OPAALS PROJECT

Contract n° IST-034824

WP10: Sustainable Community Building

Del10.8 – Report on cross-domain social networks

Project funded by the European Community under the "Information Society Technology" Programme

Contract Number: IST-034824

Project Acronym: OPAALS

Deliverable N°: D10.8

Due date: May 2009

Delivery Date: August 24, 2009

Short Description: This deliverable presents social network analysis from different domains and types, assessing the evolution and emergence of the respective communities. From the knowledge achieved in the parameters of the analysis, new features are suggested for simulation and emulation frameworks to be implemented in EveSIm. Also, in this deliverable ICT group activities are described as well as results from PIIIKO Workshop.

Author: IPTI (Fernando Colugnati, Lia Lopes), SUAS (Thomas Kurz), (LSE) Anne English

Partners contributed: IPTI, TUT, SUAS, LSE

Made available to: OPAALS Consortium and European Commission Versioning

Versioning		
Version	**Date**	**Name, organization**
V1	24/07/2009	Fernando Colugnati (IPTI)
V2	28/07/2009	Fernando Colugnati (IPTI), Thomas Kurz (SUAS)
V3 – First internal review	04/07/2009	Fernando Colugnati (IPTI), Lia Lopes (IPTI), Thomas Kurz (SUAS)
V4 edit	19/08/2009	Anne English (LSE)
V5 final	24/08/2009	Fernando Colugnati (IPTI), Lia Lopes (IPTI), Thomas Kurz (SUAS), Anne English (LSE)

Quality check

Internal Reviewers: Frauke Zeller (UniKassel) and Jaakko Salonen (TUT)

Dependences:

Achievements*	• Knowledge about collaboration process and community building • Use of new method for Social Network Analysis and Simulation based on real analysis data
Work Packages	WP10, WP3
Partners	SUAS, IPTI, TUT, LSE
Domains	Computer Science, Social Science, Social Network Analysis
Targets	All domains; OPAALS researchers.
Publications*	Dynamic social network modeling and perspectives in OPAALS frameworks. 2nd OPAALS Conference, Tampere – FI . Fernando Colugnati Dynamic social network modeling – a Forum use case, International Conference on Applied Social Modeling and Simulation 2009, Paris. Fernando Colugnati Actor based model: a Forum use case – a forum case study, 5th UK Social Network Conference 2009, London. Fernando Colugnati 'Digital Business Ecosystems', Ed. Nachira, Francesco; Nicolai, Andrea; Dini, Paolo et al: www.digital-ecosystems.org/dbe-book-2007. English, Anne with Dory, Bertrand. Chapter: DBE in Ireland: An Irish Open & Connected Digital Ecosystem
PhD Students*	None
Outstanding features*	• Knowledge about collaboration processes in web environments • Migration of Simulation Framework to the OPAALS P2P infrastructure under construction.
Disciplinary domains of authors*	Fernando A.B. Colugnati (Statistics and Social Network Analysis) T. Kurz (Information Technologies, Software and Systems Engineering, Interpersonal Communication) Anne English (Community Building and Collaboration, Knowledge and Project Management) Lia C. R. Lopes (Computer Science and Social Network Analysis)

The information marked with an asterisk (*) is provided in order to address Recommendation n. 4 from the Year 2 review report

Table of Contents

Table of Figures 6
1. Introduction 7
2. Simulation and Emulation model and EvESim collaboration 9
2.1- Integration of Guigoh, Wille Visualisation System and EvESim 12
3. Methodological aspects for longitudinal SNA: metrics and evolution assessment 13
4.1- Topological and modelling aspects in SNA 16
4.2- Dynamics and network autocorrelation 18
5. Web forum use cases 23
5.1- Conversê 23
5.2- SUPERA 27
5.2.1- Stochastic dynamic modelling of the SUPERA forum 29
6. Scientific co-authoring networks by science domains 32
6.1- Mathematics (Maths), Computer Science (CS) and Psychology (Psych) 32
6.2- Sub-areas 34
7. OPAALS Wiki 38
7.1- Stochastic dynamic modelling for the OPAALS wiki 41
8. Some remarks about the use case networks 44
9. OKS – Iterative user-led development 46
9.1- Refinement and prioritisation of user requirements 46
9.2- Testing and Validation 47
10. Phase III Kick-Off (PIIIKO) Workshop 48
11. Concluding Remarks 50
12. References 52

Table of Figures

FIGURE 2.1- MODEL FOR SOCIAL NETWORK TESTING. 9
FIGURE 2.2- EVESIM SCREENSHOT INCLUDING AGENT VARIABLE DROP-DOWN. 11
FIGURE 3.1- THIS NETWORK HAS 1 TRIANGLE AND 16 SIZE 2 PATHS. SO *CC= 6X1/16=3/8* 15
FIGURE 4.1- A CUMULATIVE DEGREE DISTRIBUTION OF A NETWORK THAT FOLLOWS, APPROXIMATELY A POWER-LAW DISTRIBUTION. 17
FIGURE 4.2- A) HIERARCHICAL CLUSTERING PROCESS GOVERNED BY *RICH GETS RICHER EFFECT*, B) HIERARCHICAL TOPOLOGY HAVING THE SAME PATTERN. 18
FIGURE 4.3- PRINCIPLES THAT MAY EXPLAIN NETWORK AUTOCORRELATION CONCEPT 20
FIGURE 4.4- NETWORK IMPORT SCENARIO WITH SNA AND EVESIM EMULATION. 20
FIGURE 4.5- BOOTSTRAP SCENARIO WITH SNA AND EVESIM EMULATION. 22
FIGURE 5.1- SOCIOGRAMS OF CONVERSÊ NETWORK 25
FIGURE 5.2- A) AVERAGE DEGREE EVOLUTION AND B) CUMULATIVE DEGREE DISTRIBUTION FOR CONVERSE DYNAMIC FORUM. 26
FIGURE 5.3 – A) LARGER CLUSTER SIZE, B) NUMBER OF CLUSTERS, ... 26
C) AVERAGE PATH LENGTH AND D) TRANSITIVITY COEFFICIENT EVOLUTIONS FOR CONVERSE DYNAMIC FORUM. 26
FIGURE 5.4 – AVERAGE DEGREE EVOLUTION AND CUMULATIVE DEGREE DISTRIBUTION FOR SUPERA DYNAMIC FORUM. 27
FIGURE 5.5- SOCIOGRAM FOR THE FINAL SUPERA NETWORK 28
FIGURE 5.6 – A) LARGER CLUSTER SIZE, B) NUMBER OF CLUSTERS, C) AVERAGE PATH LENGTH AND D) CLUSTERING COEFFICIENT EVOLUTIONS FOR SUPERA DYNAMIC FORUM. 28
FIGURE 6.1- NUMBER OF PAPERS (A), CUMULATIVE NUMBER OF PAPERS (B), NUMBER OF NEW AUTHORS (C) AND CUMULATIVE NUMBER OF AUTHORS FOR THE MAIN FIELDS (MATHEMATICS, COMPUTER SCIENCE AND PSYCHOLOGY) 33
FIGURE 6.2 – AVERAGE DEGREE EVOLUTION AND CUMULATIVE DEGREE DISTRIBUTION FOR THE THREE MAIN FIELDS 33
FIGURE 6.3 – A) LARGER CLUSTER SIZE, B) NUMBER OF CLUSTERS, C) AVERAGE PATH LENGTH AND D) CLUSTERING COEFFICIENT EVOLUTIONS FOR COAUTHORING NETWORK 34
FIGURE 6.4- CUMULATIVE DEGREE DISTRIBUTIONS FOR SUB-AREAS 35
FIGURE 6.5- SOCIOGRAMS FOR SUB-AREAS. A) APPLIES MATHEMATICS, B) INTERDISCIPLINARY (MATHS), C) ARTIFICIAL INTELLIGENCE, D) INTERDISCIPLINARY (CS), E) PSYCH. TREATMENT AND PREVENTION, F) MULTIDISCIPLINARY (PSYCH). 36
FIGURE 6.6- AVERAGE PATH LENGTH AND TRANSITIVITY COEFFICIENTS EVOLUTION FOR THE SUB-AREAS 37
FIGURE 7.1- A) AVERAGE DEGREE EVOLUTION AND B) CUMULATIVE DEGREE DISTRIBUTION FOR OPAALS WIKI COLLABORATION NETWORK 38
FIGURE 7.2- SOCIOGRAMS FOR THE OPAALS WIKI COLLABORATION EVOLUTION. YELLOW = SOCIAL SCIENCES, GREEN = COMPUTER SCIENCE, RED= NATURAL SCIENCES, BLUE = CONSULTANTS/ADM. 39
FIGURE 7.3- A) LARGER CLUSTER SIZE, B) NUMBER OF CLUSTERS, C) AVERAGE PATH LENGTH AND D) CLUSTERING COEFFICIENT EVOLUTIONS FOR OPAALS WIKI 40

1. Introduction

The notions of "**network**" and "**community**" have become central terms when dealing with contemporary societal trends. In this context OPAALS Network of Excellence (NoE) "aims to work toward an integrated theory of digital ecosystems, to create new disciplines from the merging of old disciplines, and to do so by leveraging community values and principles." (Phase II Workplan, p.5). In this NoE, the comprehension as to how networks behave and communities emerge can be considered a self reflection for the consortium's governance, as well as an important research framework for the Social and Computer Science domains since SME scenarios in the adoption and use of Information and Communication Technologies (ICTs) for collaborative and democratic business is one of the main objectives of Digital Business Ecosystems (DBEs).

ICTs have an effect on coordination, communication and control in all societal networks and communities. OPAALS developed the concept of Open Knowledge Space (OKS) that - as part of a broader definition - is technically a set of computational tools for collaboration and knowledge construction by means of virtual communities. So far there are only a few studies that analyse in detail how virtual knowledge networks grow, how the process of communication and articulation between their nodes (local actors) happens, and the dynamics that are behind those aspects. More research is needed to generate knowledge about applicable strategies and mechanisms that are essential in order to make a newly established network successful and sustainable. This knowledge can be achieved by analysing existing networks and/or simulating new ones or "forecasting" the future topology.

From the analytical point of view, community emergence processes are nowadays understood as complex emergent systems, driving quantitative methods to look at them as physical and dynamic structures. The development of Social Network Analysis (SNA) methods, in the beginning static and purely deterministic, is now moving towards dynamic stochastic processes where the role of the actors are the central point, modelling the network emergence from a bottom-up perspective that will lead to a certain topology. Both, the dynamic process led by actors by means of selection and social influence, and the topology of the network provide interpretable parameters and metrics that quantify the intensity and velocity of the emergence and formation of communities or sub-communities.

This deliverable shows the outcomes of Task 10.11, "Social networks in cross-domain collaboration". The task description states that "... Following well-known network models, it is possible to represent different sustainability models for community networks. ..." and also that "The

outcomes of this task in terms of social network analysis should enable the EveSim to configure and set up network topologies and behaviour that are very close to social networks in real life. This data leads to dynamic simulations in T3.8 and others to test the behaviour of the nodes, services and transactional models on top of this social network data."

So, those outcomes are presented here based on the analysis of data from different collaborative networks, with the modelling of the dynamics and communities detection in the social structure as the main focus. Also, as complementary results from the consortium activities, outcomes of two ICT activities are discussed. One is the set up of a questionnaire for collecting feedback and user experiences for Guigoh (IPTI) and Sironta (TI) and the second is the phase III kick-off event (PIIIKO) in Salzburg.

This deliverable is structured in three main blocks. First, an overview is given on the collaboration of IPTI and SUAS regarding the jointly developed EvESim model (Chapter 2). Chapter 3 and 4 depicts Social Network Analysis concepts and methods that will drive the EveSim model that will be implemented. Chapters 5 to 7 introduce SNA from three different types of collaboration networks, and Chapter 8 presents a comparison of these results focusing on the relevance of those parameters in the simulation and emulation of different types of networks. In that sense, the connections of SNA and the EvESim simulation cases are explained, making cross references to OPAALS infrastructure testing and other components developed in OPAALS. Beside the joint model created during a research exchange of Thomas Kurz at IPTI in February 2009 and previous and ex post online meetings, IPTI was mostly responsible for the SNA related parts, whereas SUAS adapted the EvESim framework following the requirements of the new modeling approach.

The second block brings the idea and the process followed during the creation of the testing questionnaires for Guigoh and Sironta which are summarised in Chapter 9. The quantitative and qualitative results of the questionnaire, which was set up by UniKassel, had already been forwarded to IPTI and TI. This chapter relates mainly to the work of Anne English and Thomas Kurz in the Integration Coordination Team (ICT).

The third block is a short summary of the main findings of the Phase III Kick-Off (PIIIKO) event, June 2009 in Salzburg. The design and setup of the workshop was completely different to the workshops and meetings carried out in OPAALS up to then and there was also an ICT activity of Anne English and Thomas Kurz.

2. Simulation and Emulation model and EvESim collaboration

EvESim is a distributed framework which enables disciplines to collaborate on the basis of a multi-agent simulation. Details about EvESim can be found in several deliverables and publications [1], [2], [3], [4]. The aim of this work is to summarise only the most recent modifications in EvESim and describe the design and setup of the simulation case for Phase III. A detailed description of EvESim can be found especially in [3], [4] and at http://evesim.org.

The current status of EvESim, which can be downloaded at http://sourceforge.net/projects/evesim, was extended to include the following features:

1) Improved build of EvESim using Maven and Ant Build tools

2) EvESim services running on dbe.fh-sbg.ac.at, so that the simulation can run without installing Soapod locally

3) Improved user interface, e.g. selection of simulation cases (DBE and OPAALS simulation cases are available)

4) Improved visualisation capabilities, e.g. Google earth and Google maps with animation

5) Abstract agent implementation for easy adaptation of the framework for non OPAALS simulations

6) Implementation of dynamic agent behaviour for timer events (see Figure 2.2)

7) Direct interface to extract data (nodes and topology) from Guigoh

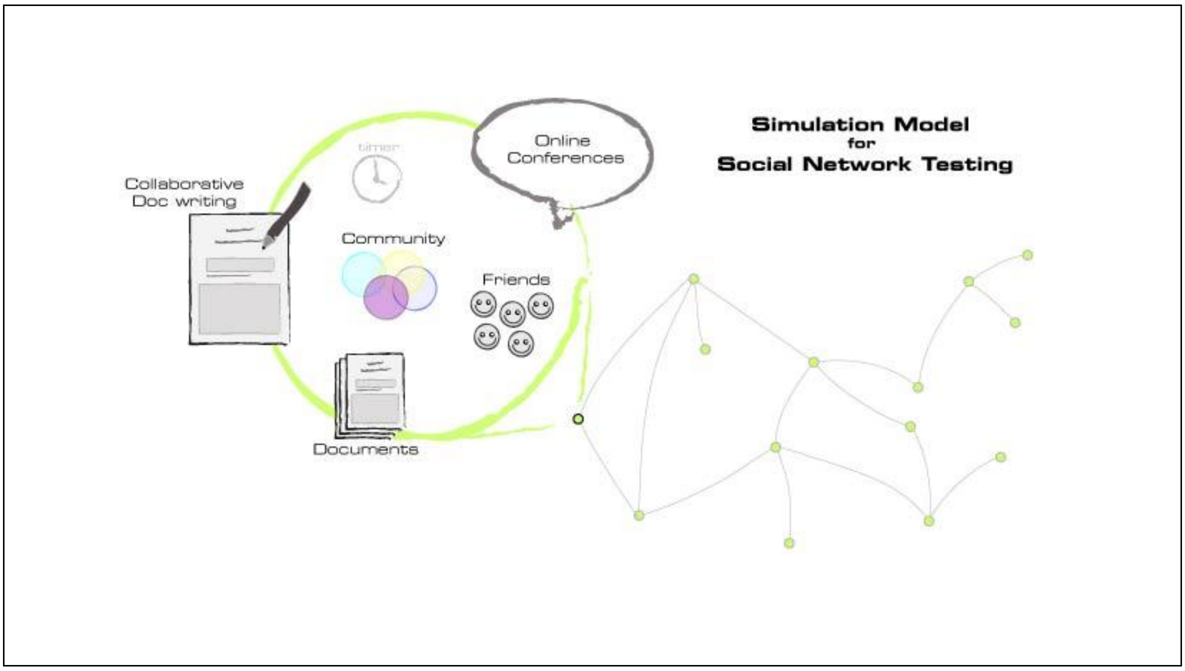


Figure 2.1- Model for Social Network Testing.

Especially item 6), the dynamic agent behaviour, is already targeted directly at the simulation model desired for Phase III of the OPAALS project. In general, the goal is to utilise the outcomes of SNA and make predictions for the evolving network. The selected model to accomplish this proposal is described in the following chapters.

Before the chapters on SNA, an overview on the desired model for implementation is given. Although, alternative scenarios and simulation cases could have been implemented, the following model was decided upon because of its similarity with the OKS architecture and the broad application area for other similar electronically supported social networks and collaboration tools, such as wikis.

As can be seen in Figure 2.1, the simulation model for Social Network (SN) testing consists of a network of independent acting agents. The detailed view of one of these agents shows the basic functionality of a P2P OKS node. In the current merely centralised implementation of OPAALS version of Guigoh, this model of a node can be seen as the profile of one of the users in the social network.

In this sense, each node or user can be connected to a couple of friends, they can also be members of communities and write and upload documents as well as create online conferences/chats with other users in the network. The collaborative document writing tool, which represents one of the first P2P services in OPAALS version of Guigoh, enables users to edit a document simultaneously..

Aside from these structural components, EvESim has implemented a timer component in each agent (in this model representing a user or node). This timer initiates actions at a given step in the simulation. In order to facilitate the changing behaviour of the users over time, the current implementation also includes the possibility to change the frequency of these actions along a given function, such as Linear, Polynomial, Exponential, Constant, Logarithmic (see the configuration drop down in Figure 2.2). This functionality will be extended continuously in order to support the specific structural and behavioural parameter changes of social networks described later in this deliverable. The arrival and departure of nodes will of course be configurable in the GUI.

In the following paragraph, we describe how EvESim bridges between SNA and utilisation for simulation, emulation and testing. The previously described model will be enhanced especially in phase III by the implementation of dynamically changing parameters, coming from SNA, described later in this deliverable. The key parameters need to be configurable by technically un-experienced people through the user interface of EvESim. For the convenience of the user, the user interface was already modified and, for example, the functionality of an editor for the network

descriptions in the UI was added. This enables a simple creation of the underlying XML file, without any knowledge of XML. These implementations enable an easy-to-use Simulation-Framework for users who do not have any knowledge of XML.

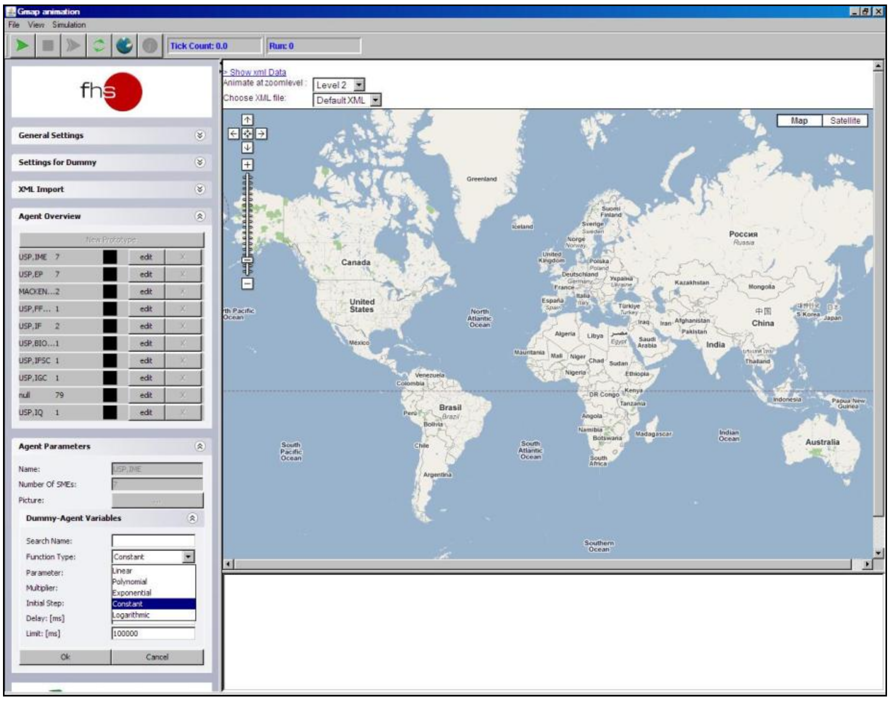


Figure 2.2- EvESim Screenshot including agent variable drop-down.

EvESim is currently running on Soapod, which is based on a Tomcat servlet-container. EvESim services, such as semantic search and agent pools are running, therefore, as Tomcat services. In August 2009, IPTI will provide the first version of a Service container for the OPAALS P2P environment. Then the EvESim services will be migrated to the OPAALS P2P network and consequently, EvESim will become one of the first distributed applications running on an OPAALS infrastructure.

By using the OPAALS infrastructure and running EvESim services in the same container as other OPAALS services, it can easily call real services instead of just simulating or emulating them and therefore use EvESim as Emulating-Framework. For example, by replacing the simple semantic search service in EvESim, documented in D3.7, by the Lucene-based semantic search from IITK, one can set up a simulation of semantic search and later on plug in the OPAALS semantic search, without changing the basic model.

By running a simulation/emulation framework on the infrastructure/network to be simulated, EvESim can act also as a Testing-Framework for OPAALS, independently of the simulation/emulation capabilities. By setting up simulations with thousands of nodes/users by the social scientists and by plugging in real components of OPAALS, like semantic search,

simultaneous editing toolor others, the infrastructure can be tested with a much higher number of users than it has at the moment. For example a conference chat can be tested by agents, setting up a couple of conferences and writing random text back and forth.

By visualising this behaviour of OPAALS components with the built in visualisation on Google maps and Google earth, the working components can be also shown to new potential users of OPAALS infrastructure.

2.1- Integration of Guigoh, Wille Visualisation System and EvESim

During the conceptualisation of EvESim, an XML schema was developed together with TUT in order to import network topologies into EvESim. The same XML schema is used for the visualisation on Google maps and Google earth. In order to develop EvESim and the Google visualisation tools independently, we extract data from TUTs Wille Visualisation System[1] and transform the data into the aforementioned XML schema.

The idea is the following: TUT already extracts data from the OPAALS wiki[2] and other tools. The data includes 1) who has set up pages, 2) who has contributed to which pages, 3) what are the relationships between the users of the system, etc. According to the previously mentioned simulation model, a wiki page will be modelled as a service of a user. It means that the user, who creates a wiki page, is the owner of the page and all contributors are "using" this service. These contributions were then modelled as connections between the users. The resulting data collected and visualised in Wille[3], will be transformed into the XML schema, EvESim and the Google maps and Google earth visualisations used.

Consequently, extensions of the data extraction, data from other networks (Guigoh, Wikis of other networks) as well as the data pre-processing from Wille can be used to provide the basis for SN simulations and consequently also emulations of similar, future tools such as the simultaneous editing tool.

[1] http://tut.fi/hypermedia/en/publications/software/wille/
[2] http://wiki.opaals.org/WP10-D10.12
[3] http://tut.fi/hypermedia/en/publications/software/wille/

3. Methodological aspects for longitudinal SNA: metrics and evolution assessment

This section introduces simple concepts for SNA that are fundamental for the complete understanding of the topologies, motifs and models described in section 4 that will provide the intended simulation set-up to be implemented in EveSim. Among the broad range of concepts related to SNA, we will focus on those related to transitivity and clustering which are essential to understand the evolution and emergency of communities. Theses concepts and others can be found in classical literature like Wasserman and Faust [7]. Further, research in this area has been growing, and more sophisticated approaches were developed using stochastic process [8], Dynamic Models [9], [10] and Bayesian statistical modelling [11] allows researchers to handle more complex and larger structures. Basically, SNA can be divided in graphical exploration (visualisation), quantification of parameters that describe some nodes and network features, and structure exploration.

Sociograms (or social graphs) are the basic visualisation format for network representation, using concepts from graph theory. Each node (vertex) represents an actor in the network, and the links among actors can be edges (ties, for some authors) or arcs (arrows), for undirected or directed networks respectively. An edge is used when the relationship is reciprocal, for example, co-authorship of scientific papers can be modelled with edges. An arc represents an action from one actor to other, and not necessarily the other way around, like scientific citation in papers. Graphical attributes are commonly used to represent some properties, for example, thickness or colour of the tie indicates the intensity or type of the connection between two actors. In the same fashion, different sizes, shapes and colours of nodes can represent specific features or metrics of the actors, generally representing the parameters estimated.

A major problem for sociograms is the optimal visualisation i.e. how best to represent the structure graphically. Some algorithms were developed which consider the network as a physical system of particles and they search for the configuration with the minimum of cross-edges in an equilibrium state for the system. These are called spring-based algorithms, because they deal with the edges as springs trying to repel each other. The most used ones are the Kamada-Kawai [12] and the Fruchterman-Reingold [13]. Kamada-Kawai is not recommended for large scale structures, and the latter can be used for large networks, but convergence for equilibrium is sometimes quite slow.

However, quantitative parameters are also used in SNA and, in general, are based on graph theory and especially in the concept of Degree and Path Length (sometimes called Geodesic Distance). These parameters can be estimated for the actors alone and for the whole network (in

general the average of actors' parameters), giving different interpretations.

An actor *Degree* is defined as the number of links that it has, that is, we say that node **A** has degree k if **A** is connected to other k actors. In the case of directed networks, Indegree and Outdegree are defined, or the number of arcs that depart from A (outdegree) or reach A (indegree). The actor degree represents the intensity with which it is connected to the network structure. The higher the degree, the more active or popular the actor is. In the same way a network has a degree measure, which is the sum of all actors' degrees, and it represents the cohesion of the network.

The first important property of a network is its *Degree Distribution*. It is defined p_k to be the fraction of vertices in the network that have degree k. This fraction is a frequentist stimate of the probability that a vertex chosen uniformly at random has degree k, so that we will use p_k as the notation for this probability without loss of generality. Then, the probability distribution is given by:

$$p_k = P(K = k) = \sum_{i=1}^{N} \delta_{ik} \Big/ N$$

Where δ_{ik} =1 if actor i has degree k and 0 if not, and N is the number of actors in the network. This distribution is usually presented as histograms. However another straightforward way to represent this property is the Cumulative Degree Distribution, given by:

$$P_k = P(K \geq k) = \sum_{i=k}^{\infty} p_i$$

This representation will be very useful in following sections, mainly in the scale-freeness assessment.

The Shortest Path Length, or the Geodesic Distance, between two actors u and v, $L(u, v)$, is the number of steps (actors) in the minimum path from u to reach v. If the network is undirected so $L(u, v)=L(v, u)$, but this is not always true in directed networks. In terms of network topology, exploration the average SPL is very convenient and very useful to assess the called *Small-World* property, as will be shown in section 3. Consider an undirected network so that the average shortest path length is given by:

$$l = \frac{1}{0.5n(n+1)} \sum_{u>v} L(u,v)$$

Concepts about the network structure help to identify substructures, that can be interpreted as communities or clusters that in turn can represent different behaviours, dynamics or other characteristics. The most frequently used ones are Triangles, Cliques, Structural Holes and Bi-components.

A *Triangle* is any formation comprising three actors that are completely connected. Figure 3.1 presents an example of a triangle (inside the circle). In case of undirected networks, just one type of triangle is possible, but for directed at least two can be distinguished. They will be presented in section 8, in the dynamic modelling using Actor Oriented Models. The triangle is a representation of the friend-of-a-friend effect, which means that if u is tied to v and u is also tied to w, so there is a good chance that v is or will be connected to w. Based on this motif, the Transitivity Coefficient, also called Clustering Coefficient gives a measure of the average probability for this (v,w) tie. This is given by:

$$CC = \frac{6 \times Number\_of\_triangles}{Number\_of\_paths\_of\_size\_2}$$

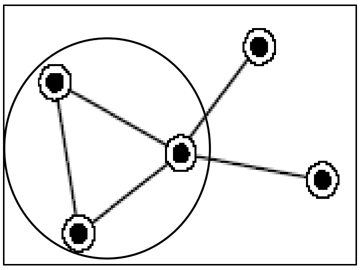


Figure 3.1- This network has 1 triangle and 16 size 2 paths. So *CC= 6*x*1/16=3/8*

A *Clique* of size c is a complete sub-structure (containing all possible edges) comprising c actors. The triangle is a *Clique* of size 3. The most common cliques are the size 3 and 4, since the probability to have a complete sub-structure tends towards 0 as c increases.

Structural Holes are characterised by weak links among clusters strongly connected, also called *Bi-components*. Most of the times, this link is a *bridge* made by only one actor called *articulator*. A *bridge* is a tie that, if removed, increases the number of isolated sub-structures in the network, because this is the unique link between two sub-structures. An *articulator* is an actor that, if removed, increases the number of components, what can be explained by the bottleneck among two or more sub-structures. It is also called a cut-vertex. Despite the almost identical definitions, it is important to note that the former is related to a tie, and the later is related to a node. Finally, a *Bi-component* is a sub-structure with a minimum size of 3 actors and does not contain an articulator. It is a closed network that cannot be divided, but can contain weak ties. In that sense, every *Bi-*component is a Clique, but the reverse is not true. This kind of structure is always related to the presence of complete sub-structures, or different communities in the main core. As it relates more and more to the *CC*, it will have higher values inside *Bi-component* compared to the network *CC*.

4. Communities in social network analysis (SNA)

In this section, models for evolution and for topology will be described giving an introductory background on the parameters, properties and other features explored in the use cases analysed here. Also, a concept of dynamics for SN is introduced in the context of *Actor Oriented Modelling*.

4.1- Topological and modelling aspects in SNA

The first model based topology defined is the Random Network, by Erdös and Rènyi in 1960 [7]. Networks that follow this topology evolve as a set of actors that build social ties randomly. For this reason, at first the actors are connected to a small number of other actors. This can be explained because when an actor is chosen randomly and a tie is established with another actor, the likelihood is that this new tie connects the actor with a low number of others, even if the actor already has some connections [14].

However, when there is a certain quantity of ties so that the mean degree is one, the fraction of the graph occupied by the largest group of the network suffers a phase shift. It means that the largest group, which was not significantly large before, now occupies almost the whole graph area. This organisation is used to explain several phenomena in nature, where small sets join together to build groups of a higher dimension, gathering every actor until the ties are built [14], [16].

In social communities, actors do not tend to establish connections in a random way. This reality influenced the development of *Small Words* model, proposed by Steven Strogatz and Duncan Watts in 1998 [15]. Any distance between two nodes are considered small in those networks comparing to the expected distances in a random graph with same dimensions. Those networks are formed by many small dense groups and the ties inside the groups are considered strong, differing from the weak ties that connect groups, in other words, structural holes are present and the *CC* for the network is usually higher than that expected in a random graph of the same dimension. This effect was characterised by an experiment made by Milgran in 1967, concluding that if each person in the world has a hundred friends, the median path length between anyone would be six.

Another topology was proposed by Barabási [16] and his team in 2000, called S*cale-Free* networks. The authors realised that the degree of distribution does not follow the Poisson or Gaussian distribution, but a Power-law distribution instead. The Power-law distribution is differentiated by its format; the maximum point is not found in the average value, the curve starts at

its maximum point and decreases towards infinity There is an exponent that can be calculated in this distribution that shows how the distribution changes along with a variable.

A function that follows a power law distribution can be described by

$$p_k \sim Ck^{\alpha} \qquad (1)$$

k is the degree, p_k is the probability of actors containing degree of value k, as defined in (1) and α is the function exponent. Another way to represent the Power-law distribution uses the Cumulative Degree Distribution, defined in (2),

$$P_k \sim Ck^{-(\alpha-1)} \quad (2)$$

And this relation will be useful for the α estimation.

So, to assess scale-freeness in a network, usually the fit of the degree distribution to a Power-law is investigated. The simplest way to do this is taking the logarithm of P_k and k and analyse the scattering of the points. A straight line is expected if the distribution approximately obeys a power law, as shown in Figure 4.1. The slope of this line is then used to estimate α, since

$$\log(P_k) = \log(C) - (\alpha - 1)\log(k).$$

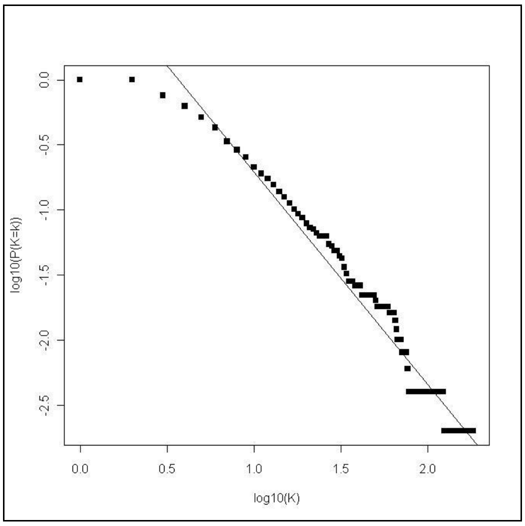


Figure 4.1- A cumulative degree distribution of a network that follows, approximately a power-law distribution. Slope estimated is 1.43, so that α=2.43

This topology also presents a few nodes with a high quantity of ties, called *hubs*. Then, the majority of the actors have a lower number of social ties and a few (*hubs*) have a higher number of ties. Those networks are more tolerant to failure and node removal. As most nodes have a few connections, the probability of removing one of them is higher than a hub, and they do not impact significantly on the network structure. However, a coordinated node removal focusing the hubs can completely break the network structure [16].

The scale-free network evolution obeys a process called *Preferential Attachment*, a.k.a. *Rich Gets Richer* effect, such that new nodes have a higher probability of connecting to the older nodes with high degree. Then, the older nodes have more chances to become hubs, but new nodes may also become hubs depending on its connectivity factor. This process usually leads to a hierarchical structure, where clusters are formed replicating the structure of predecessors, as shown in Figure 4.2 a), providing a topology like 4.2 b). Scale-free networks, having this underlying process, also presents the Small World effect.

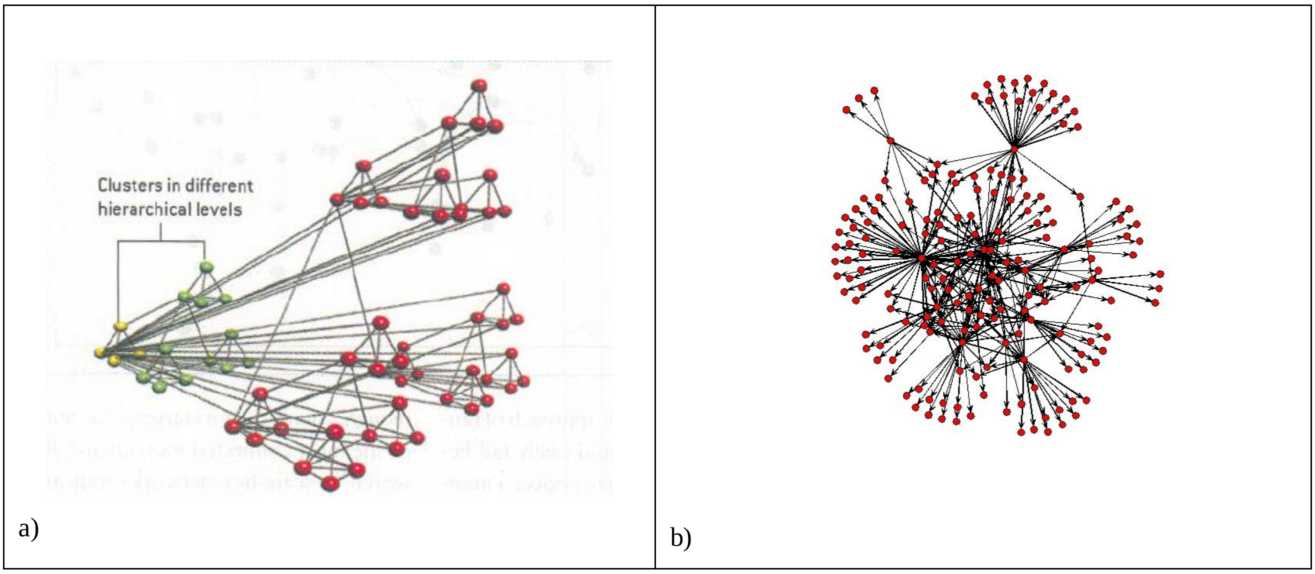


Figure 4.2- a) Hierarchical clustering process governed by *Rich gets richer effect*,
b) Hierarchical topology having the same pattern.

4.2- Dynamics and network autocorrelation

The Dynamic Network Analysis is a new framework that has arisen due to the necessity to apply SNA to other areas, the development of new methods for data collection and the growing of computational power for simulations and graphical display. It consists of a toolkit comprised by statistical modelling and visualisation techniques that used together can improve understanding and knowledge about the network process, providing information that can help in actions and decisions when this network structure should be "fostered", as in the case of a collaboration network.

This section aims to present a suggestion for SNA over panel data, based on a statistical and dynamic perspective, rather than being purely static and mathematical. The difference is in the fact that probabilistic components are included in the model enabling quantification of some parameters that can present direct or indirect influence over the network behaviour. Knowledge about those parameters could give important information on how networks behave and evolve, based on the

influence of covariates (variables), and then how it would be possible to simulate more realistic scenarios. Later, some purposes of its utilisation in OPAALS tasks are discussed.

A key feature for social network analysis is the study and interpretation of the interdependence among members from a social group, a recurring theme in the theory of formation. The concept of cohesion among those group members as an indicator of compliance with group norms is also well-known [19].

In statistical science this interdependence is called *network autocorrelation*, such that ties in the network occur more frequently among similar actors in terms of specific attributes, like geographical localisation, economic status or musical preference. Some measures from the traditional SNA methods, which lead with an instant picture of the network, give some clues about clusters or modules [7], [20], but the complete understanding about how these clusters were formed depends on the assessment of the network in a follow-up fashion. This necessity turns clearer when two principles, which are the most prominent explanations for the autocorrelation are defined.

The first principle is called *homophily* [21], a selection mechanism where actors seek relationship with similes over the time. In this case, network ties are formed according to similarity on some actor attributes, and network autocorrelation emerges as a consequence of social selection over the time. Another explanation may be the *assimilation principle*, where actors adapt their own attributes to match the group average attribute profile [22].

In this process, again over time, autocorrelation (Figure 4.3) emerges by a social influence mechanism. From these two principles, it is clear that the right understanding of the autocorrelation phenomena requires the social network to be dynamic, in its topology (selection over time) and in its endogenous attributes (influence changing behaviour). In a real social network data, it is likely that both principles can occur, and then the goal is to study how they influence the network evolution, and if one of them explain this dynamic better.

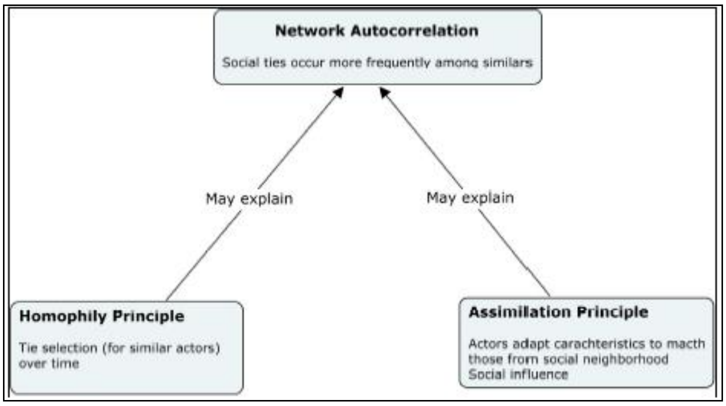


Figure 4.3- Principles that may explain network autocorrelation concept

The following is an example of how these principles and the dynamic social networks, in their topology (selection over time) and endogenous attributes (influence changing behavior) will look in a simulation scenario in EvESim. Data from the changes on pages in the OPAALS wiki were recorded over time (see t0, t1 in Figure 4.4). The user selects which pages they want to edit and with which other users they want to collaborate on a certain page or topic. This will be implemented in Phase III of OPAALS by importing data from TUTs wiki data extraction tool[4].

This selection of pages and collaboration partners leads to a collaboration topology of users, which can be directly imported into EvESim. Chapter 2.1 introduced how TUT´s Wille Visualisation System is used to import this data. The changes of topologies and the selection of collaborations indicate the behaviour of these nodes in collaborating with others. These dependencies are showed in Figure 4.4. SNA can analyse the behaviour of the users and deduce the most important behavioural variables. Clearly, the longer the time frame and the more detailed data for analysis: the better the emulation result.

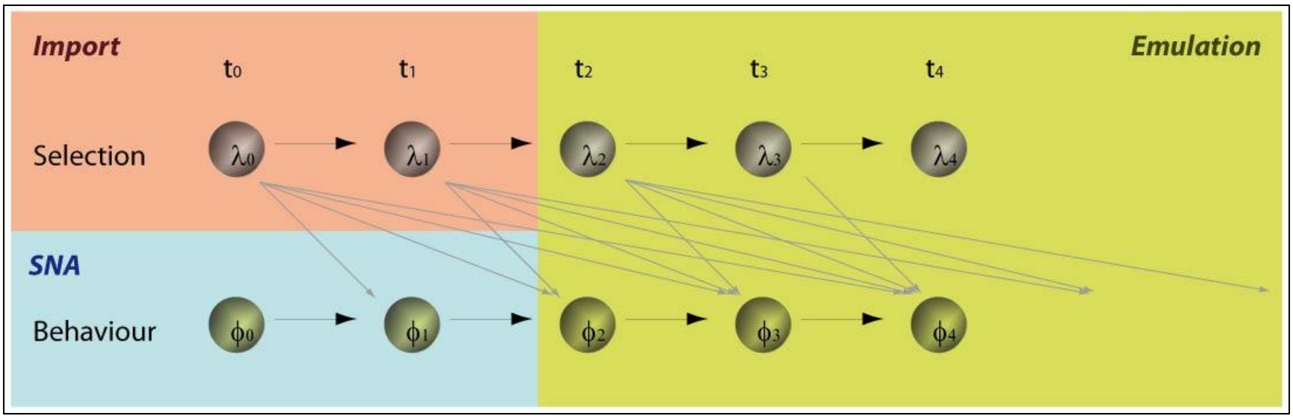


Figure 4.4- Network import scenario with SNA and EvESim Emulation.

[4] http://wiki.opaals.org/WP10-D10.12

Based on the import of the current topology and the identified behavioural variables, EvESim can simulate the evolving network in the future (t2, t3, t4 in Figure 4.5) and therefore, for example, emulate the usage of the Wiki with a larger network in 3 years time. That enables the implementation group to recognise the forthcoming computing needs and react in time on e.g. computing power or network bandwidth requirements.

It is clear that follow-up data is needed to understand autocorrelation mechanism, and in social science this kind of data comes from the so-called panel design. The group is observed in different points in time, making possible only observations of what is happening at that point. This discretisation of time observation introduces an analytical complication: when the processes of selection and influence happen in-between consecutive panels, in a continuous time spam it is not possible for them to be observed. Another complication is related to the *completeness* of the network, which can be understood as a clear definition set of actors' boundaries in the network. In general, the dynamic processes are hardly limited once those actors can enter or leave the network.

In D 3.7 [3] the *Actor Oriented Model* was presented as a way to study and simulate networks having some effects influenced by covariates which could represent the attributes of the actors. Snijders introduced stochastic actor-oriented models [23]3-[25] firstly by modelling just the evolution of the network in terms of actor selection. In other words, actor orientation means that, for each change in the network, the perspective of the actor whose tie is changing is taken. Then, this actor (let's call *i* actor) controls the whole set of tie variables included in the *i-th* row of the adjacency matrix that defines the network. This is called a *microstep*, so that the analysis of the dynamic has a bottom-up perspective, from micro changes (actor tie selection) to macro changes (change in network topology).

The observed networks are modelled by means of a parametric model for the transition probabilities between any other topology with same dimension that configures the so-called state space. In network panel data, each panel will present a configuration that pertains to the whole large set of the state space. The explanation of dynamics is formulated by means of the transition probabilities, with the first observation being that which is conditioned upon. This explanation can also be made using covariates as explanatory factors as is usually done in Regression Analysis.

In this sense, the modelling task is reduced to:

a) Modelling the change in the *microstep*

b) Modelling the occurrence of this *microstep,* over time

The first task is solved by means of a maximisation of a stochastic utility function, here called the *Objective function.* Task b) consists of a specification of a probability model for the actors' individual waiting time (usually the Exponential model), called the *Rate function*. By this

approach, the time dependence of the evolution process is implicitly modelled as an emergent consequence of model dependence on time. Both model components include dependences on previous network state (topology), time and actor, but not for the whole history of the process (Markov assumption).

Another way to study the evolution is the analysis of parameters and properties of the network panel as a time series, exploring the pattern of the evolution in terms of transitivity, clustering and size growth. By applying adequate graphical and statistical methods to these parameters, it is possible to understand relevant aspects that can complement and drive the stochastic modelling in terms of what effects the analysis.

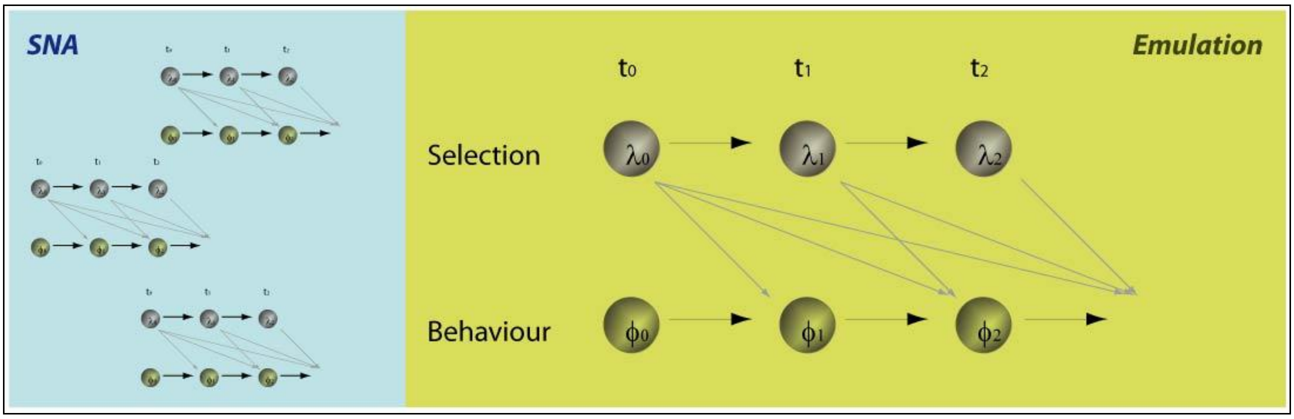


Figure 4.5- Bootstrap scenario with SNA and EvESim Emulation.

Unfortunately, during the bootstrap phase of social networks, there is not much data available for analysis, as drafted in Figure 4.5. Consequently, the following chapters analyse the data available from other social networks with similar structure and behaviour. As can be seen in Figure 4.5, the emulation in EvESim will work differently in this case.

As there is no topology to import, SNA analyses the bootstraping and dynamic behaviour of similar networks and hence, we can set up simulations and emulations of a potential bootstrap scenario for the collaborative document editing in the OPPALS version of Guigoh for example. It is important to state here, that the results as well as predictions achieved by the simulation are certainly estimations and especially in social systems it is very hard to accurately predict dynamic behaviour. Nevertheless, emulation based on detailed analysis could help the computing domain considerably in estimating at least the rough behaviour of certain user groups and components considerably.

5. Web forum use cases

This section starts the use case analysis of many different types of network. Web forums, are normally created for any kind of discussion, from sports to movies as entertainment, but also many forums are created to exchange knowledge and experience in a collaborative way. Both use cases presented here bring this collaboration characteristic, both having been constructed for Brazilian government programs aiming training in certain areas. In these cases, the areas are the use of open source and free software for drug abuse detection and therapy.

These use cases have a close relation to the OKS conceptualised by OPAALS, at least in terms of digital asynchronous collaboration and knowledge construction, so that its analysis could bring important clues about how this kind of process happens. Assuming these behaviours, it is possible to emulate similar networks, as proposed in Figure 4.5.

5.1- Conversê

Conversê (a popular denomination for “jamming” or “chat” in Brazilian Portuguese) is a platform developed for the *Cultura Digital* (Digital Culture) program from Brazilian Ministry of Culture (*MinC*). This platform aims to be an open discussion workspace for *Cultura Digital* participants (mainly) and any other person interested in the subjects posted there, as intended by the OKS, however in a more. The idea behind *Conversê* is to enhance the creation of social networks, where Brazilian cultural stakeholders can share experiences and establish collaboration and activism. Roughly speaking, it is a dynamic forum following the so-called Web 2.0 concept, where participants can exchange experiences, ask for help and discuss about politics, technology and other issues. All posts can be tagged, such that main and more frequently used and more accessed tags gain highlights in the page, generating a folksonomy.

The platform was developed using Drupal technology, and was active from 2005 to 2008. Conceived for use within the *Cultura Digital* as a way to exchange experience and offer a help desk for FLOSS users, however it became highly accessed by ‘common’ users who had found an interesting virtual space to share opinions and knowledge. Data used in this analysis pertains to the period from January to December 2006 (Table 5.1).

Table 5.1- Description for the network evolution moth-by-month along 2006

Month	*Users Registered*	*Active users*	*Cumulative (new active users only)*	*Network Degree*
January	2003	47	47	119
February	2167	53	82	259
March	2409	30	98	301
April	2975	44	123	378
May	3380	71	171	556
June	3836	108	241	767
July	4318	73	282	925
August	4813	108	334	1178
September	5430	103	387	1509
October	5451	88	434	1671
November	6783	65	473	1767
December	7516	50	494	1852

Table 5.1 presents data for registered, active users in each month, the cumulative number of users and the network degree. The estimated average rate of increase is 44 new actors each month. Figure 5.1 shows the sociograms for January, April, August and December, and Figure 5.2 presents the evolution of the average degree and the final cumulative degree distribution (CDD). The average degree presents a fast growth from month 6 to 10, intermediated by two plateaus before and after these months.

It can be explained as an effect of workshops organised by the program coordination, which took place in many different cities across all regions of the country. These workshops gathered actors to learn about and discuss the use of free software for multimedia editing, and were an important kick-off for the whole project. In October, after the project explosion, it behaviour similar to that before month 6, could be observed i.e., the network activity stabilised and actors were interacting less than before.

This effect can be also seen in the transitivity coefficient evolution, presented in Figure 5.3c. An exponential decay is observed after month 6, but in this case the transitivity continued to decrease, almost reaching 0.1, such that, the probability of an actor establishing a tie with another one 2 steps away from it is on average 10%. Even in the beginning this probability was not so high (40%). The Conversê is a very dense network, presenting a huge core with almost the size of the entire network (Figure 5.3a), and some other smaller components. For example, in the last month, the major cluster has size 460, or 93% of all actors in the network, and 16 smaller components. This could explain the low transitivity that provides a high ‘betweenness’.

This network presents the scale-free property, as can be confirmed by the CDD following

the adjusted line. As a consequence it also presents small-world property, verified by the Average Path Length, that is small, in average 2.7, and almost without variation along the period.

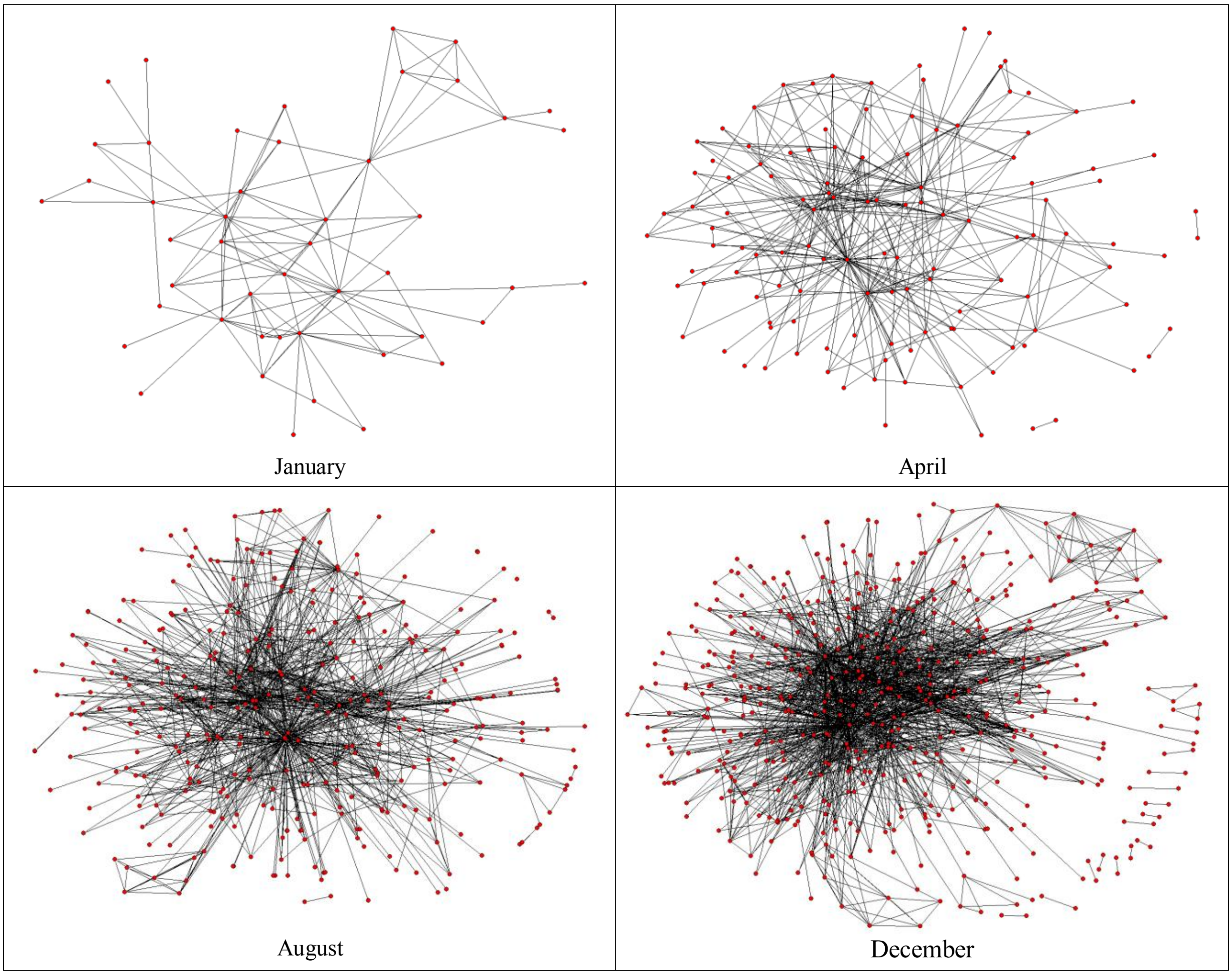


Figure 5.1- Sociograms of Conversê network

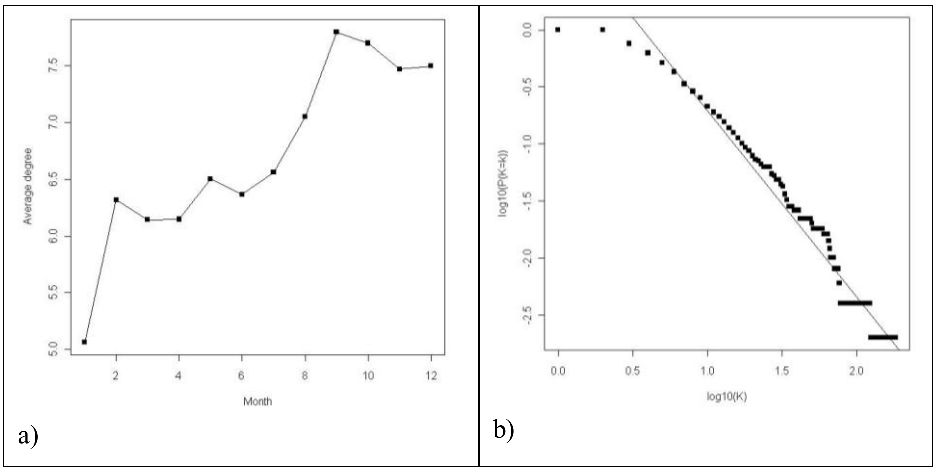


Figure 5.2- a) Average degree evolution and b) cumulative degree distribution for CONVERSE dynamic forum.

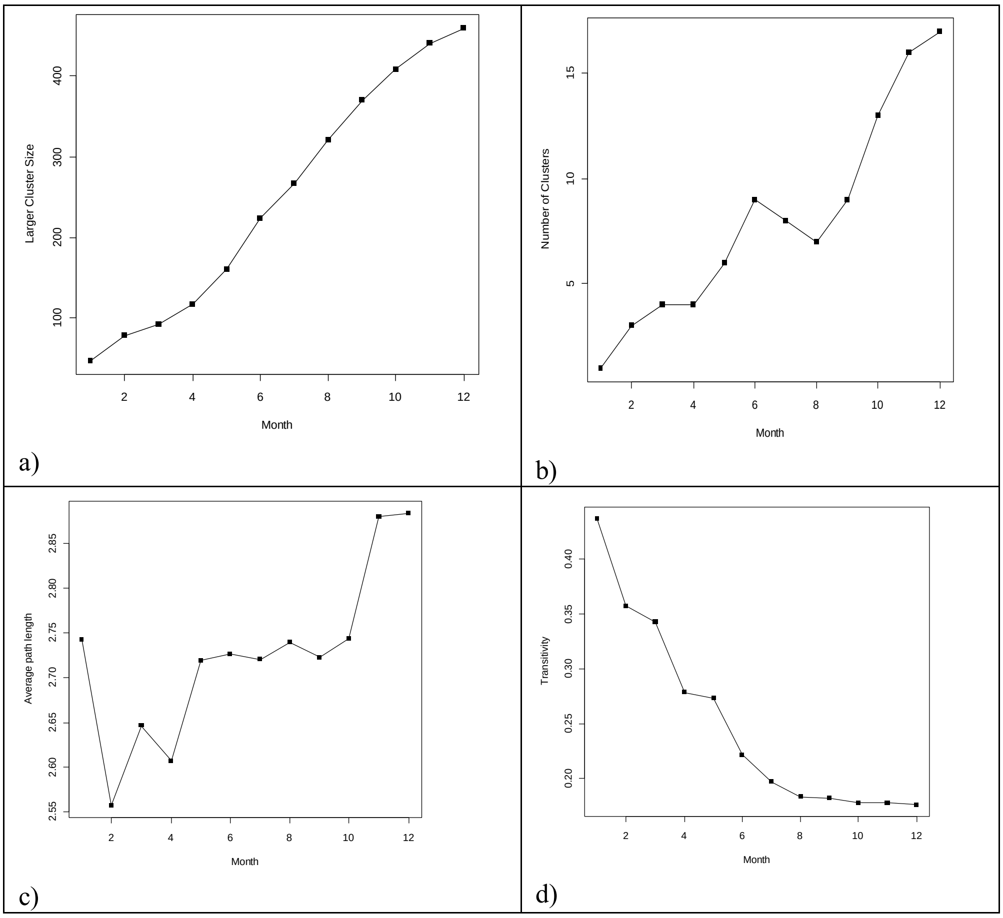


Figure 5.3 – a) Larger cluster size, b) Number of Clusters,
c) Average path length and d) Transitivity Coefficient evolutions for CONVERSE dynamic forum.

5.2- SUPERA

The National Secretary of Policy on Drugs (SENAD), of the Brazilian government, offered a distance course on drug abuse detection and therapy for health-care providers covering all the regions. In this environment, a forum was created so that attendants could discuss course contents, always mediated by the tutor responsible. The data from this forum is modelled here, using 3 panels defined by the first month (March), second month (April) and all other threads/replies after the second month (May and on), from a total of 6 months of activities.

The final number of participants in this forum was 177, such that 16 of them were the tutors, 123 were women and the average age was 41 years.

Table 4.2 **-** Network description

Month	***Active users***	***Network Degree***
March	153	132
April	170	176
May and on	177	190

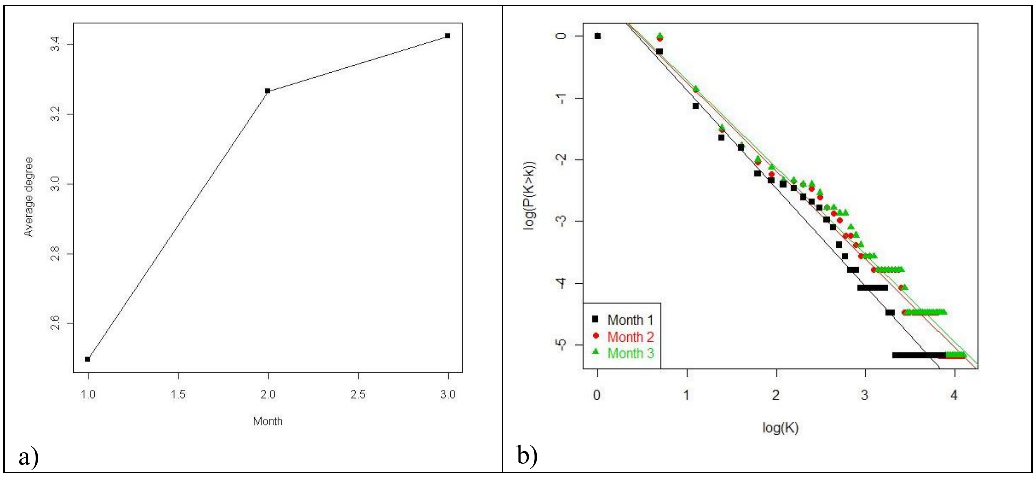


Figure 5.4 – Average degree evolution and cumulative degree distribution for SUPERA dynamic forum.

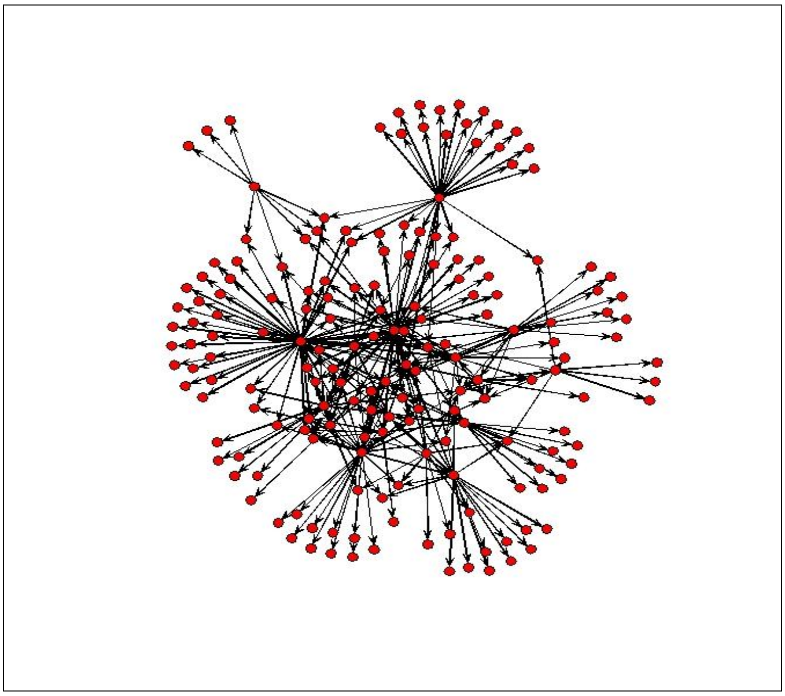


Figure 5.5- Sociogram for the final SUPERA network

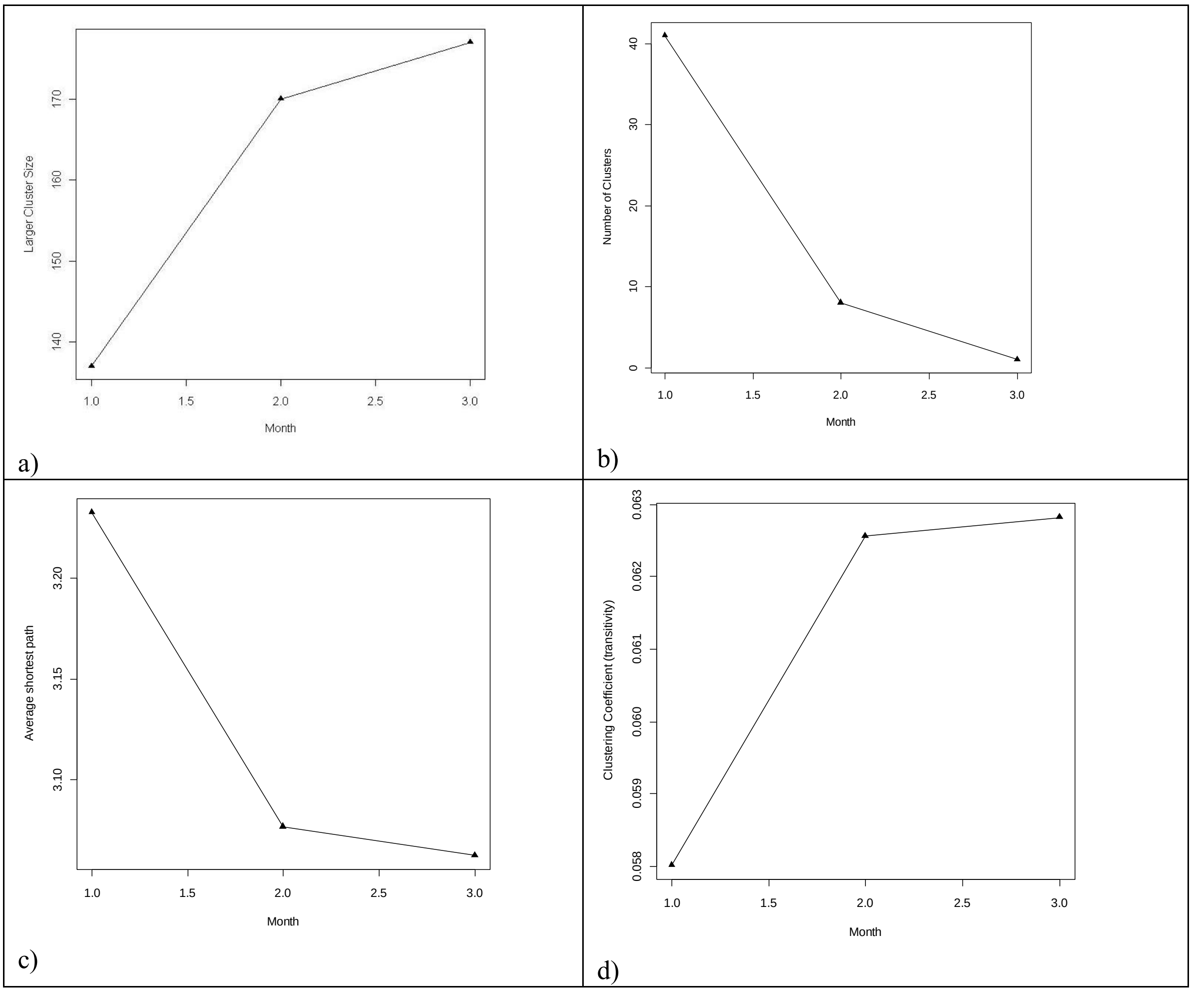


Figure 5.6 – a) Larger cluster size, b) Number of Clusters, c) Average path length and d) Clustering Coefficient evolutions for SUPERA dynamic forum.

Clearly, this network follows a power-law distribution (Figure 5.4b) and can be considered scale-free. The hierarchical structure is the most prominent characteristic, evidencing the linear preferential attachment. Actors that post later in the forum, tended to receive replies from the former actors. In terms of evolution, this network does not change all the parameters, except the number of clusters that decreases from 41 to 1. This means that this network presents high closure and it evolved quickly into a unique community, where access to information is possible to every actor, with at most 4 steps (Figure 5.6c). Transitivity is very low, and this is explained by the topology per sae, since the hierarchical structure makes it less likely that new ties happen outside the main core, since they occur almost every time in the periphery.

This Forum could be modelled as a directed network, and also with some attributes from its actors. The dynamic model could then be applied to elucidate how this closure evolutes, and if it can be explained by any of those attributes.

5.2.1- Stochastic dynamic modelling of the SUPERA forum

In this section, the forum is analysed using the Stochastic Actor Oriented Model. This modelling will concentrate in the dynamics of the network topology (objective functions) in the first place and after the influence of covariates in those metrics along the network history. The influence of covariates gender and age is very important to understand how demographic profile could influence this kind of collaboration environment. The other covariate is the actor role, if one is an attendant or tutor. This covariate helps to understand the hierarchy effect, and if it happens.

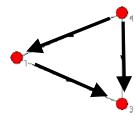
After an interactive modelling process, having tested many different effects and objective functions, 3 final models were selected for presentation. Model 1 estimates the rate transitions and objective functions (explained below) about network topology. Model 2 introduces age and gender effects in the transition rates, while maintaining the other parameters. Finally model 3 introduces the Level covariate and the final best effects that could explain the dynamics.

The density effect is high, reflecting the out-degrees, and for the modelling purpose it was considered fixed in 3.98, indicating that in the average, the number of new departing links had a fourfold increase. This constant effect is a reasonable assumption for this kind of network, since it is composed of almost all the replies from tutors to participants, and it makes the estimation of other parameters rather easy in terms of computational costs.

In Model 1 (Table 2), the estimated rates are 0.77 from first to second panel and 0.16 from second to the final one. It shows that the network had more intensive dynamics during the first 2 months, as expected.

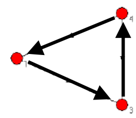
An important subject when analysing a forum is its closure, representing the way that actors behave when selecting other actors and forming communities or strong links. It is used as an approach to social capital by some authors [15]. To represent closure, two objective functions were selected. The *Transitive Triplets*, defined as:

$$s_{i4}(x_i) = \sum_j x_{ij} \sum_h x_{ih} x_{hj}$$



and the *3-cycle*s:

$$s_{i5}(x_i) = \sum_j x_{ij} \sum_h x_{jh} x_{hi}$$



.

Transitive triplets (TT) effect, 0.94, is statistically significant (t-statistic= 3.37), while the 3-cycles (3C), -1.99, can be considered significant with a significance level less than 10%. The positive value for TT shows that actors tended to increase transition motifs in the network, so the final structure presents a high transitivity. The negative sign for 3C is indicative of some hierarchical structure in the topology. Both effects have good interpretations for the network studied.

Model 2 introduces Gender and Age effects in the rates. Only the Gender effect is significant, and the negative value indicates that women (coded as 2) have had more influence in the new connections than men (coded as 1). This can be explained by the fact that men started to contribute to the forum later than women, also women are more represented in the network (71%). Interestingly, 3C had the effect increased, -2.06, and also significant (t-statistic = -1.98) with higher observed significance level.

From the statistical modelling perspective, the change in the 3C parameter when Gender was introduced is evidence that this covariate has an association with this kind of motif. Some models were adjusted to try to investigate this association; however those provided poor estimation results and even a lack of convergence for the estimation algorithm. The idea to introduce covariate for parameters related to the whole network topology was understood as the main cause for those unfortunate results. The next step is to model these covariate influence in actor-specific objective functions, such that, authority (Alter) and preference to select similar actors (Similarity). Results are presented in Model 3 (Table II). The first is related to the TT effect, and the second with the 3C. They can be seen as the *microsteps* to form the motifs TT and 3C.

Table 4.3- Dynamic models adjusted for the forum

	Effect	Std. Error	t-statistic
Rate 1 – 2	0,77		
Rate 2 – 3	0,16		
Model 1			
Transitive triplets	0,93	0,28	3,37
3-cycles	-1,99	1,08	-1,84
Model 2			
Age on rate	-0,04	0,13	-0,28
Gender on rate	-0,74	0,23	-3,21
Transitive triplets	0,87	0,26	3,30
3-cycles	-2,06	1,04	-1,98
Model 3			
Outdegree on rate	0,09	0,01	11,14
Gender on rate	0,85	0,37	2,31
Transitive triplets	-0,01	0,08	-0,18
Gender similarity	0,59	0,26	2,30
Level Similarity	-1,99	0,31	-6,33
Alter for age	-0,40	0,13	-3,13

In terms of similarities, Gender again explains a lot. This can be interpreted as women's preference to connect to other women. It is hard to state that it is a social behaviour, due to the fact that 71% of the whole network actors are women. The Level of similarity has a negative value, showing that the tendency is that Tutors communicate more with Students. This behaviour seems adequate for a mediated forum from a distance-education environment. Authority, measured by the Alter effect, appears here having an association with age. Negative sign evidences that the old-aged tends to receive more ties from younger. Actually this cannot be interpreted as authority in terms of social empowerment, but it reflects that they tend to have higher indegree from younger actors.

6. Scientific co-authoring networks by science domains

The OPAALS NoE is a multidisciplinary network with researchers mainly from three domains: Natural, Social and Computer Science. The idea of this chapter is to study collaboration of Brazilian researchers inside those domains in terms of co-authoring scientific papers from ISI indexed journals. This use case was already used in D 3.7, for the "Simulation and Emulation Framework" chapter (Chapter 4), restricted to the Physics research field.

Here, the three research areas chosen are Mathematics, Psychology and Computer Science, representing each of the former OPAALS domains. It can be argued that these areas are not the most representative for the respective domains, however the decision about the areas to be studied took into account the number of authors and papers from 2000 to 2005 as well as the dynamics observed simply by visual inspections.

Each of the three main areas cited above has many sub-areas, or specific domains, like Artificial Intelligence in Computer Science and Psychometrics in Psychology area. In the Brazilian research system, defined by the Brazilian Research Council, all areas have also the sub-area called "Interdisciplinary" or "Multidisciplinary". It was decide also to study two sub-areas from the three domains, always having the so-called interdisciplinary area. The other sub-areas chosen were: "Applied Mathematics", "Artificial Intelligence" and "Psychological Treatment and Prevention".

Firstly, the three main domains will be analysed and compared, and after those the sub-areas.

6.1- Mathematics (Maths), Computer Science (CS) and Psychology (Psych)

In the 90's, Brazilian researchers started to grow their participation in the scientific community as a result of many research enhancement programs that have substantially increased the number of doctors in the institutions. Then, the growth in the number ISI indexed papers is clear, from 2000 (Figure 6.1), while authors apparently reached a plateau after 2002 for Maths and Psych and a variation pattern for CS. Maths, a more traditional research area in Brazil presents an increase of almost 200 papers per year, followed by CS and Psych.

In terms of growth models, all areas follow a Scale-free behaviour, with linear preferential attachment, at least in part of the degree distribution. After the value 3 for the Log(k), the distribution falls rapidly, breaking the expected trend. This kind of cut-off value that can be understood as a lighter tail of the underlying Power-law, is sometimes called "Finite Population Effect". In fact, this data is about a not particularly large population and further it only analysed

papers in the ISI index, so that authors and papers that had never been published in ISI journals are not represented in this network. However, even in this "well-behaved" piece, the Psych area does not follow the power law. Looking at the sociograms, one notices that Math and CS present big main core structures those are apparently responsible for the scale-freeness, while Psych shows a spread-out network, with many small groups. This illustrates a tendency towards collaborations happening inside pre-defined groups.

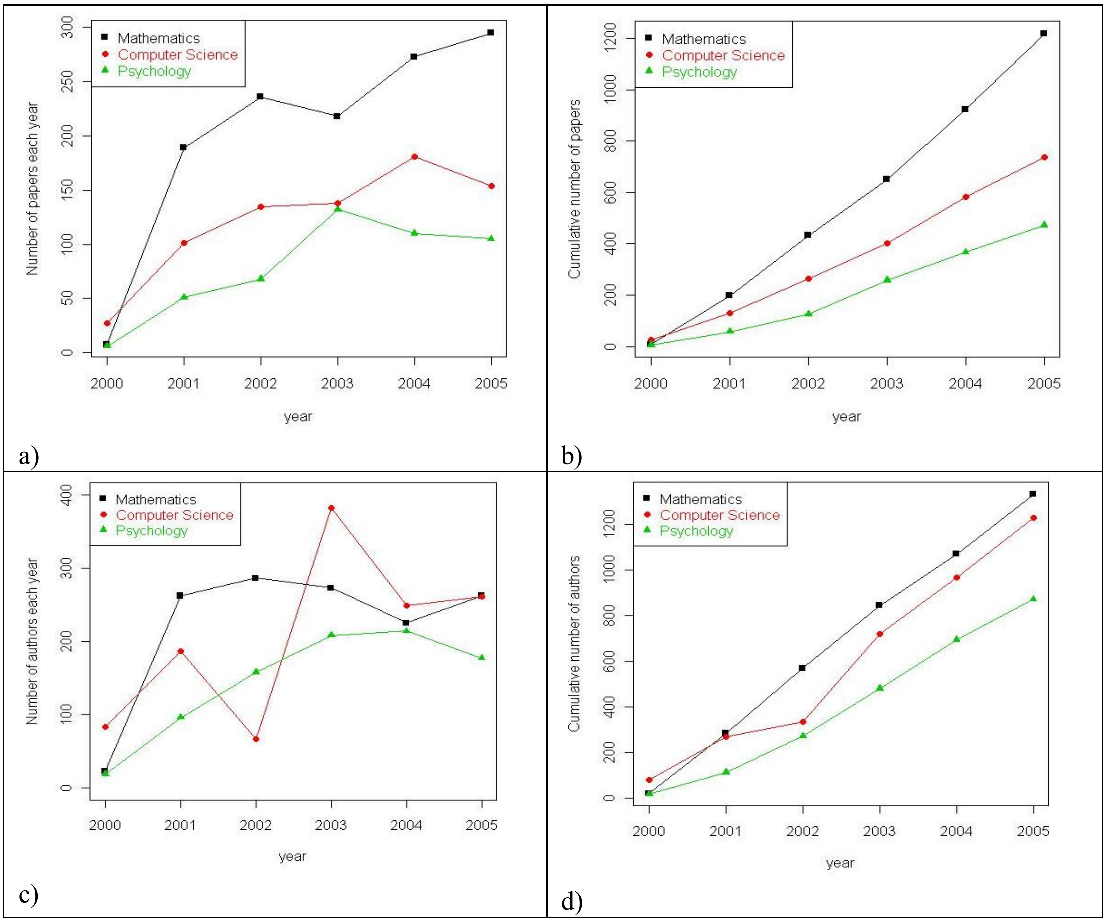


Figure 6.1- Number of papers (a), cumulative number of papers (b), Number of new authors (c) and cumulative number of authors for the main fields (Mathematics, Computer Science and Psychology)

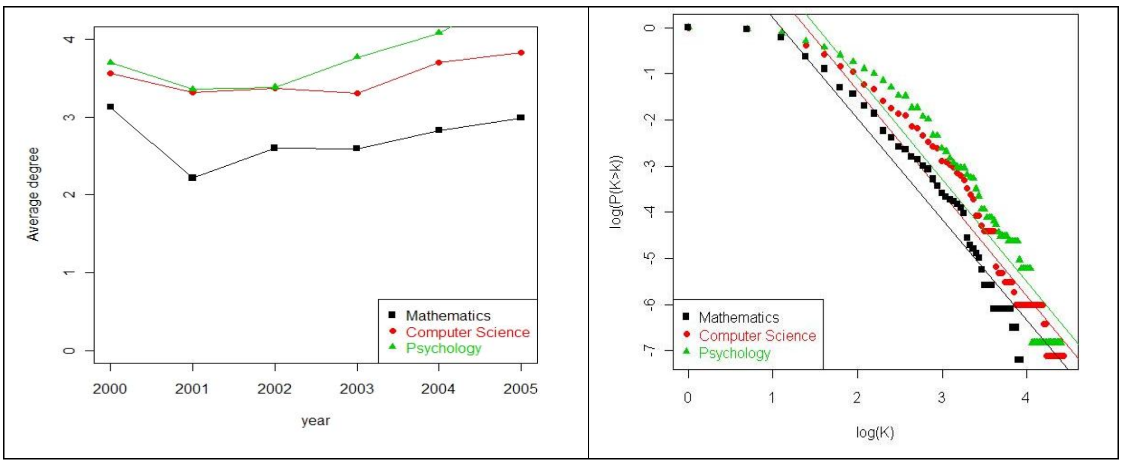


Figure 6.2 – Average degree evolution and cumulative degree distribution for the three main fields

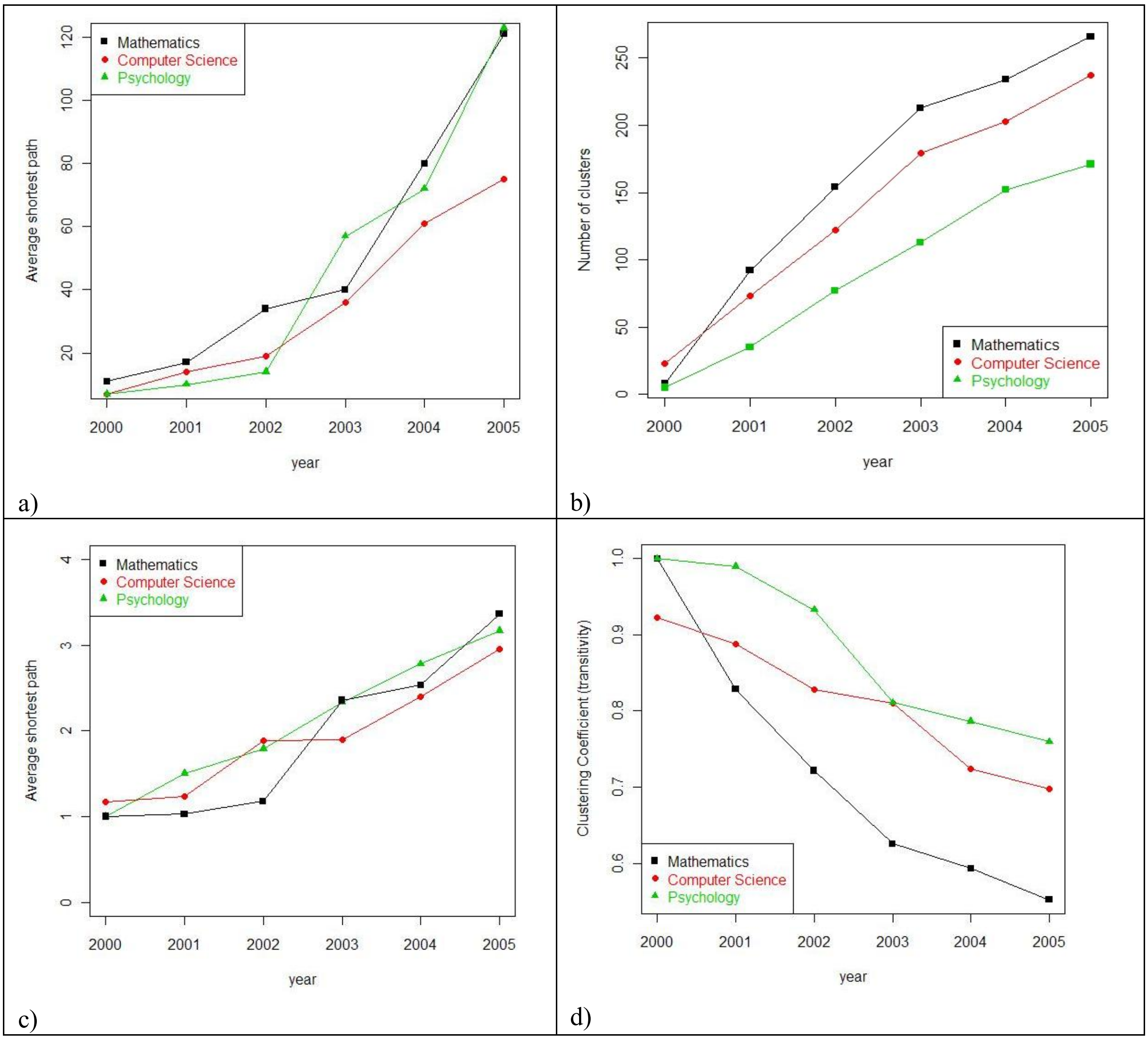


Figure 6.3 – a) Larger cluster size, b) Number of Clusters, c) Average path length and d) Clustering Coefficient evolutions for coauthoring network

6.2- Sub-areas

Looking inside each domain, it is possible to better understand some differences between them. Except for the Psych sub-areas, all others present linear preferential attachment, confirming the preceding findings.

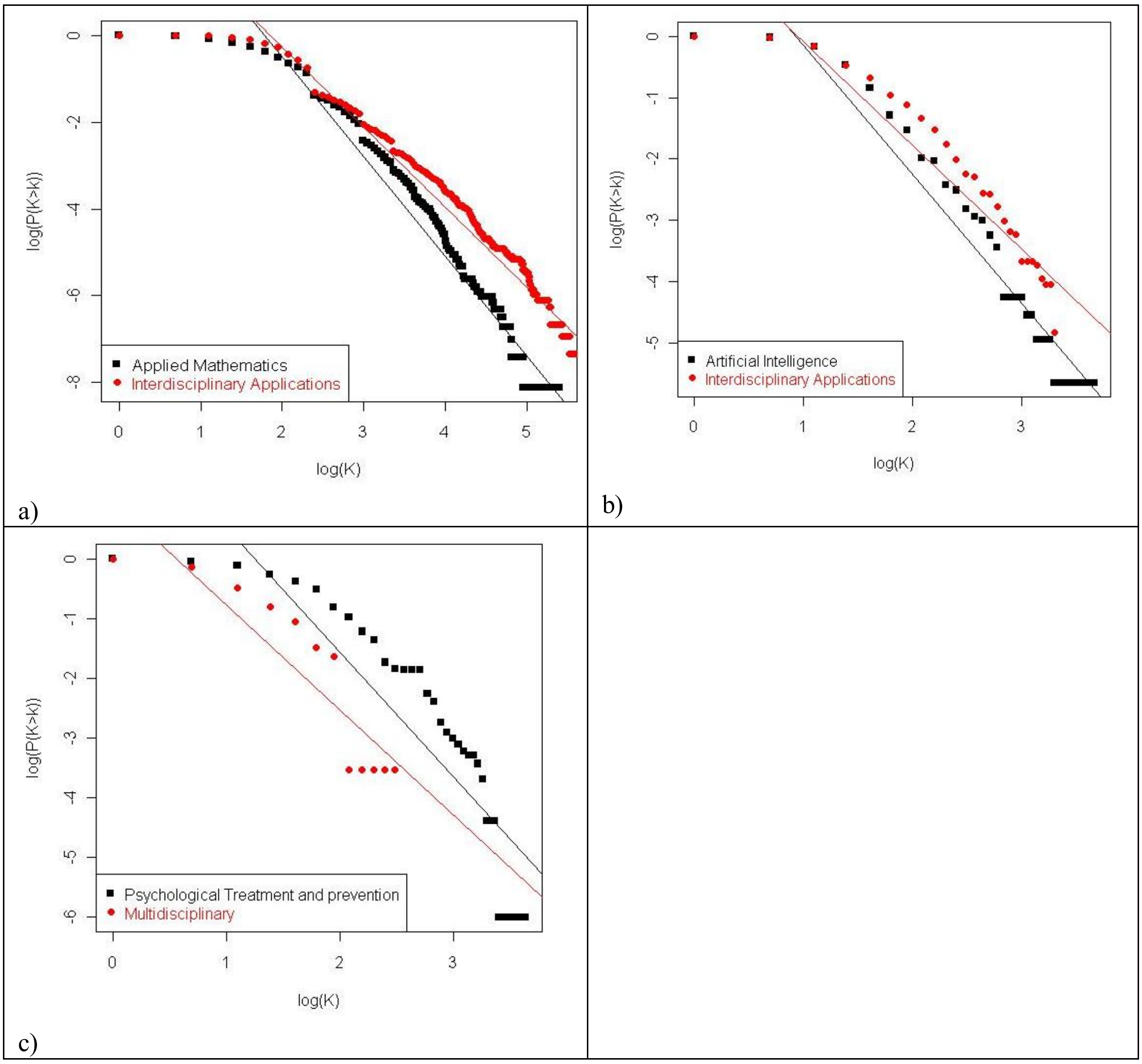


Figure 6.4- Cumulative degree distributions for sub-areas

Maths sub-areas show higher shortest paths, whereas the other sub-areas show that these values are within 1 and 2. Again this is related to the topology, since Maths has a highly connected network, with articulators linking different groups. The other sub-areas are rather spread out, with many different groups.

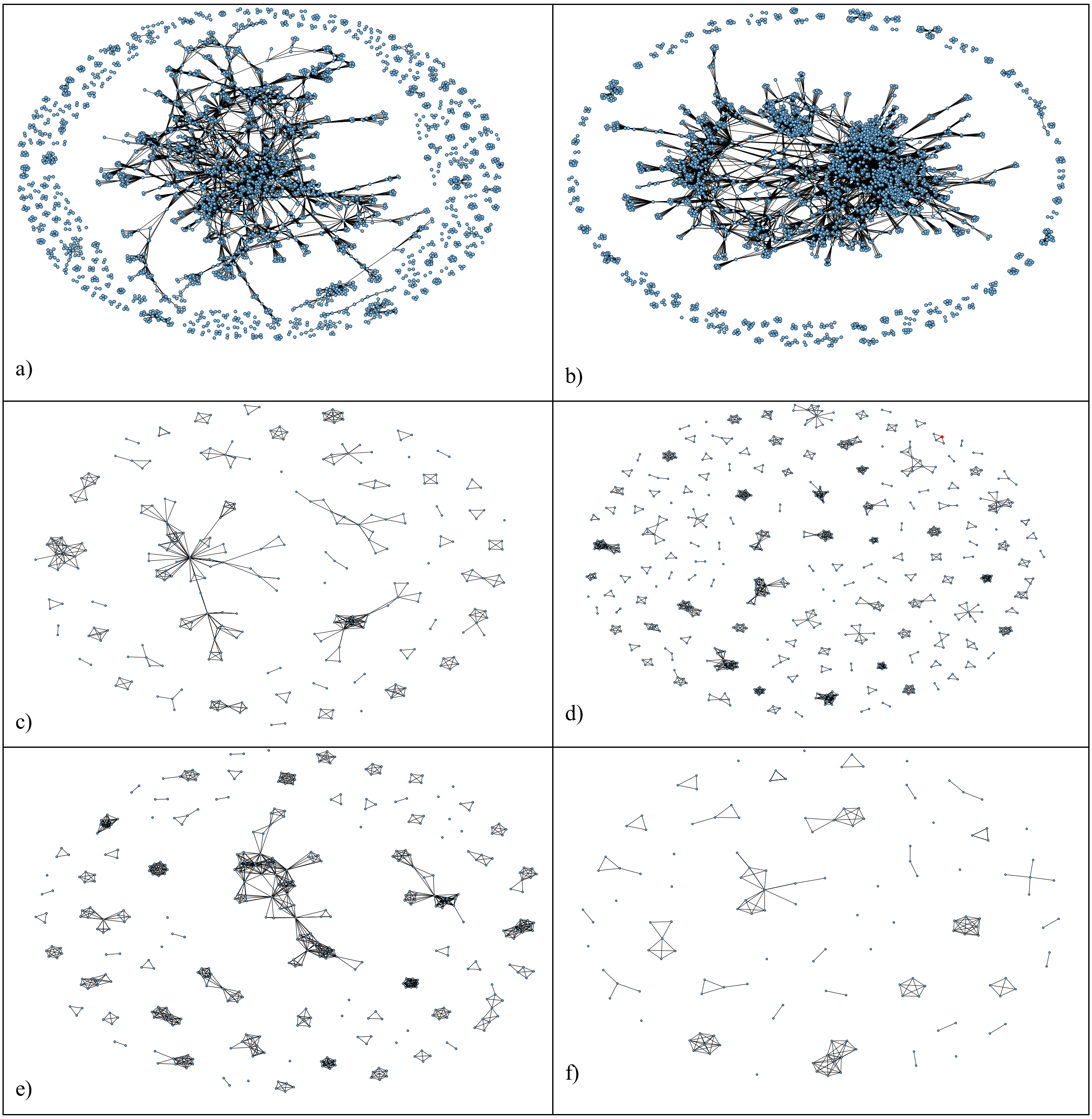


Figure 6.5- Sociograms for sub-areas. a) Applies Mathematics, b) Interdisciplinary (Maths), c) Artificial Intelligence, d) Interdisciplinary (CS), e) Psych. Treatment and Prevention, f) Multidisciplinary (Psych).

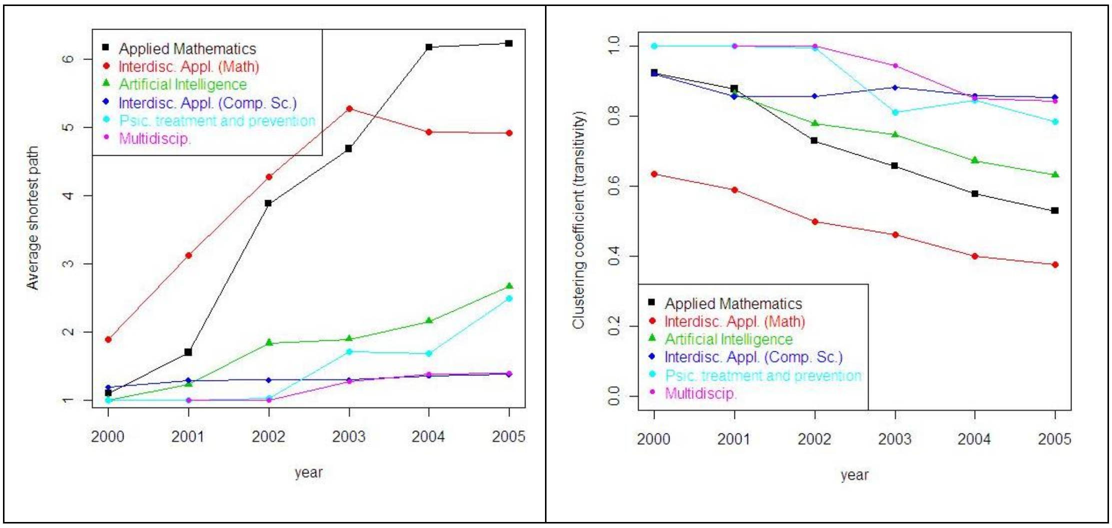


Figure 6.6- Average path length and Transitivity coefficients evolution for the sub-areas

This effect also is also reflected in the transitivity coefficient, with the CS and Psych sub-areas still presenting high probability for new collaborations, while in Maths, this coefficient decays while Average Shortest Path Length grows.

7. OPAALS Wiki

Created to be the first tool of the Open Knowledge Space (OKS), the wiki was until 2008 the main environment for collaboration in OPAALS. The main knowledge repository for the consortium is formed by: Management issues, events calendar, dissemination material, Work Packages (WP) descriptions, among others.

In this section a SNA is carried out over the 4 years of the existence of the wiki, but analysing just the WP pages, trying to assess the collaboration process over tasks defined in work packages. Data was extracted with the toolkit Wille2, which was developed by TUT.

Table 7.1- Number of actors and network degree for OPAALS wiki

Year	***Active users***	***Network Degree***
2006	15	46
2007	23	80
2008	34	222
2009	37	224

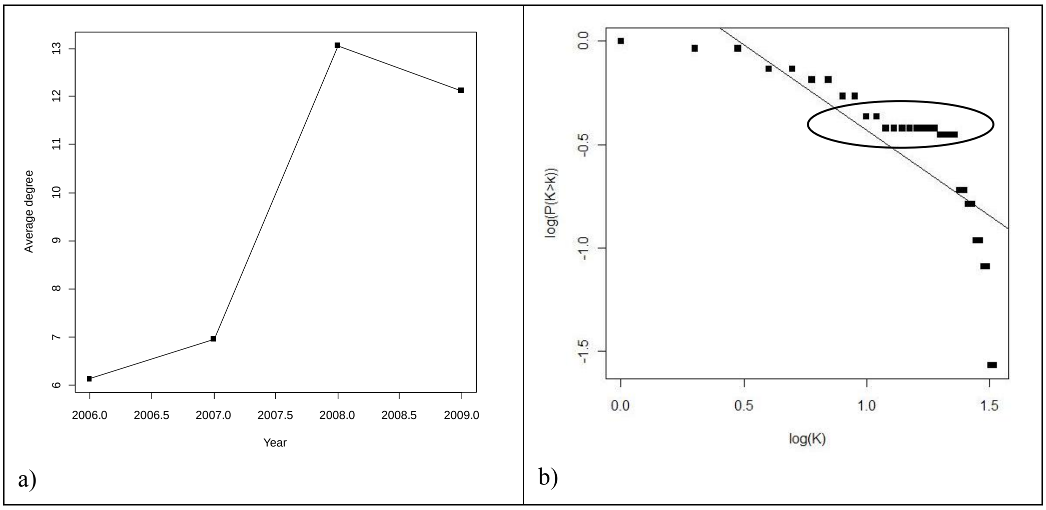


Figure 7.1- a) Average degree evolution and
b) cumulative degree distribution for OPAALS Wiki collaboration network

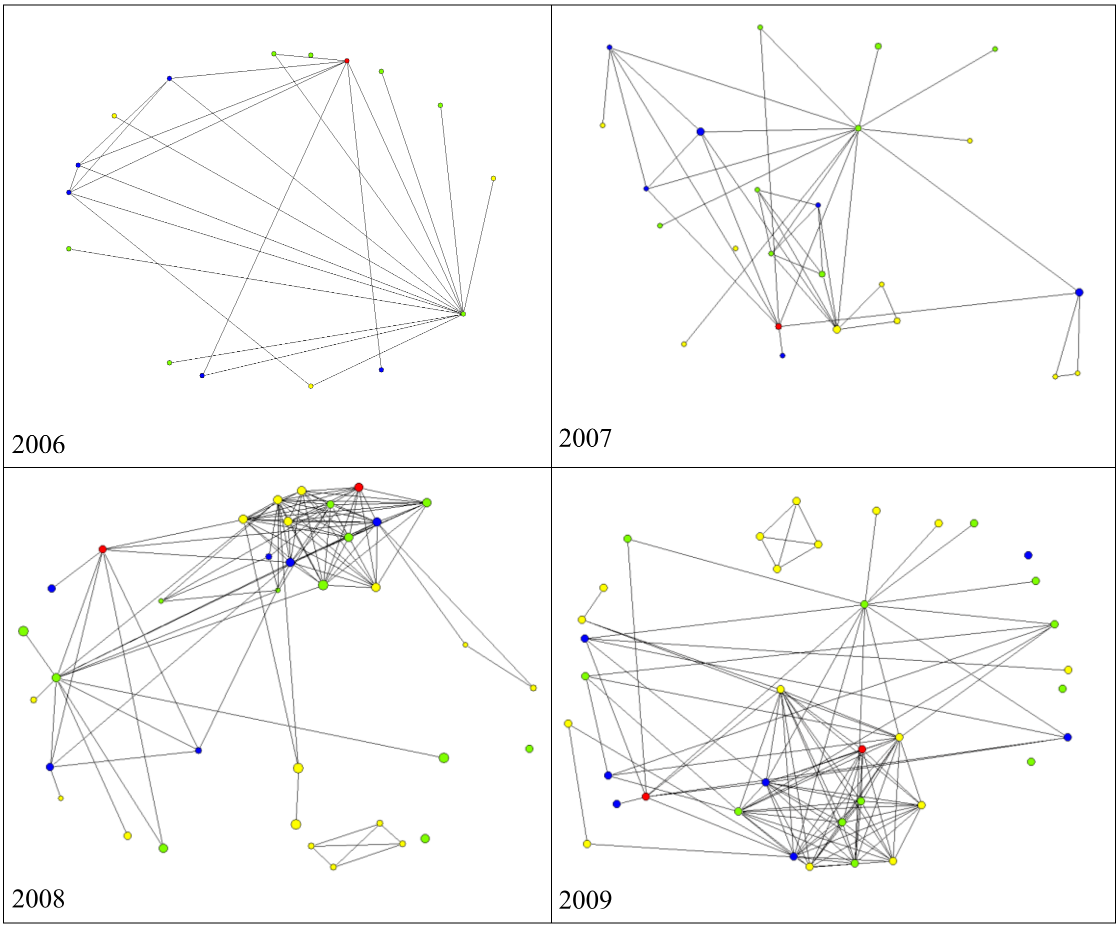


Figure 7.2- Sociograms for the OPAALS wiki collaboration evolution.
Yellow = Social Sciences, Green = Computer Science, Red= Natural Sciences, Blue = Consultants/Adm.

Table 1 describes the OPAALS wiki network, as can be seen it is not a huge network that had a major growth between 2007 and 2008, with 37 collaborating researchers. Figure 7.1 evidences that it does not follow a Power-law, but shows that there are two sets of actors in terms of degrees, above and below the horizontal line formed by the dots ("circled dots"). The set above the line are actors with few degrees, from 2 to 4. The "circled dots" comprises a few actors with degrees around from 10 to 15, and then actors with more than 20 ties.

The sociograms confirm this (Figure 7.2), and in the 2009 accumulated network, a main core is formed by 12 actors (bottom part) with high degree values and many others spread out around this core. It is noticeable also the presence of many articulators, making bridges between smaller structures with the main one. Also, the interdisciplinary characteristic is a remarkable aspect; in Fig. 7.2 each colour represents a research domain and the blue is consultants or administrative persons. This will be explored in the stochastic modelling, in chapter 7.1.

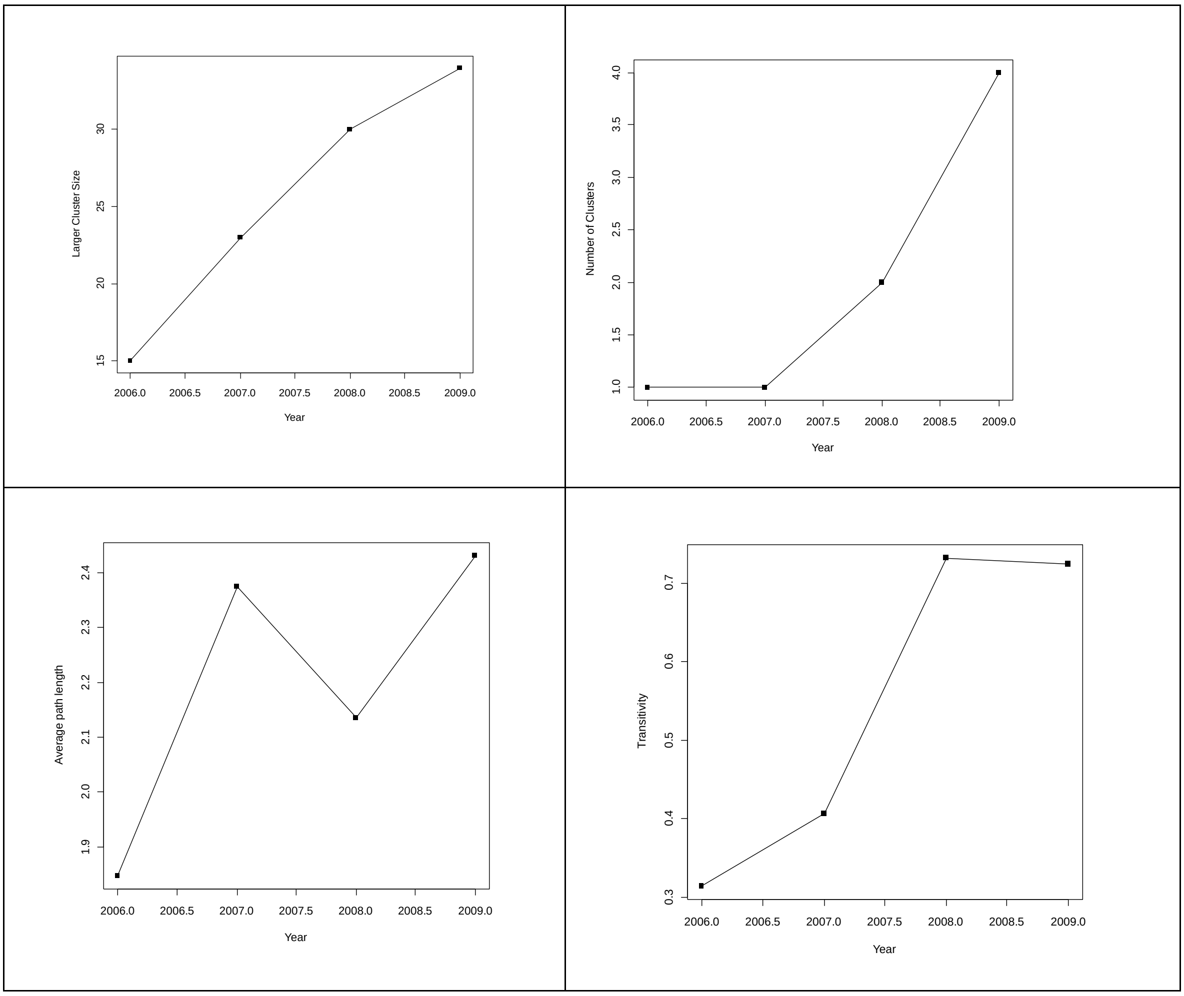


Figure 7.3- a) Larger cluster size, b) Number of clusters, c) Average path length and d) Clustering Coefficient evolutions for OPAALS wiki

In terms of evolution, the increasing transitivity (Fig. 7.3d) claims attention, since this is the unique case among all those studied here. The presence of many articulators can explain this high transitivity, increasing the probability of new ties mediated by them.

7.1- Stochastic dynamic modelling for the OPAALS wiki

Differently from the SUPERA use case, the OPAALS wiki is a non-directed network, so new objective functions are introduced, as well as their effects interpreted. A covariate was introduced representing actors' domains, as represented in the sociograms. Despite the fact that these domains have been coded as ordinal numbers, this was not taken into account in the analysis, as it was considered a categorical (nominal) variable. The rate parameters presented here, have the same meaning and interpretation as in the directed case. Four objective functions were tested for this model, firstly without any covariate effect:

1. ***Transitive ties***: is similar to the transitive triplets effect, but instead of considering for each other actor *j* how many two-paths *i* - *h* - *j* are there, we only consider whether there is at least one indirect connection. Thus, one indirect tie suffices for network embeddedness. This can be interpreted as a measure of clustering for the network, or even network closure.
2. ***Betweenness***: represents brokerage: the tendency for actors to position themselves between but not directly connected others, i.e., a preference of *i* for ties *i* - *j* to those *j* for which there are many h with *h* - *i* and *h* !- *j* ("!-"means not tied).
3. ***Number of actors at distance 2***: expresses network closure inversely: stronger network closure (when the total number of ties is fixed) will lead to smaller geodesic distances equal to 2. When this effect has a negative parameter, actors will have a preference for having few others at a geodesic distance of 2 (given their degree, which is the number of others at distance 1); this is one of the ways for expressing network closure.
4. ***Assortativity***: reflects tendencies of actors with high degrees to prefer to be tied to other actors with high degree. If this parameter is positive and significant, preferential attachment is not likely.

One covariate effect is tested, the *Identity*, in this case the Domain identity. Positive values for this effect means that actors tend to choose similes in domain, a homophily representation. Due to the small number of actors in 2006, which means that too many structural zeros are present, the analysis will comprises 2007, 2008 and 2009 only. Table 7.2 present the parameters and different models adjusted. A parameter is considered statistically significant if the *t-statistic* is bigger than |2|.

Table 7.2- Adjusted parameter for the dynamic modelling on the wiki

Effect	*Parameter*	*Std. Error*	*t-statistic*
Rate parameters			
2007 – 2008	2,77	0,33	8,34
2008 – 2009	0.02	0.02	1
Objective functions (simple models)			
Transitive triads	0.63	0.54	1.16
Betweenness	-0.56	-	-
Number of actor d2(Nad2)	-0.63	0.28	2.25
Assortativity	0.34	0.12	2.81
Multieffects Models			
Assortativity	0.34	0.12	2,81
Domain	-0.70	0.24	2,91

The estimated rates shows that in the first period, each actor had at least 3 chances to establish a tie, or at least 3 actors could collaborate with other, while in the second period this is almost 0, which is coherent with the data. Looking at the sociograms and the network evolution, few new actors entered and few new contributions were made from 2008 to 2009. So, the great part of the effects presented below happened between 2007 and 2008.

The only effects on objective functions statistically significant were the *Nad2* and the *Assortativity*. *Transitive tie* effects were not significant and *Betweenness* did not achieve numerical convergence, so that standard errors could not be computed. This situation means a lack of stability for the function used, given the dataset, i.e., it is a numerical problem rather than a network topology problem. The negative value for *Nad2* express the tendency to the network closure, such that few actors are separated at a distance two.

This is can be complementary to the positive *Assortativity*, where actors' selection is lead by the degree centrality, such that, high degree actors are more likely to tie with other high degree actors. When adjusting for *Domain*, that is negative and significant, *Nad2* looses statistical importance, but *Assortativity* remains the same. The negative sign means that ties' selection happens among different domains preferably, and this selection does not depend on the *Assortativity* since this parameter is the same in the simple and multivariate model.

To interpret this composition of effects is not an easy task, and the definition of these effects above needs to be verified. *Nad2* expresses the effect of distance 2 actors, given the degree of that actor (number of actors at distance 1). Considering *Assortativity* a degree effect, where actors tend to choose others with same degree, the chance are that distance 2 links vanishe. The interesting thing is that it does not depend on the Domain homophily; actors collaborate with others with the same intensity of collaborative activity, and not by the research domain. By having the NoE as the interdisciplinary characteristic, this is an expected effect.

8. Some remarks about the use case networks

In terms of information and communication technology, the OKS is a set of online collaboration tools such as chats, wikis, document editing tools, forums among others. The understanding of how the process of collaboration behaves in each of those virtual environments, in terms of the network formed by the researchers, can help one to follow the actual OKS dynamics, It also allows for intervention planning and strategies to keep this a sustainable community. Workshops, conferences, meetings, web-conferences are all ways to intervene in, foster or change direction in the use of the OKS. Initiatives such as PIIIKO (see Chapter 10) is a good example of how this kind of intervention, could impact behavioural dynamics as workshops influenced the Conversê network.

It is clear from table 8.1 that collaboration process differs fundamentally in the environment in which it happens. Behaviour and properties are similar for both forums and co-authoring networks. All networks that follow approximately the power-law distribution, present coefficients between 2 and 3, as found in many other studies about collaboration networks [16]. Transitivity is lower for Forums than for any other kind of text elaboration, as is suggested by the imposed and fixed treelike structure, forcing in a certain way the preferential attachment.

Then, the probability of new links among any 2 step actors is low, while for a structure free environment like a Wiki this tends to be high, as those actors have more chances to collaborate with any other actor. For the co-authoring network, the only one where ICT is not the environment, although e-mail was and still is quite useful to write in a collaborative way, the behaviour is a little different. They are truly small-worlds but not entirely scale-free networks in the examples studied. This result does not match with other findings, like Barabasi *et. al.* [18] who identified this behaviour. Here, degree distributions tended to present heavier tales due to an effect sometimes called “finite population effect”, which ca mean in this case that dataset is not complete in terms of all collaborations or authors. The dataset studied just had ISI indexed papers, so journals not indexed by this database were not contemplated.

Table 8.1- Scale-freeness and Small-worldness assessment parameters for networks in the last panel

Discipline	α	l	$R(l)$	CC	$R(CC)$
Mathematics	3.23*	3.17	0.48	0.55	**151.67**
Applied Mathematics	3.30	6.22	1.43	0.58	**237.95**
Interdisciplinary Applications	2.84	4.91	1.27	0.38	**115.23**
Computation Science	3.22*	2.95	0.54	0.70	**303.76**
Artificial Intelligence	2.76	2.66	0.59	0.63	**60.00**
Interdisciplinary Applications	2.70	1.38	0,29	0.85	**99.67**
Psychology	3.18*	3.36	0.73	0.76	**115.35**
Treatment and Prevention	3.10	2.49	0.66	0.78	**63.85**
Multidisciplinary	2.76	1.38	0.29	0.84	16.68
Supera	1.98	3.06	0.77	0.06	4.57
Converse	2.43	2.88	0.88	0.18	10.24
OPAALS Wiki	1.94	2.43	1.12	0.72	4.05

α: Adjusted exponent for the Power-law, l: Average path lenght, **R(*l*)**: ratio l / l_{random}, ***CC*:** Clustering coefficient, **R(*CC*):** ratio CC / CC_{random}
* Partially follows power-law distribution

For the two applications of stochastic models for dynamics modelling, closure was common in both cases and some covariates explained this effect. The rate parameters were very similar, around 2, which makes sense in terms of collaboration process, since opportunities to collaborate in learning and research activities are usually not so high.

9. OKS – Iterative user-led development

During the past year, we have made a concerted effort to support the partners responsible for the application layers of the OKS. By doing this, we sought to expedite the delivery of the components of the application layer that had been identified in Phase 1 and so provide for our own community the anticipated area for the capture and growth of knowledge in the community. Our approach for the implementation of the OKS has been based on the principles of online communities identified in D10.5, namely that we can only foster the use of the OKS by our own research community if components are supplied according to their specific requirements and preferences.

9.1- Refinement and prioritisation of user requirements

The activity was initiated by Thomas Kurz (SUAS) and Anne English (LSE) in Dec 2008 after agreement with the PMB. Its iterative, user-led surveying and evaluation was particularly welcomed as it upholds accountability and transparency for the entire consortium in our progression towards the opening of the OKS to the public. The action was facilitated by weekly conversations between varying constellations of Anne and Thomas with the following: Frauke Zeller (UniKassel and Social Science Coordinator), Paolo Dini (Coordinator) and Marco Bräuer and Ingmar Steinicke (UniKassel). Telephone conversations, the existing wiki and Guigoh's chat application were used as well as face-to-face meetings with the partners responsible for the application layers of the OKS namely TI and IPTI. In order to enable IPTI and TI to deliver according to these user needs, we undertook actions to:

1) identify the components already available via Sironta and Guigoh ('an inventory'),
2) ascertain the specific requirements of the community and (as they had evolved since Phase I)
3) prioritise those requirements according to the community members' needs (the latter two actions were different to the user analysis carried out in Phase 1 which was more generic in nature and not).

At least one representative from each consortium partner was contacted by Anne English and Thomas Kurz to ensure a democratic representation of the community. The output of these activities was then passed on to the relevant partners who were able to define and share implementation roadmaps for the OKS components up to the end of Phase II and for Phase III.

9.2- Testing and Validation

At the end of Phase II, a similar action supported by UniKassel was undertaken to enable users to give their feedback from testing both Guigoh and Sironta. This action was opened to Sironta as well as Guigoh so as to continue to enable the community to choose whichever environment they wished to use.

With the decision to concentrate on Guigoh for remainder of the funded lifetime of the project, efforts intensified on debugging and establishing usage on the OPAALS version of Guigoh once it was launched at the beginning of May. As the software is still beta level, a period of grace was needed to allow IPTI to stabilise the tools launched namely, chat, conference calendar, profiles networking and so on, based on feedback from various partners.

At the time of writing, preparation is underway to bootstrap community building on Guigoh at the Salzburg Phase III event. Plans have been established to continue to push and pull OPAALS community users on the OKS in Phase III and also open the OKS to SME's in order to validate the deployment of the OKS applications over the P2P architecture. These efforts will continue in Phase III to ensure continued transparency and the timely delivery of the tools required by our community in accordance with agreed timelines.

These actions have helped us identify a path for opening parts of the OKS to the public via the public website and thereby start on the path of opening the OKS to the public, also have served to help us to plan knowledge capturing and fostering actions for Phase III as well as community building actions in Phase III and indeed the use of the OKS to support integration activities via the OKS. The documentation and procedural learnings garnered in this process will ensure the availability of a reference history for future OKS users or implementers.

10. Phase III Kick-Off (PIIIKO) Workshop

From June, 24 - 25, 2009, the OPAALS Phase III kick-off event took place in Salzburg, Austria. The event was organised in the form of a workshop which enabled the participants to identify the most important objectives for Phase III of the project and to work on possible solutions in interdisciplinary groups. All preparation information, rework as well as a detailed report with the outcomes of the workshop can be found at http://dbe.fh-sbg.ac.at.

For the structure of the workshop, established interpersonal communication methods and learning techniques were chosen in order to enhance the productivity of the workshop and support the interdisciplinary knowledge transfer. The idea of the workshop was to introduce a moderation setting for the heterogeneous - interdisciplinary and/or intercultural - project team. For the basic methodologies, the Theme-Centered-Interaction (TCI) [5] and parts of the 'Reteaming' approach (see [6]) were chosen. Additionally, a main emphasis was on a common language for the interpersonal interaction in the target groups. The methodology was supported by communication exercises for making the core issues more transparent.

With the moderation of this workshop carried out by Thomas Kurz and Anne English, the goal was to have clearer leadership and facilitators for this meeting, whilst not being authorised ‘officers’. This enabled open and collaborative work across hierarchies, which was seen as very fruitful by the participants.

The main objectives of OPAALS in Phase three were identified as:

1) OKS

- Define OKS
- Functioning OKS
- Use OKS
- Content

2) OPAALS Identity

- Define ourselves (theoretical communication)

3) P2P Infrastructure

- Functioning P2P
- P2P Deployment
- Technical Integration

4) Theoretical Framework

- Interdisciplinary integration
- Glossary
- Theory of biocomputing

5) Management

- Software development coordinator
- Clarity of roles
- Improve internal communication

6) Showcase

- Field demo
- Functioning demo

7) Communication

- Strong external communication
- Teaching and training tools
- Interdisciplinary publications
- Contribution to the cluster

8) Future

- Sustainability
- Future research

For all these objectives, interdisciplinary groups worked out solution scenarios and documentation which were made available to the consortium in the reported minutes as well as in the picture gallery at http://dbe.fh-sbg.ac.at.

In order to achieve commitments also across disciplines and not only learn about but also contribute to other workpackages, each of the participants committed themselves to at least one small contribution to all eight main objectives of phase III. All contributions were sent out individually to the participants after the workshop. Additionally, the partners who did not attend the workshop were asked to commit themselves also to these contributions. The timeline and overview of the commitments can be found in the picture gallery at http://dbe.fh-sbg.ac.at.

11. Concluding Remarks

This deliverable presented the integration of two research disciplines, Social and Computer Sciences, with the Social Networks simulation and emulation as a final target, providing researchers with a new tool to study the evolution and emergency of virtual communities for collaboration and knowledge construction. Also, it delivers important analysis of reflective knowledge and self-criticism that is important for the NoE.

The integration came from the necessity to understand long-term evolution for collaborative networks, and the possibility to intervene in the evolution assessing, for example, the effects of workshops such as PIIIKO for community building i.e. in network dynamics and topology, the loss of important actors such as articulators and/or coordinators, the resilience of the network in the case of ruptures, and can be understood in terms of the sustainability of the communities, and other related opened questions. To simulate such scenarios it is necessary to understand how different kinds of networks behave, and evolve as well as the properties of specific parameters that could summarise network structures in pre-defined probabilistic models, such as the Power-Law distribution. An analogous approach is also presented in D 3.7, with the biological evolutionary processes, rather than social processes, as models for P2P infrastructure. Three types of collaboration networks were analysed and presented, and it is clear that the aim (or focus) of the network and the technology that mediates the collaboration leads to diverse scenarios that need to be studied more deeply. Once the presented pre-defined models and new features, such as the data importing, was implemented in EveSim, the consortium and the whole scientific community, will have access to a powerful toolkit to simulate and test networks and infrastructures.

In terms of collaboration, both deliverables about network models and simulations, this one and D3.7, ratify the partnership between IPTI and SUAS, as well as the OKS/ Guigoh integration.

Another important point is the knowledge generation for the self-reflection of the consortium. The analysis of the OPAALS wiki (Chapter 7) complements the analysis made by UniKassel team and presented in D10.9- “Self-reflection of community building, communication and collaboration processes in the NoE”. At the time of writing, results presented in that deliverable and the ones presented here might looks like pieces from a puzzle that still need to be matched. However, collaboration between IPTI and UniKassel in phase III is planned, and Research Exchanges of researchers from UniKassel to IPTI, will have the merging of this knowledge as the main focus. The intention is to join efforts in the understanding of the NoE, with more elaborate SNA based on content analysis from the e-mail messages and trying to include also their findings

about social relationships and technology adoption from the four panel surveys in the modelling. This collaboration was born at 2nd OPAALS Conference, in Tampere, where the Social Network Analysis Group was formed. The idea is to match the puzzle pieces and provide to the NoE a more complete and integrated way to exercise self-reflection.

12. References

[1] T. Kurz, J. Noguera, R. Eder, and T. Heistracher. Peripheral simulation framework for digital ecosystems. In 1st international OPAALS Conference on Digital Ecosystems, 2007.

[2] T. Kurz R. Eder C. Adelberger and T. Heistracher. Evaluation of two p2p approaches for a distributed simulation framework. In 2nd international OPAALS Conference on Digital Ecosystems, 2008.

[3] T. Kurz, T. J.Heistracher, F. Colugnati, A. R. Razavi, Sotiris Moschoyiannis. D3.7 – Biological and evolutionary approaches for P2P systems and their impact on SME networks. OPAALS Project, May 2008.

[4] Christoph Adelberger, Thomas Kurz, Raimund Eder, Thomas J.Heistracher. D3.5 - EvESim extensions for P2P simulation. OPAALS Project, August 2008.

[5] Schneider-Landolf, M., Spielmann J. Zitterbarth W., ed. 2009. Handbuch Themenzentrierte Interaktion (TZI). Vandenhoeck & Ruprecht GmbH&Co. KG, Göttingen.

[6] Geisbauer, W., ed. 2006. Reteaming. Methodenhandbuch zur lösungsorientierten Beratung. Carl-Auer-Systeme, Heidelberg.

[7] S. Wasserman, K. Faust, 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge, 1994.

[8] J.W. McDonald, J.W. Snith and J.J. Forster, "Markov chain Monte Carlo exact inference for social networks", Social Networks, 29, pp. 127-136, 2007.

[9] P. Sakar and A.W. Moore, "Dynamic social network analysis using latent space models", SIGKDD Explorations, 7(2), pp. 31-40, 2005

[10] Snijders, T.A.B (2005). Models for Longitudinal Network Data. Chapter 11 in P. Carrington, J. Scott and S. Wasserman (Eds.): Models and methods in social network analysis. New York, Cambridge University Press.

[11] J.H. Koskinen and T.A.B. Snijders, "Bayesian inference for dynamic social network data." Journal of Statistical Planning and Inference, 137, pp. 3930-3938, 2007.

[12] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs". Information Processing Letters, 31, pp. 7-15 1989.

[13] T.M.J. Fruchterman and E.M. Reingold, “Graph Drawing by Force-Directed Placement.” Software: Practice and Experience, 21(11), 1991.

[14] D. J. Watts (2002) Six Degrees: The Science of a Conected Age. New York: Norton & Company, Inc.

[15] R. S. Burt (2001). Structural Holes versus Network Closure as Social Capital. In: Nan Lin, Cook K, and Burt, R. S. Eds. Social Capital: Theory and Research. New York: Aldine de Gruyter, pp. 31-56.

[16] A. L. Barabási (2003) Linked. Cambridge, Plume.

[17] D. J. Watts (1999). Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton University Press.

[18] A.L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek Evolution of the social network of scientific collaborations Physica A 311, 590-614 (2002).

[19] Homans, George C. (1974). Social Behavior. Its Elementary Forms. New York, Harcourt Brace Jovanovich

[20] Newman M. E. J. & Girvan M. (2004). Finding and evaluating community structure in networks. Phys. Rev. E 69,026113.

[21] McPherson, J. Miller, Lynn Smith-Lovin and James M. Cook. (2001). Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27: 415-44.

[22] Friedkin, Noah E. (1998). A Structural Theory of Social Influence. Cambridge, Cambridge University Press.

[23] Snijders, T.A.B. (2001). The Statistical Evaluation of Social Network Dynamics. Pp. 361-395 in Sociological Methodology, edited by M.E. Sobel and M.P. Becker. Boston and London: Basil Blackwell.

[24] Snijders, T.A.B., and J.M. Huisman. (2003). Manual for SIENA version 1.98. Groningen: ICS, University of Groningen.Obtainable from http://stat.gamma.rug.nl/stocnet/ .

[25] Snijders, T.A.B (2005). Models for Longitudinal Network Data. Chapter 11 in P. Carrington, J. Scott and S. Wasserman (Eds.): Models and methods in social network analysis. New York, Cambridge University Press.