit when the query was 'miserable failure'. Since this is hardly a query that is typically used to find information, the result was a very visible form of practical joke – one which furthermore illuminated the importance of anchor text in Google's (and others') ranking algorithms.

4.5.3 Link Spam

We have already discussed link spam in Section 3.1, which treats link popularity. We recall that link popularity (i.e. simply counting the number of incoming links to a page) is vulnerable to link spam, in which UP sets up numerous dummy pages, solely for the purpose of giving more inlinks to a real page that UP wishes to boost. We also mentioned in Sec. 3.1 that insisting on assessing the *quality* of a pointing page has been viewed as a good countermeasure to link spam, and that this insistence underlies the other LA methods discussed in this article.

Here we wish to emphasize that 'countermeasure' is the appropriate term, rather than 'cure'. That is, even the circular-definition LA methods are not fully immune to link spam. There is, for instance, the (often unwitting) collective boost that comes from navigational links among a large set of pages. These may be counteracted if they can be recognized; but that is not always possible – some navigational links appear to be 'real', in that they cross domains. Also, every countermeasure has a counter-countermeasure; and there are many UPs.

5 Challenges and Future Directions

Here we depart from the well known aspects of Web link analysis, and discuss some budding or imagined future developments. Our choice of topics here is far from exhaustive; it is mainly dictated by our own interests. Also (again) our discussion will be fairly brief and superficial. Our aim is simply to offer some awareness of these questions and problems to the non-expert reader.

5.1 Clustering

The problem of finding a good *document clustering* is a relatively old one. It is also (as noted by Kleinberg [4]) a distinct problem from the problem of document ranking. Given a set of documents, a good clustering will find subsets of these documents that "belong together". The criteria for belonging together vary among the various clustering methods – for, as with document quality, there is no uniquely agreed answer to this problem.

The document ranking problem is in a sense orthogonal to the clustering problem. The ranking problem may be stated thus: given a set of documents, pick out a small (human-digestible) number which are "best" in some sense. The notion of being "best" seems fully distinct from the notion of "belonging together". Furthermore, the ranking problem is always present, whether or not a clustering is applied to the given document set – because, typically, the document set is so large that a feasible clustering will also give cluster sizes which are much larger than the human-digestible number.

In spite of this apparent orthogonality, we mention clustering here, for two reasons. First, there are good reasons to apply clustering after a hit list is generated, ranked, and truncated. Such a ranked-and-truncated list has the "best" pages for the user's query – but at the same time – barring ingenious and foolproof query analysis – there may well be more than one theme or topic embedded in this "best-hits" list. The list may further be small enough to allow for postprocessing in the form of clustering analysis. The result may be expected to be more useful and pleasing to the reader. For example, rather than being confronted with a strictly ranked hit list in which one hops from one meaning of "jaguar" to another as one scans down the list – an annoying experience – the user may instead be offered a clustered hit list, with "Jaguar the car" in one cluster, "jaguar the animal" in another, etc.

We are aware of some existing search engines which use clustering in this way; see for example Carrot [17], Clusty [18], or Grokker [19].

A second, and more far-fetched, relation of clustering to ranking is as follows. We can imagine – at least in principle – that, at some time in the future, clustering analysis could become so efficient and successful that it could replace ranking, rather than supplement it. More concretely, this idea demands both very good query analysis and very good clustering. The former will give the search engine a very good idea of what the user is actually searching for. Then the clustering analysis takes a standard, unselective hit list, and generates an increasingly refined clustering analysis, always keeping its center on the user's (analyzed) query – refining to the point that the cluster containing the query has the desired human-digestible number of pages. In short: we imagine that, some time in the future, the notion of hits "belonging together" (and with the query) may be so thoroughly refined that it becomes the same as the notion of being "best".

5.2 Multimedia Content

User-generated content on the Web is a fast-growing phenomenon, and furthermore one which shows every sign of growing further, and becoming an enduring part of our Web experience. Furthermore, a great deal of user-generated content lies in other media than text – commonly, pictures, music, and videos.

To some extent, this kind of content has managed to make itself visible in today's generation of Web search engines. This works due to several things. For one, producers of multimedia content often include metadata in the form of text describing (albeit briefly) the content. This gives conventional search engines some text which they can exploit with today's methods. Another helping factor is that popular items – for example, videos on YouTube – are quickly given hyperlinks by appreciative viewers; and those hyperlinks usually have relevant anchor text. The anchor text may be exploited to find such content in response to a search; and the hyperlinks in general are useful in boosting the LA scores for the popular content.

There remains a large amount of "less popular" content which has little or no associated metadata, and few or no hyper(in)links. Finding such content is a severe challenge for existing search engines – even though today's generation of search engines take pride in giving good results for searches in the "long tail" of queries. The "long tail" is composed of those queries which "almost no one" is interested in, and yet which are so large in number that they collectively play a huge role in the statistics.

A promising approach to the "long tail" of multimedia content is to rely on advances in machine recognition – of pictures, sound, and videos. The aim here is to get a machine to generate metadata from multimedia content. (Machine-generated hyperlinks seems very problematic – but see Section 5.4.) Computer recognition of images and sound is a huge research field, which we will not go into here – beyond mentioning some trends. For one thing, this approach is currently a priority for the European Union's IST program [20]. That is, it is very much a subject of active research. Also – in a limited way – machine recognition of pictures and music is already in use, in the form of recognizing multimedia queries (see the next subsection).

5.3 Multimedia Queries

We imagine a lazy user, and/or one with a mobile phone – in any case, one who does not want to type text into a search box. What alternative forms of query are there for such a user?

One form of query is voice: the user can speak her query to the machine. Voice recognition – that is, voice-to-text translation – is a relatively advanced technology. Furthermore, it has the advantage that its output is text – thus giving an easy interface to existing, text-based, search engine technologies.

Voice will not however always be preferred, or even possible. We mention two other examples. First, we hear some music, and want to know the song title, performers, etc. We then let our PC or mobile device hear the music: the music is the query. The reply may be obtained, with surprising accuracy, today [21]. The reason for the good performance is that the recognizer need only recognize the music with respect to a finite, known set of reference pieces. Furthermore, it must find only the best match: only the most similar item is wanted, no ranking is needed, and relevance is not important.

In another scenario, the user takes a picture of something, and wants to know something about the picture. Thus the picture is the query. It may be supplemented with voice (giving some metadata); but "a picture is worth a thousand words", and so picture recognition/analysis must be a part of the processing of the query. Furthermore, the answer may also include similar pictures. The "easy" case is (as with music) where only the best match is of interest, ie. there is a "right answer", and this is the one the user wants (a product, for example). This "easy" case is (as with music) just being implemented now, in Japan and in the US [22].

The more difficult case is also plausible: there is no single "right answer", and the user wants to see similar pictures. Here the problem is that the meaning of "similarity" will very much depend on the current state of the current user. A simple fix to this problem is to allow the user to specify (verbally) what kinds of similarity are desired. In any case, this "less easy" scenario is one with considerable challenges.

5.4 LA without Hyperlinks

There are very many documents – text or multimedia – which are not embedded in a network of hyperlinks, but which still need to be searched. Examples are: the set of documents on a user's PC; the set of documents generated by, and hosted by, a company or organization; and the "long tail" of multimedia content, mentioned above.

Why then even consider doing link analysis when there are no links? Our motivation in considering this question stems from the most basic distinction between text relevance analysis and link analysis: the former looks at the *content* of a document, while the

latter looks at its *context*. So we rephrase the question as follows: is it possible to place a given document (or electronic object), taken from a hyperlink-free environment, in a context which is both meaningful and defined by the other documents? If so, then perhaps we have a useful analog of Web hyperlink analysis for the so-called "link-poor" (and link-free) domains.

We sketch here a tentative, non-unique answer. Consider again the user who inputs a picture, and wants similar pictures as part of the reply. It is clearly essential for this enterprise that the machine be able to compute a similarity score for two objects – in this case, pictures. We then propose to take such similarity scores and view them as *weighted, symmetric links* between the pairs of objects. The resulting set of links + nodes (objects) forms a *similarity graph*. Perhaps it is possible to apply some form of link analysis to such a graph, in a useful way.

An example: take a ranked, truncated hit list, and form the similarity graph from these objects. Now one can apply a non-normalized form of link analysis to this graph, in the form of finding the principal eigenvector of the similarity matrix. This approach is known from social science – where it is called 'centrality analysis', and the resulting scores are called *eigenvector centrality* [23]. The word 'centrality' is appropriate here, for, with symmetric links, the nodes receiving the highest weight in the principal eigenvector may reasonably be regarded as being most central in the graph.

This idea amounts to a re-ranking of the ranked, truncated list: one takes the "best" hits according to other methods, and then finds the "most central" hits in this list by applying centrality analysis to the similarity graph. This is a form of postprocessing which is similar in spirit to HITS, but quite different in its methods and results.

We mention finally another application of similarity-graph analysis. Once a collection of documents is represented as a graph, various graph-theoretical methods suggest themselves for application to the clustering problem. Many approaches to the graph clustering problem are known. Here we mention one developed by two of us in the context of social network analysis [24]; this approach is immediately applicable to any graph with weighted, symmetric links. The basic idea is to pick out local maxima of the EVC, and let these maxima (each being of course a document) represent a subcluster of the graph. A reasonable rule for assigning the many remaining nodes to subclusters is to simply assign each node to the same subcluster as its most central neighbour. This most central neighbour may be reasonably viewed as being closest to the nearest local maximum of the centrality. Once subcluster membership is so determined, the 'theme' of each subcluster may be represented either by the subcluster's center (the local maximum) or by some statistical representation of all the members.

We have here implicitly generalized the application of similarity graphs from pictures – the original motivating idea – to documents in general; but the generalization is well justified. Most of the electronic information objects ("documents") that we know of lend themselves to pairwise calculation of similarities. In particular, it is straightforward to calculate the similarity between pairs of text-based documents.

5.5 Mobile Search

In this last section we wish to comment briefly on the spread of the Web search phenomenon to mobile devices. This is happening now; and we expect search to become an important part of the experience of tomorrow's mobile device users. However, it will (we believe) be rather different from today's PC-based search, due to some basic differences between a user sitting by a PC and a mobile user with a handheld device. For example, for the mobile user:

- The screen is small – so there is low tolerance for displaying results that are not highly relevant.
- Bandwidth to and from the device is often lower. This again means that information exchange must be limited to the most relevant.
- Typing on a small handset is relatively slow and inconvenient. At the same time, many handsets now have good cameras. This suggests strongly that, in mobile search, pictures (and even videos) can be a highly attractive way of sending in a search query – and also part of the reply to the query.
- Finally, the mobile user is often in a much more active state than one sitting at a PC. She wants information, often, in order to locate and/or buy something. Thus many searches from a mobile device are likely to be a prelude to action – for example, a purchase.

Thus we see that several of the technical challenges listed above are expected to be highly relevant for mobile search. Multimedia (and user-generated) content fits naturally into the mobile user's world; and multimedia queries (picture and voice) are likely to be very attractive.

Finally, we return to our main theme of link analysis, and ask what role it is likely to play in this mobile future. Such prediction is not simple – mainly due to the fact that it is not possible to simply transfer the PC Web experience, unchanged, to the mobile device. A small fraction of Web pages have been adapted for mobile terminals using WAP. Another approach is to use "transcoding" to adapt the pages [25]. In any case, such adaptation is needed – and not surprisingly: the optimal presentation for a PC is rather far from optimal for a mobile device.

It is not clear what role hyperlinks will play in a future, well-adapted, mobile Web. If they persist in a viable form – such that our statements 1–3 on page 1 hold to a useful degree – then link analysis will undoubtedly be useful for mobile search, which after all places an even higher premium than standard Web search on finding those few, best, results. In the event that hyperlinking does not survive (in sufficiently robust form) the transition, there will remain the pressure to display only a very few, good results on the small screen. In response to this pressure, we believe that it will be advantageous to exploit "link analysis without links". That is, we believe in the validity of the basic premise of link analysis: that one (machine, or human) must examine *both* the content and the context of a document, in order to best evaluate that document.

Acknowledgments. This work was partially supported by the EU within the 6th Framework Programme under contract 001907 (DELIS) and by the European Commission Sixth Framework Programme project SAPIR – Search in Audio-visual content using P2P IR.

References

1 Broder, A et al. Graph structure in the web. In: *Proceedings of the 9th international World Wide Web Conference on Computer networks: the international journal of computer and telecommunications networking*, 309–320, Amsterdam, The Netherlands, 2000. North-Holland Publishing Co.

2 *The Size of the World Wide Web*. December 5, 2007 [online] – URL: http://www.worldwidewebsize.com/

3 Adamic, L A, Glance, N. The political blogosphere and the 2004 US Election. In: *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

4 Kleinberg, J M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46 (5), 604–632, 1999.

5 The basic Kleinberg patent is US patent no. 6,112,202, *Method and system for identifying authoritative information resources in an environment with content-based links between information resources*, granted 2000. There are many related patents coming from the CLEVER project of IBM.

6 Page, L et al. *The pagerank citation ranking: Bringing order to the web*. Technical report, Stanford Digital Library Technologies Project, 1998; and Brin, S, Page, L. The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the seventh international conference on World Wide Web 7*, 1998, 107-117.

7 Berman, A, Plemmons, R J. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

8 Bjelland, J, Burgess, M, Canright, G S, Engø-Monsen, K. *Document importance scores from directed graphs*. Submitted for publication.

9 *Method for node ranking in a linked database.* US patent no. 6,285,999, granted 2001, and updated in no. 6,799,176, granted 2004, and in no. 7,058,628, granted 2006.

10 *PageRank*. December 5, 2007 [online] – URL: http://en.wikipedia.org/wiki/PageRank; and *Google searches more sites more quickly, delivering the most relevant results*. December 5, 2007 [online] – URL: http://www.google.com/technology/.

11 *Backward and forward non-normalized link weight analysis method, system, and computer program conduct*. US patent no. 7,281,005, granted 2007.

12 Lempel, R, Moran, S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*, 33, 387–401, 2000.

13 Motwani, R, Raghavan, P. Randomized Algorithms. *ACM*, New York, 1996.

14 Langville, A N, Meyer, C D. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

15 Chakrabarti, S. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan-Kaufmann, 2002.

16 *Google bomb*. December 5, 2007 [online] – URL: http://en.wikipedia.org/wiki/Google_bomb

17 *Carrot Clustering Engine*. December 5, [online] – URL: http://demo.carrot2.org/demo-stable/main

18 *Clusty*. December 5, 2007 [online] – URL: http://clusty.com/

19 *Grokker*. December 5, 2007 [online] – URL: http://www.grokker.com/

20 *Chorus*. December 5, 2007 [online] – URL: http://www.ist-chorus.org/

21 *TrackID – it's pure bloody genius*. December 5, 2007 [online] – URL: http://www.knowyourmobile.com/blog/1078/trackid_its_pure_bloody_genius.html; or
TrackID. December 5, 2007 [online] – URL: http://www.sonyericsson.com/product/trackid/

22 See *Wireless Watch Japan*, URL: http://wirelesswatch.jp/2007/08/02/bandai-adds-cameraphone-music-search/; or *Japan's Cellphone Edge*, URL: http://analytica1st.com/analytica1st/labels/Mobile%20music.html for Japan; or *See It, Snap It, Search It*, URL: http://money.cnn.com/magazines/business2/business2_archive/2006/12/01/8394981/index.htm for the US; all December 5, 2007 [online].

23 Bonacich, P. 1972. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2, 113–120.

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

25 For research and development in transcoding by IBM, see *Internet Transcoding for Universal Access*, URL: http://www.research.ibm.com/networked_data_systems/transcoding/. *InfoGin* is a commercial provider now, see http://www.infogin.com/; both December 5, 2007 [online].

Johannes Bjelland is a researcher in Telenor R&I. He holds a Master in Physics from the Norwegian University of Science and Technology (NTNU) (2003) with specialization in numerical simulations on complex systems. Before joining Telenor in 2007, he had a background from system administration, education and computer programming. His current interests in Telenor include social network analysis, datamining and knowledge discovery.

email: johannes.bjelland@telenor.com

For a presentation of Geoffrey S. Canright and Kenth Engø-Monsen, please turn to page 3.

Modelling Overlay-Underlay Correlations Using Visualization *)

VINAY AGGARWAL, ANJA FELDMANN, ROBERT GÖRKE, MARCO GAERTLER, DOROTHEA WAGNER

Vinay Aggarwal is PhD student at Deutsche Telekom Laboratories, TU Berlin, Germany

Anja Feldmann is Professor at Deutsche Telekom Laboratories, TU Berlin, Germany

Robert Görke is a PhD student at Universität Karlsruhe, Germany

Marco Gaertler is a PostDoc at Universität Karlsruhe, Germany

Overlay applications are popular as they provide high-level functionality by masking the intrinsic complexity of the underlay network. However, overlays rely on the underlay to provide them with basic connectivity. Therefore, the intrinsic features of the underlay network determine the efficiency of the overlay. Accordingly, studying the interdependency of the overlay and underlay networks leads to a better understanding of overlay application behavior. We present a visualization-driven analysis technique for evaluating the overlay architecture with respect to the underlay, driven by the goal of overlay engineering. Using Gnutella as a case study, our analysis confirms that Gnutella topology differs from a randomly generated network and that there is an implicit correlation between the overlay and underlay topologies.

1 Introduction

In recent times, the design of many real-world applications has changed from a monolithic structure to modular, yet highly customizable services. As an implementation from scratch is usually too time-consuming and expensive, these services are superimposed on an already existing underlay infrastructure as an overlay.

A well-known example arises in logistics. The highways and streets we use every day constitute a huge transport network. However, traffic in this network is far from structured. In fact, countless companies and institutions rely on this network to accomplish their regular shipping of commodities and services, and by doing so they cause the traffic on the road network to develop in certain patterns. In technical terms the road network constitutes an *underlay network* while the commodity exchange network of a set of companies implicitly building upon this network forms an *overlay network*. The overlay network uses the underlay to actually realize its tasks.

Another underlay network of prime interest is the Internet, which serves as the workhorse of countless data transfers, multimedia services and file sharing protocols. Almost any time we use the Internet, we participate in some overlay network that uses the physical Internet (comprised of routers, links, cables, wires) to actually convey the data packets. Interestingly, the Internet itself is an overlay built over the telephone network underlay. Within the Internet, a particular breed of overlays that has received a lot of attention lately are peer-to-peer (P2P) applications [17], which range from file-sharing systems like Gnutella and Bittorrent, to real-time multimedia streaming, and VoIP phone systems like Skype and GoogleTalk.

Clearly, there is a crucial interdependence between overlay and underlay networks. In particular, the emergence of overlay networks heavily affects and poses new requirements on the underlay. The major advantage of overlays is that they provide high-level functionality while masking the intrinsic complexity of the underlay structure. However, this abstraction entails a certain trade-off, namely independence versus performance. To gain a deeper understanding of the interdependency between the overlay and the underlay, this trade-off needs to be included in the corresponding analysis.

Due to the explosive growth of P2P file sharing applications with respect to total Internet traffic [17], there has been an unprecedented interest in their analysis [1, 2, 15]. There have also been attempts at investigating the overlay-underlay correlations in P2P systems. Using game theoretic models, Liu et al. studied in [10] the interaction between overlay routing and traffic engineering within an Autonomous System (AS), which is a network under a single administrative entity, normally corresponding to an Internet Service Provider (ISP). An analysis of routing around link failures [15] finds that tuning underlay routing parameters improves overlay performance. Most investigations tend to point out that the overlay topology does not appear to be correlated with the underlay (e.g. [1]), but the routing dynamics of the underlay do affect the overlay in ways not yet well understood. To address the apparent lack of overlay-underlay correlation, some schemes, e.g. [11, 12], have been proposed. More recently, [2] has made a case

*) *The authors gratefully acknowledge financial support from the European Commission within FET Open Projects DELIS (contract no. 001907) and the DFG under grant WA 654/13-3.*

Dorothea Wagner is Professor at Universität Karlsruhe, Germany

for collaboration between ISPs and P2P systems as a win-win solution for both.

In this paper, we approach the problem of modelling overlay-underlay correlations using a unique visualization-driven approach [4] which relies on the concept of cores, to analyze the overlay in the context of the underlay network. We introduce our theoretical model with examples in Section 2, followed by the introduction of a new technique for analytic visualization in Section 3. We then demonstrate the application of our technique on a case study to study the correlation of Gnutella with the AS network, as well as to compare Gnutella with a random network in Section 4. We first explain how we sample the P2P network, followed by a comparison of the P2P network with random networks. After a sensitivity analysis of the random network to generate and better understand P2P network models, we conclude in Section 5.

2 Modelling Underlays and Overlays

In this section, we introduce our model and methodology for analyzing the relation between under- and overlays as well as a first discussion about different modelling aspects.

Basically, an overlay consists of network structure that is embedded into another one. More precisely, each node of the overlay is hosted by a node in the underlay and every edge of the overlay induces at least one path between the hosting nodes (in the underlay) of its end-nodes. The formal definition is given in Definition 1.

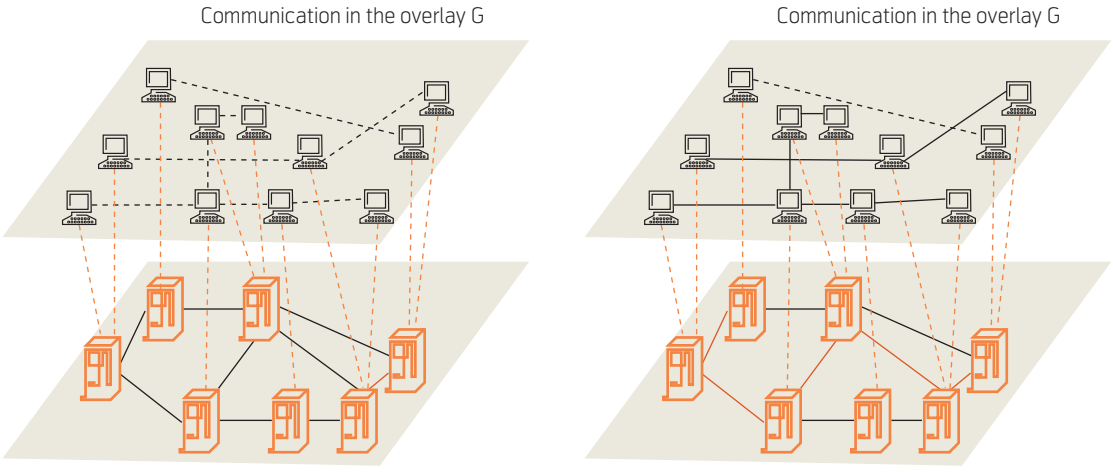
Definition 1: *An overlay is given by a four-tuple* $\mathcal{O} := (G, G', \phi, \pi)$, *where*

- $G = (V, E, \omega)$ *and* $G' = (V', E', \omega')$ *are two weighted graphs with* $\omega: E \to \mathbb{R}$ *and* $\omega': E' \to \mathbb{R}$,
- $\phi: V \to V'$ *is a mapping of the nodes of* G *to the node set of* G', *and*
- $\pi: E \to \{p \mid p$ *is a (un-/directed) path in* $G'\}$ *is a mapping of edges in* G *to paths in* G' *such that* $\{\text{source}(\pi(\{u, v\})), \text{target}(\pi(\{u, v\}))\} = \{\phi(u), \phi(v)\}$.

The interpretation of Definition 1 is that G models the overlay network itself, the graph G' corresponds to the hosting underlay, and the two mappings establish the connection between the two graphs. An example is given in Figure 1. As direct communications in the overlay, which corresponds to the edges of G, is realized by routing information along certain paths in the G', not all parts of the underlay graph are equally important. In order to focus on the relevant parts, we associate an *induced underlay* with an overlay. The corresponding definition is given in 2.

Definition 2: *Given an overlay* $\mathcal{O} := (G = (V, E, \omega), G' = (V', E', \omega'), \phi, \pi)$. *The induced underlay* $\tilde{\mathcal{O}} := H := (V'', E'', \omega'')$ *is a weighted graph, where*

- $V'' := \{v \in V' \mid \exists\, e \in E: \pi(e)$ *contains* $v\}$,
- $E'' := \{e' \in E' \mid \exists\, e \in E: \pi(e)$ *contains* $e\}$, *and*
- $\omega''(e') := \sum_{e \in E} \omega(e) \cdot [e'$ *contained in* $\pi(e)]$.



(a) Both networks G and G' with the mapping ϕ

(b) Highlighting one edge e in G and the corresponding path $\pi(e)$ in G'

Figure 1 Example of an overlay $\mathcal{O} := (G, G', \phi, \pi)$. *The mapping* ϕ *is represented by dashed lines between nodes in* G *and* G'

ISSN 0085-7130 © Telenor ASA 2008

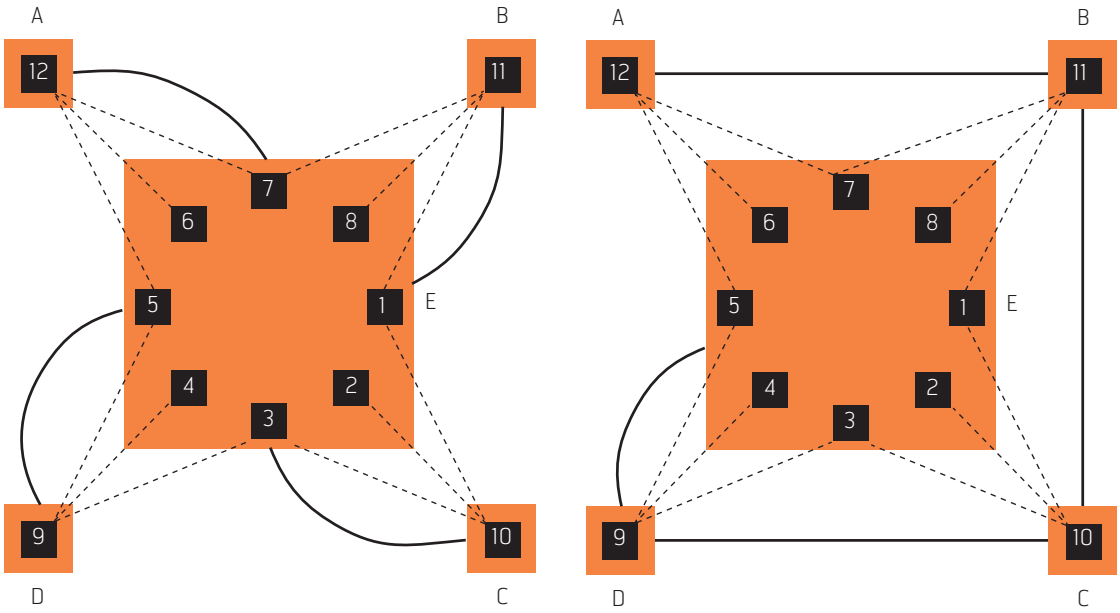


Figure 2 Examples of two overlays where only the topology in the underlay network G' changes. Nodes in the overlay network are numbered with integers, and edges are dashed, while nodes in the underlay network are labelled with characters, and edges are unbroken lines. In both cases the routing π is done via shortest-path scheme

The weight function ω'' is also called appearance weight.

The definition of ω'' is given in the Iverson Notation [9]. The term inside the squared parentheses is a logical statement and depending on its value, the term evaluates to 1, if its value is true, and to 0 otherwise. In other words, the induced underlay corresponds to the subgraph of the underlay graph that is required to establish the communication in the overlay graph. Note that the defined weight can be interpreted as the load caused by the communication and thus is per se independent of a weighting in the underlay network.

2.1 Analysis

In the analysis of overlays, we focus on two important aspects: the identification of key features with respect to the underlay and the comparison of different overlays.

property	A	B	C	D	E
number of hosting nodes	1	1	1	1	8
number of edge in the overlay network having an end-node in the node	3	3	3	3	12
UN weighted degree (star top.)	1	1	1	1	4
UN weighted degree (path top.)	1	2	2	2	1
IU weighted degree (star top.)	3	3	3	3	12
IU weighted degree (path top.)	3	9	15	21	12

Table 1 Degree information of the examples given in Figure 2. The weighted degree corresponds to the weighted degree in the underlay network (UN) and the induced underlay (IU), respectively

The first part, the identification of key features, consists of standard tasks of network analysis, e.g. determining important and relevant nodes or edges, clustering nodes with similar patterns, and detecting unusual constellations. As existing techniques can be applied to the overlay network and the induced underlay, these standard tasks are reasonably well understood in the case of the analysis of a single network. However, these techniques do not incorporate the relationship between the two networks. An example showing such dependencies is given in Figure 2 with the corresponding information about the degrees in Table 1. We use the degree, which is a popular feature, for illustration. However, in our studies we noted that these observations carry over to other characteristics. First note that the number of hosting nodes and the number of communications a node in the underlay participates in gives a first impression about its role in the network. Both pieces of information can be read off the overlying graph G. However, they are completely independent from the routing structure in the underlay. As the example illustrates, the degree of a node (in the induced underlay) heavily depends on the routing structure. In the case of the star topology, both the weighted degree in the underlying network and in the induced underlay are fairly similar, here they are even proportional and clearly identify the centre node of the star to be central for the network. The situation drastically changes when using a path topology. Although all communications

start/terminate at node E, it is not very central. The nodes C and D take on very active roles, due to the fact that most/all communication has to be routed through them. In many cases, the information provided by the induced underlay sufficiently codes the relation between the overlay and underlay networks, while still enabling us to use standard notation of network analysis. On the other hand, there are some scenarios where the provided view is too coarse. For example, it could make a difference whether a heavy edge is caused by a single heavy communication or by a multitude of small communications or, conversely, whether all communications of a node in the induced underlay have only one target in the overlay or are distributed over many targets.

One motivation for identifying key features is to build a proper model that can be used for extensive simulations. For example, simulations are used to predict scaling behaviour or to experimentally validate heuristics, enhancements, or novel techniques. As such, it is a major issue to structurally compare different overlays with each other. On the one hand, our model already reflects all dependencies between the underlay and the overlay network and, thus, it does not require the underlay network, embedding, or routing to be fixed for different instances. On the other hand, due to this elaboration of our model, a simple matching of nodes or edges will not suffice. Our idea is to match key features. For example, one can try to match the appearance weight of an edge with structural properties of its end-nodes. If both overlays have a sufficient number of such matches, it is reasonable to assume that they are created by the same mechanism.

Both parts, the identification of key features and the comparison of overlays, benefit from proper analytic visualizations that emphasize relevant aspects of the corresponding networks and make them easier perceivable. Before presenting two visualization techniques (Section 3), we briefly demonstrate our model and methodology with some experimentally generated examples.

2.2 Examples

In the following, we demonstrate our model and methodology with simple examples. Before looking at a specific overlay, we give two further intuitions.

First, assume a fixed given underlying network. The overlay communication can thus be interpreted as a sampling process of pairs in the underlay. Depending on the application, different patterns occur. For example, in services such as Internet broadcast, one can expect few highly active nodes, which correspond to the hosts of the service while the majority of nodes participate in only a few communications. Using the induced underlay, we can extract such patterns and reconstruct the sampling parameters. Second, assume the underlying network is unknown and acts as a black box, i.e. no information about routing policy and so on is available. By choosing uniformly a sample with sufficiently many communications as the overlay, we can not only discover the underlay, but also partly reverse engineer the routing mechanism of it. In the special case that the overlay network is complete, i.e. every pair of node is connected, the appearance weight of the induced underlay is highly similar to the (edge-)betweenness of the original underlying network.

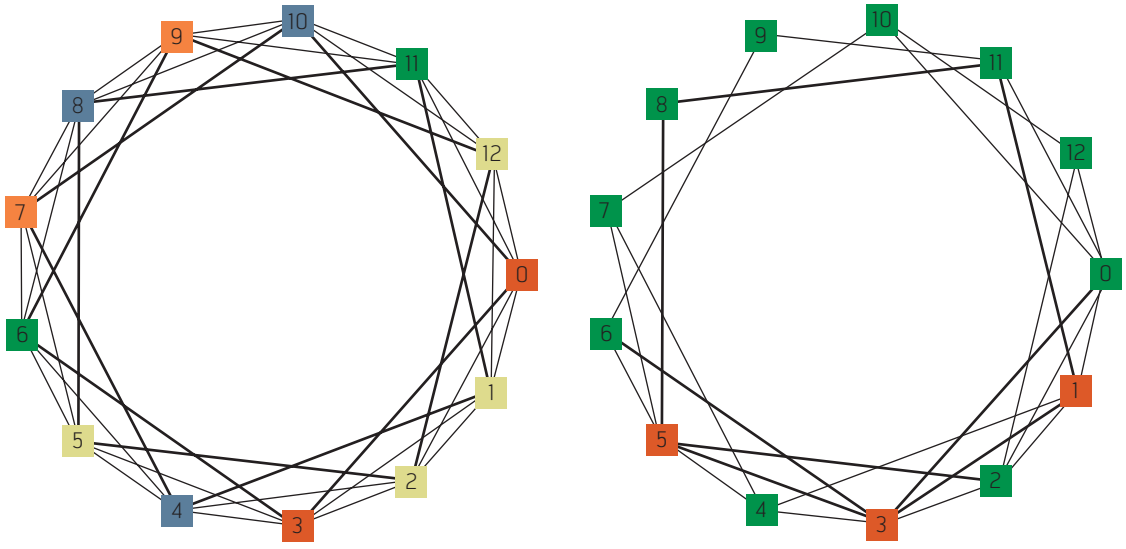


Figure 3 Example of induced underlays for different overlay networks in the same underlying network. In the left figure, the communication is uniformly at random distributed over the network and the color codes the (relative) amount of participation. In the right figure, all communications use at least one red node and select the other uniformly at random. In both cases, the thickness of an edge corresponds to the appearance weight

ISSN 0085-7130

As an example, we consider an underlying network with 13 nodes and a 3-cycle topology, i.e. nodes are cyclic-ordered and each node is connected to 3 of its immediate predecessors and successors. Traffic is routed using shortest path scheme. For simplicity, we set the node-set of the overlay network to the node-set of the underlay and thus ϕ to be the identity function. We define two overlays: the first one $\mathcal{O}_1$ (*uniform sampling*) uses uniformly at random selected pairs of nodes for communication, while in the second overlay $\mathcal{O}_2$(*star-like sampling*) the communication takes place between three predefined nodes and all other nodes chosen uniformly at random. The resulting induced underlays are displayed in Figure 3. As can be clearly seen, the short-cuts, i.e. edges that connect two nodes that have a distance of order two, have the largest appearance weight and all other edges have relatively small weights for the uniform sampling. This is not surprising as the appearance weight resembles the betweenness of edges. The situation drastically changes when modifying the sampling mechanism. As in the case of the induced underlay of $\mathcal{O}_2$, the edges relatively close to the initial set have large weights, and edges far away have small weights or do not appear at all. For example, the non-existence of the edges {9, 10} is due to the fact that no shortest path between a red node and any other node uses that edge. On the other hand, the edge {6, 7} is contained in a shortest path, namely between 3 and 7. However, its absence reveals certain aspects of the underlay routing, i.e., the routing between 3 and 7 will either use the path (3, 4, 7) or (3, 5, 7), but never the path (3, 6, 7).

3 Analytic Visualization

In the following section, we describe two visualization techniques that greatly help in the identification of key features. Both highlight a given hierarchical decomposition of the network while displaying all nodes and edges. They have been successfully applied to the network of Autonomous Systems, which is an abstraction of the physical Internet, yet are highly flexible and can be easily adjusted to other networks.

We use the concept of cores [3, 16] for the required hierarchical decomposition of the network. Briefly, the *k*-core of an undirected graph is defined as the unique subgraph obtained by recursively removing

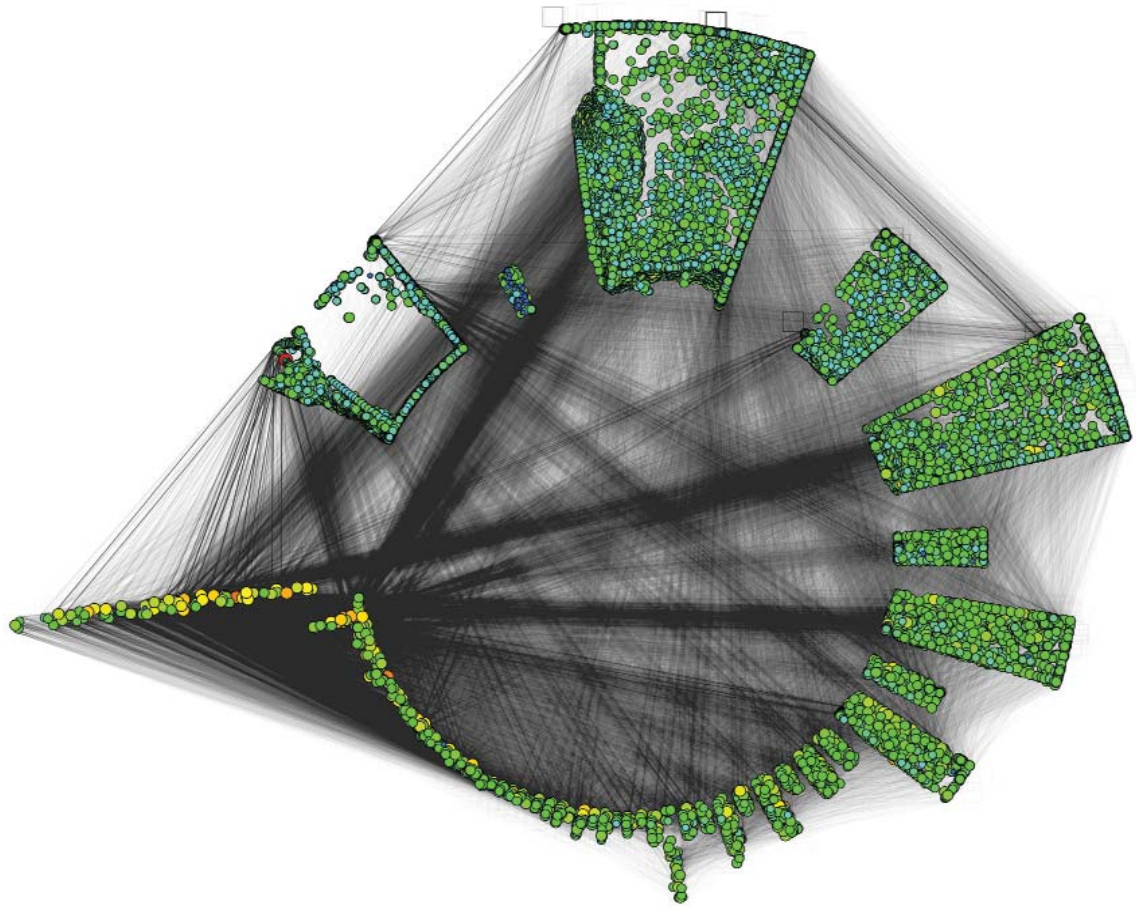


Figure 4 An example visualization of the core decomposition (segments) of the AS network using LunarVis. Each node represents an AS with size and colour reflecting the size of its IP-space. Angular and radial extent of a segment reflects the number of nodes and intra-shell edges respectively. Note the extremely large AS (upper left red node) in the minimum shell

all nodes of degree less than k. A node has coreness, ℓ, if it belongs to the ℓ-core but not to the $(\ell + 1)$-core. The ℓ-shell is the collection of all nodes having coreness ℓ . The core of a graph is the non-empty k-core such that the $(k + 1)$-core is empty. Generally the core decomposition of a graph results in disconnected sub-graphs, but in the case of the AS network we observe that all k-cores stay connected, which is a good feature regarding connectivity. Cores have been frequently used for network analysis, e.g. [5, 6].

The first technique employing the concept of cores was proposed by Baur et al. in [4]. More precisely, their algorithm lays out the graph incrementally starting from the innermost shell, iteratively adding the lower shells. Their implementation uses core decomposition and a combination of spectral and force-directed layout techniques. A successful application of this visualization technique compares actual AS graphs with generated AS graphs. The obtained layouts clearly reveal structural differences between the networks.

The nature of the above layout technique is popularly referred to as a *network fingerprint*. Such pseudo-abstract visualizations offer great informative potential by setting analytic characteristics of a network into the context of its structure, revealing numerous traits at a glance. A fingerprint drawing technique that focuses on the connectivity properties of a network decomposition has been presented in [8]. This approach, coined *LunarVis* lays out each set of a decomposition – which are the shells in our case – individually inside the segments of an annulus. The rough layout of LunarVis is defined by analytic properties of the decomposition, allowing the graph structure to determine the details. By virtue of a sophisticated application of force-directed node placement, individual nodes inside annular segments reflect global and local characteristics of adjacency while the inside of the annulus offers space for the exhibition of the edge distribution. Combined with well-perceivable attributes, such as the size and the colour of a node, these layouts offer remarkable readability of the decompositional connectivity and are capable of revealing subtle structural characteristics.

4 Case Study: Overlay Graphs of P2P systems

In this section, we exemplify our analysis technique with a case study of a P2P overlay. For our analysis we choose Gnutella [7], an unstructured file-sharing system which relies on flooding connectivity pings and search queries to locate content. Each message carries a TTL (time to live) and message ID tag. To improve scalability, nodes are classified in a two-level hierarchy, with high-performance ultrapeer nodes maintaining the overlay structure by connecting with each other and forwarding only the relevant messages to a small number of shielded leaf nodes. Responses to pings and queries are cached, and frequent pinging or repeated searching can lead to disconnection from network. More details about Gnutella can be found at [7].

4.1 Sampling and Modelling the P2P Network

In order to analyze the overlay structure, we first need to identify a representative set of connections, called edges, between nodes in the P2P network. To reduce the bias in our sample, we identify edges where neither of the two end-nodes is controlled by us. We refer to such nodes as remote neighbour servents.

Due to message caching and massive churn in P2P networks (we measured the median incoming/outgoing connection duration to be 0.75/0.98 seconds), a simple crawling approach using pings, e.g. as employed in [14], is not sufficient. However, pings identify nodes that should have been remote neighbour servents at some point.

We thus deploy a combination of active and passive techniques to explore the Gnutella network [1]. Our passive approach consists of an ultrapeer that participates in the network and is attractive to connect to. It shares 100 randomly generated music files (totalling 300 MB in size) and maintains 60 simultaneous connections to other servents. The passive approach gives us a list of active servents. The active approach consists of a multiple-client crawler that uses ping with TTL 2 to obtain a list of candidate servents. Since queries are difficult to cache, we use queries with TTL 2 to obtain a set of remote neighbour servents. These servents are then contacted actively to further advance the network exploration. This approach allows us to discover P2P edges that existed at a very recent point of time. When interacting with other servents, our crawler pretends to be a long-running ultrapeer, answering incoming messages, sharing content, and behaving non-intrusively. This pragmatic behaviour avoids bans. The client uses query messages with broad search strings, e.g. mp3, avi, rar, to obtain maximum results. We then combine active and passive approaches by integrating the crawler into the passive ultrapeer.

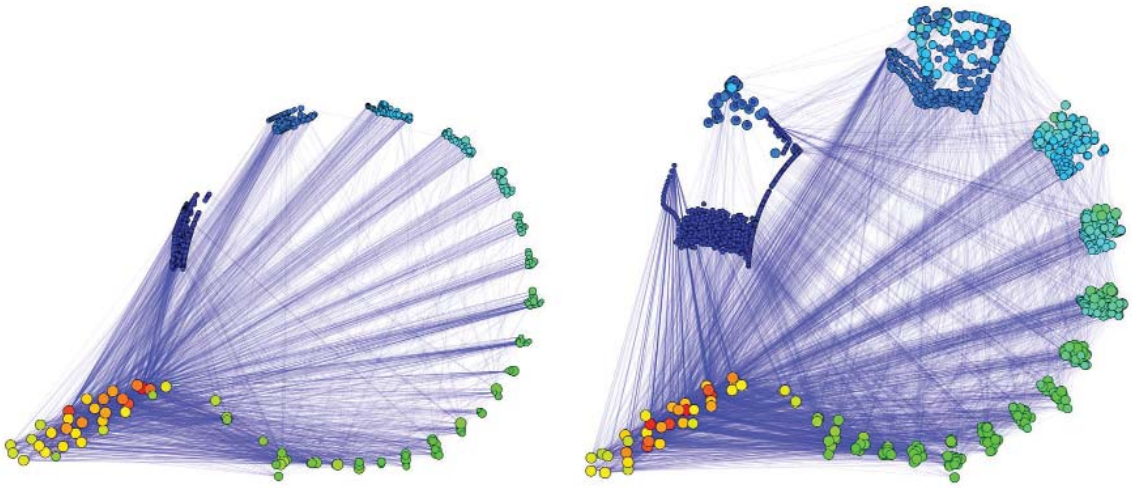
Using this setup, we sample the Gnutella network for one week starting April 14, 2005. The ultrapeer logs 352, 396 sessions and the crawler discovers 234, 984 remote neighbour servents, a figure significantly higher than most reported results during this period.

For each edge of the Gnutella network we map the IP addresses of the Gnutella peers to ASes using the BGP table dumps offered by Routeviews [13] during the week of April 14, 2005. This results in 2964 unique AS edges involving 754 ASes, after duplicate elimination and ignoring P2P edges inside an AS. For the random graph we pick end-points at the IP level by randomly choosing two valid IP addresses from the whole IP space. These edges are then mapped to ASes in the same manner as for the Gnutella edges. This results in 4975 unique edges involving 2095 ASes for the random network at the AS graph level. The different sizes of the graphs are a result of the generation process: we generate the same number of IP pairs for random network as observed in Gnutella, and apply the same mapping technique to both data sets, which abstracts the graph of IPs and direct communication edges to a graph with ASes as nodes and the likely underlay communication path as edges. This way, the characteristics of Gnutella are better reflected than by directly generating a random AS network of the same size as Gnutella network.

For our analysis, we apply the model and methodology from Section 2 as follows. The overlay $\mathcal{O} = (G, G', \phi, \pi)$ as given in Definition 1 uses the direct communication in Gnutella as graph G, the graph G' corresponds to the hosting Internet, in our case the AS level. The mapping ϕ corresponds to the IP to AS mapping, while π is the routing in the AS network. Apart from the already introduced induced underlay, we also investigate the network of direct overlay communication, yet abstracted to the level of ASes in order to be comparable to the induced underlay. Note that in a simplified model, where each communication causes uniform costs, the appearance weight in the induced underlay (ω'') corresponds to the total load caused by the overlay routing in the underlay network. As exact traffic measurements on each underlay link are non-trivial, this can be interpreted as an estimate of the actual load on underlay links due to the overlay traffic.

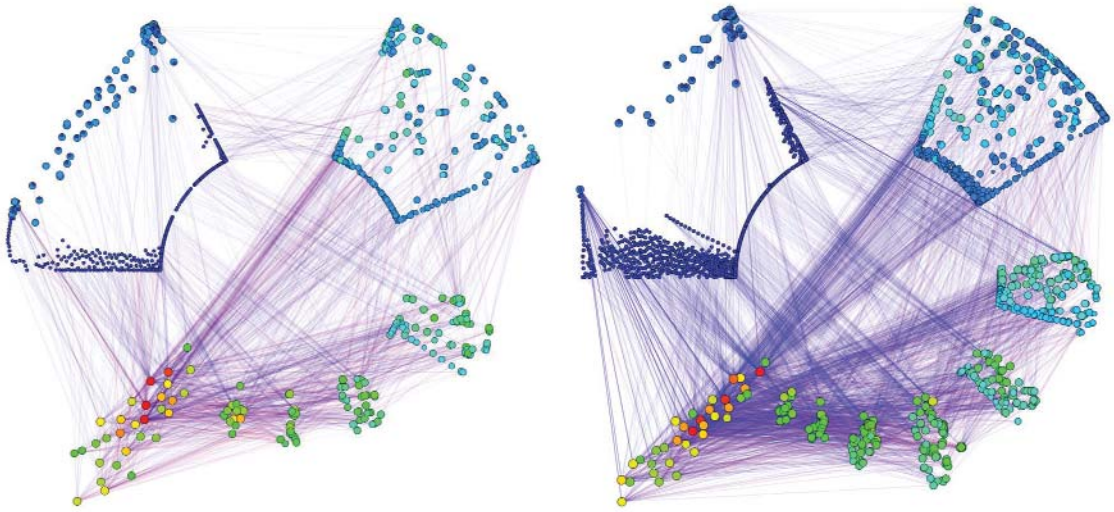
4.2 Overlay-Underlay Correlation in a P2P system

Figure 5 shows visualizations of the direct overlay communication of both the Gnutella network and a random network. Employing the LunarVis [8] technique described in Section 3, these drawings focus on the decompositional properties of the core hierarchy. Numerous observations can be made by comparing the two visualizations. Note, first, the striking lack of intra-shell edges for all but the maximum shell in the Gnutella network (small radial extent). This is also true for edges between shells, as almost all edges are incident to the maximum shell. This means that almost always at least one communication partner is in the maximum shell, a strongly hierarchical pattern that the random network does not exhibit to this degree. Note furthermore that in Gnutella, betweenness centrality (size of a node) correlates well with coreness, a consequence of the strong and deep core hierarchy, whereas in the random network the two- and even the one-shell already contain nodes with high centrality, indicating that many peerings heavily rely on low-shell ASes. The depth of the Gnutella hierarchy (26 levels) is a weighty suggestion of a strongly connected network kernel of ultrapeers,



(a) P2P network (b) Random network

Figure 5 Visualization of the core decomposition of the overlay communication networks. Core-shells are drawn into annular segments, with the 1-shell at the upper left. Angular and radial extent of a segment reflects the number of nodes and intra-shell edges respectively. Inside, each shell node is drawn towards its adjacencies. Colours represent the degree of a node while the size represents their betweenness centrality. Edges are drawn with 10 % opacity and range from blue (small weight) to red (large weight)



(a) P2P network (b) Random network

Figure 6 Visualization of the core decomposition of the induced underlay communication network. These drawings use the same parameters as Figure 5 (a) Gnutella (b) random network

(a) P2P network (b) Random network

Figure 7 Comparison of occurring communication in the P2P network and the Random network, using visualization, see Section 3

which are of prime importance to the connectivity of the whole P2P network. However, note that the distribution of degrees (node colours) does not exhibit any unusual traits and that no heavy edges are incident to low-shell ASes, in either network.

Figure 6 visualizes the induced underlay communication of both the Gnutella network and a random network, employing the same technique and parameters as in Figure 5. The drawings immediately indicate the much smaller number of ASes and overlay nodes in the Gnutella network. As a consequence, more heavy edges (red) exist and the variance in the appearance weight (edge colour) is more pronounced. This is because of the fact that not all the ASes host P2P users (this is in accordance with our measurements in Section 4.1), as is the case for the random network. Again, the distributions of degrees do not differ significantly.

For a closer comparison, Figure 7 shows a top-down view of the visualizations of communication edges in Gnutella and random network. The visualization technique places nodes with dense neighbourhoods (tier-1 and tier-2 ASes) towards the centre, and nodes with lesser degrees (tier-3 customer ASes) towards the

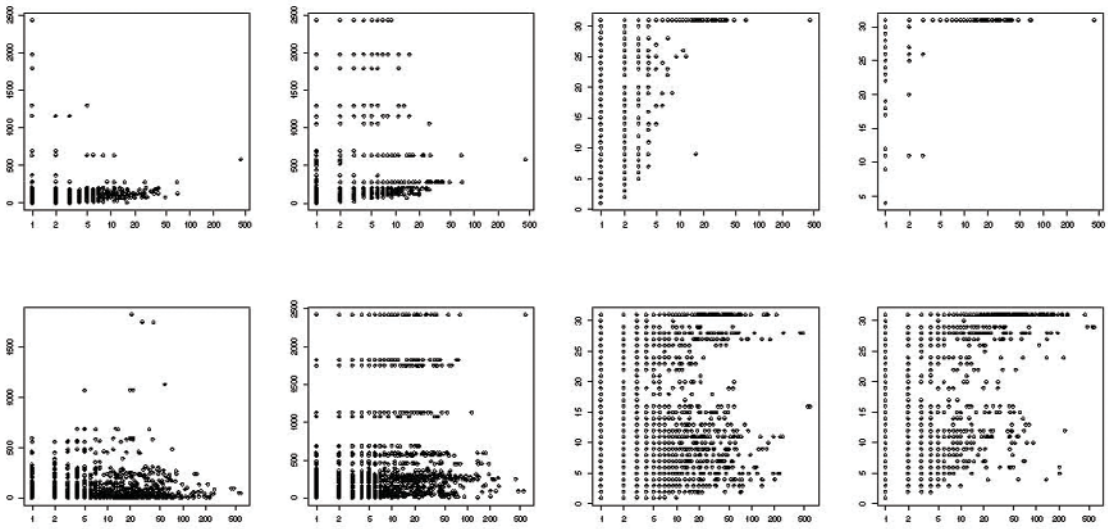


Figure 8 Comparing appearance weight with minimum (first column) and maximum (second column) degree and coreness (third and fourth columns) of the corresponding end-nodes in Gnutella (top row) and the random network (bottom row). Each data point represents an edge, the x-axis denotes the appearance weight and the y-axis reflects the degrees (coreness) of the end-nodes. All axes use logarithmic scale

periphery. We can observe that while both networks have many nodes with large degrees in the centre, the random network possesses several nodes with large degree in the periphery. Gnutella, on the other hand, has almost no nodes with large degree in the periphery in both models. Moreover, this pattern is more pronounced for Gnutella in the direct overlay communication model, while the random network is largely similar in both models. In other words, it appears that Gnutella peering connections tend to lie in ASes in the core of the Internet where there may be high-bandwidth links available.

To further corroborate our observations, we investigate structural dependencies between the induced underlay communication model and the actual underlay network, by comparing the appearance weight with node-structural properties of the corresponding end-nodes in the original underlay. We focus on the properties degree and coreness, as both have been successfully applied for the extraction of customer-provider relationship as well as visualization [18,5], due to the ability of these properties to reflect the importance of ASes. We systematically compare the weight of an edge with the minimum and maximum degree and coreness of its end-nodes. Figure 8 shows the corresponding plots.

From the plots of minimum and maximum degree it is apparent that the appearance weight of an edge and its end-nodes' degrees are not correlated in both the Gnutella and the random network, as no pattern is observable. Also, the distributions are similar as the majority of edges are located in the periphery of the network where the maximum degree of the end-nodes is small. We thus hypothesize that the relation of load in the P2P network and node degree in the underlying network is the same in both the Gnutella and the random network. In other words, the Gnutella network does not appear to be significantly affected by the node degree of underlay nodes.

However, considering the coreness reveals interesting observations. From the networks of minimum and maximum coreness in Figure 8, we can observe that although there is no correlation in either of the two networks, their distributions are different. In the random network the distributions are very uniform, which reflects its random nature. But in the case of Gnutella almost no heavy edge is incident to a node with small coreness, as can be seen in the minimum coreness diagram. Positively speaking, most edges with large appearance weights are incident to nodes with large minimum coreness. Interpreting coreness as importance of an AS, these Gnutella edges are located in the backbone of the Internet, an important observation. The same diagram for the random network does not yield a similar significant distribution, thus denying a comparable interpretation. For instance, in the random network, there exist edges located in the periphery that are heavily loaded. As an aside, backbone edges need not necessarily be heavily loaded in either network.

All these observations and analysis show that the Gnutella network differs from random networks and there appears to be some correlation of Gnutella topology with the Internet underlay.

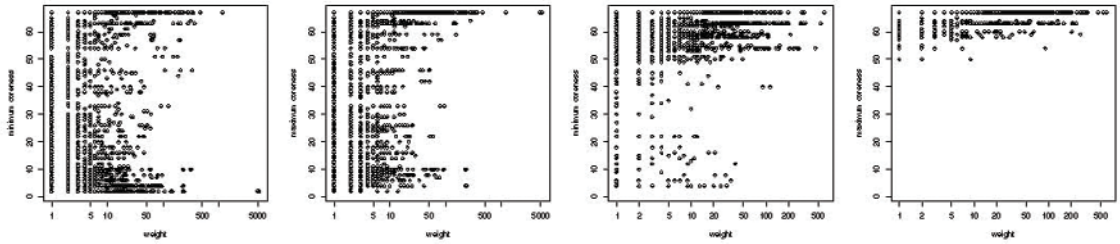


Figure 9 Comparing appearance weight with minimum and maximum coreness of the corresponding end nodes in certain random networks. In the first and second pairs of figures, the communicating IPs that are hosted by ASes with low coreness (≤ 4) and with high coreness (≥ 50), respectively

4.3 Sensitivity Analysis to Refine the Model

The analyses conducted in Section 4.2 and the conclusions drawn, solely rely on analytic visualizations. Based on these we now aim at a deeper understanding of the properties of the underlay communication the P2P network induces. Modifying the generation process for the random networks in ways suggested by our observation, we are now able to conduct a sensitivity analysis, in order to find parameters for the random network that lead to a more aligned edge coreness distribution with the observed P2P network.

It is reasonable to assume both that many nodes are in lower shells (customer ASes) and that heavy nodes (ultrapeers) are in higher shells. Therefore we consider two modifications: The *low coreness communication* restricts the IP-spaces that are available for communications to those hosted by ASes with low coreness. Analogously the *high coreness communication* uses only IPs located in ASes with high coreness. For reasons of space and simplicity we present only the plots of two of our various experiments. In order to model the routing in the Internet more accurately, we considered the AS network as directed and thus had to adjust the coreness calculation properly. As a rule of thumb, the values roughly double compared to the original scenario described in Section 4.2. Figure 9 shows the plots that correspond to the right four diagrams in Figure 8. Again a data point is plotted for each edge in the induced underlay, with coordinates that correspond to its appearance weight (x-axis) and to its minimum/maximum incident node coreness (y-axis). The corresponding plots of the degree distributions are omitted as they did not differ much.

At first glance we can observe that the restriction to low coreness communication does not yield a significant difference to the corresponding plot of our initially unrestricted random network (Figure 8 lower right). Although the distributions are shaped in a highly similar manner, they differ in the maximum occurring appearance weight. On the other hand, the high coreness communication exhibits a very different pattern. Its distributions are more similar to those from Gnutella than the random ones. A very interesting observation is that although communicating IPs are located in ASes with high coreness, some routing path uses low-coreness ASes.

Interpreting these findings, we conclude that the observed part of Gnutella mainly corresponds to a large part of the network spanned by the ultrapeers and only few leaf nodes are included. Typically ultrapeer nodes maintain a connection to a certain (small) number of leaf nodes. On the other hand, the leaf nodes possess only slow Internet connections and connect to the well performing ultrapeers, who in turn shield them from a large amount of P2P traffic, yet enable them to locate and share content. The well-know effect of rampant free-riding corroborates our interpretation. More precisely, the phenomenon refers to the fact that a large number of nodes remain online for very short durations, share no content, and are only interested in finding content, while a small percentage of nodes participate in the network for very long durations, and provide most of the content sought in the network. Hence, they participate in much more communications as compared to the other P2P nodes.

5 Conclusion

In this paper, we present a novel model and technique to analyze the overlay in the context of the underlying network. The major focus of our analysis is the identification of key features as well as the structural comparison between different overlays. More precisely, we transform the overlay to a corresponding subgraph in the underlying network that is crucial for the functionality required by the overlay.

The driving force behind this work is the engineering of overlays which is demonstrated using a case study of the real-world Gnutella network. On the one hand, our analysis reveals differences between it and experimental mimics that are founded on the same principles and prerequisites. On the other hand, by repeatedly modifying and adjusting the corresponding generation process, based on the insights obtained

through detailed analysis and visualization, we are able to deepen our understanding of the real-world instance. In addition, we identify certain artefacts that incite further research. More precisely, our extensive case study incorporates existing visualization techniques for the underlying Internet and establishes that while overlay networks like Gnutella use an arbitrary neighbourhood selection process, their topology differs from randomly generated networks.

Our methodology of analyzing the overlays and underlays supported by analytic visualizations, offers a powerful and flexible tool in the general engineering process of overlays.

References

1 Aggarwal, V, Bender, S, Feldmann, A, Wichmann, A. Methodology for Estimating Network Distances of Gnutella Neighbors. In: *INFORMATIK 2004 – GI Jahrestagung*, 219–223, 2004.

2 Aggarwal, V, Feldmann, A, Scheideler, C. Can ISPs and P2P Systems Cooperate for Improved Performance? In: *ACM SIGCOMM Computer Communication Review*, 37(3), 2007.

3 Batagelj, V, Zaversnik, M. Generalized Cores. Preprint 799, *IMFM Ljublana*, Ljubljana, 2002.

4 Baur, M, Brandes, U, Gaertler, M, Wagner, D. Drawing the AS Graph in 2.5 Dimensions. In: *Proceedings of the 12th International Symposium on Graph Drawing (GD'04)*. Springer-Verlag, January 2005. Lecture Notes in Computer Science, 3383, 43–48.

5 Gaertler, M, Patrignani, M. Dynamic Analysis of the Autonomous System Graph. In: *IPS 2004 – Inter-Domain Performance and Simulation*, 13–24, March 2004.

6 Gkantsidis, C, Mihail, M, Zegura, E W. Spectral Analysis of Internet Topologies. In: *Proceedings of Infocom'03*, 2003.

7 *Gnutella*. Wikipedia, the free encyclopedia. [online] – URL: http://en.wikipedia.org/wiki/Gnutella. Accessed 31 August 2007

8 Görke, R, Gaertler, M, Wagner, D. LunarVis - Analytic Visualizations of Large Graphs. In: *Proceedings of the 15th International Symposium on Graph Drawing (GD'07)*. Springer-Verlag, 2008. Lecture Notes in Computer Science. To appear.

9 Knuth, D E. Two notes on notation. *American Mathematical Monthly*, 99, 403–422, 1990.

10 Liu, Y, Zhang, H, Gongi, W, Towsley, D. On the interaction between overlay routing and traffic engineering. In: *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, 4, 2543-2553. IEEE Computer Society Press, March 2005.

11 Nakao, A, Peterson, L, Bavier, A. A Routing Underlay for Overlay Networks. In: *SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, 11-18. ACM Press, 2003.

12 Ratnasamy, S, Handley, M, Karp, R M, Shenker, S. Topologically-aware overlay construction and server selection. In: *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, 1, 1190-1199. IEEE Computer Society Press, 2002.

13 *University of Oregon Routeviews Project*. October 17, 2007 [online] – URL: http://www.routeviews.org/.

14 Saroiu, S, Gummadi, P K, Gribble, S D. A measurement study of peer-to-peer file sharing systems. In: *Proceedings of SPIE/ACM Conference on Multimedia Computing and Networking (MMCN) 2002*, January 2002.

15 Seetharaman, S, Ammar, M. On the interaction between dynamic routing in native and overlay layers. In: *Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, 1-12. IEEE Society Press, April 2006.

16 Seidman, S B. Network Structure and Minimum Degree. *Social Networks*, 5, 269–287, 1983.

17 Steinmetz, R, Wehrle, K (Eds.). Peer-to-Peer Systems and Applications. *Lecture Notes in Computer Science*, 3485. Springer, 2005.

18 Subramanian, L, Agarwal, S, Rexford, J, Katz, R H. Characterizing the Internet Hierarchy from Multiple Vantage Points. In: *Proceedings of Infocom'02*, 618–627, 2002.

Vinay Aggarwal received a Bachelor of Engineering (BE) degree from Gujarat University in 2001, India, and his Masters in Computer Science at the Max-Planck-Institut für Informatik (MPI), Universität des Saarlandes, Saabrücken, Germany. Today he is a PhD student at the Deutsche Telekom Laboratories, TU Berlin, researching innovation solutions that will enable ISP performance improvements, cost savings and end-user experience.

email: vinnay.aggarwal@telekom.de

Anja Feldmann is a Professor at the Deutsche Telekom Laboratories at the Technical University Berlin. Previously she was a Professor in the computer science department at the Technical University of Munich and at the computer science department at University of Saarbrücken in Germany. Before that, she was a member of the IP Network Measurement and Performance Department at AT&T Labs. Her current research interest is network performance debugging.

email: anja.feldmann@telekom.de

Robert Görke studied technical mathematics at Universität Karlsruhe, Germany, and received his Diplom in 2005. Since then he has been a PhD student at the chair of Algorithmics I of the Faculty of Informatics in Karlsruhe. His research focuses are the clustering of graphs both in a static and in an evolving environment and the analytic visualization of networks.

email: robert.goerke@kit.edu

Marco Gaertler studied mathematics at Universität Konstanz, Germany, and received his Diplom in 2002. In 2007 he received his PhD in computer science at the chair of Algorithmics I of the Faculty of Informatics in Karlsruhe. His research focuses as a PostDoc are fingerprinting and the analytic visualization of static and evolving networks as well as graph clustering.

email: marco.gaertler@kit.edu

Dorothea Wagner received the Diplom and PhD degrees in mathematics from the Rheinisch-Westfälische Technische Hochschule at Aachen, Germany, in 1983 and 1986, respectively. In 1992, she received the Habilitation degree from the Department for Mathematics of the Technische Universität Berlin. Until 2003 she was a Full Professor of computer science at the University of Konstanz, and since then holds this position at Universität Karlsruhe. Her research interests include discrete optimization, graph algorithms, and algorithm engineering. Alongside numerous positions in editorial boards she is vice president of the German Research Foundation (DFG) since 2007.

email: dorothea.wagner@kit.edu

Terms and Acronyms in Network Analysis

Acronym/ Term	Definition	Explanation	Web Resources
ACE	Agent-based Computational Economics	ACE is the computational study of economic processes modeled as dynamic systems of interacting agents.	
ACO	Ant Colony Optimization	ACO, introduced by Marco Dorigo in 1992 in his PhD thesis, is a probabilistic technique for solving computational problems which can be reduced to finding good paths through graphs. They are inspired by the behaviour of ants in finding paths from the colony to food.	http://www.aco-metaheuristic.org/
AS	Autonomous System network	Term used for large national networks – the backbone of Internet.	
ATM	Asynchronous Transfer Mode	A high bandwidth, low-delay, connection-oriented, packet-like switching and multiplexing technique. ATM allocates bandwidth on demand, making it suitable for high-speed connections of voice, data and video services. Access speeds are up to 622 Mb/s and backbone networks currently operate at speeds as high as 2.5 Gb/s. Standardised by ITU-T.	http://www.itu.int
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 IETF RFC 1771.	http://www.ietf.org /rfc/rfc1771.txt
BISON	Biology-Inspired techniques for Self-Organization in dynamic Networks	European research project in the 5th Framework Programme IST running from 2003 to 2005. The aim of BISON was to explore the use of ideas derived from complex adaptive systems (CAS) to enable the construction of robust and self-organizing information systems for deployment in highly dynamic network environments.	http://cordis.europa.eu
CC	Connected Component	In an undirected graph, a *connected component* or *component* is a maximal connected subgraph. [Source: Wikipedia]	
CE	Cross Entropy	In information theory, the cross entropy between two probability distributions measures the average number of bits needed to identify an event from a set of possibilities, if a coding scheme is used based on a given probability distribution q, rather than the "true" distribution p.	
CEAS	Cross-Entropy Ant System	CEAS is a distributed version of Reuven Rubinstein's popular cross entropy method for optimization. CE-ants have been shown to find near-optimal solutions to NP-complete path-finding problems of large scale.	
CPLEX		Optimization Software Package from ILOG which can be used to solve linear and quadratic problems.	http://www.ilog.com /products/cplex/
DAG	Directed Acyclic Graph	A DAG is a directed graph with no directed cycles; that is, for any vertex v, there is no nonempty directed path that starts and ends on v. [Source: Wikipedia]	
DELIS	Dynamically Evolving, Large-scale Information Systems	European research project in the 6th Framework Programme IST running from 2004 to 2008 with the objective to develop methods, techniques and tools to cope with challenges imposed by the size and dynamics of today's and especially future information systems, in an interdisciplinary effort of Computer Science, Physics, Biology, and Economy.	http://delis.upb.de, http://cordis.europa.eu
DFG	Deutschen Forschungsgemeinschaft (German Research Foundation)	The central, self-governing research funding organisation that promotes research at universities and other publicly financed research institutions in Germany. The DFG serves all branches of science and the humanities by funding research projects and facilitating cooperation among researchers.	http://www.dfg.de

ISSN 0085-7130 © Telenor ASA 2008

Acronym/ Term	Definition	Explanation	Web Resources
HITS		HITS is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. US patent 6 112 202.	http://www2002.org /CDROM/refereed /643/node1.html
ICT	Information and Communication Technology	The technology required for information processing. In particular the use of electronic computers and computer software to convert, store, protect, process, transmit, and retrieve information from anywhere, anytime.	
IM	Instant Messaging	An instant messaging service is reached by the use of an instant messenger client. Instant messaging differs from e-mail in that conversations happen in real-time. Also, most services convey an "online status" between users, such as if a contact is actively using the computer. Generally, both parties in the conversation see each line of text right after it is typed (line-by-line), thus making it more like a telephone conversation than exchanging letters. Instant messaging applications may also include the ability to post an away message, the equivalent of the message on a telephone answering machine. Popular instant messaging services on the public Internet include Jabber, AOL Instant Messenger, Yahoo! Messenger, .NET Messenger Service and ICQ. These services owe many ideas to an older (and still popular) online chat medium known as Internet Relay Chat (IRC).	
IS-IS	Intermediate System to Intermediate System	Protocol for routing in an Internet domain, based on Dijkstra's algorithm, standardised as ISO10589. Specified in IETF RFC 1195.	http://www.ietf.org /rfc/rfc1195.txt
ISP	Internet Service Provider	A vendor who provides access for customers to the Internet and the World Wide Web. The ISP also typically provides a core group of internet utilities and services like e-mail and news group readers.	
IU	Induced Underlay network		
LA	Link Analysis	The analysis of networks through graph theory. The networks may be social, transportation, electrical, biological, Internet networks, etc. Analysis includes descriptions of structure, such as small-world networks, social circles or scale-free networks, optimisation, such as Critical Path Analysis and PERT (Program Evaluation & Review Technique), and properties such as flow assignment.	
MATLAB	MATrix LABoratory	MATLAB is a numerical computing environment and programming language. (Not to be confused with Matlab Upazila in Chandpur District, Bangladesh.)	http://www .mathworks.com/
MMS	Multimedia Message Service	MMS is a communications technology developed by 3GPP (Third Partnership Project) that allows users to exchange multimedia communications between capable mobile phones and other devices. An extension to the Short Message Service (SMS) protocol, MMS defines a way to send and receive, almost instantaneously, wireless messages that include images, audio, and video clips in addition to text.	http://www.3gpp.org
MODAGENT	Modelling System for Agent-based Computational Economics	Agent-based Computational Economics is the computational study of economic processes modeled as dynamic systems of interacting agents.	
MPLS	Multi Protocol Label Switching	An IETF standard intended for Internet application. MPLS has been developed to speed up the transmission of IP based communications over ATM networks. The system works by adding a much smaller "label" to a stream of IP datagrams allowing "MPLS" enabled ATM switches to examine and switch the packet much faster. It is specified in IETF RFC 4368.	http://www.ietf.org /rfc/rfc4368.txt
NDD	Node Degree Distribution	In graph theory, degree distribution gives the probability distribution of degrees in a network. Its use originates from the study of random graph by Paul Erdös and Alfréd Rényi, and it has become an important concept which describes the topology of complex networks.	
NO	Network Operator	A network operator, also known as carriage service provider (CSP), wireless service provider, wireless carrier, mobile phone operator, or cellular company, is a telephone company that provides services for mobile phone subscribers.	

ISSN 0085-7130 © Telenor ASA 2008

Acronym/ Term	Definition	Explanation	Web Resources
NOK	Norwegian Kroner		
OSPF	Open Shortest-Path First	A link-state, hierarchical Interior Gateway Protocol (IGP) routing protocol. The shortest path tree is calculated using cost as its routing metric. A link state database is constructed of the network topology which is identical on all routers in the area.	
P2P	Peer-To-Peer	A computer network that does not rely on dedicated servers for communication but instead mostly uses direct connections between clients (peers). A pure peer-to-peer network does not have the notion of clients or servers, only equal peer nodes that simultaneously function as both "clients" and "servers" to the other nodes on the network.	
POTS	Plain Old Telephone Service	A very general term used to describe an ordinary voice telephone service. See also PSTN.	
PSTN	Public Service Telephone Network	Common notation for the conventional analog telephone network.	
QAP	Quadratic Assignment Problem	QAP is one of fundamental combinatorial optimization problems in the branch of optimization or operations research in mathematics, from the category of of the facilities location problems.	
QoS	Quality of Service	The "degree of conformance of the service delivered to a user by a provider, with an agreement between them". The agreement is related to the provision/ delivery of this service. Defined by EURESCOM project P806 in 1999 and adopted by ITU-T in recommendation E.860.	http://www.itu.int, http://www.eurescom.de
SALSA	Stochastic Approach for Link-Structure Analysis	This is a link analysis algorithm that examines random walks on two different Markov chains which are derived from the link structure of the WWW. An approach for finding hubs and authorities in a network. First described by R. Lempel and S. Moran in 2001.	
SBPP	Shared Backup Path Protection		
SCC	Strongly Connected Component	A directed graph is called *strongly connected* if for every pair of vertices *u* and *v* there is a path from *u* to *v* and a path from *v* to *u*. The SCC of a directed graph are its maximal strongly connected subgraphs. C18	
SDH	Synchronous Digital Hierarchy	SDH (Synchronous Digital Hierarchy) is a standard technology for synchronous data transmission on optical media. It is the international equivalent of North Amercan SONET (Synchronous Optical Network). Both technologies provide faster and less expensive network interconnection than traditional PDH (Plesiochronous Digital Hierarchy) equipment. It is a method of transmitting digital information where the data is packed in containers that are synchronized in time enabling relatively simple modulation and demodulation at the transmitting and receiving ends. The technique is used to carry high capacity information over long distances. SDH uses the following Synchronous Transport Modules (STM) and rates: STM-1 (155 megabits per second), STM-4 (622 Mb/s), STM-16 (2.5 Gb/s), and STM-64 (10 Gb/s). SDH is specified by ITU-T G.707.	www.itu.int
SMS	Short Message Service	A means by which short messages can be sent to and from digital cellular phones, pagers and other handheld devices. Alphanumeric messages of up to 160 characters can be supported.	
SNA	Social Network Analysis	SNA is the mapping and measuring of relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships. [Source: orgnet.com]	
SQG	Stochastic programming software environment	Software for solving stochastic programming problems with stochastic gradient methods.	

Acronym/ Term	Definition	Explanation	Web Resources
T-RANK		This is a link analysis algorithm giving a single score to each web page. US patent no. 7 281 005.	
TTL	Time-To-Live	A limit on the period of time or number of iterations or transmissions in computer and computer network technology that a unit of data (e.g. a packet) can experience before it should be discarded. In IPv4, time to live is an 8-bit field in the Internet Protocol (IP) header. It is the 9th octet of 20. The time to live value can be thought of as an upper bound on the time that an IP datagram can exist in an internet system. The TTL field is set by the sender of the datagram, and reduced by every host on the route to its destination. If the TTL field reaches zero before the datagram arrives at its destination, then the datagram is discarded and an ICMP (Internet Control Message Protocol) error datagram is sent back to the sender.	
UN	Underlay Network		
VC	Virtual Connection, Virtual Circuit	A connection oriented communication service that is delivered by means of packet mode communication. After a virtual connection is established between two nodes or application processes, a bit stream or byte stream may be delivered between the nodes. A virtual circuit protocol hides the division into segments, packets or frames from higher level protocols.	
VNO	Virtual Network Operator	A VNO is a company that provides mobile phone services but does not have its own frequency allocation of the radio spectrum, nor does it have all of the infrastructure required to provide mobile telephone service.	
VoIP	Voice over Internet Protocol	Voice over Internet Protocol is the routing of voice conversations over the Internet or any other IP-based network. The voice data flows over a general-purpose packet-switched network, instead of traditional dedicated, circuit-switched voice transmission lines. Several standards exist to support VoIP, like H.323 from ITU-T and SIP (IETF RFC 3261).	http://www.itu.int, http://www.ietf.org
WS-graph	Watts-Strogatz-graph	The Watts and Strogatz model is a random graph generation model that produces graphs with small-world properties, including short average path lengths and high clustering. It was proposed by Duncan J. Watts and Steven Strogatz in their joint 1998 *Nature* paper. The model also became known as the (Watts) beta model after Watts used β to formulate it in his popular science book *Six Degrees*.	
WWW	World Wide Web	An international, virtual network based information service composed of Internet host computers that provide on-line information. A hypertext-based, distributed information system/service created by researchers at CERN in Geneva, Switzerland in 1991. Users may create, edit or browse hypertext documents. The clients and servers are freely available.	http://www.w3c.org
XPRESS	Optimization Software Package	Optimization Software Package. A Linear Programming (LP) and Mixed Integer Programming (MIP) solver from Dash. XPRESS can also be used to solve Quadratic Program (QP) and Mixed Integer Quadratic Program (MIQP) problems.	http://www .dashoptimization .com/

ISSN 0085-7130

Status

Introduction

PER HJALMAR LEHNE

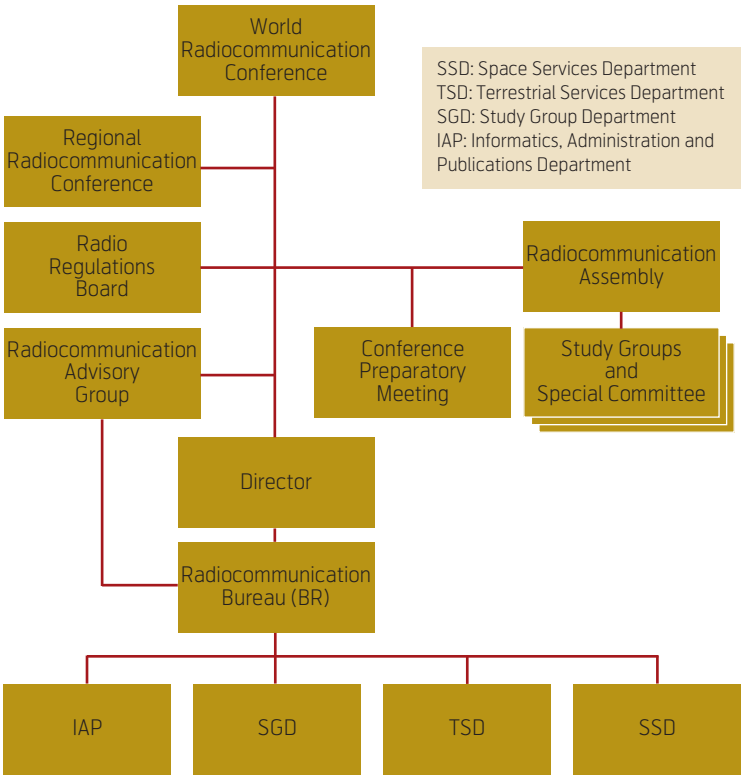
Per Hjalmar Lehne is Researcher in Telenor R&I and Editor-in-Chief of Telektronikk

This issue of *Telektronikk*'s status section contains three articles, all related to the *International Telecommunications Union – ITU*.

Two important global events for the wireless community took place in Geneva in October – November 2007. The *ITU-R Radiocommunication Assembly – RA-07* and the *World Radiocommunication Conference – WRC-07* take place very three to four years. Both these events are important because they lay the foundations for global common use of frequencies and wireless standards. *Anne Lise Lillebø*, *Terje Tjelta* and *Erik Otto Evenstad* have written two comprehensive reports from these two events.

The Radio Assembly provides the necessary technical basis for the WRC. One of the important decisions that were taken on the RA-07 was to approve Mobile WiMAX as a member of the IMT-family of 3G technologies.

The World Radiocommunication Conference addressed the worldwide use of radio spectrum. The pressure is high for new frequencies for IMT, especially in the UHF and C-band.



The ITU-R organization

The 470 – 862 MHz UHF band is currently used for terrestrial analogue broadcast services, but with the introduction of new digital technologies also here, the spectrum efficiency improves a lot. This releases a so-called "digital dividend" which is very attractive both for providing an increased amount of broadcast services and for terrestrial mobile services. The frequencies are very attractive due to the favourable properties for coverage and implementation. Parts of the UHF band were globally harmonized for IMT at WRC-07.

Other issues were on the satellite component of IMT and Fixed Satellite Services (FSS). Climate issues are of course also influencing ITU, both for Earth observation, disaster prediction and mitigation. Thus, the WRC adopted a new resolution on the use of ICT in emergency and disaster relief.

The third article is related to ITU work on Identity Management, in which the *X.500 Directory standard* is a key component. *Erik Andersen* has written about the X.500 in the context of Next Generation Networks (NGN). Readers will recall that the previous issue of *Telektronikk* (3/4.2007) was themed *Identity Management*, and we recommend reading more about the subject here.

Telektronikk has a long history in printing reports from ITU events and from different ITU study groups since 1927. Consult the web, *http://www.telektronikk.com* for an overview of reports since 1992.

Per Hjalmar Lehne is Researcher at Telenor R&I 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&I 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

ISSN 0085-7130

ITU-R Radiocommunication Assembly 2007

ANNE LISE LILLEBØ, TERJE TJELTA

Anne Lise Lillebø is Director in Group Regulatory, Telenor ASA

Terje Tjelta is Senior Research Scientist in Telenor R&I

The main objective of the Radiocommunication Sector of the International Telecommunication Union (ITU-R) is to ensure rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including those using satellite orbits. ITU-R held its Radiocommunication Assembly (RA-07) in Geneva, 15-19 October 2007. A major outcome was the change of Study Group (SG) structure by reducing the number of SGs from seven to six, for example creating one large new SG (SG 5) on terrestrial services by merging fixed and mobile services, previously SGs 9 and 8, respectively. RA-07 also decided to include a sixth terrestrial air interface in the International Mobile Telecommunications (IMT) family by including the OFDMA TDD WMAN technology, a WiMAX derived technology. Mobile telecommunications were at centre stage at this RA as the Assembly adopted flexibility by choosing IMT as the root name for IMT collectively. IMT-2000 technologies as well as coming evolutions of the future IMT are now referred to as IMT. This had a major consequence for the following World Radiocommunication Conference (WRC-07) as spectrum both previously allocated to IMT-2000 and future allocations to IMT-Advanced in the Radio Regulations (RR) will now be collectively referred to as IMT. The ITU-R and its RAs play a major role for the development of global radiocommunications and make very important decisions for international telecommunications business.

1 Introduction

The International Telecommunication Union (ITU) is an intergovernmental organisation and a specialised agency of the United Nations for telecommunications with 191 Member States and some 600 Sector Members. ITU carries out its tasks in three sectors; i.e. the Radiocommunication Sector (ITU-R), the Telecommunication Standardization Sector (ITU-T), and the Telecommunication Development Sector (ITU-D).

The mission of the ITU-R is to ensure the rational, equitable, efficient and economical use of the radio frequency spectrum by all radiocommunication services and to carry out studies and approve Recommendations on radiocommunications matters. One of the primary objectives is to ensure interference free operation of radiocommunications systems. The Radiocommunication Sector plays a central role in the technological progress of telecommunications. In ITU-R standardisation is made through the ITU-R Recommendations and the objective is to ensure the necessary performance and quality in operating radiocommunication systems. Recommendations are also intended to provide flexibility for future expansion and new technological developments. At the same time considerations of regulatory and procedural issues cannot be separated from the technological studies. ITU-R needs to balance technological needs against those of compatibil-

Centre International de Conférences de Genève (CICG), Geneva, venue of RA-07

ISSN 0085-7130 © Telenor ASA 2008

ity between the various services, at the same time seeking innovative means to optimise the use of the frequency spectrum for the good of all.

ITU-R comprises several parts, most importantly the World Radiocommunication Conference (WRC), the Radiocommunication Assembly (RA), and the ITU-R Study Groups (SG), see illustration in Figure 1 and below for further details.

Whereas the WRC treats issues of a global nature covering all three Radio Regions of the ITU, the Regional Radicommunication Conference (RRC) has a more limited task and deals with topics for one or two Radio Regions. The Radio Regulations Board (RRB), the Conference Preparatory Meeting (CPM), the Radiocommunication Advisory Group (RAG) and some other functions are set up to assist ITU in performing its work and preparing for the World Radiocommunication Conference. The Radiocommunication Bureau (BR) with its director is responsible for coordinating and organising the work in the R-Sector.

In the R-Sector, the Radiocommunication Assembly (RA) normally convenes in conjunction with a World Radiocommunication Conference (WRC). Since 1992, the RA has met every three to four years associated in time and place with the WRC to improve the efficiency and the effectiveness of the R-Sector.

The Radiocommunication Assemblies provide the necessary technical bases for the work for the WRCs and respond to all requests from the WRCs. The role of the RA is a vital part of the structure of ITU-R, fulfilling a number of the essential purposes of the Union. One of the objectives of the RA-07 is to ensure that the structure and terms of reference of the Study Groups keep pace with technological developments and associated spectrum issues and to optimise the efficiency of the Radiocommunication Sector.

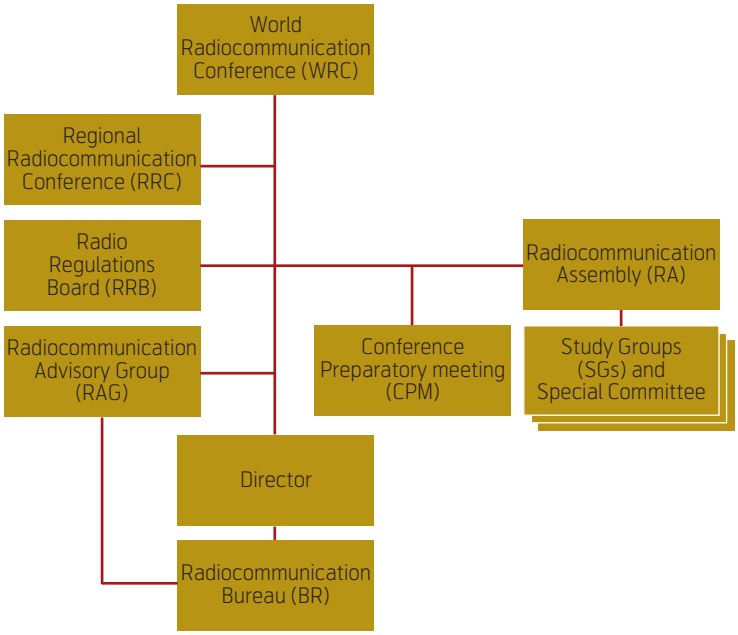


Figure 1 ITU-R organisation

This paper is organised in six sections: Introduction, Radiocommunication Assembly main duties, New study group structure, International Mobile Telecommunications (IMT) development, Future work and other issues, and Conclusions.

2 Radiocommunication Assembly Main Duties

RA-07 met in Geneva from 15 to 19 October 2007 followed by the WRC-07. The RA-07 was chaired by Bruce Gracie, Canada. Issues high on the agenda were the approval of Draft Recommendations enabling WiMAX derived technology to become part of the IMT family, Resolutions on IMT as the root name and the use of ICT in disaster mitigation as well as reviewing and updating of ITU-R Resolution 1: *"Working methods"* with the detailed working procedures of the ITU-R.

ITU

Kevin Hughes, ITU-R Councillor; Bruce Gracie, Canada, Chairman of RA-07; Valery Timofeev, BR Director

The principal duties of the Radiocommunication Assembly are:

- To provide the necessary technical basis for the work of World Radiocommunication Conferences and to respond to all requests from such Conferences;
- To establish future work programmes for the Study Groups;
- To review the Study Group structure;
- Appoint chairmen and vice-chairmen of the Study Groups;
- Update and review the working methods and procedures and relevant resolutions for the work in the ITU-R;
- Approve ITU-R Recommendations developed by the Study Groups.

The main issues at the RA-07 were:

- Approve Draft Recommendations on new radio interface for IMT (WiMAX derived technology);
- Adopt an updated SG structure merging former SG8 and SG9 into an SG for terrestrial services;
- Establish and agree on the work programme for the next study period;
- Review ITU-R Resolution 1 – Working methods Approve Resolutions pertinent to future work of ITU-R such as a naming of IMT, principles for the development of IMT-Advanced and emergency communications and disaster relief.

3 New Study Group Structure

RA must ensure that the structure and terms of reference for the SGs keep pace with technological developments and associated spectrum issues and optimise the efficiency of the radiocommunication sector. The technical, operational and procedural work is carried out in the various SGs where Member States, Sector Members and Associates of the ITU-R develop Recommendations and reports on the basis of the study of Questions (areas of study).

Eirik Bliksrud (NPT), Head of the Norwegian Delegation

The ITU-R SGs carry out the work in the R-Sector. The SGs develop Recommendations for the various fields of international radiocommunications on the basis of the study of Questions, as well as reports and handbooks. The recommendations are regularly updated in accordance with the developments in the field. The reports and handbooks are also revised and updated, but not so often.

Members of the ITU-R have discussed a restructuring of the SGs for a number of years with the objective of arriving at a more efficient distribution of the work load and a structure which better reflects today's radiocommunication technologies. Two options were developed for a potential new structure that would include only six Study Groups as compared to seven Study Groups in the last study period.

Eirik Bliksrud, Head, Norwegian Delegation, and Anne Lise Lillebø, Terje Tjelta, Telenor Delegation

ISSN 0085-7130 © Telenor ASA 2008

From RA-07 onwards the ITU-R Study Groups are:

Study Group	Scope
SG1 Spectrum management	Spectrum management principles and techniques, general principles of sharing, spectrum monitoring, long-term strategies for spectrum utilisation, economic approaches to national spectrum management, automated techniques and assistance to developing countries in cooperation with the Telecommunication Development Sector.
SG3 Radiowave propagation	Propagation of radio waves in ionised and non-ionised media and the characteristics of radio noise, for the purpose of improving radiocommunication systems.
SG4 Satellite services	Systems and networks for the fixed-satellite service, mobile-satellite service, broadcasting-satellite service and radiodetermination-satellite service.
SG5 Terrestrial services	Systems and networks for fixed, mobile, radiodetermination, amateur-satellite services.
SG6 Broadcasting service	Radiocommunication broadcasting, including vision, sound, multimedia and data services principally intended for delivery to the general public. Broadcasting makes use of point-to-everywhere information delivery to widely available consumer receivers. When return channel capacity is required (e.g. for access control, interactivity, etc), broadcasting typically uses an asymmetrical distribution infrastructure that allows high capacity information delivery to the public with lower capacity return link to the service provider. This includes production and distribution of programmes (vision, sound, multimedia, data, etc) as well as contribution circuits among studios, information gathering circuits (ENG, SNG, etc.), primary distribution to delivery nodes, and secondary distribution to consumers. The Study Group, recognising that radiocommunication broadcasting extends from the production of programmes to their delivery to the general public, as detailed above, studies those aspects related to production and radiocommunication, including the international exchange of programmes as well as the overall quality of service.
SG7 Science services	1 Systems for space operation, space research, Earth exploration and meteorology, including the related use of links in the inter-satellite service. 2 Systems for remote sensing, including passive and active sensing systems, operating on both ground-based and space-based platforms. 3 Radio astronomy and radar astronomy. 4 Dissemination, reception and coordination of standard-frequency and time-signal services, including the application of satellite techniques, on a worldwide basis.

A vast majority of the RA-07 supported the option which merges former SG8 (mobile) and SG9 (terrestrial fixed services) into a Terrestrial Services Study Group, whereas SG4 (Satellite), the satellite service aspects of former SG6 and SG8 were merged into a new Satellite Services Study Group. This solution was supported both by the Norwegian Post and Telecommunications Authority (NPT) representing Norway and Telenor. The current scope and responsibilities of SG1 Spectrum management, SG3 Radio wave propagation, and SG7 Science services were maintained.

Before RA-07 there were seven SGs, but in order to economise, mainly due to translation costs as ITU works in six languages, the RA-07 decided to reduce the number of SGs to six as listed above. The transition was basically done as sketched in Figure 2 where the solid lines represent the main part of transfer and the broken lines a minor part of an SG's responsibility.

The scope and responsibilities of Study Group 6, Broadcasting Services, remained almost intact, but the activities on satellite broadcasting undertaken in

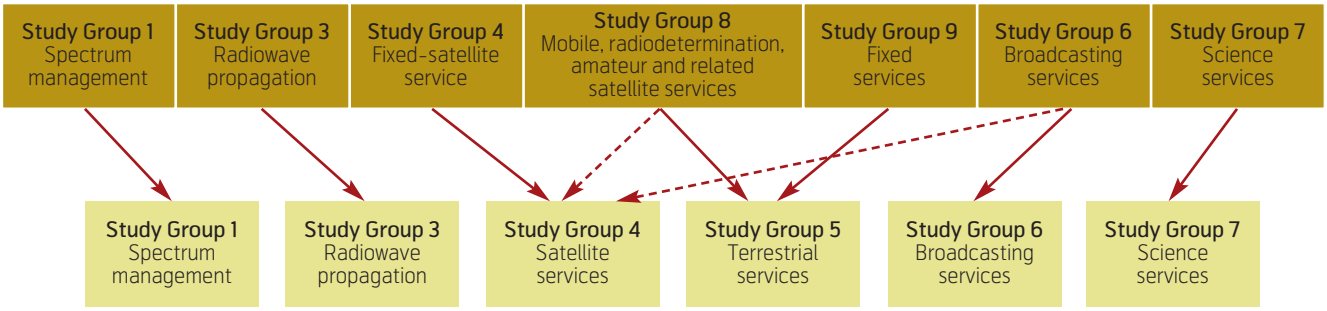


Figure 2 Change of SG structure from RA 2003 to RA 2007 reducing total number of SGs from seven to six SGs. The solid lines represent the transfer of an SG's main responsibility and the broken lines a minor part

ISSN 0085-7130

Ross Bateson and Robert Ercole, GSM Association

Working Party 6S were aligned with other activities relating to satellite matters in SG4 Satellite Services.

The ITU-R community represents a world leadership in spectrum management and radiowave propagation. For the other areas, the service groups in the ITU-R gather a global community for international harmonisation of technical issues. This is very important for many areas, in particular satellite communication that is global in nature, e.g. geostationary locations and satellite orbits. In addition, the future development needs for various large regions, and globally, are also very well dealt with in the ITU-R. An example is the study of necessary bandwidth for future mobile broadband communication systems [4] that played such an important role for the WRC-07.

4 International Mobile Telecommunications (IMT) Development

There is a major interest world-wide in the use and further developments of IMT. The growth of mobile telecommunication has been remarkable in all parts of the world, and now it is noted that IMT is increasingly being adopted. It is of vital importance for the mobile telecommunications business area to prepare for new generations of systems, in particular with respect to identification of type of systems and radio spectrum resources [5].

4.1 New Generic Root Name IMT

In its Resolution ITU-R 56 *Naming for International Mobile Telecommunications* RA-07 agreed on a common naming scheme for the present IMT-2000 technologies (3G) and those being developed for the 4th generation (4G) of mobile telecommunications – IMT-Advanced.

- The term "IMT" should be the root name to cover the capabilities of "IMT-2000, future development of IMT-2000 and systems beyond IMT-2000" collectively;
- That the term "IMT-Advanced" should be applied to systems that include new radio interfaces that support the new capabilities of systems beyond IMT-2000 (4G);
- That "IMT-2000" encompasses its enhancements and future developments (3G).

This decision gives increased flexibility and will enable refarming for spectrum previously allocated to 3G technologies in all bands identified for IMT. It implies that spectrum identified for 3G technologies can be used in the future for 4G technologies and spectrum identified for 4G can be used for any IMT technology.

Resolution ITU-R 57 *Principles for the process of development of IMT-Advanced* establishes guiding principles that underpin the process for specifying the radio interfaces for IMT-Advanced such as developing Recommendations and Reports for IMT-Advanced including Recommendations for radio interface specifications.

4.2 WiMAX – New Air Interface for IMT-2000

Most Draft Recommendations are normally approved according to a written procedure based on a two-step process whereby the Study Group concerned adopts the draft Recommendation and the Member States approve the Draft Recommendation by written consultation, the so-called Traditional Approval Procedure (TAP) or the faster Procedure for Simultaneous Adoption and Approval (PSAA). By exception, some of the Draft Recommendations are forwarded to RA for approval, either because of lack of time in the Study Group or because the Draft Recommendations in question are of a contentious nature requiring important policy decisions.

The Recommendations submitted to RA-07 for approval were directly linked to the discussions at the WRC-07 and constitute the technical basis for decisions at the WRC-07 which started immediately after the RA-07.

Three Draft Recommendations regarding the inclusion of WiMAX derived technology based on IEEE 802.16e in the IMT 2000 family were submitted by Study Group 8 to the RA-07 for approval. Since many members considered that approval of these recommendations should be done before the WRC, there was not enough time to use the written approval procedure.

RA agreed to expand the IMT-2000 3G radio interface with OFDMA technology and the three recommendations were approved on a consensus basis:

- Recommendation ITU-R M.1457-6 Detailed specification of the radio interfaces of International Mobile Telecommiunications-2000 (IMT-2000);

- Recommendation ITU-R M.1580-1 Generic unwanted emission characteristics of base stations using the terrestrial radio interfaces of IMT-2000;

- Recommendation ITU-R M.1581.1 Generic unwanted emission characteristics of mobile stations using the terrestrial radio interfaces of IMT-2000.

China was initially opposing the revision of Recommendation ITU-R M.1457, but accepted the new revision after the inclusion of a footnote in the Recommendation stating that China did not agree to the approval of the revision of the recommendation for inclusion of OFDMA TDD WMAN during RA-07.

Germany and Sweden had concerns with the approval of the Draft Recommendations ITU-R.M.1580-1 and ITU-R M.1581-1, but accepted the Recommendations after the inclusion of noting C in both Recommendations. Noting C emphasises that additional urgent study on WiMAX TDD is needed and should start urgently.

Both the Norwegian Administration (NPT) and Telenor endorsed the approval of the recommendations. This approval means that the WiMAX technology is recognised as the sixth terrestrial IMT-2000 radio interface.

With the decision to include OFDMA TDD WMAN in the IMT-2000 family, WiMAX derived technology enjoys the same spectrum access as any other IMT technology in bands identified for IMT in the Radio Regulations. With an ITU IMT family consisting of six technologies it can be expected that regulators will open future spectrum intended for IMT to a number of IMT technologies and there will be an increased competition for spectrum among operators and vendors. ITU agreements are normally very influential although an ITU Recommendation does not compel national administrations to grant spectrum access to WiMAX. However, there is a global manufacturing market and many countries follow the ITU Recommendations.

The six IMT-2000 terrestrial radio interfaces are now:

- IMT-2000 Direct Spread (WCDMA/UTRA FDD)
- IMT-2000 Multi Carrier (CDMA2000)

Anne Lise Lillebø and Terje Tjelta, Telenor Delegation

- IMT-2000 Time Code (UTRA TDD, TD-SCDMA)
- IMT-2000 Single Carrier TDMA (UWC-136)
- IMT-2000 Frequency Time (DECT+)
- IMT-2000 OFDMA TDD WMAN (Mobile WiMAX).

5 Future Work and Other Issues

RA-07 dealt with several other issues in addition to IMT and the SG structure. Most importantly among these were the future Work Programme, updating of Resolutions, adoption of new Resolutions and the approval of some other proposed Draft Recommendations.

5.1 Work Programme in the Study Period 2007 – 2011

The work programme for the next study period encompassing some 300 Questions was approved and the complete list of questions and their priority is included in Resolution ITU-R 5-5, Annex 1.

The Questions are grouped according to subject matter within the remit of the individual Study Group and are categorised according to the priority and urgency

- "C" indicates the highest priority and is granted to Questions associated with tasks related to preparations for World and Regional Radiocommunication Conferences (WRC/RRC);

- "S" indicates Questions that are referred to the RA by the Plenipotentiary Conference, any other conference, the Council or the Radio Regulations Board (RRB).

Questions identified as being suitable for approval by the alternative approval process (AAP) must be within the category "S" and are marked as AP.

Regarding future IMT work the new SG5 *Terrestrial services* (former SG8 Mobile services and SG9 Ter-

ISSN 0085-7130 © Telenor ASA 2008

restrial services) will be responsible. This work will be performed by SG5 WP5D (former WP8F). Based on the RA-07 Resolution on principles for developing IMT-Advanced, ITU-R will invite relevant organisations external to ITU to propose candidate radio interface technologies for IMT-Advanced. ETSI – the European Standards Institute – is now developing a plan enabling the institute to make technical contributions to the ITU-R process.

5.2 Resolutions of ITU-R

In his opening speech to the RA-07, Mr Valery Timofeev, Director of the Radiocommunication Bureau, noted that the ITU-R Resolutions are key to increasing the efficiency both in the working methods of the Study Groups and when planning future work programmes and that it is of great importance that the ITU-R resolutions provide clear guidance on which the SGs can pursue their activities without imposing unnecessary bureaucracy which might retard progress in the studies.

Through the instrument of ITU-R Resolutions RA provides future orientation on specific topics such as spectrum management, IMT, propagation, disaster relief etc. The RA-07 agreed on a number of ITU-R Resolutions which are also of high relevance for the WRC-07.

Use of Radiocommunications in Disaster Mitigation and Relief – ICT in Disaster Relief

Following the tsunami in Asia in 2004 there is now increased attention on the important role that telecommunications can play in disaster relief and emergency situations. Resolution ITU-R 53: *The use of radiocommunications in disaster response and relief*, and Resolution ITU-R 55: *ITU studies of disaster prediction, detection, mitigation and relief*, urge the ITU-R Study Groups to undertake studies regarding management of radiocommunication in disaster prediction, detection, mitigation and relief within ITU and with organisations external to ITU. The BR Director is instructed to cooperate with the directors of the T- and D-Sector to assist Member States in their emergency radiocommunication preparedness activities.

Harmonisation for SDRs

Resolution ITU-R 54 *Studies to achieve harmonisation for short-range radiocommunication devices (SRDs)* recognises the importance of short range devices (SDRs) which in general utilise the unlicensed bands. The Resolution encourages ITU-R to continue its studies to enable implementation of advanced technologies for SDRs, focusing in particular on a strategy for the future.

Fatih M. Yurdal, European Radiocommunications Office (ERO)

5.3 Other Issues

Earth Stations onboard Vessels (ESV)

SG 4 had worked out a revision to Recommendation S-1487-1 "Technical characteristics of earth stations on board vessels communicating with FSS satellites in the frequency bands 5 925–6 425 MHz and 14–14.5 GHz which are allocated to the fixed-satellite service". The revisions had not yet been agreed due to opposition from the Arab States. With some changes suggested to the RA-07 and additional remarks that further studies on the topic of ESVs should be undertaken in ITU-R, the revision was adopted.

Influence of Power Line Telecommunication (PLT) on Broadcasting Systems

SG 6 presented a draft new Recommendation on protection requirements for broadcasting systems operating in the LF, MF, HF and VHF bands below 80 MHz against the impact of power line telecommunication (PLT). RA-07 decided to send the material back to SG 6 for further consideration and to consider the possibility of developing a Report on the subject.

6 Conclusions

ITU-R Radiocommunication Assembly (RA) is the highest level with respect to studies and organisation of the work that forms the basis for a large number of important recommendations within radio based telecommunication business. Most importantly, RA-07 prepared the ground for the World Radiocommunication Conference 2007 held right afterwards, especially as regards mobile communications.

The RA-07 was a breakthrough for the WiMAX derived technology which was acknowledged as a

new member of the IMT-2000 family of terrestrial air interfaces. WiMAX technology now enjoys the same access to spectrum as the other technologies recognised by ITU-R. The ensuing WRC-07 endorsed RA-07's Resolution to use the term IMT generically as root name both for IMT-2000 and IMT Advanced. This provides a large degree of flexibility and all the six members of the IMT family can be used in any Recommendation referring to IMT or in a reference in the Radio Regulations.

The RA-07 reduced the number of Study Groups from seven to six, by merging the former SG8 Mobile services and SG9 Fixed services into a new SG5, Terrestrial Services. This may enable ITU-R to carry out its work within this field in a more efficient way and will save costs relating to translation and interpretation.

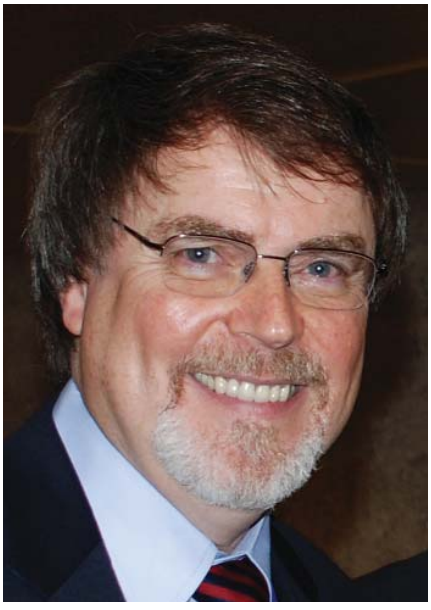
7 References

1 ITU. *ITU-R – Radiocommunication : The future is wireless*. October 2007. Available from: http://www.itu.int/dms_pub/itu-r/opb/gen/R-GEN-OVW-2007-E10-PDF-E.pdf

2 UN. *UN Charter, Article 57*. Nov 2007 [online] – http://www.un.org/aboutun/charter/index.html

3 UN. *UN in brief – The specialized agencies*. United Nations. Nov 2007 [online] – http://www.un.org/Overview/uninbrief/agencies.html

4 ITU. *Estimated spectrum bandwidth requirements for the future development of IMT-2000 and IMT-Advanced*. ITU-R, 2006. (Report ITU-R M.2078)



Bruce Gracie, Chairman of RA-07

5 Tjelta, T, Lillebø, A L, Evenstad, E O. ITU-R World Radiocommunication Conference 2007. *Telektronikk*, 104 (1), 144–159 (this issue).

6 ITU-R. *Final Acts*. WRC-03, Geneva, 2003.

7 ITU-R. *Final Acts of the Regional Radiocommunication Conference for planning of the digital terrestrial broadcasting service in parts of Regions 1 and 3, in the frequency bands 174–230 MHz and 470–862 MHz*. RRC-06, Geneva, 2006.

8 ITU. *The ITU Telecom Information Exchange Services (TIES)*. Nov 2006 [online] – http://www.itu.int/TIES/index.html)

9 ITU. *ITU Publications*. Nov 2006 [online] – http://www.itu.int/publications/default.aspx

Annex A RA-07 Organisation and ITU-R Working Methods

RA – Organisation and Management

The Assembly was chaired by Bruce Gracie from Canada and the work was organised in five committees:

Committee 1: Steering committee
Committee 2: Budget control
Committee 3: Editorial committee
Committee 4: Structure and work programme of the Study Groups
Committee 5: Working methods of the Study Groups

Updating of ITU-R Working Methods

A large majority of ITU-R's detailed working methods are contained in Resolution ITU-R 1-5 *Working methods for the Radiocommunication Assembly, the Radiocommunication Study Groups, and the Radiocommunication Advisory Group*. The RA-07 reviewed a number of provisions and updated them in order to make ITU-R's working methods more efficient.

In addition, Resolution ITU-R 1-5, para 8 refers to Guidelines issued by the Director of the BR which provide additional information on contribution to the ITU-R SGs and on aspects of the current working procedures, particularly those relating to meetings and documentation. The guidelines also address practical matters concerning the effective distribution of documents by electronic means

AAP (Alternative Approval Procedure)

The Alternative Approval Procedure was agreed by RA in 2003 as a permanent procedure in Resolution ITU-R 45-1: *Application of an alternative approval procedure (AAP) for Recommendations* and enables

Terje Tjelta, Telenor and Håkan Lilja, TeliaSonera

Sector Members to be consulted and take part in the approval procedure for ITU-R recommendations. This procedure is a "fast track procedure" which saves a lot of time and has become a big success in the ITU-T Sector. At the RA-07 there was a formal proposal to suppress this Resolution as it was claimed that it had not been used since 2003. This may well be the case since any Member State may argue that the Draft Recommendation contains issues of a regulatory nature and in such cases the procedure in Resolution 45-1 cannot be used.

Since the decision of the Plenipotentiary Conference in 1998 which opened the possibility of changing the approval procedures of Recommendations in the Sectors, CEPT and Sector Members have fought for many years to have the so-called AAP procedure in place. A number of CEPT countries and Sector Members such as Telenor did not support the proposal to suppress the AAP in ITU-R altogether. Suppressing the possibility of using AAP in the R-Sector would be considered as a very negative signal from ITU to the industry. In our opinion, the R-Sector should keep the procedure as an alternative for approval of recommendations.

After a CEPT consultation where Norway raised the question, CEPT members were encouraged to support the retention of the AAP and Resolution 45. At the last plenary meeting of the RA-07 the proposal was withdrawn, but it is noted in the minutes that if the procedure is not used until the next RA Resolution 45 should be suppressed. Reconsidering and possibly changing or suppressing ITU-R Resolutions is the prerogative of any RA and it is up to the next RA to reconsider the matter.

Telenor is pleased with this result and appreciated the active support from the Norwegian Administration during the CEPT consultation.

Number of Working Parties in SGs

A proposal to limit the number of Working Parties established by a Study Group was not supported. Many members found that limiting the number of WPs that an SG might like to establish in order to perform its duties was unnecessary micro management and should in principle be left to the SG to decide taking into account the views of the members of the SG. It is important for the SG to develop an efficient internal structure in order to perform its tasks in the best way possible. The provision now states that Study Groups should establish only the minimum number of Working Parties, normally three or four Working Parties.

Term of Office for Vice-Chairmen in SGs

The term of office for chairmen and vice-chairmen of the Study Groups was aligned and now cover chairmen and vice-chairmen of Study Groups, the coordination committee for Vocabulary and the Radiocommunication Advisory Group. The maximum term of office is two consecutive terms.

Annex B Appointment of Chairmen and Vice-Chairmen

Appointment of Chairmen and Vice-Chairmen

The candidatures for chairmen and vice-chairmen of the Study Groups, the Coordination Committee for Vocabulary (CCV), the Radiocommunication Advisory Group (RAG), the Conference Preparatory Meeting (CPM) and the Special Committee for Regulatory/Procedural matters are discussed at the meetings of Heads of Delegation of Member State where Sector members are not authorised to take part. The RA finally appoints the list of candidates.

Anders Frederich, Sweden, Vice-Chairman of CPM and Head of Swedish Delegation

According to the provisions in the Resolution ITU-R 15-4, candidates for these posts should be identified by Member States and Sector Members of the ITU-R preferably three months before the opening of the Assembly, and their nominations should be accompanied by a biographical profile highlighting the qualifications of the individuals proposed which will be circulated to the Heads of Delegation present at the Assembly. Despite these clear procedures, the selection of Chairmen and Vice-Chairmen has become an overwhelmingly political and sensitive issue and there were many changes of candidates up to the very last minute, and the RA-07 saw a record number of Vice-Chairmen for a number of Study Groups with SG5 Terrestrial services having a total of 10 Vice-Chairmen.

Anders Frederich, National Post and Telecom Agency, Sweden, was appointed Vice-Chairman of the Conference Preparatory Committee (CPM). It should be noted that Mr Frederich is the only representative from the Nordic countries among the Chairmen and Vice-Chairmen of ITU-R.

In the next study period (2007 – 2011) the ITU-R will have six Study Groups with the following scope and Chairmen:

Study Group (SG)	Chairman
SG 1 "Spectrum management"	R Haines (USA)
SG3 "Radiowave propagation"	B Arbesser-Rastburg (European Space Agency)
SG 4 "Satellite services"	V Rawat (Canada)
SG 5 "Terrestrial services"	A Hashimoto (Japan)
SG 6 "Broadcasting services"	C Dosch (Germany)
SG 7 "Science services"	V Meens (France)
CCV "Coordination committee for vocabulary"	N Kisrawi (Syrian Arab Republic)
Radiocommunication Advisory Group (RAG)	J-B Yao Kouakou (Côte d'Ivoire)
CPM – Conference Preparatory Meeting	A Nalbandian (Armenia)
Special Committee	M Ghazal (Lebanon)

Anne Lise Lillebø is Director in Group Regulatory, Telenor ASA. Her main responsibilities include spectrum management and policy matters related to international telecommuncations organisations. She has represented Televerket/Telenor at the ITU Radiocommunication Assemblies in 1993, 1995, 1997, 2000, 2003 and 2007. She holds a Master of Arts degree from the University of Oslo.

email: anne-lise.lillebo@telenor.com

Terje Tjelta is Senior Research Scientist in Telenor R&I. He received the MSc degree in physics from the University of Bergen, Norway, in 1980, and Dr.Philos. from the University of Tromsø in 1997. He joined Telenor Research and Innovation in 1980 and has been there since except for one year (1984/85) as visiting researcher at Centre Nationale des Études des Télécommunications (CNET) in France. His research covers radio communication systems, in particular high capacity links and broadband wireless access. He has experience from several international co-operative research projects and standardisation activities for the International Telecommunication Union.

email: terje.tjelta@telenor.com

ISSN 0085-7130 © Telenor ASA 2008

ITU-R World Radiocommunication Conference 2007

TERJE TJELTA, ANNE LISE LILLEBØ, ERIK OTTO EVENSTAD

Terje Tjelta is Senior Research Scientist in Telenor R&I

ITU-R's World Radiocommunication Conference (WRC) is the international forum for revising the Radio Regulations, the international treaty governing the use of the radio-frequency spectrum and satellite orbits. The WRC in 2007 (WRC-07) addressed the worldwide use of radio frequencies and sought solutions on how to exploit the limited resource of the radio frequency spectrum in the most rational and efficient way in order to meet the global demand for spectrum generated by rapid technological developments and growth in the information and communication technology (ICT) sector.

The highlights of the WRC-07 included the issues regarding more spectrum for International Mobile Telecommunications (IMT), especially in the UHF-band and the C-band, and the revision of the regulatory procedures applicable to satellite services, including the Plan for fixed satellite services that provides spectrum and orbit resources to the Member States of the ITU.

Anne Lise Lillebø is Director in Group Regulatory, Telenor ASA

WRC-07 agreed on the following decisions of importance to telecommunications operators and vendor industry:

- New allocations of additional spectrum for mobile services and IMT;
- A revised framework of procedures for the fixed satellite service (FSS);
- Resolution advocating the use of ICT in emergency and disaster relief including the prevention of climate change.

The decisions of the WRC-07 in the form of Final Acts will be part of the international treaty "Radio Regulations" (RR).

Geir Seljeseth

Erik Otto Evenstad is Senior Adviser in Telenor Satellite Broadcasting AS

1 Introduction

ITU-R's World Radiocommunication Conference (WRC) is an intergovernmental conference where ITU's Member States participate, and the WRC meets every three to four years and reviews and revises the Radio Regulations (RR), the international treaty governing the use of the radio frequency spectrum and the geo-stationary satellite and non-geo-stationary satellite orbits, see the companion article on RA-07 [1].

The radio frequency spectrum is a finite resource increasingly in demand through the growth of information and communication technology services (ICT) and the development of wireless broadband. The decisions of the WRC aim to reach efficient use of the radio frequency spectrum based on global harmonised allocations of frequency bands as well as common sharing and protection criteria to ensure seamless access to common services, global interoperability and international roaming. The WRCs seek to produce timely and effective international rules for the establishment of advanced new wireless services and applications, and at the same time taking care of interests and rights of existing radiocommunication users.

ITU

Geneva, ITU Headquarters

Communication with other people and other countries outside own national system requires mutual technical agreements to make it function in an optimal way. Global agreements may create large markets as well as facilitating the take up of new services by the user. Harmonised regulation at national, regional, and global levels would make it considerably easier for global business. Technical compatibility, interoperability and vast market penetration could increase

ISSN 0085-7130 © Telenor ASA 2008

Venue of WRC-07, Centre International de Conférences de Genève (CICG)

business and create economies of scale for widespread services.

A majority of Telenor businesses are based on wireless technologies, and access to radio spectrum is an essential input when establishing wireless networks. For mobile communications as a mass market product, it is of crucial commercial importance that global and regional agreements exist between countries regarding the harmonised use of radio spectrum. In our opinion the ITU process and the RR will continue to play a vital role in fostering the rapid evolution and expansion of wireless systems on a global scale.

1.1 Predictability and Economies of Scale

Allocation and identification of spectrum for advanced mobile services at a global level such as the WRC send a very clear signal to the industry about the direction that future product development might take. Global harmonisation of spectrum provides a certain degree of predictability for the mobile industry, manufacturers and operators and is of vital importance in order to obtain economies of scale to offer mobile devices and services at affordable prices.

ITU agreements are normally very influential and indicate that there will be potential for a global manufacturing market and many countries follow the ITU provisions in the RR and the ITU-R Recommendations.

1.2 Mission of the ITU Radiocommunication Sector (ITU-R)

The mission of the Radiocommunication Sector (R-Sector) is laid down in Article 1 of the ITU Constitution [2] and states i.a. that ITU-R should

- "effect allocation of bands of the radio frequency spectrum, the allotment of radio frequencies and the registration of radio frequency assignments and, for space services, of any associated orbital position in the geostationary-satellite orbit or of any associated characteristics of satellites in other orbits, in order to avoid harmful interference between radio stations of different countries";

- "coordinate efforts to eliminate harmful interference between radio stations of different countries and to improve the use made of radio-frequency spectrum for radiocommunication services and of the geostationary-satellite and other satellite orbits".

1.3 WRC-07

The WRC-07 met in Geneva in the period 22 October – 16 November 2007 and was chaired by François Rancy, France. There were a total of around 2800 participants at the WRC-07 representing 164 Member States and 104 observers from i.a. the European Commission, European Broadcasting Union (EBU), European Radiocommunications Office (ERO), GSM Association (GSMA), Universal Mobile Telecommunications System (UMTS) Forum, Worldwide Interoperability for Microwave Access (WiMAX) Forum, Asia-Pacific Telecommunity (APT), and Inter-American Telecommunications Commission (CITEL).

The agenda of the WRC-07, established by the 2004 session of the ITU Council [3], contained about 30 items related to almost all terrestrial and space radio services and applications. It comprised mobile communications, including 3G and future generations, aeronautical telemetry and telecommand systems,

ITU

Valery Timofeev, Director, ITU Radiocommunication Bureau

ITU

François Rancy, France, Chairman of WRC-07

satellite services, including meteorological applications, maritime distress and safety signals, digital broadcasting, and the use of radio in the prediction and detection of natural disasters.

One of the major issues at the WRC-07 was Agenda Item 1.4 concerning the allocation of additional spectrum for third generation mobile services (3G) and identification for International Mobile Telecommunications (IMT) to ensure there is sufficient bandwidth to provide mobile broadband services as well as suitable spectrum for covering rural areas to help combat the digital divide in developing markets.

The Norwegian delegation and Telenor followed closely the debate on additional spectrum for IMT and Agenda Item 1.10 regarding the updating of procedures for fixed satellite services.

This article is organised in seven main sections, including introduction and conclusions. Section 2 covers the RR, Section 3 focuses on new spectrum for IMT, Section 4 presents results on fixed service satellite replanning, Section 5 covers some other decisions, and Section 6 presents the agenda for the next conference, i.e. WRC-11.

Members of the Norwegian Delegation: Erik Jørol, Deputy Head, and Geir Jan Sundal, Head of Delegation, Norwegian Post and Telecommunications Authority

2 The Radio Regulations

The Radio Regulations (RR) [4] is an international treaty governing the use of the radio frequency spectrum and the satellite orbits. ITU's RR form the international framework for global harmonisation of radio applications and services. The RR provides a large degree of flexibility allowing individual Member States to develop their own national legislation and regulation for radio spectrum use. Observing and conforming to the provisions in the RR are based on goodwill from ITU's Member States as compared to regulatory sanctions imposed on a national level. The RR ensures that the services allocated in a frequency band can be operated by various countries through cross-border coordination. The RR are updated regularly to keep track with the development and expanding use of wireless services through decisions made by WRCs.

The objective of the management of the radio regulations is to ensure worldwide harmonised use of telecommunications services, establish common sharing and protection criteria to avoid harmful interference, and agree on common global allocations. International harmonisation is key to mobile radiocommunication applications that are used across borders and enables interoperability and roaming. Harmonisation will also provide possibilities of obtaining reduced equipment cost through economies of scale.

In 2006 ITU celebrated the 100th anniversary of the RR [5]. The International Radiotelegraph Convention was signed in Berlin in 1906, and its annex contained the world's first set of rules which were later called the Radio Regulations.

In the RR various parts of the spectrum are allocated to specific services such as broad-

François Rancy, Chairman of WRC-07, Chris van Diepenbeek, Chairman of CEPT ECC, Dutch Delegation, Anne Lise Lillebø and Geir Jan Sundal, Norwegian Delegation

casting, mobile, fixed, satellite etc. in accordance with the service definitions of the RR.

2.1 ITU's Radio Regions of the World

In the RR the world has been divided into three so-called "radio regions":

- Region 1: Europe, Africa and some countries in the Middle East, also including Armenia, Azerbaijan, the Russian Federation, Georgia, Kazakhstan, Mongolia, Uzbekistan, Kyrgyzstan, Tajikistan, Turkmenistan, Turkey and Ukraine
- Region 2: The Americas
- Region 3: Asia/Oceania (except the Asian countries included in Region 1).

See Figure 1 for an illustration of the three geographical regions.

2.2 Primary and Secondary Services

The RR Article 5, Section II, contains the global Table of Frequency Allocations. A primary service allocated in the table is printed in capitals, for example "MOBILE". A secondary service allocated in the table is printed in normal characters, for example "mobile". There might also be more detailed written provisions in the form of a footnote. These footnotes are referred to in the table of frequency allocations. Footnotes will often also refer to resolutions for further details and requested further work, see illustration in Figure 2.

A secondary service shall not cause harmful interference to a primary service. A secondary service cannot claim protection from harmful interference from stations of a primary service. However, a secondary service can claim protection from harmful interference from stations of the same or other secondary services to which frequencies may be assigned at a later date.

A WRC often agrees on Resolutions which give instructions to the Radiocommunication Sector and the Director of the Radiocommunication Bureau (BR Director).

2.3 WRC Preparations

The preparatory work for the WRCs takes place at the global, regional, and national level. At the global level, the Conference Preparatory Meeting (CPM) of the ITU-R normally convenes right after the completion of a WRC to assign responsibility for specific studies to the appropriate ITU-R Study Groups to be finalised before the next WRC. The ITU-R Study Groups prepare the technical basis for the work at the subsequent WRC and report to the second session of

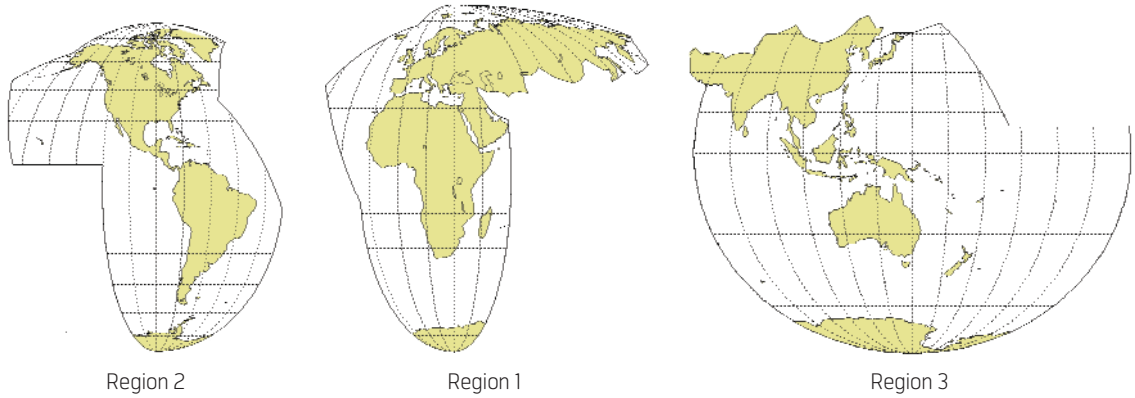


Figure 1 ITU-R Regions

ISSN 0085-7130 © Telenor ASA 2008

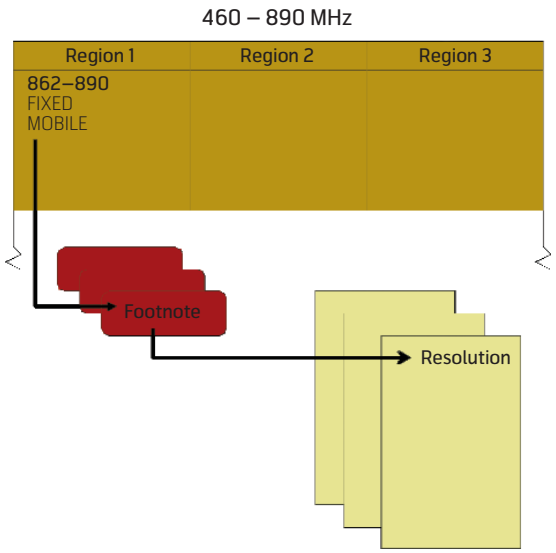


Figure 2 Illustrative example of Table of Frequency Allocations with footnotes and Resolutions in the RR

the CPM which draws up a consolidated report forming the basic support of the work of the WRC.

3 New Spectrum for IMT

Long before the conference started mobile vendor industry and operators, regulators, and several organisations had spent considerable effort in studying spectrum requirements for future IMT. Other business sectors such as broadcast and satellite, had prepared positions to protect existing non-IMT applications. On the one hand the mobile telecommunications business and interest organisations did push hard for a significant amount of new bands allocated on a primary basis for mobile and identified for IMT in footnotes. On the other hand, many broadcasters and satellite interest organisations objected to most of the proposals that had been put forward. At the conference itself all parties were present either as part of national delegations, as observers or at stands outside the conference area providing information and arguments, including lunches and receptions.

One of the first decisions of the WRC-07 was to agree to use the name IMT generically. The root name IMT now encompasses both IMT-2000 and IMT-Advanced collectively. There is now only one single IMT classification in the RR. RR Article 5 "Frequency allocations" now only refers to IMT. This decision gives increased flexibility and will enable refarming of spectrum previously allocated to 3G technologies in all bands identified for IMT in the RR. Spectrum currently identified for 3G technologies in the RR can be used in the future for 4G technologies and when frequencies are identified for 4G they can also be used for any IMT technology.

3.1 Additionally Required Bandwidth for Future IMT

It is a challenge to accurately estimate required new bandwidth for the fast expanding mobile services and IMT applications. The IMT long-term goals are very ambitious, as expressed in ITU-R Recommendation M.1645 [6], mentioning an end-user requirement of up to 100 Mbit/s access capacity for a full mobility mode and up to 1 Gbit/s for a fixed mode application. Cleary, this sets a significant pressure on radio frequency allocations as the common view seems to be that the useful mobile spectrum is below 5 GHz. To be more specific, a study was performed where a number of countries across the world identified how much new bandwidth would be necessary for them [7]. The demand was in the range from 1280 MHz based on a low market setting, to 1720 MHz for a higher market setting by year 2020. This would imply an additional spectrum requirement of from 700 to 1200 MHz of spectrum. However, not even Europe managed to propose this amount of spectrum in their common proposals to the conference [8].

3.2 Identification of Candidate Bands

The ITU-R Conference Preparatory Meeting (CPM) had identified the following bands as possible candidates for the terrestrial component of IMT to be considered by WRC-07 [9]:

- 410 – 430 MHz
- 450 – 470 MHz
- 470 – 806/862 MHz
- 2.3 – 2.4 GHz
- 2.7 – 2.9 GHz
- 3.4 – 4.2 GHz
- 4.4 – 4.99 GHz

The bands below 1 GHz are sometimes presented as good for coverage, whilst the bands with the highest frequency as good for capacity. The reasoning behind this is that radio waves at low frequencies propagate much longer and penetrate buildings and vegetation better than at high frequencies. The lower frequency bands cannot carry as much traffic as the higher end spectrum. Regarding the propagation issues it is well accepted that lower frequencies offer longer reach. For the capacity issue it has to be noted that it is the available bandwidth more than the radio frequency that matters. A number of Member States were of the opinion that at the higher frequency bands there would also be larger bandwidths available for IMT than in the lower bands.

At the time CPM issued its final report there were several options suggested on the identification for IMT, in particular whether IMT-2000 and IMT-advanced should be treated separately or not. By

the decision made by RA-07 [10] it was possible for WRC-07 to handle both IMT-2000 and IMT-advanced in a similar manner and existing and new identifications for IMT cover both types of systems, e.g. the existing IMT-2000 and IMT systems coming in the future.

3.3 Conflicts with Existing Services

There were two major conflicting issues with respect to the candidate bands as listed in the previous section. The spectrum below 1 GHz, the UHF band (470 – 862 MHz), is largely used for broadcasting services. The bands in the range 3.4 and 4.2 GHz are largely used by satellite services in many parts of the world.

3.3.1 UHF Band

Particularly the band 470-862 MHz, which was of great interest for mobile services, had for Region 1 and Iran just been replanned at the Regional Radiocommuncation Conference (RRC) in 2006, resulting in the Geneva 06 agreement GE06 [11]. This is tightly connected to the digital switchover, i.e. switch off of analogue broadcasting transmission and moving to fully digital broadcasting with the benefit of much better spectrum utilisation. If the same number of channels is broadcast with the same quality, the digital systems will only use a smaller part of the spectrum needed by the analogue systems. However, high-definition television (HDTV) is coming. HDTV requires significantly more spectrum than the currently used standard definition television, but still a noticeable portion of spectrum will be left over, called the digital dividend. The debate has been going on for several years on how to make use of the digital dividend; broadcasters state that there is no dividend as they wish to transmit more, e.g. several camera angles or additional pictures for the users to choose when watching a programme; others see an opportunity to offer broadband access utilising the spectrum dividend. The European Commission (EC) clearly recommended that this band be identified for IMT [12], whilst a majority of the individual EU-members opposed the same idea [8]. CEPT had submitted a European Common Proposal (ECP) on the UHF band supporting no change (NOC) at the WRC-07 and that the matter should be dealt with in a new agenda item at the subsequent conference – WRC-11 [8].

3.3.2 C-band

The global satellite business quickly established a strong opposition to any change in the band 3.4 – 4.2 GHz; the C-band [13]. These frequencies are used as downlink by satellite systems offering both broadcasting and broadband services. A significant number of satellites are already in orbit or planned making use of these bands. The terrestrial business, in particular the worldwide interoperability for microwave access (WiMAX) forum has been issuing profiles and equipment for the C-band. But the band was thought to be very useful for emerging IMT systems and the objective was to achieve a global spectrum allocation with identification for IMT in the C-band. In a way this conflict also turned into different interests by the developing world and the developed world.

CEPT had developed an ECP supporting the allocation to mobile on a primary basis and identification for IMT in a footnote in the frequency band 3400 – 3800 MHz for Region 1. This position was supported by the EC in their Communication on WRC-07 [12].

3.4 Bands Identified for IMT

One of the most contentious agenda items at the WRC-07 was the identification of future spectrum for IMT to ensure that there is sufficient bandwidth to provide mobile broadband services both to rural and densely populated areas. WRC decisions often have a long term perspective; some have a typical time span of 10 to 15 years from the date of the decision till practical implementation. For example, the first bands to meet the needs of the third generation mobile (3G) were already identified in 1992, at WARC-92 [14], for the future public land mobile telecommunication systems (FPLMTS). Having made such decisions the ground was prepared for both vendor industry to develop the systems and for the operators to strategically prepare for business to come. About ten years later 3G solutions started to come onto the market and are now a fast growing business. In November 2007, the bands 2.5 – 2.6 GHz allocated for mobile service on a primary basis and identified for IMT at the WRC-2000 [15] were auctioned in Norway and will be extensively used for 3G applications.

One of Telenor's most important objectives at the WRC-07 was to earmark additional spectrum in suitable bands for harmonised worldwide use and roaming to facilitate the development of mobile broadband which would provide higher transmission rates and enable new mobile applications.

3.4.1 Terrestrial Component of IMT

WRC-07 agreed that no identification of spectrum for IMT should be made in the frequency bands 410 – 430, 2700 – 2900 MHz and 4400 – 4990 MHz. Although these bands had been identified as possible candidate bands by ITU-R, see Section 3.2 , they did not get sufficient support at the conference.

The first agreements on a global primary allocation for mobile services and identification for IMT were reached in the following bands:

- 450 – 470 MHz
- 2300 – 2400 MHz

Both industry and operators had made substantial preparations ahead of the conference and several candidate bands below 5 GHz had been supported. At the WRC-07 the issue causing the highest level of disagreement was the possible allocation of new frequencies for mobile services and identification of IMT in the two candidate bands UHF (470 – 862 MHz) and the C-band (3400 – 4200 MHz).

WRC-07 agreed on the following globally harmonised spectrum bands identified for use by IMT:

- 450 – 470 MHz
- 698 – 806 MHz in Region 2
- 698 – 790 MHz in nine countries in Region 3 (Asia/Oceania) (Bangladesh, China, Korea, India, Japan, New Zealand, Papua New Guinea, Philippines and Singapore)
- 790 – 862 MHz in Region 1 (Europe, Africa) and Region 3 (Asia/Oceania)
- 2300 – 2400 MHz
- 3400 – 3600 MHz (no global allocation, but accepted by many countries).

Figure 3 illustrates per ITU-R region the various bands ranging from 450 MHz to 3.8 GHz, where there is a primary new or existing allocation to mobile services and identification to IMT. The allocation is either done in the main RR table or in a footnote to the table. For Region 1 (Europe, Africa) this means a total of 392 MHz new identification for IMT in bands allocated to mobile services on a primary basis.

The allocation of mobile service in the band 790 – 862 MHz in Region 1 will come into effect from 17 June 2015. This date is linked to the date 16 June 2015, which is the cut off date for transition from analogue to digital broadcasting in Region 1 according to the GE06 agreement. However, countries in this region that complete the transition from analogue to digital broadcasting before this date have the flexibility to introduce IMT in this band at an earlier date.

3.4.2 Satellite Component of IMT

For the satellite component of IMT the WRC-07 agreed to an allocation to the mobile satellite service on a primary basis in the frequency bands 1518 – 1525 MHz and 1668 – 1675 MHz in all three radio regions without extensive discussions. This was in line with the CEPT ECP on this issue [8] and in accordance with EC's Communication on WRC-07 [12].

Overall the WRC-07 result is good both for the industry, mobile operators, and consumers, especially given the position prior to the conference where CEPT supported no change for the UHF band and both CEPT and APT supported a postponement of

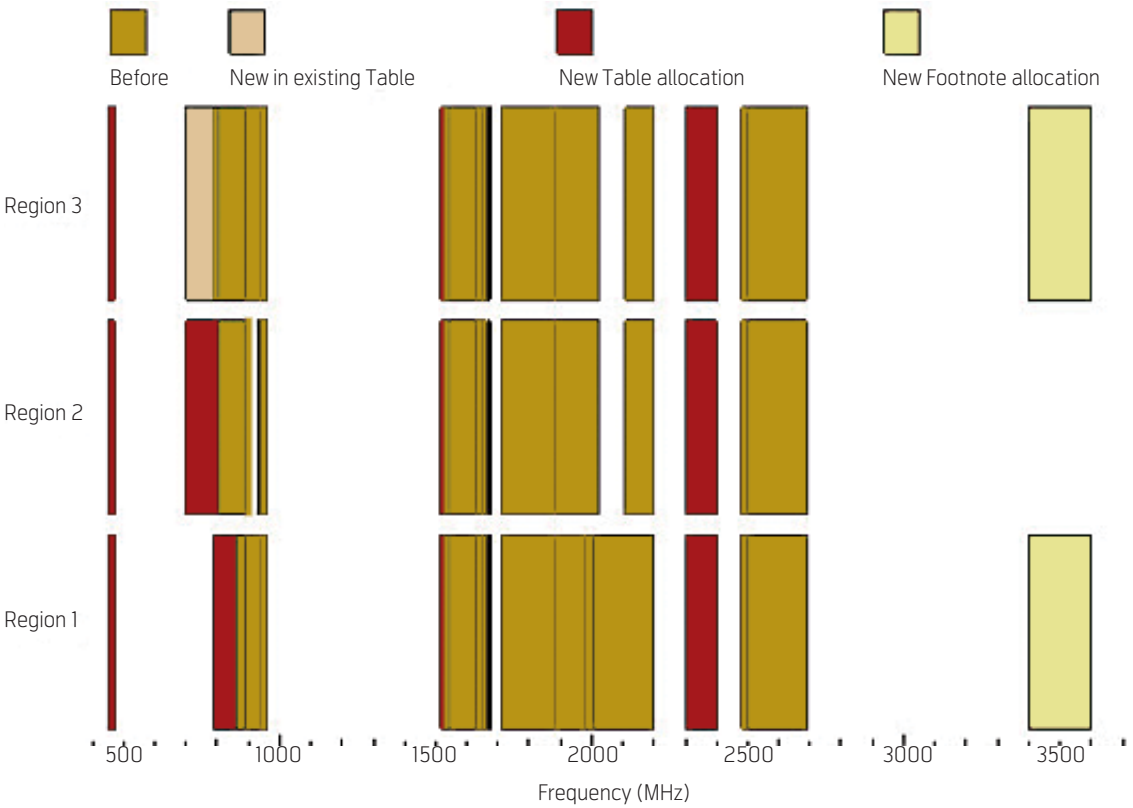


Figure 3 Radio spectrum allocated to mobile service on a primary basis and identified for IMT

WRC-07 outcome for 470 – 862 MHz, the UHF band

- 698 – 806 MHz: Primary allocation to the mobile service in the table and footnote identification of IMT in Region 2 (Americas)
- 698 – 790 MHz: Footnote identification of IMT in nine countries in Region 3 (Asia/Oceania): Bangladesh, China, Korea, India, Japan, New Zealand, Papua New Guinea, Philippines and Singapore
- 790 – 862 MHz: Primary allocation in table and footnote identification to IMT in Region 1 (Europe and Africa) and Region 3 (Asia/Oceania).

Comes into effect in Region 1 from 17 June 2015.

WRC-07 outcome for 3400 – 4200 MHz, the C-band

Regions 1 and 2

- 3400 – 3600 MHz in Region 1 (Europe and Africa)
- Allocation to the mobile service on a primary basis in footnote and identification for IMT for 82 countries (incl. all Nordic countries) in Region 1 (secondary allocation to mobile service in the table)
- 3400 – 3500 MHz in Region 2 (Americas): Allocation to the mobile service on a primary basis in footnote in 13 countries – not in US and Canada (no IMT identification) (secondary allocation to the mobile service in the table)

Region 3

- 3400 – 3500 MHz in Region 3 (Asia/Oceania): In Bangladesh, China, India, Iran, New Zealand, Singapore and French Overseas Communities, footnote allocation on a primary basis to mobile service and identification for IMT. Allocation efffective from 17 November 2010 (secondary allocation in the table)
- 3400 – 3500 MHz in Region 3: In Korea, Japan and Pakistan, this band is identified for IMT in a footnote (secondary allocation to the mobile service in the table)
- 3400 – 3500 MHz in Region 3: In Korea, Japan, Pakistan, Dem. People's Republic of Korea, and Singapore, footnote allocation to mobile on a primary basis (secondary allocation to mobile in the table)
- 3500 – 3600 MHz in Region 3: In Bangladesh, China, Korea, India, Iran, Japan, New Zealand, Pakistan and French Overseas Communities, this band is identified for IMT in a footnote (primary allocation to mobile service in the table)

consideration of this band till WRC-11. The allocation to the mobile service on a primary basis and identification to IMT in the band 790 – 862 MHz in Region 1 was in line with the Telenor position before the conference. The WRC-07 results form a good basis for further product developments of technologies for high data rate broadband systems. The development of IMT Advanced is a long term process where ITU-R and industry will cooperate to complete the technical specifications in the form of ITU-R Recommendations.

4 Plan for Fixed Satellite Service (FSS)

The WARC ORB-88 conference adopted in 1988 a plan giving in principle all ITU Member States a future possibility to operate geostationary satellites in the so-called Planned Fixed Satellite Service (FSS) bands [16]. The plan entered into force on 16 March 1990. The Thor II and THOR 5 satellites operated by Telenor Satellite Broadcasting AS are using these bands today. In the RR the regulatory and technical provisions relating to these bands are covered in Appendix 30B (AP30B) to the RR.

In the regulatory world there has been a trend towards simplification and improvement to certain procedures, and Agenda Item 1.10 of WRC-07 was established at WRC-03 to work out and implement regulatory and technical improvements in the bands covered by AP30B of the RR. The aggregate number of operational geostationary satellites has passed 330 [17]. Looking at Figure 4 showing several hundred operational geostationary satellites, the importance of planning and coordination is clearly understood.

The GSO is an orbit in which a satellite orbits the earth at exactly the same angular speed as the earth turns and at the same latitude. A satellite orbiting in GSO appears to be hovering in the same spot in the sky, and is directly positioned over the same patch of the earth at all times. As a result, an antenna can point in a fixed direction and maintain a link with the satellite. The satellite orbits in the equatorial plane in the direction of the earth's rotation, at an altitude of approximately 35 786 km above the earth. Hence, GSO becomes a natural resource having strategic value for all countries in the world.

An important issue related to coordination is to prevent loss of investments, customers and revenue from unstable capacity due to interference from other radio sources (e.g. other satellites operating in services using the same frequency bands).

There are in principle two mechanisms for sharing satellite orbits and frequency spectrum:

ISSN 0085-7130

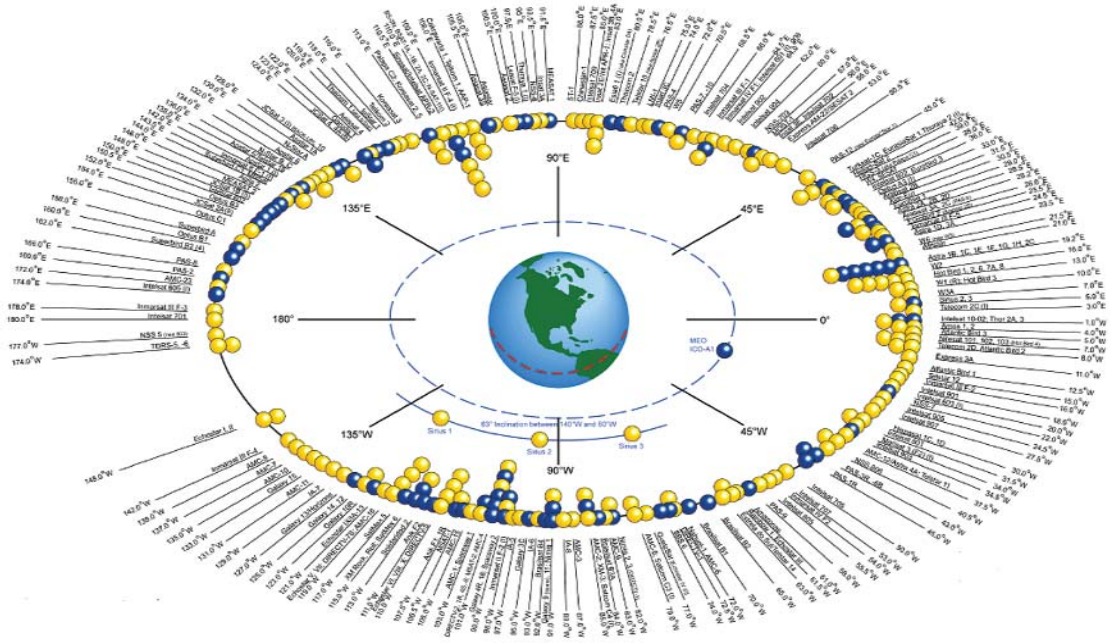


Figure 4 Existing satellites orbiting the earth in Geostationary orbit (GSO) [18]

- First come first served approach (Efficiency: actual requirements)
  - The legal right is acquired through coordination with administrations/satellite operators for actual usage.
  - Efficient orbit/spectrum usage
  - Homogenous orbital distribution of space stations
- Planning approach (Equitable access: Plan for future use)
  - Congestion of GSO by operators today having technical and financial skills to launch and operate a real satellite
  - Frequency/orbital plans
  - Guarantee for equitable access to the spectrum/orbital resources
    - Spectrum set aside for future use for all countries
    - Predetermined orbital position for each country and frequency spectrum

In this article the focus is on the planning approach.

It is important to be able to launch future spacecraft having legal rights to operate in the FSS bands. The planned bands have an inherent quality aiming at giving all ITU Member States a legal right to operate a geostationary satellite using certain technical and regulatory provisions, as opposed to the non-planned bands where the first come, first served principle is used. The objective of the Radio Regulations covering the Planned FSS bands (AP30B) is therefore to guarantee, for all countries, equitable access to the geostationary satellite orbit in the frequency bands of the fixed-satellite service covered by AP30B.

The Planned FSS bands cover 800 MHz for uplink and 800 MHz for downlink in the frequency spectrum:

- 6/4 GHz band (300 MHz):
  - 6725 – 7025 MHz (uplink)
  - 4500 – 4800 MHz (downlink)
- 13/10 -11 GHz band (500 MHz):
  - 12.75 – 13.25 GHz (uplink)
  - 10.70 – 10.95 GHz and 11.20 – 11.45 GHz (downlink)

Today, relatively few operators make use of the AP30B bands due to complicated regulatory provisions. The application of this band could either be based on national or international exchange of information between end users, disaster relief, telemedicine, Internet, and television.

The service area of an allotment in AP30B is inside national borders, and consequently the AP30B plan is meant for national coverage. However, submitting a filing under another category could make it possible

to extend the service area. The legal rights to operate are assured by giving frequency band protection at specific geographical test points on the earth within the coverage area of a satellite beam. A maximum of 10 test points apply for allotments and 20 test points for assignments.

The impractical processing method of filings covered by the RR has been the basic problem with the regulation regime in the Planned FSS bands. Prior to WRC-07, the filings or networks in this frequency band were treated sequentially; i.e. one-by-one in time. The result was lack of processing efficiency, a huge back-log of filings and the fact that the notifying administration did not know until a few weeks before publication of the processed information which networks it had to coordinate with. Consequently, it was difficult to coordinate in a satisfactory manner. Given a list of affected administrations/networks, the notifying administration only had 30 days to complete coordination. As a result administrations introduced holes in coverage diagrams to avoid conflicts.

4.1 WRC-07 Results for FSS Replanning

The new provisions for the Planned FSS bands have adopted some of the vocabulary from the existing Planned Broadcasting Satellite Service (BSS) regulatory provisions in AP30 (BSS downlink) and AP30A (BSS uplink). The backlog of Planned FSS networks counted more than 100 prior to WRC-07, giving a worst case period of about four years in the backlog queue before processing by the ITU. The result of WRC-07 is that the processing and publication of networks will be non-sequential to reduce the backlog and give more predictability for the administrations and operators involved in the coordination process. The legal status of a satellite filing in AP30B band will in the future be given by:

- Date of receipt (by the ITU)
- Date of "Part A" publication
- Date of "Part B" publication
- Date of notification

"Part A" publication contains the assignments and all findings for the technical analysis, and listing of all networks which the system in question must be coordinated with. Hence, the notifying administration knows exactly which networks it has to coordinate with.

Generally, a frequency filing expires after eight years if a satellite network using the filing is not brought into use. The notifying administration has several years to coordinate a satellite network. "Part B" publications could therefore, in principle, be published several years after "Part A" and will contain the same filing with technical parameters as a result from the coordination process. In general, a filing may only reduce parameters being a source of radio transmitting interference, i.e. coverage diagram, service area, and power density. The network will enter the List of assignments after successful coordination and "Part B" publication. The notification process can start after "Part B" publication, for example using the same parameters as specified in "Part B". Following the notification process, the assignments contained in the filing will be recorded in the ITU Master International Frequency Register (MIFR).

AP30B containing the Planned FSS bands is an Allotment plan, and the filing has to be entered into the List before it can be notified. This is in opposition to Planned BSS, which is a plan for BSS assignments giving administrations or operators the ability to notify the filing without modifications to the Plan.

Other changes to the Planned FSS bands after WRC-07 are:

- Introduction of a "Coordination arc", i.e. only necessary to coordinate within +/- 9 degrees from nominal orbital position in 13/10 – 11 GHz band and +/- 10 degrees in 6/4 GHz band.

- Pre Determined Arc (PDA) concept is removed. Previously it was possible to move the orbital position of neighbouring satellites up to +/10 degrees from its nominal orbital position in order to achieve coordination profits. All administrations will be assigned a fixed orbital location. The Norwegian orbital position is fixed to 0.8 degrees west longitude.

- The "existing systems", i.e. systems received by the ITU before WARC-ORB-85 were supposed to expire in March 2010. However, these networks may continue to operate; this is covered in a new Resolution 148.

Anne Lise Lillebø, Telenor ASA, and Kjersti Hamborgstrøm, Telenor Satellite Broadcasting, Delegates, Norwegian Delegation

- Updated technical criteria by using C/I criteria to protect networks inside the coordination arc and power flux density (pfd) limits to protect networks outside the coordination arc. Aggregate carrier-to-interference ratio (C/I)agg is reduced from 23 to 21 dB while single-entry carrier-to-interference ratio (C/I)s.e is reduced from 27 to 25 dB. The reference antenna diameter is reduced to 5.5 m in 6/4 GHz band and 2.7 m in 13/11 – 10 GHz band.

- New ITU Member States have priority in the processing of Planned FSS networks. This is a new regulatory concept in this band. A natural consequence here is the necessity for administrations having filings in the backlog to coordinate their networks with new Member State(s) networks. Member States having sent request for an allotment to ITU before WRC-07 were Uzbekistan, Kazakhstan, Azerbaijan, Belarus and Lithuania. In addition, the Czech Republic had submitted a filing with given priority in January 2008.

- The changes for Planned FSS bands adopted at WRC-07 entered into force on 17 of November 2007 according to Resolution 149.

- Transitional arrangements have been given for networks already in the backlog queue before WRC-07 since all networks (pending and new) will be treated according to the new regulations (also applies to networks in the backlog having been submitted under the "old regime").

- Satellite networks previously named "additional use" and "subregional systems" have been merged into a category called "additional systems". There is no time limit (expiry) for networks in the Planned FSS bands having already been brought into use, as opposed to the Planned BSS bands (not an agenda item at this conference) where networks are given 15 years of life with an option of another 15 years provided that no modifications apply to the BSS filing.

- It is still necessary to obtain bilateral agreements with administration(s) inside the service area of a satellite beam. Such agreements must be completed before "Part B" publication in the coordination process.

4.1.1 Due Diligence

WRC-03 adopted a revision to Resolution 49 in order to get administrations to provide due diligence information about the satellite, launch vehicle, etc for the spacecraft intended to use a satellite filing [19]. If Resolution 49 information is not submitted to the ITU in time, the filing will expire after eight years. This was an important result aiming at reducing the backlog of satellite filings. Resolution 49 is still valid for all satellite networks with the exception of networks belonging to new ITU Member States.

4.1.2 Cost Recovery for Satellite Network Filings

Resolution 88 of the ITU Plenipotentiary Conference [20] decided to implement cost recovery for satellite filings. Decision 482, modified in 2005 [21], describes the cost recovery fees for satellite filings, and is still valid except for new Member States.

4.1.3 Categories of Filings

Following WRC-07, a filing can be in one or up to two of the following categories under the Planned FSS bands:

- *Allotments* – predetermined orbital position and frequency resources for national coverage/use;

- *Assignments in the List* – coordinated frequency rights for planned or operational spacecraft:

- *Additional systems* – systems with coverage/use beyond national coverage.

A real satellite may apply its planned or operational frequency rights from one or several frequency filings; i.e. there is normally not one filing corresponding to one satellite.

When an administration intends to utilise its national allotment it may convert the allotment to an assignment with parameters given in the Plan without any modification. If one or several modifications are included, the filing must be coordinated according to the existing reference situation of AP30B containing all other allotments and assignments in the List. By the end of WRC-07, the number of networks forming reference situation in 13/10 – 11 GHz band of AP30B was more than 305.

Telenor Satellite Broadcasting AS operates the Thor II and THOR 5 satellites using the AP30B band. However, Thor II will be taken out of service when THOR 5 is brought into operation during April 2008. THOR 5, launched in February 2008, uses 250 MHz in uplink and 250 MHz in downlink, and THOR 6 (planned for launch in May/June 2009) will use 250 MHz in uplink and 250 MHz in downlink of the Planned FSS bands.

5 Some Other WRC-07 Decisions

5.1 Climate Issues and Disaster Prediction

The WRC-07 adopted a new Resolution on "Radiocommunications use for Earth observation applica-

tions" where the availability of radio frequencies to carry out earth observation is identified to be of vital importance. The earth observation data are essential for monitoring and predicting climate changes, for disaster prediction and for increasing the understanding of all aspects of climate change. The importance of earth observation radiocommunications applications has been stressed by a number of international bodies such as the Group on Earth Observation (GEO), the World Meteorological Organization (WMO) and the Intergovernmental Panel on Climate Change (IPCC). The Resolution invites ITU-R to carry out studies to improve the recognition of the essential role and global importance of Earth observation radiocommunications applications.

This WRC-07 Resolution was considered in the ensuing meeting of ITU-T's Telecommunications Standardization Advisory Group (TSAG) in December 2008, and ITU-T will organise two workshops on ICTs and climate change in 2008 in cooperation with ITU-R and ITU-D to consider how ICT can help to contribute to reduce the climate change.

5.2 Updated Radio Regulations (RR) – Entry into Force

The entry into force of the updated WRC-07 RR is generally 1 January 2009. However, for certain services some specific decisions for other entry dates apply. In addition, there are specific dates for entry into force in individual country footnotes.

6 WRC-11 Agenda

The WRC-07 agreed on an agenda for the next WRC which will take place in 2011. It should be noted that there is no single agenda item dedicated to additional spectrum for IMT. For Telenor, the following agenda items would be of interest:

6.1 Selected Agenda Items (AI)

AI 1.2 *"taking into account the ITU R studies carried out in accordance with Resolution 951 (Rev. WRC-07), to take appropriate action with a view to enhancing the international regulatory framework;"*

This agenda item opens the possibility of revising the current service definitions in the RR which do not reflect the ongoing convergence between the different radio services. ITU-R has outlined a number of options including a review and possible revision of some of the service definitions, introduction of a new provision in the RR enabling substitution between assignments of specific services and the introduction of so-called "composite" services in the Table of Frequency Allocations.

Erik Jørol, Deputy Head, Norwegian Delegation, depositing the signed Final Acts on behalf of Norway

AI 1.13 *"to consider the results of ITU-R studies in accordance with Resolution 551 [COM6/13] (WRC-07) and decide on the spectrum usage of the 21.4 – 22 GHz band for the broadcasting-satellite service and the associated feeder-link bands in Regions 1 and 3."*

This agenda item proposes to take into account results from ITU-R studies for a possible use of BSS (downlink) for HDTV in the band 21.4 – 22.0 GHz in Regions 1 and 3. This band was allocated at WARC-92 and came into force on 1 April 2007. The band is currently allocated on a primary basis to the Fixed Service and Mobile Service for Regions 1, 2 and 3. In Regions 1 and 3 the band is also allocated on a primary basis to Broadcasting Satellite Service.

AI 1.17 *"to consider results of sharing studies between the mobile service and other services in the band 790 – 862 MHz in Regions 1 and 3, in accordance with Resolution [COM4/13] (WRC-07), to ensure the adequate protection of services to within this frequency band is allocated, and take appropriate action."*

This agenda item proposes a follow up of the decision regarding the UHF band at the WRC-07. WRC agreed that studies should be carried out to solve any interference problems between other services includ-

ing broadcasting and IMT in the UHF band in time for the next WRC in 2011. WRC-07 adopted a Resolution asking for further sharing studies between the mobile and the broadcasting service in the UHF band in Regions 1 and 3 and requests WRC-11 to take appropriate action.

> ***AI 1.19*** *"to consider regulatory measures and their relevance, in order to enable the introduction of software-defined radio and cognitive radio systems, based on the results of ITU-R studies, in accordance with Resolution [COM6/18] (WRC-07)".*

The RA-07 Res ITU-R 54 recognises the growing importance of short range devices (SRDs) which generally use the unlicensed bands and calls for continuing ITU studies in this field. SRDs use a wide range of frequencies around the world and the Res 54 will also study whether spectrum management techniques like Cognitive Radio, Software Defined Radio and "polite" protocols can help overcome these inconsistencies.

6.2 Follow-up

All preparatory work as agreed by the first session of CPM-11 will be performed within the framework of the planned work programme and within the new organisation of the ITU-R Study Groups adopted by the RA-07. The only exception is the handling of the new Agenda Item 1.17, regarding the complex issue related to the use of the band 790 – 862 MHz in Regions 1 and 3.

6.2.1 CPM Follow-up – Sharing Studies in the UHF Band

The Conference Preparatory Meeting (CPM) which met directly after the WRC-07 decided that the sharing studies between the mobile service and other services in the band 790 – 862 MHz in Regions 1 and 3 will be conducted by a CPM Joint Task Group 5/6 (ITU-R Study Group 5 – Terrestrial services and Study Group 6 Broadcasting services).

During a GSMA debrief in November following the WRC-07 GSMA members agreed that industry and operators should monitor and participate in this work to strike a balance between mobile and broadcasting interests. GSMA considers taking an initiative in this matter.

6.2.2 Technical Studies in ITU-R Study Groups

ITU-R Study Group 5 Terrestrial services is i.a. responsible for mobile services and IMT, and Study Group 5 Working Party 5D will start studies on band plans and other relevant technical parameters for the additional bands identified for IMT by the WRC-07. The first meeting of this WP is scheduled to take place in Geneva in January/February 2008.

7 Conclusions

The article has presented a number of main results from the World Radiocoummucation Conference 2007 (WRC-07). In particular the primary allocation of new spectrum for mobile communications, also identified for International Mobile Telecommunications (IMT) received a lot of attention at the conference as well as through the whole period of preparations since 2004. Indeed, a significant amount of spectrum was allocated and identified, up to 400 MHz for many countries. This provides a good basis for future development of the mobile industry and operators, towards both global coverage and broadband services, where a 10-15 year perspective often is necessary. Both new bands and global harmonised frequencies are expected to become of great importance.

WRC-07 revised the technical and regulatory provisions for Fixed Satellite Service (FSS) for various conditions and types of services. The result is an improved utilisation of the natural resources and better effectiveness of the planning procedures.

A number of other issues were covered as well, of interest to telecommunication operators and others. These covered areas such as emergency and disaster relief, technical issues of importance for climate and earth exploration, and ideas for enhancement of radio services using short range devices that use radio spectrum locally.

Telenor, as a telecommunication operator with a major interest in services making use of radio spectrum, is very dependent on both enough spectrum resources and effective and interference-free access to the spectrum. The WRC-07 results were satisfactory from this perspective.

References

1 Lillebø, A L, Tjelta, T. ITU-R Radiocommunication Assembly 2007. *Telektronikk*, 104 (1), 134–143, 2008 (this issue)

2 ITU. *Constitution of the International Telecommunication Union*. Geneva, ITU, 2004.

3 ITU. ITU Council. *Agenda for the World Radiocommunication Conference (WRC-07)*. 18 June 2004. (Document C04/84)

4 ITU-R. *Radio Regulations*. Geneva, ITU, 2004.

5 ITU. *ITU News. Commemorative edition: The centenary of the international Radio Regulations*. 3, April 2006.

6 ITU-R. *Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000*. Geneva, ITU, 2003. (Recommendation ITU-R M.1645)

7 ITU-R. *Estimated spectrum bandwidth requirements for the future development of IMT-2000 and IMT-Advanced*. Geneva, ITU, 2006. (Report ITU-R M.2078)

8 ITU-R. *European common proposals for the work of the conference. Part 4*. Geneva, ITU, 5 June 2007. (WRC-07 Doc. 10 Add. 4)

9 ITU-R. *CPM Report*. Geneva, ITU, 2007.

10 ITU-R. *Radiocommuncation Assembly. Final Acts*. Geneva, ITU, 2007.

11 ITU-R. *Final Acts of the Regional Radiocommunication Conference for planning of the digital terrestrial broadcasting service in parts of Regions 1 and 3, in the frequency bands 174-230 MHz and 470-862 MHz*. RRC-06, Geneva, 2006.

12 European Commission. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions. *The ITU World Radiocommunication Conference 2007 (WRC-07)*, COM(2007) 371 Final, Brussels 2 July 2007.

13 *No Change Initiative. NOC 3.4-4.2*. February 2008. http://www.no-change.info

14 ITU-R. *Final Acts WARC-92*. Geneva, ITU, 1992.

15 ITU-R. *Final Acts WRC-2000*. Istanbul, 2000.

16 ITU-R. *Final Acts WARC ORB-88*. Geneva, ITU, 1989. (ISBN 92-61-03931-6)

17 TBS. *The Satellite Encylopedia*. TAG's Broadcasting Services (TBS). http://www.TBS-internet.com

18 Boeing. *Commercial communication satellites. Geostationary orbit*. February 2008, http://www.boeing.com/defense-space/space/bss/launch/980031_001.pdf

19 ITU-R. *Final Acts, WRC-03*. Geneva, ITU, 2003.

20 ITU. *ITU Plenipotentiary Conference*, Marrakech 2002.

21 ITU-R. Decision 482 (modified 2005). *Implementation of cost recovery for satellite network filings*. Geneva, ITU, 25 July 2005. (Document C05/106)

Annex

WRC Preparations

The preparatory work for the WRCs takes place at the global, regional and national level. At the global level, the Conference Preparatory Meeting (CPM) of the ITU-R normally convenes right after the completion of a WRC to assign responsibility for specific studies to the appropriate ITU-R Study Groups to be finalised before the next WRC. The ITU-R Study Groups prepare the technical basis for the work at the subsequent WRC and report to the second session of the CPM which draws up a consolidated report forming the basic support of the work of the WRC.

The regional organisations play an instrumental role in facilitating discussions of national proposals and contribute to an important harmonisation of national positions before the conference. Most decisions at WRCs are taken by consensus and the regional preparations around the world contribute significantly to the consensus building both before and during the conference. All the ITU Member States also have the right to submit formal proposals directly to the WRC.

ITU-R organised an information meeting on WRC-07 for African countries in Geneva in August 2007. Both ATU (African Telecommunications Union), APT, CEPT, CITEL, RCC (Regional Commonwealth in the field of Communication), GSMA and UMTS Forum took part in this meeting which proved very useful for

Anne Lise Lillebø, Telenor ASA, Delegate, Norwegian Delegation, and Moriani Mohamed, DiGi, Delegate, Malaysian Delegation

ITU

GSMA's stand at WRC-07

explaining common proposals and exchanging points of views.

In Asia, preparatory work for the WRC is undertaken by the Asia Pacific Telecommunity (APT) which normally establishes a special group, the APT Preparatory Group (APG), to harmonise views of APT members and to develop common proposals from the Asia-Pacific region for submission to the WRC. The APT submitted a total of 28 common proposals to the WRC-07. The Telenor subsidiary DiGi in Malaysia is an Affiliate Member of APT and participated in the preparations.

In Europe, CEPT – Conférence Européenne des Administrations des postes et des télécommunications – with its 48 members, is an association of telecommunication administrations of all European countries, both EU Member States and non EU members. CEPT and its Electronic Communications Committee (ECC) are responsible for the European preparations to the WRC through its Conference Preparatory Group (CPG). Telenor was allowed to take part in the Norwegian delegation to the meetings of the CPG. The CPG-07 was chaired by Anders Frederich, Sweden, and agreed on 25 Common European Proposals (ECPs) which were submitted to the WRC as formal proposals to the conference.

Anders Frederich, Deputy Head, Swedish Delegation, Vice-Chairman of WRC-07 and Chairman of CEPT ECC CPG, and Marianne Treschow, Head, Swedish Delegation

Both APT, CEPT and CITEL (Inter-American Telecommunication Commission) exchange observers on a mutual basis and this ensures a large degree of transparency in the preparations to the conference and helps to build consensus across regions of the world.

GSM Association (GSMA), where Telenor is a member, rallied their members to a common lobbying activity during the last session of the CPM-07. GSMA organised meetings with members and representatives from the various radio regions to promote the GSMA view for more spectrum for IMT at the WRC-07.

In Norway the Norwegian Post and Telecommunications Authority (NPT) set up a special preparatory group called NORWRC, and all stakeholders in Norway were invited to participate. NPT provided valuable information on the European preparations within the framework of CEPT and agreement was achieved on which European Common Proposals that Norway should support. Telenor and NPT were in agreement both on Agenda Item 1.4 and 1.10, and had a very good cooperation both before and during the WRC-07.

Consultations During WRC

During the conference both APT and CEPT held frequent coordination meetings with their members to discuss the progress of the conference and how to arrive at possible compromises. This process is extremely valuable as positions tend to shift during the WRC and sufficient support for a change of direction is necessary to carry the discussions forward to a final result.

As a member of the Norwegian delegation Telenor was invited to take part in CEPT consultations on agenda items of importance to our company, and our views were respected and served as valuable input in the discussions.

During the WRC-07 the GSMA set up a separate booth in the vicinity of the conference venue where delegates received information about the GSMA positions concerning future need for spectrum for mobile services. GSMA actively lobbied for their positions by inviting delegates to the booth and to mutual dialogues to exchange views.

Norwegian Delegation to WRC-07

Geir Jan Sundal, Director of the Radio Department, Norwegian Post and Telecommunications Authority, was head of the Norwegian Delegation to WRC-07, whereas Erik Jørol and Eirik Bliksrud, Norwegian Post and Telecommunications Authority, were Deputy Heads of the delegation. Telenor was represented by Kjersti T. Hamborgstrøm and Erik Otto Evenstad, Telenor Satellite Broadcasting AS, Stein Gudbjørgsrud, AeroMobile, Terje Tjelta, Telenor R&I and Anne Lise Lillebø, Group Regulatory, Telenor ASA. All Telenor representatives were delegates and members of the Norwegian delegation.

Erik Jørol, Deputy Head, Norwegian Delegation and Vice-Chairman of WRC-07 Committee 5, and François Rancy, Chairman of WRC-07

Conference Structure

François Rancy, Director of the Agence Nationale des Frequences, France, was appointed Chairman of the WRC-07. In addition seven Vice-Chairmen were appointed representing the five administrative regions of the ITU. Anders Frederich, Sweden, was appointed Vice-Chairman and represented Western Europe. Erik Jørol, Norway, was appointed Vice-Chairman representing Western Europe in Committee 5.

There were a total of seven committees:	
Committee	Chairman and composition
Committee 1 – Steering	Composed of the Chairman and Vice-Chairmen of the Conference and of the Chairmen and Vice-Chairmen of the Committees
Committee 2 – Credentials	Chairman: Mr S. Coulibaly (Mali)
Committee 3 – Budget Control Committee	Chairman: Mr Carlos A. Merchan (Mexico)
Committee 4 – Specified agenda items, including AIs 1.4 and 1.9	Chairman: Mr M. Dupuis (Canada)
Committee 5 – Specified agenda items including AI 1.10	Chairman: Mr A. Hashimoto (Japan)
Committee 6 – Future agenda and work programme	Chairman: Mr A. Nalbandian (Armenia)
Committee 7 – Editorial	Chairman: Mr F. Sillard (France)

Erik Otto Evenstad is senior adviser in Telenor Satellite Broadcasting AS, Space Systems Division. He is working with interference analysis and frequency coordination aiming at maintaining existing and achieving future legal frequency rights for satellite systems operated by Telenor. He received his Master of Electrical Engineering in 1986 from NTH (now NTNU), Trondheim, Norway, and his Master of Management in 2003 from the Norwegian School of Management. He has been employed in several departments in Telenor, i.a. Telenor R&I from 1995 to 2000. Erik Otto Evenstad has been working with satellite communications in Televerket/Telenor for about 20 years.

email: erik-otto.evenstad@telenor.com

For a presentation of Terje Tjelta and Anne Lise Lillebø, please turn to page 143.

ISSN 0085-7130 © Telenor ASA 2008

The X.500 Directory Standard: A Key Component of Identity Management

ERIK ANDERSEN

Erik Andersen is an independent consultant with the company Andersen's L-Service

New things generally fascinate people. This is especially true within the field of Information and Communications Technology (ICT). In science, it is good practice to make a literature search when starting a research project. This is a less common practice within ICT. Right now Identity Management (IdM) and Next Generation Network (NGN) are attracting great interest. Before developing directory like specifications within these areas, current works should be considered. It appears that the X.500 standard could play a major role here.

Introduction

X.500 is a directory standard and is therefore a specification for how information about entities (objects) is stored, retrieved, updated, deleted, managed and protected. X.500 has been developed and continuously extended for the last 23 years and this process continues. Many highly skilled people have contributed to the work over the years. The X.500 standard is an extensive specification consisting of ten documents (X.500, X.501, X.509, X.511, X.518, X.519, X.520, X.521, X.525 and X.530). X.509 is widely known as the basis for digital signatures, public-key infrastructure (PKI), etc. X.500 is an introductory document, but the term X.500 is here used as a synonym for the whole series. The X.500 standard is developed jointly between ITU-T and ISO/IEC. ISO/IEC labels the standard ISO/IEC 9594, Parts 1 to 10.

The Internet Engineering Task Force (IETF) has developed the Lightweight Directory Access Protocol (LDAP) for accessing X.500 directories. It is part of the X.500 family.

Identity Management (IdM) and Next Generation Network (NGN) use the term *entity* about things that need to be identified and named. X.500 uses the term *object*. This is the term used in the following.

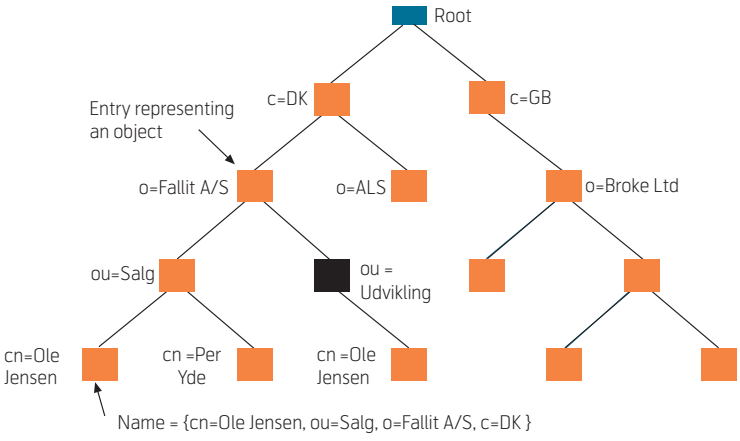


Figure 1 Small sample DIT

You can find more information about X.500 and the standardization process with links to documents, etc. at http://www.x500standard.com/. It is a good site to visit.

X.500 is many things. It defines protocols, procedures, replication, etc., and it provides models for naming structure, information structure, protection, management, etc. Modelling is a major part of the X.500 standard.

Directory Information and Naming Model

An object, say, a human being, only has an identity if it has a name (and possibly other identifiable attributes) that uniquely identifies that object, at least within a specific context. An object may have several names, but no two objects should have the same name. X.500 has a naming model that assures unique naming if proper naming authorities are in place. The naming model recognises that objects are hierarchically organised by having a hierarchical naming structure. As an example, a person within an organisation may have a name consisting of a sequence of name components, like a country name component, an organisation name component, an organisational unit name component and finally a personal name component. We call such a name a *Distinguished Name*.

An *entry* within a directory represents an object and entries are viewed as having the same hierarchy as the corresponding objects. This hierarchy of entries is called the *Directory Information Tree* (DIT). A rather simple DIT is shown in Figure 1.

Information about an object is modelled as *attributes* within its entry. Attributes are of different types according to the kind of information.

Figure 2 illustrates the entry and attribute concepts. An attribute is of a specific type, for example, a telephone number. An attribute may have multiple values.

ISSN 0085-7130 © Telenor ASA 2008

Recent editions of the X.500 standard allow an entry to contain so-called child entries, which allow attributes to be grouped according to different purposes. A person having different roles may have its attributes grouped according to those roles. Identity Management talks about different contexts for an entity. Child entries are the support for this concept.

Directory Schema

It is important that the information in a directory be orderly organised. An administrator can impose a directory *schema* that determines the types of objects for which the directory may hold entries, where such entries should be placed in the hierarchy, what kind of attribute types that may be associated with each type of object, etc.

Directory Components and Protocols

Let us get some terminology in place. A Directory System Agent *(DSA)* is a directory server holding directory information. Several interconnected DSAs may form a distributed directory. A Directory User Agent *(DUA)* is a client system supporting the user access to a directory by connecting to a DSA within the directory. X.500 defines several protocols. The Directory Access Protocol *(DAP)* supports communication between a DUA and a DSA. This protocol is used to invoke directory operation The Directory System Protocol *(DSP)* is use between DSAs when an operation requires the involvement of more than one DSA. The Directory Information Shadowing Protocol *(DISP)* is used when replicating information from one DSA to another DSA. The Directory Operational Binding Management Protocol *(DOP)* is used to establish a common understanding between two DSA, for example, about knowledge of each other.

The Lightweight Directory Access Protocol *(LDAP)* was initially developed within the Internet Engineering Task Force (IETF) as an X.500 access protocol. Later it has also developed LDAP directory server specifications based on the X.500 model.

Directory Database

A DSA needs some kind of underlying database to support directory information handling. X.500 does not specify any particular database technique, as the type of database does not affect interoperability among different systems. How entries and attributes are reflected in the database is a pure implementation issue. However, an efficient database technology with an efficient data structure is essential for performance. Database performance is one of the more critical issues.

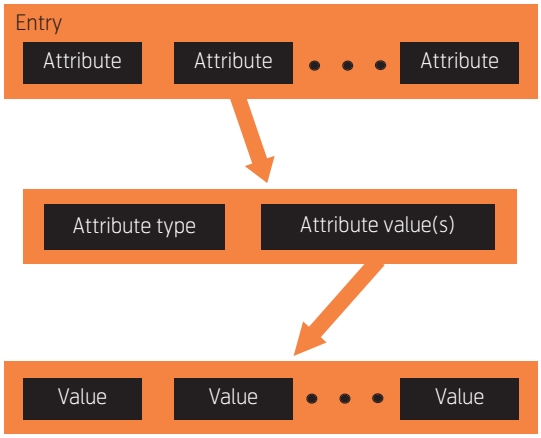


Figure 2 Entry and attribute structure

Distribution and Replication

From the very beginning, it was realised that the X.500 standard should allow *distribution* of information among different systems. Later, it also became evident that *replication* is a major requirement. Distribution is important, as it allows a service provider to maintain its own directory system and then interconnect that system with systems of other service providers to establish an integrated global system. This is illustrated in Figure 4. Users may utilise the combined service of all the interconnected DSAs.

Replication of directory information is illustrated in Figure 5. It provides availability and load sharing. Replication is also called *shadowing*. An administrator may decide that only a part of the information held by a DSA may be replicated to another DSA. Even individual entries and attributes may be excluded. The DSA receiving the shadow may in addition hold information of its own.

Replication and in particular distribution require elaborate procedure specifications to allow proper interworking among different vendor implementations to provide a seamless service.

The X.500 naming structure allows navigation within a distributed and replicated directory providing the

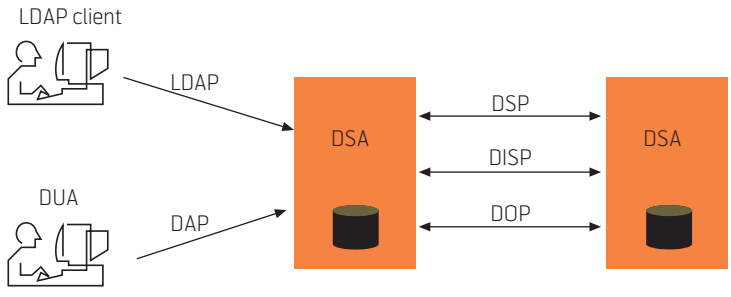


Figure 3 Distributed directory

ISSN 0085-7130

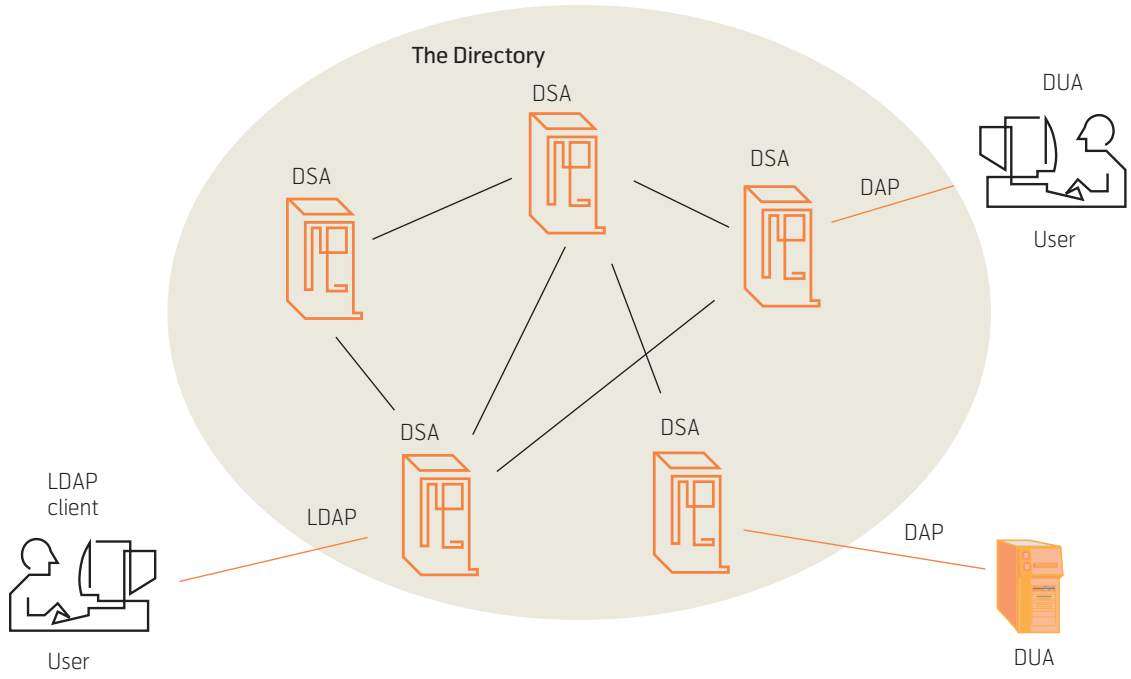


Figure 4 Distributed directory

different systems maintain knowledge of each other's network addresses and name domains. Procedures are defined for how such knowledge can be maintained. X.500 is the only directory standard that has specifications for distribution and replication.

Deployment of Directories

An X.500 directory may be employed whenever it is possible to establish a consistent naming structure. Network Directory Server is one such important deployment of an X.500 directory, and in more simple cases by an LDAP directory (Microsoft's Active Directory is an important example).

Directory Operations

A user may invoke directory *operations*. There are operation types for reading a named entry, searching, adding, updating and deleting entries, etc.

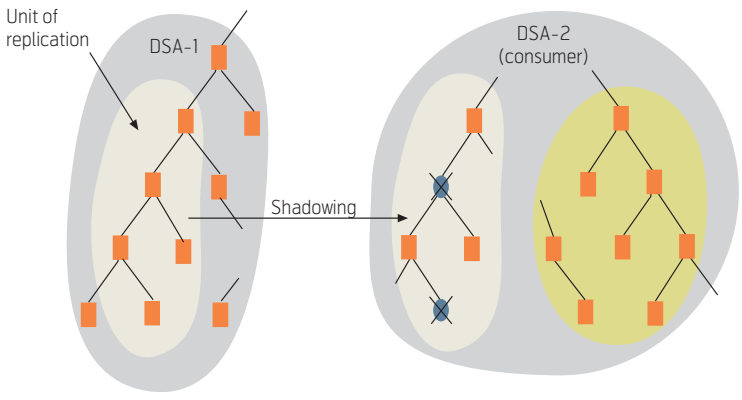


Figure 5 Replication

The search operation is the primary operation for retrieving information. It is the most complex operation, as it requires many entries to be checked and it requires matching between the information supplied by the DUA (search criteria) and the information within entries. Matching can be rather straightforward or it can be quite complicated. X.500 defines a range of matching rules from quite simple rules to very sophisticated rules.

Data Protection

Data protection is much in focus at the time being. It is a major part of Identity Management. Data protection is primarily a privacy issue, but may also be a way to protect the assets the data may represent. Almost from the beginning, data protection features have been an important part of the X.500 standard. X.500 is the only directory specification having these important features. A common data protection model is important for several reasons. It gives a consistent service to users, as a systems reaction is not depending on the brand of system. In a replicated multi-vendor environment, it is necessary that data protection information be replicated together with the data allowing the receiving system to protect such information.

Different levels of data protection are required for different types of accessing users. The authentication level of a user also affects the access right. X.500 provides for several levels of authentication, ranging from none, only name, name and password, name and encrypted password, and finally strong authentication based on digital signatures.

An elaborate access control has been part of the X.500 standard almost from the beginning. Access control is a question about who may do what or may not do what based on the level of authentication. Retrieving sensitive information or updating entries may require higher level of authentication than just retrieving non-sensitive information.

Access control is about the *right to know*. However need-to-know goes beyond access control, as the right-to-know should not be sufficient to retrieve information if a need-to-know is not established. For this reason, a service administration concept has been developed. A directory system without service administration could be compared to a buffet restaurant, while a directory system with service administration could be compared with an à la carte restaurant. An administrator may define a final set of distinct services, specify what kind of input is required and what type of output is relevant. Such services may be tailored differently for different user groups. In addition, the input restriction may be tailored to prevent devious searches and to prevent data trawling.

Distributed Management

An X.500 directory is designed to be distributed and to be owned by different service providers and organisations. The administration of the directory can be distributed by defining *autonomous administrative areas*. Such areas could follow DSA boundaries, but may not necessarily do so. Several organisations or organisational units could share a single DSA by dividing the DSA into multiple autonomous administrative areas. An autonomous administrative area may also span DSAs.

Within an autonomous administrative area, an administrator may define its own directory schema, its own access control, its own service administration, etc. An autonomous administrative area may be subdivided for different purposes, for example, to support fragmented access control. When service administration is applied, an administrative area acts as a mini directory. A search initiated outside the administrative area will not migrate into that area. Likewise, a search initiating within the area will not migrate out of the area.

X.509 and PKI

X.509 is a subject of its own. It also has a life outside the remainder of X.500. It is the basis for many other specifications, such as the Secure Socket Layer (SSL). X.509 is about message integrity, authentication and authorisation. X.500 protocols use the features of X.509 to provide security functions not available in other types of directory specifications.

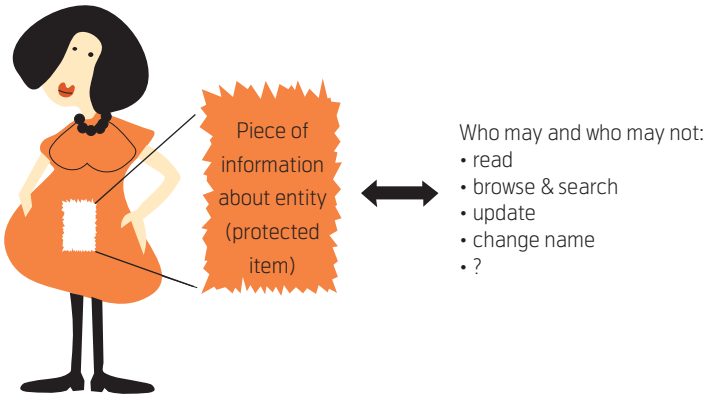


Figure 6 Access control

Authentication requires the use of hashing and asymmetric encryption. Hashing is the transformation of a message into a usually shorter fixed-length value string that represents the original string using some kind of algorithm. The algorithm must have the characteristic that it is virtually impossible to create a message resulting in a given hash value. A hash value will typically change considerably if just one bit is changed in the original message. This allows for message integrity. If the hash value is attached to the message when transmitted, the receiver can create its own hash value and compare it to the one attached. If different, the message has been changed and should be discarded.

Asymmetric encryption requires the use of an encryption key pair consisting of a private key and a corresponding public key. A message encrypted using one of these keys can only be decrypted using the other key. The owner of the key pair is in the position of the private key. Copies of the public key may be distributed to several parties. A message encrypted by a public key can only be decrypted by the holder of the private key. This can be used, for example, to encrypt e-mails sent to the holder of the private key.

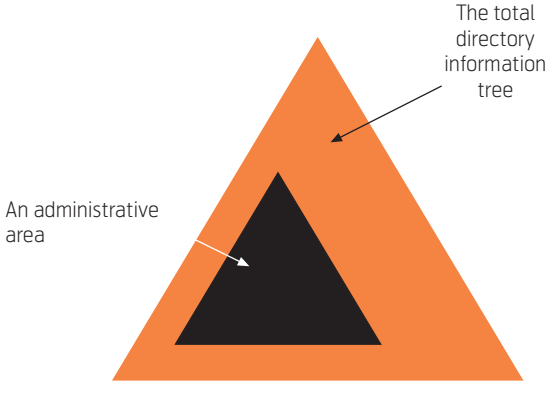


Figure 7 Administrative areas

ISSN 0085-7130 © Telenor ASA 2008

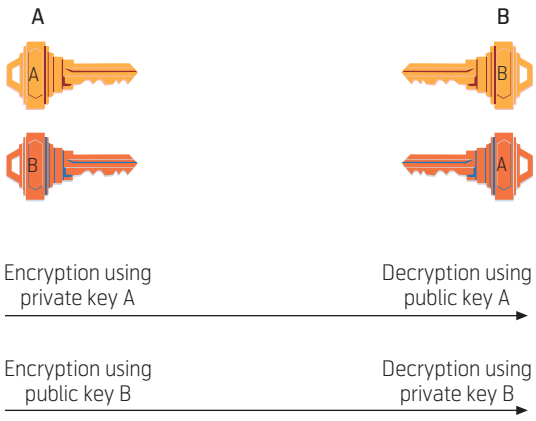


Figure 8 Asymmetric key encryption

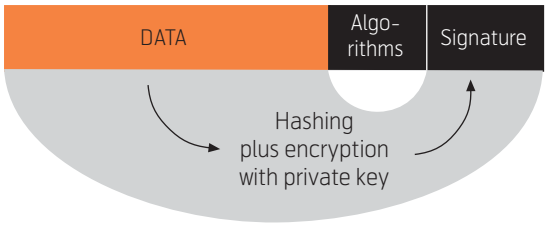


Figure 9 Digital signature

On the other hand, a message encrypted by the private key can be decrypted by anyone holding a copy of the public key. If decryption is possible, only the holder of the private key could have sent this message. This technique is used to create digital signatures as shown in Figure 9.

When a message is to be digitally signed, a hash of the message is created. The hash is encrypted using the private key and appended to the message as a digital signature. As indicated in Figure 9 also information about the used algorithms for hashing and encryption is attached. The receiver decrypts the signature using the public key. It then creates its own hash of the message. If the two hashes are identical, the receiver knows that the message has been transmitted unchanged and that the sender's identity is known with a high level of certainty. This gives an end-to-end security also in a distributed environment.

This sounds quite easy, but is actually quite complicated. The holder of a public key needs to be reasonably sure about the identity of the holder of the corresponding private key. The binding between a public key and the identity of the holder of the corresponding private key is provided in a so-called *public-key certificate*. For it to be trustworthy, a *Certification Authority* (CA) needs to verify the content of the public-key certificate and sign it. If the public key of the CA is known and the CA is trustworthy, we should be safe. However, there are further complications. The owner of a private key must protect it. If somebody else gets hold of the private key, it can be misused. If suspension of something like that has happened, the certificate must be revoked and put on a revocation list to be accessed. As a further complication, there may be many CAs out there. All this requires the establishment of a Public Key Infrastructure (PKI). X.500 or LDAP directories are typically applied to establish a PKI.

An *attribute certificate* is a way to bind a privilege to a specific object. Such an attribute certificate is signed by an Attribute Authority. As for public-key certificates, attribute certificates require the establishment of a Privilege Management Infrastructure *(PMI)*.

X.500 products

Several vendors have implemented the X.500 standard. A list of X.500 vendors and short product descriptions may be found at http://www.x500standard.com/. For each vendor there is a link to more detailed product information. Many vendors refer to their systems as Information Management systems.

Summary and Conclusions

X.500 is a very comprehensive directory standard having excellent support for the requirements of our time. It is open-ended and extensible. It is under continuous development to cope with future challenges. When it comes to features, distribution, replication and security, it is the only game in town.

Erik Andersen holds a Master of Science in Electronics and Physics from the Danish Technical University. He has worked for IBM for 27 years as an expert on data communications and software architecture. He was IBM's representative in data communication standardisation work, including X.500. Since 1995 he has been the owner of Erik Andersen's L-Service consultancy company with continued participation in the X.500 standardisation work. He is now the main project editor for X.500 and the ITU-T Rapporteur for the Directory question. For the Association for the Directory Information Industry (EIDQ) he has developed several ITU-T Recommendations for the Directory Assistance area. He has been member and chair-person of European workshops on Directory.

email: era@x500.eu

Terms and Acronyms in Status

Acronym/ Term	Definition	Explanation	Web Resources
AAP	Alternative Approval Procedure	A permanent procedure agreed by ITU RA in 2003 in Resolution ITU-R 45-1: "Application of an alternative approval procedure (AAP) for Recommendations" and enables Sector Members to be consulted and take part in the approval procedure for ITU-R recommendations. This procedure is a "fast track procedure".	
BR	Radiocommunication Bureau	The ITU Radiocommunication Bureau organises and co-ordinates the work of the Radiocommunication Sector (ITU-R).	
CCV	Coordination committee for vocabulary	Committee under ITU-R responsible for coordination and approval in close collaboration with the Radiocommunication Study Groups, the General Secretariat (Conferences and Publications Department) and other interested organizations concerning vocabulary, including abbreviations and initials, and related subjects (quantities and units, graphical and letter symbols).	http://www.itu.int/ITU-R/index.asp?category=study-groups&rlink=rccv&lang=en
CPM	Conference Preparatory Meeting	Function set up to assist ITU-R in its work and prepare for the World Radiocommunication Conference.	
IMT-2000	International Mobile Telecommunications 2000	The global standard for third generation (3G) wireless communications, defined by a set of interdependent ITU Recommendations. IMT-2000 provides a framework for worldwide wireless access by linking the diverse systems of terrestrial and/or satellite based networks. It will exploit the potential synergy between digital mobile telecommunications technologies and systems for fixed and mobile wireless access systems.	http://www.itu.int/home/imt.html
ITU	International Telecommunication Union	On 17 May 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.	http://www.itu.int
ITU-R	International Telecommunication Union Radiocommunication Sector	A sector of the ITU whose mission is, inter alia, to ensure rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including those using satellite orbits, and to carry out studies and adopt recommendations on radiocommunication matters. It was created on 1 March 1993, replacing the former International Radio Consultative Committee (CCIR).	http://www.itu.int/ITU-R
ITU-T	International Telecommunication Union – Standardization Sector	A sector of the ITU, which mission is to ensure an efficient and on-time production of standards (Recommendations) covering all fields of telecommunications. It was created on 1 March 1993, replacing the former International Telegraph and Telephone Consultative Committee (CCITT).	http://www.itu.int/ITU-T/
PP-02	Plenipotentiary Conference 2002, Marrakesh	ITU's Plenipotentiary Conference – PP – is the top policy-making body of the ITU and meets every four years. The PP is the key event at which ITU Member States decide on the future role of the organisation, drawing up a strategic plan and deciding on the budget. The PP is an intergovernmental conference where only sovereign Member States of the ITU have the right to send delegations. Each Member State has one vote. A number of international organisations and Sector Members may attend the PP as observers. The PP-02 was hosted by Morocco and was held in Marrakesh from 23 September till 18 October 2002. See also report in *Telektronikk* 99 (1), 2003.	http://www.itu.int/plenipotentiary/index.html, http://www.itu.int/plenipotentiary/2002/PP-02.html, http://www.telenor.com/telektronikk/volumes/pdf/1.2003/Page_138-153.pdf

ISSN 0085-7130

Acronym/ Term	Definition	Explanation	Web Resources
PSAA	Procedure for Simultaneous Adoption and Approval	"Fast track" procedure for adopting and approving ITU-R Draft recommendations. See also TAP.	
RA	Radiocommunication Assembly	Radiocommunication Assemblies (RA) are responsible for the structure, programme and approval of radiocommunication studies. They are normally convened every three or four years and may be associated in time and place with World Radiocommunication Conferences (WRCs). The Assemblies assign conference preparatory work and other questions to the Study Groups; respond to other requests from ITU conferences; suggest suitable topics for the agenda of future WRCs; approve and issue ITU-R Recommendations and ITU-R Questions developed by the Study Groups; set the programme for Study Groups, and disband or establish Study Groups according to need.	http://www.itu.int/ITU-R/index.asp?category=conferences&rlink=ra&lang=en
RAG	Radiocommunication Advisory Group	According to the ITU Constitution and Convention, the Radiocommunication Advisory Group (RAG) is tasked to review the priorities and strategies adopted in the Sector; monitor progress of the work of the Study Groups; provide guidance for the work of the Study Groups; recommend measures to foster cooperation and coordination with other organizations and with the other ITU Sectors. The RAG provides advice on these matters to the Director of the Radiocommunication Bureau. Radiocommunication Assemblies may refer specific matters within its competence to RAG.	http://www.itu.int/ITU-R/index.asp?category=conferences&rlink=rag&lang=en
RRC	Regional Radiocommunication Conference	Regional Radiocommunication Conferences (RRC) are conferences of either an ITU Region or a group of countries with a mandate to develop an agreement concerning a particular radiocommunication service or frequency band. Such conferences cannot modify the Radio Regulations, unless approved by a WRC, and the Final Acts of the conference are only binding on those countries that are party to the agreement.	http://www.itu.int/ITU-R/conferences/rrc/index.html
SG	Study Group	Groups with specialists, from telecommunication organizations and administrations throughout the world concerned with drafting Technical bases for Radiocommunication Conferences; developing Draft Recommendations and compiling Handbooks. The ITU-R has currently defined six Study Groups.	http://www.itu.int/ITU-R/index.asp?category=study-groups&rlink=rsg&lang=en
TAP	Traditional Approval Procedure	A written approval procedure of the ITU-R based on a two-step process whereby the Study Group concerned adopts the draft Recommendation and the Member States approve the Draft Recommendation by written consultation.	
WRC	World Radiocommunication Conference	ITU-R's World Radiocommunication Conference (WRC) is an intergovernmental conference where ITU's Member States participate. The WRC meets every three to four years and reviews and revises the Radio Regulations (RR), the international treaty governing the use of the radio frequency spectrum and the geo-stationary satellite and non-geo-stationary satellite orbits.	http://www.itu.int/ITU-R/index.asp?category=conferences&rlink=wrc&lang=en