Project Number 001907

DELIS

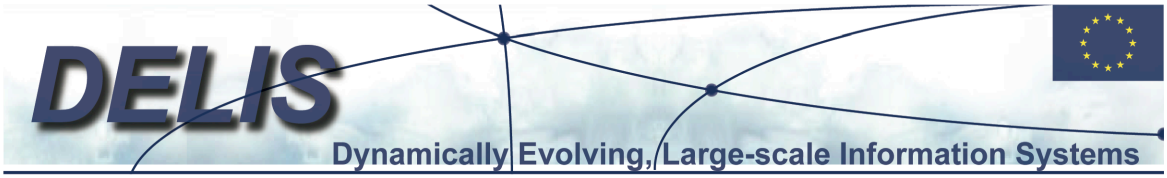
Dynamically Evolving, Large-scale Information Systems

Integrated Project

Member of the FET Proactive Initiative **Complex Systems**

Deliverable D5.5.1

Identifying and promoting industrial applications and knowledge transfer



Start date of the project:	January 2004
Duration:	48 months
Project Coordinator:	Prof. Dr. math. Friedhelm Meyer auf der Heide Heinz Nixdorf Institute, University of Paderborn, Germany
Due date of deliverable:	December 2006
Actual submission date:	January 2007
Dissemination level:	PU – public
Work Package 5.5:	Identifying and promoting industrial applications and knowledge transfer
Participants:	Telenor AS, Norway Universita di Bologna (UniBO), Italy Universitá Pompeu Fabra (UPF), Barcelona, Spain
Authors of deliverable:	Geoffrey Canright (geoffrey.canright@telenor.com) Kenth Engo-Monsen (kenth.engo-monsen@telenor.com) David Hales (dave@davidhales.com) Ricard V. Solé (ricard.sole@upf.edu) Sergi Valverde (svalverde@imim.es)

1 Introduction

This document discusses the potential applicability of a number of research directions which are being pursued in Subproject 5 (SP5) of DELIS. These research themes, and our ideas for how they may be applied in a commercial or non-commercial setting, will be discussed in separate sections. In this Introduction, in contrast, we wish to discuss some more general problems associated with the aim of taking research out into practical application. Our point of view will be highly personal, since a thorough, general discussion is beyond the scope of this Deliverable (and also of the authors). Hence we will start with what we know best, namely, our own experience with regard to the complex problem of seeking and developing useful applications for good research.

We will however attempt to extract some general themes from our own experience. We are all aware of a tension between the poles of "pure research" on the one hand and "bottom-line business" on the other. Furthermore, we view the EU IST (Information Society Technologies) program as seeking to build bridges between these two poles. Hence our discussion and focus here are, we believe, highly relevant for the goals of the EU IST program. We speak here as a group of committed researchers who believe both in the value of research and in the IST goal of rendering the achieved research results useful.

1.1 Experiences with exploitation

Telenor is the only industrial partner in SP5, and, not surprisingly, the partner with the most extensive experience in commercial exploitation of research. Here we summarize the activities and experiences of two of us, Geoff Canright (GSC) and Kenth Engø-Monsen (KEM). We note that our discussion will not be limited solely to experience with SP5-related research, as we see no reason for such a limitation. Our aim instead is to give a more comprehensive overview.

- *Patents*: Telenor actively supports the generation of patents from research at Telenor R&I. KEM and GSC have between them 2 patents granted, and 7 patents pending. These patents are filed in Norway, in the US, and in the PCT (Patent Cooperation Treaty) process. Almost all of these patents are built upon work by the two of us, on problems related in some way to network analysis. In particular, several are in the area of link analysis as applied to the ranking of documents, and others are relevant for social networks (and related), and to information spreading on the same.

  It is clear to us, from our own experience, that the patenting process requires a quite large input of resources, on the part of ourselves, Telenor, and the larger society. The writing and submission processes, for one patent, are roughly equivalent in time demand to that required to write and publish a paper. In addition there is the highly elaborate legal machinery for testing, challenging, and protecting IPR; this machinery mostly lacks a counterpart in the world of research publishing.

  On the positive side, we find that writing patents is often very useful for stimulating ideas. Of course, all researchers have experienced similar effects when writing a research paper: formulating the ideas in a coherent written form is a challenge which tends to illuminate weaknesses in the writers' thought processes. However, we would say that the writing of a patent application stimulates thinking in a different way, precisely because of the focus on applicability. An invention should solve a technical problem in a novel way, which is plausibly or demonstrably better (in at least some circumstances) than existing solutions. These requirements are not pertinent for a research paper, and they tend to stimulate new ideas, even after one thinks that an invention has come to the stage where it can be written up.

- *Applied projects*: Some of our ideas are now being actively evaluated for use in the daily operation of the Telenor concern. While we do not feel that it is appropriate here give a

detailed description of these projects, we want to emphasize that the fact of doing research for a large telecommunications firm can be both motivating and intellectually stimulating. That is, once again we are continually challenged to "solve a technical problem in a novel way". Furthermore, in the Telenor environment, meeting this challenge is just one step in a process, which can often lead to internal implementation of the novel solution. We have some limited experience in this direction—again, we are at the testing stage. In short, our point here is simply that working with this kind of challenge on a daily basis is something that we experience as a quite positive influence on our work as researchers.

- *Commercialization*: Our ideas on link analysis are not suitable for internal implementation at Telenor. Instead, the decision has been made to spin off a small company. The aim of the company is to seek to commercially exploit the IPR represented by a set of patents which are relevant to search and ranking. This company was formally founded in May of 2006. Both of us (KEM and GSC) are active in working with the company.

  Since the company is quite new, our experience is again very limited. However, we would already say that working in this situation gives us valuable and unique feedback on our ideas. That is, we find ourselves very close to the "bottom-line business" pole of the continuum; and yet, at the same time, that fact that the company is based on innovative technical ideas means that the research element, with its intellectual and creative challenges, is never out of the picture.

1.2 Barriers and opportunities

Now we want to offer some more general comments on things which we perceive as standing in the way of effective exploitation of research, as well as things which facilitate such exploitation. These comments are of course in some sense distilled from our experience; but we seek here to speak more generally.

We begin with the academic world. Successful application of one's research is, in many academic environments, not highly rewarded. Instead, a successful researcher is one who has earned the esteem of his peers. Already one can see that this criterion can quickly lead to a rather closed and even fragmented world, in that the opinions of "outside" persons are not important. Furthermore, this kind of peer influence can extend so far as to resist or "punish" work which is viewed as outside the boundaries of the accepted—for instance, precisely because it is "applied" and therefore subject to inlfuences by the non-initiated. In short, we simply are pointing out the well known prejudice *against* applications which is found, to varying degrees, in many academic environments. In its milder form, this prejudice can simply amount to a strong pressure to do work which impresses the right peers—with the pressure so great that the researcher finds no remaining resources for anything else, such as applied work.

The counterpart of this kind of prejudice is of course found in the business world. Here we restrict the discussion to commercial enterprises which support research as an explicit item in their budget (since companies that do not include any research in their activities are basically irrelevant to this discussion). Thus we consider a company which includes a research unit, and which, naturally, seeks to get the best return from this expense. The danger here is that truly novel ideas may not be supported. Novel ideas involve, after all, a high degree of uncertainty in terms of whether, and by how much, they will ever give a financial return on the investment required to generate and develop them. More specifically: if units whose "purpose in life" (ie, principal performance criterion) is a good bottom line are involved in supporting, steering, and evaluating research, they will normally exert an extremely conservative pressure on that research, pushing it towards very short-term work whose likely outcomes can be bounded before the work is even begun. In other words, they will run the research as one runs a business. Even work whose motivation is explicitly towards an important

application may not be supported in such a conservative environment, if that work is too far from the incremental and predictable.

Finally we want to comment briefly on the open-source phenomenon. We feel that open source fits neither of the above (somewhat stereotypical) pictures. Open source software is not research (although it may be the result of research)—it is instead a product, ie, an artifact which is built to solve a technical problem. However, this product is produced for free. Often the labor involved is unpaid; and in any case the resulting product is not in itself saleable for profit. There is a well known sort of peer pressure in the open source community; but one is evaluated (by one's peers) in terms of *performance*, ie in terms of solving practical problems. Also, since the peers need not have paid positions, the barrier to joining the community is much lower than that for joining academic research.

The aspect of being not for profit has however a downside, since most programmers need salaries. Thus the entire open source phenomenon is dependent upon some segments of society—individual programmers, research institutes, or companies—being in a position to (and willing to) give away products of their labor. Furthermore, the focus on products (code) means that open source work (again) is not research. This distinction is not just semantic hair splitting. For example, the kinds of research directions which we discuss in the remainder of this document may (in some cases) be implemented in code, which in turn may (or may not) be open source. But it is clear that the open source community—a community of programmers—does not of itself produce *ideas*, of the kind that exemplify good research. Instead, the open source community produces *implementations* of ideas.

Thus, after discussing these three (somewhat idealized) communities, we retain two as important sources of innovative research: the academic and the industrial. We then ask, how can the barriers to good exploitation of good research be lowered, or even replaced with facilitation?

We can cite one example from the example of Telenor. Telenor has, at the very top level of management, made an explicit and real commitment to the closely related concepts of research and innovation. (Hence the new name: "Telenor R&I".) Furthermore, there is an awareness that innovation cannot be run precisely like a business. Therefore, the R&I unit is positioned so as not to be directly under any of the business units: it reports directly to the top management of the company. In fact, we (KEM/GSC) feel that research at Telenor has, to some extent and at some times, been dominated by the kind of short-term bottom-line approach described above, but that the company is now consciously and vigorously taking steps to provide the kind of support that is needed for genuine innovation. Based on our discussion of the business culture above (and on our experience), we feel that this can only be done if the governance of the R&I unit is uncoupled, to a significant extent, from the priorities of those business units who struggle to survive on a daily basis. This decoupling has occurred, and we view it as a sign that Telenor has recognized (as telecoms should!) both the need *for* and the needs *of* innovation.

2 Epidemic spreading

2.1 Overview

Telenor has worked on understanding epidemic spreading on networks. Most of this work [7, 8, 9, 10] has been based on spreadng over *symmetric* networks—that is, networks for which the probability for spreading between two nodes A and B is the same in each direction ($A \to B$ and $B \to A$). Some preliminary ideas about extending the Telenor approach to the asymmetric case are given in [10].

The basic idea is to measure 'well connectedness' of nodes, and then to equate spreading power with well connectedness. More precise mathematical definitions are given in [8]. Furthermore, the Telenor analysis breaks down the network into *regions*. We find that spreading can be understood in terms of regions, in the sense that spreading within a region is fast and fairly predictable, while spreading between regions is slower and hard to predict.

These ideas have obvious implications for design, modification, and protection of networks, with regard to spreading. That is, for a given (static) network, the regions analysis leads to clear ideas [9, 10] for how to modify the network—towards the goal of hindering spreading via inoculation, *or* towards the goal of enhancing spreading. The latter case also makes sense, because the analysis applies, not just for diseases, but for any proliferative spreading process: for example, gossip, innovation, or information.

2.2 Applications

Now we discuss a number of possible applications for the regions analysis approach. In each case we will give a rough assessment of the potential for application. In this Deliverable, we will include for-profit, not-for-profit, and even free applications. In short, we will consider any application which takes the ideas beyond research and to the point of being used.

2.2.1 Innovation spreading.

We have said above that the regions analysis applies to innovation spreading [12]. Clearly, there are commercial actors who have a strong interest in *accelerating* the spread of an innovation, for which these actors receive income. Spreading in this context is also known as *viral marketing*—"viral" because it depends on the network to do the spreading (as with a disease). In other words, the idea of deliberately enhancing network effects in marketing is not new—but it is then all the more clear that there is great interest in any novel approaches to this kind of (viral) marketing. The regions analysis may be used to this purpose, with the caveat that one must be able to map out (or at least estimate) the network topology.

This requirement is far from trivial. In fact, further research is needed to determine to what extent it is possible to estimate the topology of network of potential customers in various circumstances.

Supposing however that such mapping *is* possible, we believe that this application of regions analysis has both commercial and noncommercial potential. Most marketing activity is commercial; but there is also considerable activity in the form of noncommercial marketing (spreading). That is, the innovation or product being marketed may be free; or it may cost money, but support a not-for-profit enterprise.

In any case, we regard the ideas as being sufficiently technical that they require a significant investment of learning, software, and possibly hardware, up front. This means that likely users of technologies based on this approach are either marketing firms, or very large firms.

2.2.2 Information flow.

Now we let the quantity which spreads over the network be *information*. The obvious connection is that there are many organizations which are interested in "improving" internal information flow. We use quotation marks here because the word "improvement" may mean different things at different times and in different circumstances. However, *understanding* information flow is clearly of value, regardless of the type of improvement which is desired.

We think a promising scenario is as follows: a consulting firm C performs an analysis of information flow in an organization O. This give C a map of the communications network of O—with weights on the links which give a measure of the amount of flow. The consulting firm C can then perform a regions analysis of the network, and from this, make recommendations for improvement. The improvement, we imagine, would typically be of the form of *better* (ie, more efficient) spreading. In such cases, the regions analysis offers clear guidelines as to where new links (or stronger links) should be placed. We can also imagine cases in which it is desired to *limit* information flow—say, from one department to one or more others. Again the regions analysis will be of use in such cases.

This application has a clear commercial slant—as is implicit in our scenario, where a consulting firm is paid to do the analysis.

2.2.3 Social networks.

This application is related to the previous one—which also involves a form of social network. Here we consider mainly *online* social networks. These networks have the advantageous property that they are readily mapped out and hence analyzed. Thus an application of centrality and regions analysis is very straightforward for such networks.

Next we note that members of such networks are often very motivated (if not fascinated) by their *status* and/or *place* in the network. The regions analysis gives information of both kinds: eigenvector centrality is the social-network analog of PageRank, and thus gives a measure of status; also, perhaps many members would be fascinated to know in which region they lie, which regions are close, etc. Such information could be conveyed in a manner which respects privacy concerns, just as current online social networks (such as LinkedIn) do: the regions' identities need not include protected personal information. We speculate further that statistical analysis could be used to give a *profile* for each region—that is, a suitable aggregation of the profiles of the region members. Members would thus acquire a further sense of their "identity" in the network, through their belonging to a region with a "personality".

Furthermore, the network service could use the regions analysis to suggest new links to members. Members wanting to know more people in their "neighborhood" (defined by our analysis) could be given suggestions—as could members seeking new contacts from "foreign" regions of the network.

All of these possibilities amount to "value-adding" features of an online social network. Thus, the business models for these applications are essentially the same as those for the online social networks. Different networks have different business models, and the market is in a strong state of testing, evolution, and flux.

We note finally that, if the analysis were covered by a patent, then the owner could license the analysis to any interested online networking firm. This comment of course applies to all applications in this section. However, in this case, we imagine that many social-networking firms are likely to wish to license or otherwise outsource the analysis than to apply it themselves—due, again, to the rather technical nature of the analysis.

2.2.4 Biological diseases.

Here it may seem that we have an obvious application of an understanding of epidemic spreading. The catch is, of course, that the regions analysis is based on a complete and static picture of the network of infecting links. Few, if any, real biological infections allow for the possibility of defining and measuring such a network.

Nevertheless, we can envision some applications. At the 2006 Sunbelt Social Network Analysis conference, two of us met a researcher who had mapped out a limited but detailed network of human sexual contacts. This network changes on a relatively long time scale. Telenor is currently collaborating with this researcher to analyze the sexual network. The analysis can give the likely course of any STD infection, and also recommendations for most efficiently hindering this spread. We regard this work as on the boundary between research and application. If the results are good, then one might expect broader applications to arise from them.

2.2.5 Electronic viruses.

Here we find a close analogy to the biological case—with the nice difference that many infectible networks of machines are fairly static, and readily mapped out.

There is also another important difference—one that has not been so important in previous subsections. The most common form of virus propagation in computer networks is via email. The virus cannot send itself to addresses which are not found in the infected host's address list. This means that, in many cases, computer A (if infected) can infect B, but B cannot infect A—because B does not have A's address. In other words, the infecting links should be regarded as *directed* for this case [16, 17].

A similar reasoning holds for viruses on mobile phones. Here the phone's list of callable numbers plays the role of the computer's email address list.

The Telenor regions analysis, being based on obtaining the eigenvector centrality (EVC) for the nodes in the network, is only applicable for graphs with symmetric (undirected) links. Thus, for these typical cases of electronic virus, the analysis cannot be applied without making the rather large approximation that the links are symmetric. Furthermore, both the mathematics, and the behavior of the epidemic spreading, are rather different for the case of a directed graph [16, 17].

Telenor has published some preliminary steps [10] towards an extension of the regions analysis to the case of directed graphs. This work is thus clearly still in the research phase, and not ready for application yet.

Summing up, we find that the Telenor regions analysis may be applied to computer networks and virus infection only if one ignores the asymmetry of the links. We believe that this is a nontrivial approximation, which limits the applicability of the approach.

3 Distributed Power Method

3.1 Summary of the research

Telenor and Bologna [14] have developed several forms for finding the dominant eigenvector of a matrix, using a distributed Power Method. The Power Method is very simple in principle: one iterates the operation (matrix)×(vector) many times, until the eigenvector corresponding to the largest eigenvalue comes to dominate (since its growth rate is given by its eigenvalue). Also, at first glance, running the Power Method in distributed form also seems simple, since the operation (matrix)×(vector) only involves "local" operations (in a space whose coordinates are defined by the matrix coordinates).

In fact (as discussed in [13]) there can arise the need for performing "global" operations in the course of a distributed Power Method. First: if the dominant eigenvalue is not 1 (typically it ia greater than one), then repeated (matrix)×(vector) operations give a vector that grows exponentially large. For offline, centralized computations, this is not a problem: one simply rescales the vector periodically (eg, at each iteration). However, finding the length of a vector requires global knowledge—ie, knowledge of all components of the vector.

A second need for global information arises when the matrix is not irreducible. This means that its corresponding graph is not strongly connected—which means, in turn, that there will be many zero elements in the dominant eigenvector [11]. These zero elements are undesirable when the matrix represents a Web graph, and the eigenvector is to be used for scoring and ranking the pages of the Web graph. The PageRank [6] solution is to add a "random surfer" operator, which is simply a complete graph (all-to-all), with some weight ϵ. The random surfer (RS) operator makes the graph strongly connected and so gives nonzero weight everywhere. However, implementing the random surfer operator requires that all nodes (pages) know about all others: again, global information.

This background information will be useful when we discuss the various cases for the distributed Power Method (DPM).

3.1.1 Normalized, undirected graphs

A "normalized" graph is weight conserving, that is, multiplying by the matrix leaves the sum of the weights (ie, the L1-norm of the vector) unchanged. At the same time we know that the dominant eigenvalue is 1. Thus, one avoids (at least, in the noise-fee case) any need to rescale the vector in the DPM. Also, an undirected graph is strongly connected if it is not disconnected. Hence one avoids the need for the RS operator. This makes this case very simple. However, it is only a test case, since the dominant eigenvector can be found analytically [5].

3.1.2 Non-normalized, undirected graphs

This case corresponds to finding the eigenvector centrality (EVC) vector for an undirected graph. Typically, the dominant eigenvector λ_1 is greater than one—and in any case, it is seldom exactly one, so that vector rescaling is needed.

3.1.3 Normalized, directed graphs

This case corresponds to the PageRank approach of Google [6]. Hence one needs the RS operator—and so the distributed form of the Power Method needs a distributed implementation of the RS operator.

3.1.4 Non-normalized, directed graphs

This case has not received much attention. Telenor has studied this case [11] in its non-distributed form, and found novel methods for solving the "sink problem" (that is, the problem of many zeroes in the dominant eigenvector for non-strongly-connected graphs). These novel methods do not however give any escape from the need for global information; we claim in fact that it is impossible in general to solve the sink problem without global information. The Telenor/Bologna work [13, 14] thus implemented only the standard RS solution. Also, since in this case the graph is non-normalized, one needs to rescale the vector periodically. Thus, this case must implement *both* kinds of global operations in order to function. It is, in this sense, the most difficult case.

Now we discuss applications.

3.2 Applications

3.2.1 Peer-to-peer search

The applications of DPM to directed graphs were directly motivated by their possible utility for distributed link analysis and page ranking in distributed, peer-to-peer (P2P) search engines (SP6). Thus we get the two cases for directed graphs: normalized, and non-normalized.

- *Distributed PageRank.* The distributed RS operator worked well, so that this case gave good performance. Also, we were able to prove stable convergence for a broad (but not complete) range of conditions. On the other hand, in collaboration with MPII, we compared their JXP [15] algorithm (which is a quite different approach to distributed PageRank), and found the JXP approach to be even more stable—in fact, it converges monotonically.

  The JXP approach allows each peer to have an arbitrary view (set of known pages), as does the Telenor/Bologna approach. So, in this regard, we find no important difference between the two approaches.

  In sum however the JXP approach has the advantage of greater stability and good smooth convergence, without any apparent disadvantages. Hence we, at least tentatively, favor the JXP approach to distributed PageRank.

- *Distributed T-Rank.* Here the term 'T-Rank' refers to the approach, developed and studied by Telenor, for performing link analysis without normalizing the outlinks of the Web graph. (This term has been used for other things; but it is convenient here.)

  There is no obvious way to generalize the JXP approach to the case $\lambda_1 \neq 1$; and in fact this may be impossible. In any case, our implementation of the DPM for $\lambda_1 > 1$ is, to our knowledge, the only such implementation. Furthermore, we know of no other groups studying the corresponding centralized approach to link analysis. Hence we will examine the applicability of both the centralized and the decentralized versions here.

  The centralized version, while giving results which are distinct from PageRank, appears from simple tests [11] to give results which are at least as useful. Hence we find that T-Rank is worth further investigation as a possible competitor to PageRank for link analysis, as used in search engines for ranking of hits. Telenor is in fact actively investigating the possible commercialization of (centralized) T-Rank.

  The decentralized version [14] is less stable than the $\lambda_1 = 1$ case—not surprisingly, since for $\lambda_1 > 1$ one must also rescale the vector at each iteration. We have not been able to find any proofs of stability for the asynchronous power method with $\lambda_1 > 1$; nor are we aware of any such proofs in other work. Hence this is a rather unexplored case. Even though our results show good convergence in many cases, we would say that the approach is still rather firmly grounded in the research phase.

  If we suppose that good stability is (some day) proven or demonstrably achieved, then we see distributed T-Rank as a viable alternative to distributed PageRank for P2P search services. We doubt that distributed T-Rank will ever show the same degree of stability as that 1achieved by JXP. However, it is possible that the non-normalized approach may reveal advantages, other than strong stability, which will render it attractive for both centralized and distributed applications. Furhter research is needed however to test these suppositions.

3.2.2 Self-mapping networks

Now we come to the case of undirected graphs. We ignore the normalized case, since it is solved analytically, and remind the reader that the elements of the resulting eigenvector, for the non-normalized case, give the 'eigenvector centrality' or EVC for each node in the graph.

We have already discussed, in the previous section, a number of interesting applications of the Telenor 'regions' analysis—which is built upon a computation of the EVC for the entire graph. Furthermore, we pointed out a *disadvantage* of the regions analysis, namely, that it relies on global information: a complete map of the network topology.

Our point should now be clear: implementation of a distributed EVC calculation points the way towards removing this need for global information. We say "points the way" for two reasons. First, proof of stability for the $\lambda_1 > 1$ case for undirected graphs is just as elusive ss it is for directed graphs. We can only prove stability if one knows λ_1 in advance—itself a nontrivial task, involving global knowledge.

Secondly, the regions analysis requires more than just the EVC. In addition, one must compute something which may be called the 'steepest-ascent graph' or SAG. We find however (unpublished) that computing the SAG, and using it to find the regions, appears to be very straightforward: one requires only a finite number of steps, using only local exchanges, to obtain exact convergence.

Thus we again assume that stability of a distributed EVC calculation is 'demonstrably achieved'—at least for a range of cases of practical interest. This gives the result that any network of symmetric links can perform regions analysis 'on itself': the nodes can find their own eigenvector centrality, their region, the Center of their region , etc. What then are the practical implications of this?

To answer this question we revisit the applications which we discussed in section 2. In each case we attempt to evaluate the potential of allowing the network to 'map itself'.

- *Innovation spreading.* For this case (viral marketing), we believe that there is no good mechanism for getting the network (of potential customers) to map itself. Hence we see little promise for the self-mapping approach here.

- *Information flow.* In many organizations, a great deal—although certainly not all—information flow is mediated electronically. This suggests that one can 'piggyback' a low-overhead, distributed EVC+SAG calculation on top of standard elecronic communication. The result would be indeed that the network 'maps itself'—and furthermore does so in a way which is continually self-updated.

  A further step is to allow the network to generate 'its own' suggestions for *improving* the network topology, towards the goal of improved information flow. We note that, somewhere in this process of increasing decentralization, the need for a 'human in the loop will arise': not because the entire application cannot be decentralized, but because one presumably wants to inject some human judgement before implementing changes in the information flow.

  In section 2 we envision a consulting firm C carrying out the regions analysis for an organization. Here, we can imagine that C simply downloads the application into the organization's intranet. The role of C thereafter appeasr to be small, as long as the application works as planned; management in O can read the steadily updated results, and approve or not any proposed changes in network connections.

- *Social networks.* If the social network is online, and in addition managed by a central provider (Orkut, LinkedIn, MySPace, etc), then the practical distinctions between centralized and decentralized regions analysis seem small.

  If the social network is online, but not so managed—for example, a P2P network where the members require only shared downloaded protocols to participate—then, as in the previous case, we see the possibility of piggybacking a network mapping application on top of normal communications. A possible problem here is that many such networks make no provision to ensure two-way connections; hence the distributed application of a regions analysis based on EVC would not be possible.

  If the social network is 'offline' (relative to the Internet), but still connected electronically (eg, by telephones), then we believe that such a distributed mapping operation is possible. The principle problem here is to define network membership. The network of telephone users, for example, spans the globe.

- *Biological diseases.* Given that biological transmission channels are never (to our knowledge!) electronic, we see no application for the self-mapping idea here.

- *Electronic viruses.* Here we avoid the preceding (biological) objection. However, in most cases (as discussed in section 2), the network which supports spreading is directed—built up from one-way links. Furthermore, it is not possible for the distributed approach—which actually uses the existing links to perform the calculation—to make the approximation that the links are symmetric.

  Hence, we believe that an application of the self-mapping approach to this problem must await a generalization of the regions analysis to directed graphs. This generalization must include a well-motivated approach to the 'sink problem' (see [11] for some ideas in this regard). Hence we see the distributed approach to the computer-virus (and related) problems as lying clearly in the research phase. The problem is clearly of great interest, and hence worth more attention;

after all, a network that can map itself, analyze its own susceptibilty to infection, and generate its own initiatives for improving resistance, would be a wonderful thing. It would of course also have to protect itself from 'smart' viruses that find ways to compromise the self-mapping or self-protecting functions ... so we reserve this dream for future work.

4 Open-source network structure

The analysis of e-mails exchanged during software development reveals that successful Open Source Communities require a core of stable members (i.e., they are centralized instead of self-organized). These methods can be extended to advanced management of open source teams. For example, we can detect "hot-spot" members without the need to deal with any member of the community.

5 Motifs and software graphs

5.1 Complex Software Queries.

We have conjectured that source code search engines (one of these engines is found at http://www.koders.com/) can be greatly enhanced by the so-called "motif analysis" of software networks. These search engines enable the programmer to find all the source code files containing a given keyword (i.e., the name of a software component like a class or a subroutine). However, this is a rather limited way to surf large source code databases. Instead we propose more advanced code searches by specifying a complex query pattern defined by a small set of software components and their relationships (a motif). Given this information, the search engine will return all the source code files where the previous pattern was instantiated.

5.2 Reverse Engineering.

The previous technique analyzes a single snapshot of a given software system. However, complementary information can be obtained if we analyze many different versions of a software system, simultaneously. In particular, CVS repositories provide an invaluable window into processes of software evolution. Our methods enable us to detect what subsets of source code files are more likely change together. We envisage a useful and novel method for clustering software based on the analysis of development activity.

6 Cooperation and trust in P2P networks

Several simulation models have been developed within SP5 and SP4 that apply a novel socio-inspired approach for promoting cooperation and coordination in a robust and distributed way [1, 2]. These models show that there can emerge a form of distributed trust between nodes, and that this trust is robust to certain kinds of selfish and cheating behavior. Interestingly, the mechanism is based on a simple node level protocol which updates (re-wires) neighbor links within the overlay network based on local performance measures (a utility value). This produces a dynamic method of supporting trust between neighbor nodes without central control which is highly scalable (up to millions of nodes) and robust to churn - where nodes dynamically join and leave the network. There are a number of possible exploitable applications for this kind of technology which we summarize in the follow sections. Finally we will conclude with a discussion of open issues and potential for progress towards these applications.

6.1 Cooperative Resource Replication

Web based content providers need to asses demand for particular resources (say media items) and allocate appropriate bandwidth and server space. However, demand is highly dynamic and may change quickly or be hard to predict. Given this it would useful for a set of servers to dynamically coordinate content provision cooperatively. A distributed dynamic approach in which server nodes cooperate to achieve this could be valuable since latent resources can be utilized. A distributed trust mechanism would allow for multiple administrative entities (different organizations) to pool their server resources such that each benefits without any individual organization exploiting the system. This could be a simple alternative to more complex and centralized trading mechanism which generally require centralized and secure trusted authorities.

6.2 Spam and Malware Protection

Most current approaches to Spam and Malware (Spyware, Adware, Viruses etc.) prevention rely on a single trusted authority. Either the individual node develops a private filter from experience (adaptive spam filters in e-mail applications for example) or a central global database is queried at regular intervals. Such central solutions often require a financial subscription, while the individual approach requires each node to rediscover the same threats. A fully distributed system that shares information could potentially offer a more robust and faster system. A network of trusted nodes is a prerequisite for distributed approaches to Spam and Malware protection. Essentially, if a node can rely on its neighbors to advise it concerning a "blacklist" of items identified as risky by others, then a highly robust distributed collective "filter" can be produced. Although it has been suggested that existing human friendship relations could be used, this is only a partial solution—because, although a friend may be trusted, their node may become infected or hijacked by malicious code or individuals. By utilizing the distributed trust mechanism we have developed it could be possible to automatically produce and maintain such a network automatically. We have argued this elsewhere and developed a protocol variant with desirable properties for this task [2].

6.3 Cooperative Broadcasting

Recently we have produced a further variant of our socio-inspired approach for self-organising, robust and efficient broadcasting protocols [3]. Broadcasting involves one node distributing a message to the entire network via message passing. For all nodes to receive the message, intermediate nodes need to cooperatively pass the messag—even when they may have an incentive to not do so. We believe we have identified a novel approach which reduces the number of messages required to be passed over a simple approach in which all nodes pass to all other nodes (the so-called flood-fill approach). This approach exploits an emergent process which has direct links to percolation theory developed within physics. Interestingly, we believe that, the mechanism may be general and could have implications for percolation theory within physics itself.

6.4 Looking forward

We believe that we have identified a number of promising applications for our approach to trust and cooperation in P2P networks. However, there are still open issues and further work required. Currently our protocols only exist as simulation models (implemented within the PeerSim system). Since we don't have proofs of the protocols, only with full implementation could their effectiveness be tested. Another aspect, especially related to Malware protection, is the issue of security. These protocols are "open" in the sense that they are intended to be deployed in environments where a node can join the network without identity checks or central administrative control. This leaves them

open to certain kinds of malicious attacks. We have evaluated performance under certain kinds of attacks but this aspect requires further work.

7 Summary

This deliverable is a collaborative effort of the three partners in SP5. We have not attempted to force the different contributions to adhere to a common style. Instead, we believe that it is useful to allow each participating partner to employ its own preferred style. We feel that the result makes the document more interesting and illuminating, expressing as it does the above-discussed tensions between two cultures: the academic culture and the business culture. Telenor has clearly (and not surprisingly) played a rather dominant role in this document; nevertheless, we feel that all the partners' inputs represent worthwhile contributions to the Deliverable and towards its goals.

We will not attempt here to pluck out the "most promising" applications from the above discussion. The reason is that we regard this entire document as a sort of summary, and furthermore, a summary which includes a strong speculative component. Therefore we would rather allow the above sections, with their summaries and speculations, to speak for themselves. Put another way: *time* is needed to pluck out the most promising applications; and we will not attempt here any shortcut for that process.

We believe that writing this Deliverable has been a useful exercise. We support the IST goal of building bridges between research and industry, and view this document as a modest, but consciously designed, contribution towards this goal. We hope that this document will stimulate further thoughts and ideas, perhaps along new lines, in many readers. In any case, we can confirm that the writing of this document has indeed helped to stimulate our own thinking about the promise and likelihood of future applications of our own work.

References

[1] Hales, D. (2006) Emergent Group-Level Selection in a Peer-to-Peer Network. Complexus 2006;3.

[2] Hales, D. and Arteconi, S. (2006) SLACER: A Self-Organizing Protocol for Coordination in P2P Networks. IEEE Intelligent Systems, 21(2):29-35.

[3] Arteconi, S. and Hales, D. (2006) Broadcasting at the Critical Threshold. University of Bologna, Dept. of Computer Science, Technical Report UBLCS-2006-22.

[4] Arteconi, S. and Hales, D. (2005) Greedy Cheating Liars and the Fools Who Believe Them. University of Bologna, Dept. of Computer Science, Technical Report ULCS-2005-21.

[5] Rajeev Motwani and Prabhakar Raghavan, Randomized Algorithms. Cambridge University Press, Cambridge, UK, 1995, 132.

[6] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, The PageRank Citation Ranking: Bringing Order to the Web. 1998, citeseer.nj.nec.com/page98pagerank.html.

[7] Geoffrey Canright and Kenth Engø-Monsen, Roles in networks. Sci. Comput. Program. 53(2), 2004, 195-214.

[8] Geoffrey Canright and Kenth Engø-Monsen, Spreading on networks: a topographic view. Proceedings, ECCS05, Paris, 2005.

[9] Geoffrey Canright and Kenth Engø-Monsen, Epidemic spreading over networks: a view from neighbourhoods. Telektronikk 101, 65-85 (2005).

[10] Geoffrey S. Canright and Kenth Engø-Monsen, Some relevant aspects of network analysis and graph theory, to appear as a chapter in Handbook of Network and Systems Administration, Jan Bergstra and Mark Burgess (eds), Elsevier, 2007.

[11] M. Burgess, G. Canright, and K. Engø-Monsen, Importance-ranking functions derived from the eigenvectors of directed graphs, submitted for publication.

[12] E. M. Rogers, Diffusion of Innovations, 3rd ed. Free Press, New York (1983).

[13] Márk Jelasity, Geoffrey Canright, and Kenth Engø-Monsen, Efficient and Robust Fully Distributed Power Method with an Application to Link Analysis. DELIS Tehcnical Report DELIS-TR-0320, 2006.

[14] Márk Jelasity, Geoffrey Canright, and Kenth Engø-Monsen, Fully Distributed Timed Asynchronous Power Iteration. DELIS Technical Report DELIS-TR-0326, 2006.

[15] Josiane Xavier Parreira and Gerhard Weikum, JXP: Global Authority Scores in a P2P Network. WebDB (2005), 31-36.

[16] Jeffrey O. Kephart and Steve R. White, Directed-Graph Epidemiological Models of Computer Viruses. Proceedings of the IEEE Computer Symposium on Research in Security and Privacy, pages 343–359, May 1991.

[17] M. E. J. Newman, Stephanie Forrest, and Justin Balthrop, Email networks and the spread of computer viruses. Phys. Rev. E 66, 035101 (2002).