Project Number: 7PM- 266588
Project Title: SISOB
(An Observatorium for Science in Society based in Social Models)
Deliverable Type: Report
Deliverable Number: D 2.1
WP: 2
Version: 1
Date: April 15, 2011
Title of Deliverable: Literature Review
Editor: Richard Walker
Authors: Richard Walker (FrontiersIn), Judith Alexander (FrontiersIn), Beatriz Barros (UMA), Eduardo Guzman (UMA), Melissa Cochrane (FrontiersIn), Aldo Geuna (FR), Ulrich Hoppe (UDE), Raimondo Iemma (FR), Soos Sandor (MTA KSZI), Sam Zeni (UDE), José del-Campo, (UMA), Gonzalo Ramos (UMA)
Dissemination level: PU
Keywords:
Literature review, conceptual model, social network, social network analysis, social network indicators, data mining, text mining, mobility, knowledge sharing peer review, open peer review, old boy network, cronyism

Abstract:
This report provides an overview of the literature in the main areas of research relevant to the SISOB project. In each case, the review is informed by the basic conceptual model underlying the project. The paper provides a preliminary outline of the model, going on to review the concepts and conceptual tools that will drive the project: social networks, Social Network Analysis and related tools and technologies. The final section provides a review of the concepts, theories and experimental work that underlie the three SISOB case studies, dedicated to researcher mobility, knowledge sharing and peer review respectively, and illustrates how the research described relates to the SISOB conceptual model.
## Version history

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.03.2011</td>
<td>Preliminary version for revision by partners</td>
</tr>
<tr>
<td>2</td>
<td>14.04.2011</td>
<td>Pre-Final version</td>
</tr>
<tr>
<td>3</td>
<td>20.04.2011</td>
<td>Final version</td>
</tr>
</tbody>
</table>
Executive Summary

The social impact of science depends on a broad range of factors, some linked to the way scientific knowledge is produced, some to the way it is distributed to actors outside the knowledge production system, and some to the way it is received, exploited and consumed. All involve social interactions among different actors (individuals, groups, institutions) working in different contexts or settings. These interactions constitute a complex social system whose dynamics unfold on a global scale over long periods of time.

The SISOB case studies will focus on just three aspects of the system – mobility, knowledge sharing and the role of the peer review system. However, the goal is to develop concepts and tools that can also be applied outside the specific areas of research covered by the case studies.

The SISOB conceptual model conceptualizes the chain of events leading from discovery to impact as follows:

- **Actors** and Knowledge Production Networks produce Artifacts in a Context that affects what they produce and the efficiency with which they produce it;
- Distribution Actors and Distribution Networks distribute these artifacts to knowledge users;
- Knowledge users use the knowledge, directly or indirectly producing outcomes.

The general goal of SISOB is thus to develop measurements of production networks and/or production contexts and/or distribution networks and/or distribution contexts and to relate these measurements to outcomes.

The knowledge production and knowledge distribution networks cited in the SISOB conceptual model are examples of social networks, a subject of social science research since the pioneering work of Moreno in the 1930s. This and other related work contributed to the birth of a new academic discipline: Social Network Analysis.

Later in the 1960s, studies of citation networks made social networks a popular theme in scientometrics. Following up on this work, Crane used social network concepts to study informal pathways for the diffusion of knowledge through scientific communities. More recently, many web-based communities have become social networks in their own right. These developments have inspired scholars to apply computerized Social Network Analysis to scientific communities.

The goal of this report is to provide an overview of relevant concepts, techniques and tools and to relate them to the case studies. It begins by reviewing the main concepts and indicators used to analyze generic social networks and to collect the necessary data and goes on to present taxonomy of the software tools used in this work. These include tools to explore and visualize social networks, tools to extract data from the web (and other sources) and tools to extract information from this data (Knowledge Discovery from Databases).
It concludes with a review of literature relevant for the case studies. The main results are summarized below

1. **Researcher Mobility** is an important element in the formation of knowledge production and knowledge distribution networks. Movement of scientists and scientific knowledge between academic institutions, universities and society, and between different scientific fields is widely believed to be vital to further scientific quality and research development and mobility of academics has thus become a major policy goal for the European Union. The report provides a detailed review of the experimental and observational evidence underlying these assumptions and policies.

2. **Knowledge sharing** plays a key role in knowledge creation and knowledge distribution networks and can be studied using tools from Social Network Analysis. Supporters of the concept of “Mode 2 science” argue that knowledge production is a transdisciplinary enterprise involving heterogeneous groups of actors. This means that to understand the production of knowledge we need to take account of the context in which it is produced. In this setting, scientific results and products can be seen as *boundary objects*- abstract or concrete objects that are vague enough to allow collaborating actors from different social worlds to interpret them from their own perspectives but robust enough to keep their own identity despite these differences in interpretation. As such, they are important not only to scientists but also to investors, political actors, journalists etc. Against this background the report reviews, concepts and techniques for the role of these objects and the role of social networks in their generation and transmission.

3. Peer review can be defined as a system of formal evaluation in which scientific research is subjected to the scrutiny of others who are experts in the relevant field. It is commonly used to evaluate scientific papers, contributions to conferences, requests for funding, and on-going projects and, sometimes, labs, departments, and entire universities. As a result, it plays a hugely important role in determining what science is published and funded. In terms of the SISOB conceptual model, peer reviewers are *brokers* controlling access to knowledge distribution networks. The report summarizes the history of peer review and its alleged failing and biases. These include high cost, lack of transparency in the choice of editors and reviewers, cognitive biases, sexism, nationalism, “institutionalism”, nepotism and so-called cognitive cronyism. Two final sections look at attempts to reform the system (e.g. proposals for *open review*) and examine how SISOB can analyze peer review with the tools of Social Network Analysis.
# Table of contents

1. Objectives and structure of this document ........................................................................... 1
2. The social impact of science and the SISOB conceptual model .............................................. 2
3. Network science as a new source of insight ............................................................................. 4
   3.1 Social networks, semantic networks etc. ........................................................................... 4
      3.1.1 Introduction ........................................................................................................... 4
      3.1.2 Semantic Social Networks .................................................................................... 4
      3.1.3 Network Text Analysis ......................................................................................... 5
      3.1.4 Pattern based Logical Approaches ....................................................................... 6
   3.2 Network analysis .............................................................................................................. 6
      3.2.1 Introduction ........................................................................................................... 6
      3.2.2 Levels of analysis and measurements .................................................................. 6
   3.3 Methods, Tools and technologies ..................................................................................... 11
      3.3.1 Methods for identifying social networks ............................................................... 11
      3.3.2 Software for Social Network Analysis ................................................................. 12
      3.3.3 Graphical Techniques for Exploring Social Networks ........................................... 15
      3.3.4 Web Data Extraction Tools .................................................................................. 15
      3.3.5 Data mining – general concepts ......................................................................... 17
4. Applying network analysis to science – the SISOB case studies ............................................. 20
   4.1 Network Analysis and Researcher Mobility ...................................................................... 20
      4.1.1 Introduction ........................................................................................................... 20
      4.1.2 Spillover effects in science .................................................................................... 20
      4.1.3 Role of prestige .................................................................................................... 21
      4.1.4 Mobility between Academia and Industry .............................................................. 22
      4.1.5 Obstacles to mobility ........................................................................................... 22
   4.2 Network Analysis and Knowledge Sharing ...................................................................... 23
   4.3 Network Analysis and peer review ................................................................................... 24
      4.3.1 Introduction ........................................................................................................... 24
      4.3.2 The history of peer review .................................................................................... 25
      4.3.3 Strengths of current models of peer review ........................................................... 25
      4.3.4 Weaknesses in current models of peer review ....................................................... 26
      4.3.5 Reforming the peer review process ...................................................................... 30
      4.3.6 Peer review and SISOB ....................................................................................... 31

REFERENCES .......................................................................................................................... 33

Appendix A: Common Network Indicators ................................................................................. 41
   Introduction ......................................................................................................................... 41
   Basic network indicators ...................................................................................................... 41
      Indicators for the network (subnetwork) level ................................................................. 41
      Indicators of overall structure ......................................................................................... 42
      Indicators for the node/edge level ................................................................................... 42
   Indicators of community structure ...................................................................................... 43
      Subgraph statistics ........................................................................................................... 43
      Structural equivalence statistics ...................................................................................... 43
   Combined indicators .......................................................................................................... 43
      Correlates of positional features ..................................................................................... 43
      Tie or attachment types and diversity ............................................................................. 44
      Network diversity measures ........................................................................................... 44
Community profiles ...................................................................................................................................................... 44
Appendix B: Software tools for KDD ........................................................................................................................................ 45
  Open Source........................................................................................................................................................................ 45
  Commercial (DM, KDD and BI – business intelligence --) ................................................................................................... 47
  Other software relevant to SISOB (Web Reviews) .................................................................................................................. 48
1 Objectives and structure of this document

This report is the first deliverable from WP 2 (Conceptual Model). The goal of the report is to provide an overview of the literature in the main areas of research relevant to the project. Of necessity, the literature reviews for each area are informed by the basic conceptual model underlying the project. Chapter 2 thus offers a preliminary outline of the model. Chapter 3 provides a general introduction to the concepts and conceptual tools that will drive the project. Finally, chapter 4 provides a review of the concepts, theories and experimental work that underlie the three SISOB case studies, dedicated to researcher mobility, knowledge sharing and peer review respectively. In each case, the review will illustrate how the research described in the literature relates to the SISOB conceptual model.
The social impact of science and the SISOB conceptual model

The social impact of science depends on a broad range of factors, some linked to the way scientific knowledge is *produced*, some to the way it is *distributed* to actors outside the knowledge production system, and some to the way it is *received*, *exploited* and *consumed*. All involve social interactions among actors (individuals, groups, institutions) working in different contexts or settings. These interactions constitute a complex social system whose dynamics unfold on a global scale over long periods of time.

The SISOB case studies will focus on just three aspects of the system – mobility, knowledge sharing and the role of the peer review system. However, the goal is to develop concepts and tools that can also be applied outside the specific areas of research covered by the case studies. The first step is to create a common vocabulary and a common conceptual framework. SISOB thus proposes a shared *conceptual model* that will inform the research on all the case studies. At the time of writing, work on the details is still in progress. However, it is already possible describe of some of the key entities in the model.

- **Production Actors.** *Production Actors* are the agents that produce scientific/technological knowledge. An actor can be an individual, an institution or a *production network* (see below).
- **Production network.** A production network is a network of production actors who collaborate to produce scientific/technological knowledge (e.g. the members of a lab, the members of an institution, the members of a collaboration, a co-authoring network).
- **Artifact.** An artifact is a measurable output produced by an actor (e.g. a scientific paper, a patent, a prototype, a therapeutic protocol etc.).
- **Production context.** A production context consists of a set of conditions, external to the production actors, that influence the kind of artifacts actors produce and the efficiency with which they produce them (e.g. availability of financial resources, know-how, incentives for innovative behavior, career structures etc.).
- **Distribution Actors.** *Distribution actors* are agents (e.g. reviewers, writers of review articles, journalists) involved in filtering and distributing scientific/technological knowledge to *knowledge users* (see below). By their actions, distribution actors can facilitate or hinder the distribution process.
- **Distribution network.** A distribution network is a network of distribution actors (e.g. journalists who cite each other's work, reviewers who often work together)
- **Distribution context.** A distribution context is a set of conditions, external to the distribution actors, that influences the way scientific/technological knowledge is distributed to *knowledge users* (e.g. interest in science, scientific literacy, etc.)
- **Knowledge users.** Knowledge Users use scientific/technological knowledge (e.g. as the basis for new products, new techniques of production, new medical techniques)
- **Outcomes.** An outcome is an effect of new scientific/technological knowledge likely to be of interest to policy makers (e.g. improved scientific knowledge, new products/services etc., employment, improvements in public health, prestige etc.)

The model conceptualizes the chain of events leading from discovery to impact as follows:

- **Actors** and **Knowledge Production Networks** produce **Artifacts** in a **Context** that affects what they produce and the efficiency with which they produce it;
- **Distribution Actors** and **Distribution Networks** distribute these artifacts to knowledge users;
- **Knowledge users** use the knowledge, directly or indirectly producing **outcomes**.

Given this basic model we can redefine the goals of SISOB as follows.

*The goal of SISOB is to develop measurements of production networks and/or production contexts and/or distribution networks and/or distribution contexts and to relate these measurements to outcomes.*
3 Network science as a new source of insight

3.1 Social networks, semantic networks etc.

3.1.1 Introduction

The knowledge production and knowledge distribution networks cited in the SISOB conceptual model are examples of social networks, a subject of social science research since the pioneering work of Moreno[1] in the 1930s. This and other related work contributed to the birth of a new academic discipline: Social Network Analysis.

In the 1960s, studies of citation networks [2] made social networks a popular theme in scientometrics. Following up on this work, Crane [3] used social network concepts to study informal pathways for the diffusion of knowledge through scientific communities. More recently, many web-based communities have become social networks in their own right. For instance, the Twitter community uses follower/following relations to create a social network. Facebook structures information spaces through the use of ego-centered networks. These developments have inspired scholars to apply modern techniques of Social Network Analysis to scientific communities (i.e. to networks of actors engaged in the production and distribution of scientific knowledge). Mendeley[4], for example, combines citation networks and reference management software to investigate personal relations between researchers.

3.1.2 Semantic Social Networks

Mature Social Network Analysis (SNA)[see 5] uses standard mathematical definitions and techniques to characterize so-called two-mode networks or “affiliation networks”. These are essentially bipartite graphs with two types of nodes, known as actors and affiliations.
Typical examples include, e.g., network relating movie actors (actors) to the films in which they have played (affiliations), and networks relating employees (actors) to the companies that employ them (affiliations) (see paragraph 3.2). Another example is the relation between people who make posts to blogs (actors) and the threads to which they post (affiliations).

Malzahn et al.,[7]and Harrer et al.,[6]extend bipartite social networks with a knowledge dimension represented as an ontology. Peter Mika [8]begins with semantic descriptions of social/personal relations and induces relations based on shared personal interests (affiliations). In both cases, the goal is to integrate semantic and social relationships. Networks constructed in this way can be used to identify people who are interested in similar topics (which they may describe in different terms) and to uncover latent relationships between concepts and instances.

Researchers have proposed extensions to standards to accommodate the needs of this new approach. Michael Galla [9], for example, extends the Web Ontology Language OWL with social relationships and Peter Mika [10] extends W3C’s RDF standard with a social dimension. Although the authors are unaware of any Web 2.0 application that applies this kind of technology to scientific communities, it would be easy to apply in services such as researchGATE [11].

### 3.1.3 Network Text Analysis

When the SISOB conceptual model is applied to the case studies, it is often necessary to identify actors who share common interests. This for instance could be a symptom of the kind of cognitive cronyism that has been alleged to impede effective peer review (see paragraph 4.3.4). One way of identifying such shared interests is through Network Text Analysis.
This discipline originally emerged from the field of machine learning and content analysis [12] which researchers enhanced with a network dimension [13]. Text analysis identifies networks by measuring co-occurrences of concepts within a moving time window. Concepts are generalized at a low level by stemming and use of thesauri. A meta-matrix then assigns them to higher-level classes such as People, Knowledge/Resources, Events/Tasks, and Groups/Organizations.

Dynamic Text Analysis [14] is a dynamic extension to Network Text Analysis that takes account of changes in the meta-matrix. This technique can be used to track and show changes in the structure of scientific communities e.g. by comparing publication records for different years [15].

### 3.1.4 Pattern based Logical Approaches

Another way of identifying similarities among actors is to generate logic based inferences on top of lower level analyses [16]. For instance, it is possible to identify a troll in a blog from the logical pattern of always starting threads and never answering posts [17]. This is of course a very basic example. More advanced applications of logical patterns could make it possible to generate dynamic indicators representing changes in the structure of scientific communities.

### 3.2 Network analysis

#### 3.2.1 Introduction

Social network analysis (SNA) is a rapidly progressing branch of research, which is developing on the interface between several different fields of the natural and social sciences. In a recent paper in *Social Networks*, the core journal of SNA, Leydesdorff and Shank [18] demonstrates that papers in the journal are cited by papers in a broad range of disciplines including sociology, physics, computer science and mathematics. However, the same study shows that *Social Networks* is primarily a sociology journal. This conclusion suggests the need to distinguish between network analysis (NA), in the sense of a toolbox of mathematical techniques for analyzing networks, and Social Network Analysis: the application of these techniques to the study of social relations.

In what follows we provide a brief overview of key methodological issues in NA, and its application to the analysis of social networks (SNA).

#### 3.2.2 Levels of analysis and measurements

The most general questions in network analysis address the structural, graph-theoretic properties of the network. In the literature, such properties are characterized on three levels of aggregation.
3.2.2.1 Node-level measures

Most measures in NA focus on the structural features of individual elements (nodes) of the graph. Some of the most popular are concepts of node centrality that quantify the connectedness and relative position of elements. The best-known measures are degree centrality, betweenness centrality, closeness- and Eigen-centrality (in the order of observable popularity).

3.2.2.2 Global network-level measures

In many cases, structural features of the whole network are encapsulated in quantitative indices. Genuine network-level measures rely on the edge set of the whole graph (network density, number of components, isolates, path lengths). Other measures provide statistical descriptions of node-level properties characteristic of the network (degree distribution, network centrality, average path lengths etc.).

3.2.2.3 Local cluster-level measures

Several algorithms (n-cliques or clans, k-cores etc.) make it possible to detect and characterize cohesive subgroups (subgraphs, groups, modules etc.) in a larger network. Such structures are often described in terms of subgraph-level measures. Studies of sub-structures in complex networks use a range of topological measures. The clustering coefficient, for instance, expresses the group-level structural of nodes.

3.2.2.4 Positions and block models

Whereas cohesive subgroups are characterized by their local, internal connectivity, positions or blocks are subsets of nodes (actors) are characterized by the similarity of their relations to other actors or categories. This kind of positional analysis often uses so-called blockmodels. In these models, clusters of actors are located in different positions in an image matrix.

Another important notion is the concept of “brokers”[19]. Brokers are nodes in sparsely populated regions of a network that act as a bridge between clusters of densely connected components and control the flow of resources and information between these components. Brokers usually have a high betweenness centrality. In scientific communities, brokers can play a critical role in knowledge sharing. Their positional dimension is strongly related to mobility effects.

Peer reviewers connecting knowledge production networks to knowledge distribution networks can be seen as brokers (see paragraph 4.3.1)

3.2.2.5 Descriptive vs. predictive analysis

The literature on network models can be divided into descriptive and predictive approaches. Descriptive approaches model and analyze observed relational patterns describing the relevant structural properties of the model in terms of a set of indices and measures (network statistics). Predictive approaches estimate structural
features (the probabilities of links between actors in the network, structural indices) of a partially observed network from empirical data or a sample network.

3.2.2.6 Network dynamics and evolving networks

Recent years have seen a change in perspective in the NA-community, with the emphasis gradually shifting from the structure of networks towards changes in the structure of networks, that is, to dynamic aspects of network analysis. Since “real networks are dynamic” (see dynanets.org), changing and evolving through time, realistic models should be capable of coping with these phenomena. This poses a number of challenges.

Time window

Network analyses need to take account of time even when models are cross-sectional, and are not intended to express temporal changes in the domain. To extract a network structure from a real-life domain (be it social communities or scholarly communication) it is necessary to decide the time window to be sampled. For example, a single co-author, citation network, or a document similarity graph may cover publication periods of different lengths in the same analysis. In many cases, the resulting structures are dependent on the window chosen. In rapidly evolving structures, like scientific communities, it may be possible to observe multiple networks with very different densities, distributions, and community structures.

Levels of change

Other models may be genuinely dynamic or longitudinal in type. In this case, the goal is to measure changes in the relational structure of a network. This is often achieved by considering a time series of (cross-sectional) networks, representing network states at consecutive time intervals. The extent to which such series can be characterized by classic NA-concepts and measures depends on the components of the network affected by its change:

- In the “first-order” case (for example, a longitudinal study of relationships within the same group of people), the set of actors in the network remain intact, while their connections change. In such situations, changes can be characterized naturally by applying traditional measures to each member of the network series, and operationalizing change as a time series of the resulting values for each measure.
- In the “second-order” case, actors and connections vary over time and the size of the network changes over time. In these situations, network-level measures, such as density or the number of connected components, can be traced through time as before. However, continuous changes in network membership means that changes in node- and cluster-level properties are difficult to formalize. One of the most demanding challenges is tracking the dynamics of the network’s community structure and the trajectory of individual groups. For
example, in a co-word map, where groups represent topics or research directions, each consecutive time slice might be made up from different, or partially overlapping lexicons, each with its own cluster structure. In these conditions, it can be very difficult to identify trends relating initial clusters to subsequent ones and to measure within-cluster changes. For a recent approach see [15].

**Roles and positions over time**

One technique increasingly used to trace changes over time is positional analysis. Stegbauer and Rausch [20] have applied block modeling with moving time slices to online communities. In bibliometrics, Kronneger, Ferligoj and Doreian [21] have compared longitudinal co-authorship networks for biotechnology, mathematics, physics and sociology for the period 1986-2005. In contrast with Stegbauer and Rausch [20], they use sequences of time slices. Changes are visualized on image matrices. This approach is intermediate between the first-order and second-order cases discussed earlier. While the number of the blocks remains constant over time, the number of actors can change.

**3.2.2.7 Types of models: unipartite vs. bipartite networks**

Probably the most relevant dimension in practical network analysis is the selection of the type of model best suited for a particular research question. A first step is to decide whether to use directed or an undirected network, and whether to use weighted or unweighted edges. Directed networks are capable of expressing the directionality of relations (e.g. information flow in citation networks), while undirected models show symmetric relations (e.g. similarity networks of documents). Weighted graphs make it possible to express degrees of relatedness, in cases where this may be useful (e.g. in a network of documents linked by document similarity).

In modeling complex phenomena, there is a clear distinction between unipartite and bipartite graphs. Informally speaking, bipartite graphs depict affiliation networks with two kinds of actors and, implicitly, relations, while unipartite graphs involve one kind of actor and one type of relation. The scheme of a bipartite graph is A–B–C. This translates into “A and C is affiliated with B”, with A and C representing one of the kinds, and B the other. A unipartite network would encode the same relationship as A–C, read as “A is in relation with C”. The bipartite model, therefore, says two things, namely, that (1) A and C are connected, and (2) they are connected by virtue of their affiliation to B. The unipartite model only captures relation (1), without exposing the underlying factor B.

As the above outline suggests, bipartite graphs are generally considered to be of great expressive power, making it possible to represent multiple relations in a single model. Project-participant affiliation networks provide a good example [see 22, 23]. In these models the vertices represent a set of projects (type 1 vertices) and a set of participants (type 2 vertices) while the edges connect individual participants to
projects (and vice versa). It can be argued that this kind of graph is more effective than unipartite graphs in expressing complex patterns of collaboration between agents. In brief, affiliation networks are extremely useful for modeling multiple relations in complex systems [see the concept of untidy networks in 24]

3.2.2.8 Complex networks and community detection

An important distinction between Network Analysis and traditional graph theory is that the former deals with real-life phenomena, many of which involve large and complex networks with a community structure. In such models, it is possible to identify groups of actors and their associated semantics from the network architecture. An important theme of research in the NA-community is the development of criteria, methods and algorithms to achieve this. The goal is to find coherent sub-graphs in networks representing structural units (e.g. author communities in a co-author network, topics formed out of mutually related terms in a co-word map etc.).

**Partitions**

Most community detection methods are designed to partition networks – that is to cut them into pieces. These algorithms yield pairwise distinct or mutually exclusive subgraphs representing non-overlapping communities (sets of nodes) in which each actor belongs to exactly one community. The general goal is to identify a community structure in which connections between actors are far denser within than between communities. The most popular method is the so-called modularity-maximizing procedure [25], in which modularity measures to what extent different node types (clusters) are separated from each other. Another established family of approaches is based on the eigen-decomposition of the graph (its underlying adjacency matrix and its derivatives), often referred to as its spectral decomposition [A handy review of existing approaches and new directions, in the case of bibliometric networks, can be found in 26].

**Overlapping communities**

In real-life situations communities often overlap. Detecting groups with shared members requires an algorithm that is able to assign network actors to more than one subgraph in the model. The most successful solution to date is called clique percolation[27], and is implemented in the free software Cfinder. CP operates by mapping potentially overlapping, cohesive parts of the graph at different levels of integration (k-cliques). This makes it possible to detect communities with different levels of cohesion ranging from loosely connected communities up to highly integrated, dense subgroups of actors. Cfinder further elaborates the picture by generating a next-order network, in which each community identified by the software is identified as a node and overlaps between communities are represented as edges. This approach makes it possible to observe and analyze relations (overlaps) between groups.
Community detection over time

The CP method has also been applied in to the temporal dynamics of communities. Wang & Fleury [28] propose a framework addressing the influence of temporal variation by “regularizing communities by initializing sub-communities corresponding the cores and proportionality”. Palla et al., [29] formulate an extension on the CP method that uses the size of communities and their age as heuristics. Greene et al., [30] develop a model for tracking communities in dynamic social networks based on events. As in some of the cases cited earlier, the choice of time window is critical. This result is confirmed by Chakrabarti et al., [31].

3.3 Methods, Tools and technologies

3.3.1 Methods for identifying social networks

3.3.1.1 Full network Methods

The goal of full network methods is to collect information about each actor’s ties with all other actors. Full network data gives a complete picture of relations in the population. Most techniques of network analysis reply on this kind of data.

Full network data is necessary to properly define and measure many of the structural concepts of network analysis and allows very powerful descriptions and analysis of social structures. Unfortunately, it can also be very expensive and difficult to collect. Obtaining data from every member of a population, and having every member rank or rate every other member are challenging tasks for any but the smallest groups. One way of making it more manageable is to ask respondents to identify a limited number of specific individuals with whom they have ties.

3.3.1.2 Snowball methods

Snowball methods begin with a focal actor or set of actors. Actors are asked to name some or all of the other actors with which they have ties. These actors are then tracked down and asked for their ties. The process continues until no new actors are identified, or until we decide to stop (usually for reasons of time and resources, or because it is only discovering actors who are very marginal to the group we are trying to study).

The snowball method can be particularly helpful for tracking down “special” populations (small sub-sets of people mixed in with large numbers of others). However, it has two major potential limitations. First, it cannot identify actors who are not connected to other actors (i.e. “isolates”). Second, the result depends on where the snowball starts "rolling". This means the method will not necessarily find all the connected individuals in a population. Snowball approaches can be strengthened by taking special care with the selection of the initial nodes.
3.3.1.3 Ego-centric networks (with alter connections)

In many cases, it is not possible (or necessary) to track down full networks beginning with focal nodes (as in the snowball method). An alternative approach is to begin with a selection of focal nodes (egos), and identify the nodes to which they are connected. We then determine which of the nodes identified in the first stage are connected. This approach can be effective in collecting relational data from very large populations, and can be combined with attribute-based approaches. It can also yield information about the network as a whole, though not as much as snowball or census approaches.

Ego-centric networks (ego only)

Egocentric methods focus on the individual, rather than on the network as a whole. By collecting information on the connections among actors connected to each focal ego, we can get a pretty good picture of individuals’ "local" networks or "neighborhoods". Such information is useful for understanding how networks affect individuals, and can give a (incomplete) picture of the network as a whole.

Analyzing Network Data

Network data can be used to calculate properties of network positions, dyads, and networks as a whole. Properties of network positions include the number of edges associated with a node and the extent to which the node is a bridge between other nodes [32]. Dyads can vary in the strength or reciprocity of their ties, the similarity between the nodes (homophily), the content of the nodes, the number of relation types shared (multiplexity), and the number of communication media used (mediamultiplexity).

When studying properties of networks as a whole, researchers can look at such things as the proportion of dyads connected to one another (density), the average path length necessary to connect pairs of nodes, the average tie strength, the extent to which the network is dominated by one central actor (centralization)[32], and the extent to which the network is composed of similar nodes (homogeneity) or of nodes with particular characteristics, (composition). Networks can also be studied in terms of the number of ways they can be divided into subgraphs. For example, networks may consist of multiple components: sets of nodes that are tied directly or indirectly to one another but are not tied directly to nodes in other components. They may also include cliques, in which every node is tied directly to every other node.

3.3.2 Software for Social Network Analysis

Table 1[33] presents a list of software tools frequently used for Social Network Analysis and provides an outline of their key characteristics.
<table>
<thead>
<tr>
<th>Program</th>
<th>Version</th>
<th>Objective</th>
<th>Data Type</th>
<th>Missing values</th>
<th>Visualization</th>
<th>Analyses (Non-statistical)</th>
<th>Analyses (Statistical)</th>
<th>Support</th>
<th>Availability</th>
<th>Manual Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agira (Rome, 2001)</td>
<td>2.1.1</td>
<td>General</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Sequential</td>
<td>Free</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Basile (Hart et al., 2002)</td>
<td>2.0.4</td>
<td>Network dynamics</td>
<td>Complete</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Free</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>dFinder (Fahl et al., 2003)</td>
<td>2.1.4</td>
<td>Clique subgroups</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Consens (Flem, 2004)</td>
<td>2.0</td>
<td>Visual analysis</td>
<td>Complete</td>
<td>-</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>FADAT (Fischer, 1992)</td>
<td>2.0</td>
<td>Visual analysis</td>
<td>Complete</td>
<td>-</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>HipH (Benson, Heymann, and Spence, 2009)</td>
<td>2.0</td>
<td>HipH</td>
<td>Complete</td>
<td>-</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>GraVAD (Springer et al., 2000)</td>
<td>2.0.4</td>
<td>Graph analysis</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>Free</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>H㎏low (Contractor et al., 1997)</td>
<td>2.1.4</td>
<td>Knowledge network</td>
<td>Complete</td>
<td>-</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>illev (Kreis, 2002)</td>
<td>2.0</td>
<td>Network mapping</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Asses and positions</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Merged (Kreis, 2002)</td>
<td>2.0</td>
<td>Merged</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Motif (Richards et al., 2000)</td>
<td>2.0</td>
<td>Motif</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Roles and positions</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MIO (Richards, 1996)</td>
<td>2.0</td>
<td>Gilead</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Asses and positions</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>NetDraw (Jorgensen, 2002)</td>
<td>2.0.89</td>
<td>Visualization</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>NetDraw III (Furlan, 2008)</td>
<td>3.5.1</td>
<td>Visual analysis</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Roles and positions, Dyadic and tridac methods,少しと puisoは, Statistical</td>
<td>Free</td>
<td>Partial</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Netica (Venners, 2003)</td>
<td>2.0</td>
<td>Visual exploration</td>
<td>Complete</td>
<td>Yes</td>
<td>No</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Network Workbench (1999)</td>
<td>2.0</td>
<td>Knowledge network</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Dyadic and tridac methods</td>
<td>Free</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Reve (Schmutte et al., 2005)</td>
<td>2.0.1</td>
<td>Reve</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Roles and positions, Statistical</td>
<td>Free</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Papk (Chapin et al., 2005)</td>
<td>2.0.1</td>
<td>Papk</td>
<td>Complete</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>ResNet (Akkerman et al., 2004)</td>
<td>2.0</td>
<td>ResNet</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Shap (Gould, 2001)</td>
<td>2.0</td>
<td>Shap</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SI-Miner (Jin et al., 2002)</td>
<td>2.0.1</td>
<td>SI-Miner</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SPIN (Hagen, 2003)</td>
<td>2.0</td>
<td>SPIN</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>TIDE (Jorgensen, 1999)</td>
<td>2.0</td>
<td>TIDE</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VINCINT (Heer et al., 2005)</td>
<td>1.0.1</td>
<td>VINCENT</td>
<td>Complete</td>
<td>Yes</td>
<td>No</td>
<td>Descriptive, Dyadic and tridac methods</td>
<td>Free</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>STRUCLIQUE (Krum, 2001)</td>
<td>4.0</td>
<td>Structural analysis</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Roles and positions</td>
<td>Free</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>UCONNECT (Bouchet et al., 2006)</td>
<td>3.0.1</td>
<td>UCONNECT</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Roles and positions, Statistical</td>
<td>Free</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>e-motion (Erents et al., 2003)</td>
<td>2.0.1</td>
<td>e-motion</td>
<td>Complete</td>
<td>No</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Roles and positions</td>
<td>Free</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>PREPITAB (Crouch et al., 2004)</td>
<td>1.0.7</td>
<td>PREPITAB</td>
<td>Complete</td>
<td>Yes</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 1: SOCIAL NETWORK ANALYSIS TOOLS [33]

<table>
<thead>
<tr>
<th>Program</th>
<th>Version</th>
<th>Objective</th>
<th>Data Type</th>
<th>Missing values</th>
<th>Visualization</th>
<th>Analytical (Non-Statistical)</th>
<th>Analytical (Statistical)</th>
<th>Availability</th>
<th>Manual Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>AntNet</td>
<td>0.5.4</td>
<td>Tools: General</td>
<td>Complete</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNO (Bitha et al., 2003)</td>
<td>2.0.1</td>
<td>Tools: Modeling graphs</td>
<td>Complete</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Visualization</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NetworkX (Hinckley, 2002)</td>
<td>1.3</td>
<td>Tools: Structural analysis</td>
<td>Complete</td>
<td>No</td>
<td>Descriptive, Structure and location, Ethological</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Networks</td>
<td>1.4</td>
<td>Tools: General</td>
<td>Complete</td>
<td>Yes</td>
<td>Descriptive, Structure and location</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pygraphviz</td>
<td>1.0</td>
<td>Tools: Graph visualization</td>
<td>Complete</td>
<td>Yes</td>
<td>Visualization</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA (Sune, 2002)</td>
<td>0.4</td>
<td>Tools: General</td>
<td>Complete</td>
<td>No</td>
<td>Descriptive, Structure and location, Divide and Conquer methods, Relate and position, Visualization</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNAP (Leskovec, 2004)</td>
<td>11.05.06</td>
<td>Tools: General</td>
<td>Complete</td>
<td>No</td>
<td>Descriptive, Structure and location, Divide and Conquer methods, Relate and position</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vFEX (Atmokos, 2005)</td>
<td>2.8</td>
<td>Tools: Visual exploration</td>
<td>Complete</td>
<td>Yes</td>
<td>Descriptive, Structure and location, Visualization</td>
<td>Free</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.3.3 Graphical Techniques for Exploring Social Networks

There are two main approaches to constructing graphical images of social networks [34]. The first is based in a search algorithm called Multi-Dimensional Scaling (MDS). MDS requires that the investigator specify a desired dimensionality - typically, one, two or three. It then uses a search procedure to find optimal locations at which to place the points. Optimal locations are either (1) those that come closest to reproducing the pattern of the original N-dimensional social proximities contained in the data matrix (metric MDS), or (2) those that come closest to reproducing the order, but not necessarily the exact magnitudes, of the original proximities (non-metric MDS).

The second approach is based on, singular value decomposition (SVD). SVD transforms the Noriginal variables into Nnew variables, or dimensions. The variable with the most variance is always associated with the first dimension, the variable with the second most variance with the second dimension and so on.

Using MDS or SVD together with graphics display programs like MAGE or MOVIEMOL, researchers can determine whether a data set contains interesting structural features.

Several tools for the analysis of social network data (e.g., Pajek and Visone) incorporate one or more algorithms for the visualization and visual exploration of social network data. Most of these algorithms are in the MDS category.

Recently there have been several attempts to work with alternative visualization techniques, incorporate additional data, and to include temporal development of networks into the visualization. One technique is to render the state of the network at different points in time into single images which then can be viewed and explored like movies [35]. Another is to render the network information at a point in time as a two dimensional image and use the third dimension to show the development over time [36, 37]. Examples for embedding additional information into classical network views and new approaches for spatial arrangement algorithms can be found in the work of Lothar Krempel [7].

3.3.4 Web Data Extraction Tools

The SISOB case studies (see section 4) will make intensive use of data on scientific communities and artifacts, collected from the web. This will require the use of Web Data Extraction tools.

Current tools are divided into two categories: tools with a query search interface allowing them to search the “deep web” (e.g., DBMS-based sites such as Amazon) and tools that carry out “traditional” data extraction.

There are many techniques for developing Web Data Extraction Tools, some of which are summarized in Figure 2 [38].
FIGURE 2: TECHNIQUES FOR DATA EXTRACTION[38].

Figure 3 provides taxonomy of some of the most common tools.

FIGURE 3: CLASSIFICATION OF TOOLS [38]
3.3.4.1 Data extraction tools specialized in SNA

Some Social Network Analysis tools include data capture capabilities relevant to SNA. One such tool is Netminer. This is a commercial tool whose data capture features are provided by external modules. The software makes it possible to define association rules for mining relational patterns from existing data bases and provides additional analysis features.

www.netminer.com

Another commercial tool is Commetrix. Commetrix extracts data from mails, newsgroups, and discussion boards. The extraction process explicitly supports time events to extract data for dynamic analysis. Analysis and visualization components are built into the tool.

www.commetrix.de

Another tool which can extract data from web is NodeXL[39]. NodeXL is free software that can extract data from Twitter, flickr, and Youtube by using the APIs of these websites. It also provides analysis and visualization features.

http://nodexl.codeplex.com

The Data-Multiplexer-Demultiplexer (DMD) is a research prototype developed by COLLIDE research group[37] which may well be suitable for use in SISOB. DMD can extract data from mailing lists, BibTex bibliographies, discussion boards, and wikis as well as transform them to common formats for tools like Pajek and UCINET.

3.3.5 Data mining – general concepts

Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns from large datasets. At the heart of KDD is Data Mining (DM) - the use of pattern recognition and statistical techniques to discover meaningful new correlations, patterns and trends in large amounts of data stored in repositories [40, 41].
There are many different approaches in the KDD process. Maimon[40] summarizes the main issues. The knowledge discovery process is iterative and interactive, consisting of the nine steps illustrated in Figure 4. Note that the process is iterative at each step, meaning that it may be necessary to move back and forward between steps.

**FIGURE 4: THE KDD PROCESS**

### 3.3.5.1 STANDARDS

Data mining and statistical models generated by commercial data mining applications are often used as components systems for Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), risk management, and intrusion detection. In the research community, data mining is used in systems processing scientific and engineering data. Data mining standards make it much easier to integrate, update and maintain such systems[42, 43].

### 3.3.5.2 MODELS

There are two main standards for defining the models generated by data mining techniques. They are PMML and CWM-DM. The most popular is PMML. Below we summarize key features of each standard.
**PMML (Predictive Model Markup Language) [Pecther-2009][Guazelli-2009b]**

- [http://www.dmg.org](http://www.dmg.org)
- XML language for describing statistical and data mining models in an application and system independent fashion. With PMML, models can be exchanged easily between systems and persisted to repositories.
- Version 4.0 was released in May 2009 (currently v.4.0.1)
- Is being used to do extensions: An Extended Predictive Model Markup Language for Data Mining (EPMML) [Zhu-2010]
- Increasing number of applications that support PMML (see table 4.4.2)
- There exists tools to validate and convert models:
  - Zementi’s PMML converter ([http://www.zementis.com/pmml.htm](http://www.zementis.com/pmml.htm)). You can use the to validate your PMML file against the specification for versions 2.0, 2.1, 3.0, 3.1, 3.2, and 4.0. If validation is not successful, the converter will give you a file back with explanations for why the validation failed

**CWM-DM (Common Warehouse Metamodel)[Poole-2003]**

- Open industry standard defining a common metamodel and eXtensible Markup Language (XML) based interchange format for meta data in the data warehousing and business analysis domains
- Version 1.1 was released in April 2003 (currently v.1.1) [OMG-2003]

3.3.5.3 TOOLS for Statistics, Machine Learning, Data Mining or Knowledge Discovery in Databases

There are many tools and libraries for KDD. These can be grouped into two main categories: open source and commercial. Appendix B provides a list and a brief description of these tools.
4 Applying network analysis to science – the SISOB case studies

4.1 Network Analysis and Researcher Mobility

4.1.1 Introduction

Mobility is an important element in the formation and maintenance of knowledge production and knowledge distribution networks. Movement of scientists and scientific knowledge, and hence social capital, between different academic institutions, university and society, and between different scientific fields is believed to be vital to further scientific quality and research development. Mobility of academics has thus become a major policy goal for the European Union.

It is generally believed that mobility of scientists facilitates knowledge and technology transfer, creation of networks and productivity[44-46]. These assumed positive effects are related to the embedded character of scientists’ human and social capital[47, 48]. When they move, scientists increase their own value as human capital[49-54] and contribute to the receiving institution's stock of social capital [55-57]. In this perspective, it is assumed that mobility benefits both the research system and individual researchers.

The benefits of mobility for academic careers are largely due to enhancement of social capital. Mobile researchers gain access to academic networks, develop scientific contacts and widen their communication channels. Mobile researchers, moreover, receive intellectual stimuli from their new environment and enhance their personal skills. Early studies in the US suggested that mobile academics have higher levels of scholarly productivity than academics employed by their PhD awarding institution [58].

Many authors have suggested that in the US, mobility is strongly associated with scientific merit and encouraged by universities. By contrast, most countries in Europe are characterized by academic inbreeding and a reluctance of academics to move [59].

4.1.2 Spillover effects in science

Mobile researchers have been shown to create positive spillovers by enabling knowledge flows and exchange of expertise. Studies on social capital of mobile inventors have shown that links to the original location are maintained and that knowledge flows are deeply embedded in labor mobility [60-62]. Similarly, Azoulay et al., [63] find that researchers moving to a new institution increase their citations from the destination university noticeably, while citation rates from the origin institution are not affected. The evidence shows that researchers can increase their visibility and credibility by moving to a different academic environment.
A move to a new environment is likely to have a positive effect both on the productivity of a researcher and on that of the receiving department. Literature on mobile inventors has indeed found a positive link between inventor mobility and productivity [64]. On the other hand, there is very little evidence on the effect of mobility on research productivity. Some weak evidence suggests that immobility has a negative impact [58] and that post-doctoral research abroad may be especially important [65]. However, other evidence indicates that mobility may be an intrinsic characteristic of productive researchers rather than a direct contributor to productivity [66]. Some studies have shown that researchers tend to adjust their productivity to that of the department where they work [67]. However, Kim et al. [68] have demonstrated that this effect was less important in the 1990s than in earlier periods. Nonetheless, top scientists still tend to agglomerate in high ranked universities. Looking at the mobility and promotion patterns of a sample of 1,000 top economists, Coupé et al. [69] suggest that the sensitivity of promotion and mobility to production diminishes with experience, indicating the presence of a learning process.

Cañibano et al. [70] argue that it is the qualitative dimension of mobility impact that is most important. Most internationally mobile researchers are embedded in larger networks, co-operating with foreign researchers and gaining access to international sources of funding.

The positive effects on their new institutions have become more evident in recent years. In the UK, university departments have used tactical hiring as a mechanism to increase their credibility. In the case of short-term research visits, receiving institutions hope to indirectly increase their reputation in the home environment of the visiting scientist.

4.1.3 Role of prestige

Though mobility has become an important element in US and UK academic careers, it is driven by reputation and largely limited to a small group of elite universities [71]. Several papers have considered the relationship between the prestige of PhD granting and hiring institutions and researchers’ early careers. Most find that for young scientists looking for their first position, the prestige of the university where they obtained their doctorate is more important than their productivity during their PhD training [72-75]. Crane [72, 73] studied the probability of young scholars being hired by one of the top 20 departments in their field and found that the best predictor was the prestige of their doctoral program. A study of neural network researchers by Debackere and Rappa [76] suggests that prestige matters most early on in a scientist’s career, but that at later stages it no longer plays a significant role. Chan et al. [77] find that scientists coming from highly ranked PhD granting institutions have a better chance than others of finding a position in such an institution. A scientist’s choice of where to take her Ph.D thus involves tactics of professional socialization with the right choice providing access to elite networks and career advantages [78].

At later stages, the impact of institutional prestige declines and productivity becomes at least equally important [74]. Nonetheless, academia’s prestige culture still
favor elite institutions. Chan et al. [77] find that very few researchers are able to move from a lower to a higher ranked institution and that these few exceptional scientists are two times more productive than the average academic at the destination university. In a recent paper Kim et al. [68] find that in the US prestige culture and its agglomeration effects remain very strong.

4.1.4 Mobility between Academia and Industry

A special case of researcher mobility is the transition between the private and public sector. Intersector job changes are encouraged by policy makers. Dietz and Bozeman [79] have found the career patterns of academics coming from industry are very different than that of their peers.

Zucker et al. [80] have studied the probability of an academic star moving into industry and Crespi et al. [81] have investigated similar moves by inventors. Both these studies show the relevance of embedded human and social capital for knowledge and technology transfer and identify key conditions that influence academia to industry and firm-to-firm mobility.

4.1.5 Obstacles to mobility

One of the main obstacles to the mobility of scientists is academia’s elite culture. Given that researchers aim to enter the best departments and that departments try to hire the most promising staff, choice and opportunities are limited [71]. In the US and the UK, selection starts at an early stage. The best students enroll in the best undergraduate programs, gain the best degrees and in turn have the best opportunities to choose the best PhD program [71]. In continental Europe, mobility is particularly low due to tenured academic staff’s civil servant status and lifelong appointments. Different organizations tend to be very similar and competition among them is weak. As a result, a move is not seen as a possibility for advancement [59]. Opportunities for promotion depend to a large extent on social ties, further limiting opportunities for mobility [82, 83].

Intersector mobility poses a different set of problems. Marcson [84], Krohn [85], Kornhauser [86] and Hagstrom [87] have analyzed the ‘role strain’ involved in job transitions between academic and business environments. The focus of these studies is on difficulties in adaptation to new norms and patterns of behavior. In general it appears that researchers in Europe have far fewer opportunities than their US peers. The R&D labor market is not very strong [88]. While a transition to industry is possible, a return to academia is very difficult.

A short essay by Melin [89] points out an additional obstacle to mobility. In interviews with Swedish postdoctoral returners he found that 10-20% were unable to exploit the knowledge they had gained abroad and that many did not receive the recognition they had expected. This failure in knowledge transfer may be an intrinsic characteristic of Europe’s highly inflexible academic market.
4.2 Network Analysis and Knowledge Sharing

Knowledge sharing, including knowledge transfer and knowledge creation in and between teams and communities, plays a key role in knowledge creation and knowledge distribution networks and can thus be studied using tools from Social Network Analysis. Valente[90] provides an overview of existing work. One approach is to compare the evolution of knowledge within networks using methods similar to those of Rogers [91] used to trace the diffusion of innovations. For example, these methods make it possible to identify early adopters with an influential position in the network.

Recent research on knowledge creation and sharing has introduced other innovative concepts. One important approach centers on the concept of “Mode 2 science” [92]. In this view, knowledge production is a transdisciplinary enterprise involving heterogeneous groups of actors. This means that to understand the production of knowledge we need to take account of the context in which it is produced and in the particular the market for knowledge.

In this approach scientific results and products are boundary objects[93] - abstract or concrete objects that are vague enough to allow collaborating actors from different social worlds to interpret them from their own perspectives but robust enough to keep their own identity despite these differences in interpretation. As such they are important not only to scientists but also to investors, political actors, journalists etc.

In this setting, transdisciplinarity may be interpreted as multidisciplinarity within a broader field. One example of a transdisciplinary field is nanotechnology, which integrates contributions from physics, chemistry, and electrical engineering, as well as from a broad range of application areas [94]. It is obvious that this kind of field involves significant transfers, integration and re-definition of knowledge. From a methodological point of view, disciplinary heterogeneity can be addressed by mapping different networks of collaboration and cooperation, and analyzing them individually and comparatively. Heimeriks et al.[95] have demonstrated the effectiveness of this approach in biotechnology research.

Another factor that needs to be considered is the role of context in the knowledge creation process. The tacit dimension of knowledge and the influence of the context in which knowledge is externalized have been described by Nonaka and Takeuchi [96]. The recent network research literature suggests that the influence of the context, and especially institutional context, can be observed at the macro level. Wagner and Leydesdorff [97] compare the evolution of co-authorship networks in six distinct fields with a known model for network evolution (“preferential attachment”, as described by Albert & Barabasi [98]). They explain the divergences between the observed data and the model in terms of institutional constraints, going on to categorize authors into continuants, transients, newcomers and terminators. These categories, first introduced by Braun et al. [99], help to explain actors’ productivity and patterns of cooperation between them.

Nonaka and Takeuchi’s proposal leads to a model of knowledge and actor heterogeneity which can be addressed by social network analysis [100]. In this setting it is possible to model informal knowledge like “who knows what” and to
generate weighting indicators like the level of trust associated with authorities [6]. The analogy with implicit and explicit knowledge leads to models, which make a distinction between individual knowledge and common knowledge. These models allow the generation of affiliation networks in which network properties like holes or measures of individual entities like betweenness or brokerage can be used to identify factors facilitating or hindering knowledge flow. Recent studies show that there are relations between network structure and knowledge production although these relations are specific to cases and cannot be easily generalized ([101][102]).

Other recent developments include the work of Van den Besselaar & Heimeriks [103]who show that a method based on combined use of title words and cited references can provide promising two dimensional indicators representing the position of research topics on a network map for a discipline. Another approach is to apply techniques developed for the representation and analysis of semantic networks to social networks. Mika et al. [10] demonstrate that semantic technologies can be used to aggregate and analyze data from different heterogeneous sources. In a study of semantic web research, they confirm the positive effect of structural holes [19] on the innovativeness of researchers. Other insights into knowledge sharing and creation may come from studies of technologies for information exchange and communication. For instance, Bullinger et al. [104] have used qualitative analysis techniques to provide a first tentative taxonomy of Social Research Network Sites (SNRS).

4.3 Network Analysis and peer review

4.3.1 Introduction

Peer review can be defined as a system of formal evaluation in which scientific research is subjected to the scrutiny of others who are experts in the relevant field. It is commonly used to evaluate scientific papers, contributions to conferences, requests for funding, and on-going projects and, in some countries, labs, departments, and even entire universities. As a result, it plays a hugely important role in determining what science is published and funded. In terms of the SISOB conceptual model, peer reviewers act as brokers controlling access to distribution networks. While a positive peer review is no guarantee of success (see below) negative reviews block access to the network. In other words, science that fails to pass peer review is unlikely to have any impact. To the extent that careers are determined by publication and ability to attract funding, peer review determines the course of scientists’ careers, and the success of their labs and institutions. In summary, peer review is an essential element in the organization of modern science. As such it has itself become an object of study.

In the period from 1970 to 2000, scholars generated a steadily expanding literature describing its history and failings, characterizing its underlying mechanisms and making suggestions for improvement. Much of this work is ably summarized in major review articles by Campanario[105] and Armstrong[106]. Since 2000 the number of studies of peer review has tended to fall – perhaps because many originally controversial results have won wide acceptance in the community. In what follows
we refer to this later literature, wherever useful. However, we find that many recent papers do little more than recapitulate and replicate earlier findings. We also note that the majority of studies focus on reviews of journal papers and that there have been far fewer analyses of peer review by funding agencies. This is a general weakness of the literature.

### 4.3.2 The history of peer review

Several authors have written about the history of peer review [107-110]. The main facts are not in dispute and can be summarized as follows:

1. **Peer review is ancient.** Examples have been documented as far back as the 9th century C.E, when Ishap bin Ali al Rahwi, a Syrian physician, called on his colleagues to keep notes of the condition of their patients. If the patient later died the notes would be examined a by a committee of physicians which would decide if he/she had received adequate care. Forms of peer review more familiar to the modern scientists are documented as far back as 1732, when, the Royal Society of Edinburgh had instituted a select committee of scientists to decide which papers to publish in its *Philosophical Transactions*[107].

2. **Peer review only attained its current importance in the second half of the twentieth century.** Between the mid-19th century when the number of scientific papers produced first began to exceed the ability of journals to publish them, and the mid-20th century when modern practices first began to come into force, publication decisions usually depended on the policies and decisions of the editor. When, in 1905, Max Planck and Wilhelm Wien decided to publish five extraordinary papers by Albert Einstein that were to revolutionize modern physics, they made their decision without any form of external review [111]. The Journal of the American Medical Association (JAMA) did not use outside reviewers until the 1940s [107]. Lancet did not institute systematic peer review until 1976[110].

3. **The pervasive application of peer review is associated,** on the one hand with technological change (the introduction of the photocopier which made it easy to distribute papers to reviewers), on the other with the exponential growth in scientific production that began in the period after the end of World War II[107].

Today, the exponential growth in the production of scientific papers continues but new web technologies creates opportunities for new forms of peer review that conserve the strengths of the common system, while remedying its well-recognized weaknesses.

### 4.3.3 Strengths of current models of peer review

All of the papers reviewed in this study support the use of some kind of peer review to select papers for publication and (in some cases) projects for funding. The main argument is that peer review improves the quality of scientific publications [112]
Armstrong reports that “journal peer review is commonly believed to reduce the number of errors in published work, to serve readers as a signal of quality and to provide a fair way to allocate journal space”[106]. Surveys of authors and expert reviewers [113-115] confirm this widely held view.

4.3.4 Weaknesses in current models of peer review

Despite widespread belief in the essential role of peer review, and its contribution to quality, all the papers reviewed point to weaknesses in the current model. In many cases, authors use evidence from experimental and observational studies. Below we summarize the main arguments.

4.3.4.1 Time and money

The first, very common objection is that peer review is slow and expensive. Editing a journal is a time-consuming process that absorbs a huge effort from editors and their assistants. Although much of this work is unremunerated or badly paid, it nonetheless gives rise to significant administrative costs. As early as the 1970s the New England Journal of Medicine’s review process required 6-7 person years of effort and more than $100,000 of office expenditure per year[116]. Scanlan [117] reported in the early 1990s that reviews cost biomedical journals an average of $75-150 per article.

Obviously peer review process involves the work not only of editors and their assistants but that of reviewers. Reports of the mean time a reviewer takes to review a paper range from 2 to 6 hours[109]. In some disciplines, well-known reviewers may review as many as 15 manuscripts in a year. Unsurprisingly, editors report that delays in returning reviews are common. Further delays are introduced by requests for revision, which do not always contribute to the quality of the paper (see below). Although we were unable to find a systematic study of times to publication, anecdotal evidence suggests that the usual lag is more than six months and sometime more than a year. For many commentators[108, 110], this represents an unacceptable slowing down of the scientific process.

4.3.4.2 Choice of editors and reviewers

Once the editor of a journal sends an article for review, the ensuing process is highly formalized. Many commentators observe, however, that the choice of editors and reviewers is less transparent. Given that editors decide which papers are given out for review this is an important issue. To date, however, there seems to have been relatively little research on the way editors are selected.

A study by Yoels [118] suggests that editors from top American universities tend to recruit editors from the same universities. Other studies (none recent) confirm that at least in some disciplines nearly all reviewers come from a small group of top universities. It is not known however how far this tendency persists today.

A second linked issue is the choice of reviewers. Many observers have suggested that the power to choose reviewers gives editors undue influence over the outcome of
reviews[119]. Several studies have suggested that referees tend to come from the same institutions as editors[120]. Another so far unstudied issue is to what extent reviewers and authors form “cliques” with a small group persistently reviewing each other’s papers.

But, even in the absence of obvious sources of bias, there is no guarantee that the reviewers chosen are those most qualified to do the job. Campanario reports that much reviewing work is done, not by senior scientists, but by their more junior colleagues, sometimes by direct decision of the editor, sometimes, as a form of hidden “subcontracting” [109].

### 4.3.4.3 Cognitive Biases

Studies over many decades have shown that human reasoning is subject to systematic cognitive biases. Peer review process is not exempt from this tendency. Below we describe some of the reported sources of bias.

#### Preference for positive results

In a classical study, Mahoney[121] asked four groups of reviewers to referee a manuscript. The first group received manuscripts that described positive results, the second group papers with negative results; the third and fourth groups papers with mixed and no results respectively. Experimental procedures were identical in all papers.

The results were surprising. Referees who reviewed papers with positive results gave a higher score to the papers they reviewed than the other referees. Conversely, referees showed a clear bias against papers reporting negative or mixed results. Interestingly many of their criticisms focused on methodology, which, in reality, was identical in all the papers. If we accept the idea that scientific progress depends on the falsification of accepted hypotheses[122], this tendency is clearly counter-productive.

#### Difficulty in publishing replication studies

Surveys of reviews and of journal publications suggest that reviewers are unwilling to approve the allocation of scarce space and funds to studies whose main goal is to replicate a previous study [123, 124]. Many journals and funding agencies, including the European Union’s Framework Program [125] include novelty among their selection criteria for papers and articles. The result is that many results with far-reaching theoretical or practical implications have never been replicated in the open literature. Reviewer practices thus create a paradoxical situation in which replicability is considered essential for good science but actual replication is actively discouraged.
False cues to quality

An extremely useful review article by Armstrong[108] suggests that many reviewers use “false cues” in judging the quality of a paper. His examples include the following:

- **Use of statistics.** A substantial body of experimental and observational studies suggest that studies reporting tests of statistical significance are significantly more likely be published that studies that do not. For instance Atkinson, Furlong and Wampold [126] report an experiment in which reviewers were asked to evaluate three versions of a bogus manuscript that differed only in the reported statistical significance of their results. The papers with non-significant results were rejected three times more often than those whose results were reported as significant. Interestingly reviewers mostly based their rejection decisions on the experimental design, which was actually identical in all the papers.

- **Obscure writing.** An earlier paper by Armstrong[127] reports an experiment in which referees were asked to evaluate a short (fake) text, apparently extracted from the conclusions of a manuscript. In one version the text used long words and complicated sentence structure. The other used simpler vocabulary and less convoluted structures. Unfortunately for the prejudices of this author, the reviewers systematically preferred the more obscure versions of the text, which they apparently interpreted as a sign of “intelligence” and “depth” [108].

- **Other false cues.** Armstrong[108] points to other “false cues” that reviewers may use to judge the quality of a paper, in particular the use of sophisticated statistical methods and large sample sizes. Unfortunately he provides no experimental evidence that these factors actually influence acceptance or rejection of papers.

- **Confirmatory bias.** An important issue in the peer-review process is so-called confirmatory bias: the strong human tendency to give greater weight to evidence that confirms our own views than to evidence that contradicts them. Given the operational difficulty in identifying reviewers’ opinions, there have been relatively few studies of this potential source of bias. Initial evidence is presented in Mahoney [121]. A rare ethnographic study of peer review in a funding agency by Travis [119] confirms that reviewers tend to accept proposals that agree with their own scientific opinions and to reject those that do not.

### 4.3.4.4 Sexism, Nationalism, “Institutionalism” and Nepotism

- **Sexism.** Lloyd [128] shows that manuscripts with female author names have a far higher acceptance rate when they are reviewed by female rather than male reviewers (62% vs. 21%). A widely quoted study of grant awards in Sweden [129] suggests that proposals from male candidates receive systematically higher evaluations than those from female candidates. A recent follow-up study [130] shows no significant change in this tendency.

- **Nationalism and “institutionalism”:** A study by Link [131] suggests that American authors are significantly more likely to accept a paper by another American author than a paper by an author of a different nationality. Peters
and Ceci [132] report a quasi-experiment showing that papers by authors from high-prestige institutions have a significantly higher chance of acceptance than similar papers by authors with less prestigious affiliations.

- **Nepotism.** Some of the earliest criticisms of the peer review process claimed that it systematically favors some authors over others – a tendency sometimes referred to as “old-boyism”, “cronyism” or “particularism” [119]. Wade [133], cites a widely held opinion that “the peer review system consists of a group of committees whose members hand out grants to each other and to their friends”. The two studies of grant awards in Sweden cited earlier [129, 130] show that grant applicants sharing an affiliation with a member of the selection committee have a significantly higher chance of receiving an award than other applicants of the same gender with the same reported level of scientific productivity.

4.3.4.5 Ability to detect errors, plagiarism and fraud

In addition to the suggestion that peer review is subject to systematic bias, many authors have suggested that it is unable to reliably detect errors, plagiarism and fraud. Armstrong [108] quotes a long series of studies showing that a high proportion of published papers contain serious errors in bibliographical references, in interpretations of the work by other scientists, and in the use of statistics. He cites a study by Epstein (Epstein 1990) showing that in many cases, reviewers do not detect obvious cases of plagiarism. Recent cases of scientific misconduct [134] have shown that it is relatively easy for authors to deliberately mislead reviewers. This has led authoritative commentators to conclude that “clever fraud can only be detected by people working in the lab who have access to the raw data or by other labs who try to replicate (...) later” [135].

4.3.4.6 Reliability of peer review

In view of the many weaknesses described in the previous section, it is not a surprise that many studies cast doubt on the reliability of the peer review process.

**Inter-referee agreement**

It is a common experience for authors to receive radically different assessments of their work from different reviewers. Experimental studies of the peer review process show very low levels of agreement among referees. For example, in the study by Mahoney [121] cited earlier, there was very little correlation between rankings by different reviewers. Cole, Cole and Simon submitted a series of grant proposals previously reviewed by the NSF to a second panel of independent reviewers. In a large number of cases, the second review gave the opposite result to the first (acceptation of a previously rejected proposal, rejection of a previously accepted proposal). Subsequent analysis showed that most of the variance in proposal attributes could be attributed not to the inherent quality of the paper but to differences in opinion among the reviewers. In other words, “getting a research grant depends to a significant extent on chance” [136].
**Ability to predict success – inhibition of innovation**

Campanario [109] distinguishes errors by reviewers into two classes. Type I errors – where reviewers accept papers that do not merit publication and Type II errors – where they reject papers that should have been published.

Analyses of citations show that a large proportion of all scientific papers receive less than 10 citations and that only very few receive more than 50[137]. This suggests that type I errors are extremely common. Evidence for Type II errors is obviously harder to come by. Nonetheless, authors of important scientific discoveries and highly cited papers cited in Campanario [109] often report encountered serious obstacles during the review process. In some cases resistance from reviewers was such that very important discoveries (e.g. the First Law of Thermodynamics) were published in very obscure journals. These findings confirm suggestions by Mahoney [121] and later authors that the peer review process can create serious obstacles for innovative thinking and results.

---

**4.3.5 Reforming the peer review process**

The alleged weaknesses of the peer review system, have inevitably given rise to proposals for reform, several of which have been adopted by reputable journals and publishers. Below we briefly summarize some of the most popular.

**4.3.5.1 Result-blind review**

In view of reviewers’ observed “confirmatory bias”, several authors [105, 121] have argued that acceptance or rejection of a study should not depend on the results themselves but exclusively on the quality of the methodology and (possibly) the importance of the findings. This implies that reviewers should be blinded to the results section of the papers they are reviewing. To the knowledge of the authors, no journal has attempted to put this suggestion into practice. Interestingly however, the guidelines for reviewers for Frontiers Open Access journals state that reviewers should reject papers containing factual errors but should not attempt to evaluate authors’ interpretation of their results[138].

**4.3.5.2 Author-blind review**

Given the evidence for nepotism, sexism, “institutionalism” and nationalism there have been several proposals that reviewers should be blinded to the names of authors and their affiliations. Other commentators have observed, however, that reviewers familiar with their field can often identify anonymous authors from their approach, ideas, style and references[112, 139, 140]. A survey of experimental studies [141] suggests that this kind of blinding has little effect on the outcome of reviews.

**4.3.5.3 Removing the anonymity of reviewers**

Another common suggestion for reforming the review process is to remove the anonymity of reviewers. However, a review of several studies by Benos et al., [110] concludes that that the removal of reviewer anonymity makes little difference to the
quality of reviews. Understandably, authors of favorable reviews are reported to be more willing to give up their anonymity than authors of negative reviews. A number of publishers have instituted systems of open review that take this preference into account. For example, the Frontiers open review system, referred to earlier, only publishes the names of reviewers for papers that are accepted for publication. Reviewers of rejected papers remain anonymous. This policy encourages reviewers to take responsibility for the papers they select while allowing them full freedom to express negative criticism.

4.3.5.4 “Open review”

Several online journals (such as Biology Direct; Atmospheric Chemistry and Physics; Electronic Transactions on Artificial Intelligence) are experimenting with systems of “open review” [110]. To the knowledge of the authors, there has been no formal evaluation of the effectiveness of these systems.

Another common practice (especially in the physics and mathematics communities) is to publish preliminary papers on pre-print servers, where other scientists can comment and make suggestions for improvement, prior to formal submissions for publication. Several major publishers have adjusted their rules for publication to accommodate this practice (for example, Nature [142])

An alternative model of open review, introduced by the Frontiers publishing group, is to open papers for community comment after publication.

4.3.6 Peer review and SISOB

The central theme of SISOB is the way social networks influence the social impact of science. The critiques of the peer review system described in the previous paragraphs include numerous examples.

Following Travis et al.,[119] we distinguish two ways in which social networks can influence the peer review process.

1. The first is “old boy cronism”: the tendency of editors to choose reviewers of their own gender and national origin, or coming from the same institution. It is possible (though unproven) that the formalization of review procedures and modern rules on conflict of interest have reduced the importance of this “traditional” form of bias.

2. The second is so-called “cognitive cronism”: the tendency of reviewers to favor authors whose work confirms their own theories and ideas. Travis et al. suggest that this tendency may be especially pernicious in highly innovative areas of science, in which leading researchers disagree about fundamental principles.

Hypotheses about “old boy cronism” and “cognitive cronism” can be formulated in terms of hypotheses about the operation of social networks. Such hypotheses can be formulated in the general form:
The probability of outcome x is influenced by membership in network y (1)

A more specific hypothesis might be formulated as follows:

The probability of publishing a paper in journal x is increased by affiliation to institution y (old boy cronyism)

or

The probability of publishing a paper in journal x is increased by membership of a network of authors and reviewers sharing belief y (cognitive cronyism).

Or

An open review process reduces certain kinds of bias

The group that will conduct the SISOB study of peer review has access to an unprecedented corpus of data on papers submitted for publication, the origins of authors and reviewers, the results of the review process, and subsequent post-publication evaluations by readers.

Using results from this literature review, the group will formulate hypotheses concerning the way “old boy” and “cognitive cronyism” biases the review process and test these hypotheses using the data it has collected. It will go on to look at the way other aspects of relevant social networks can influence the peer review system favoring or hindering the publication of high impact work.
REFERENCES


44. OECD, Mobilising human resources for innovation, 2000, Paris: OECD.


Appendix A: Common Network Indicators

Introduction
The aims of this appendix are (1) to list the most common measures and indicators used in complex network analysis at the technical level (2) to impose a taxonomy on these concepts and (3) to prepare an overview for a subsequent Deliverable, concerning the actual meaning and use of these indicators in research on science and technology.

The appendix is organized in the following sections:

Basic network indicators. This section contains a structured list of the most common network indicators in the literature.

Combined indicators. This section proposes taxonomy of measures constructed by combining basic measures. Here we describe the typical ways in which basic network measures are combined in the literature. The interpretation and use of these tools in STS and policy studies will be discussed in a future Deliverable.

In what follows we use the following abbreviations:

<table>
<thead>
<tr>
<th>(W)</th>
<th>Weighted graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D)</td>
<td>Directed graphs</td>
</tr>
</tbody>
</table>

Basic network indicators

Indicators for the network (subnetwork) level

Distributional indicators
Statistics (distributions) of structural node, edge (link), properties and communities at the network level

Degree distribution
Centralization
Edge weight distribution (W)
Size distribution of connected components
Size distribution of Detected communities

Distribution of individual node degree values (cf. degree centrality)
Distribution of individual node centrality values (cf. various centrality indices)
Distribution of edge weight values
Cf. connected components
Cf. community structure
### Indicators of overall structure

**Single-valued indicators of overall structure**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Density</strong></td>
<td>Ratio of the number of potential and actual ties in a graph. The length of the maximum path in a graph, whereby a path between two nodes a and b is a series of ties jointly connecting a and b.</td>
</tr>
<tr>
<td><strong>Diameter</strong></td>
<td>The mean value of path lengths for a graph.</td>
</tr>
<tr>
<td><strong>Average path length</strong></td>
<td>The mean length of the shortest paths for any node, the latter calculated according to the element subgraphs, respectively.</td>
</tr>
<tr>
<td><strong>Global and average clustering</strong></td>
<td>Network-level measures of the degree to which nodes in a graph form fully connected groups (cf. density).</td>
</tr>
<tr>
<td><strong>Coefficient/Transitivity</strong></td>
<td>For a graph, the degree of containing mutually isolated subgraphs (connected components).</td>
</tr>
<tr>
<td><strong>Connectedness</strong></td>
<td>Number of dyads/triads in a graph as that of two- and three-element subgraphs, respectively.</td>
</tr>
<tr>
<td><strong>Dyadic/Triadic census</strong></td>
<td>Similarity measures for network structure in terms of matrix similarity.</td>
</tr>
</tbody>
</table>

### Indicators for the node/edge level

**Positional (centrality) measures**

<table>
<thead>
<tr>
<th>Centrality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Degree centrality</strong></td>
<td>For any node, the number of edges connecting with the node/the number of direct neighbors.</td>
</tr>
<tr>
<td><strong>Indegree centrality (D)</strong></td>
<td>For any node, the number of incoming edges/links.</td>
</tr>
<tr>
<td><strong>Outdegree centrality (D)</strong></td>
<td>For any node, the number of outgoing edges/links.</td>
</tr>
<tr>
<td><strong>Betweenness centrality</strong></td>
<td>For any node, the number of shortest paths the node is crossed by.</td>
</tr>
<tr>
<td><strong>Closeness centrality</strong></td>
<td>For any node, the mean length of the shortest paths from the node to all other nodes reachable from it.</td>
</tr>
<tr>
<td><strong>Bonacich/Eigenvector centrality</strong></td>
<td>For any node, the sum of recursively assigned scores of its ties, the latter calculated according to the principle that ties to well-connected nodes score higher than ties to nodes with fewer connections.</td>
</tr>
</tbody>
</table>

**Path length measures**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shortest path between nodes</strong></td>
<td>(cf. diameter)</td>
</tr>
</tbody>
</table>

**Node Attributes**

Non-relational properties (variables) assigned to individual nodes
Edge Attributes

Non-relational properties (variables) assigned to individual edges

Indicators of community structure

Subgraph statistics

Number, size distribution of

Connected components

Clques

(k-cliques, k-communities etc.)

Neighborhoods

Detected partitions

Detected overlapping communities

Structural equivalence statistics

Number of equivalence classes: equivalence classes are sets of nodes with identical structural features (being connected to the same nodes, having identical centrality values etc.)

Combined indicators

Combined indicators in the literature most often relate node attributes to structural node properties. Major types are:

Correlates of positional features

Measures of the associations between positional node properties (e.g. betweenness centrality of a node) and node attributes (some non-relational variable value of the same node)
**Tie or attachment types and diversity**

Measures of the relations between node properties of structurally related nodes (neighbors, nodes related by different types of paths etc.)

**Network diversity measures**

- Diversity exhibited by networks measured as some function of (1) node attributes [e.g. size] and (2) some indicator of node relatedness (e.g. distance).

**Community profiles**

- Distributional properties or characterization of all types of communities measured in terms node values
Appendix B: Software tools for KDD

Open Source

- Febrl (Freely Extensible Biomedical Record Linkage) [Christen-2009]
  - Preprocessing. It includes **a variety of functionalities required** for data cleaning, deduplication and record linkage, and it provides a graphical user interface that facilitates its application for users who do not have programming experience.

- R: Statistics
  - [http://www.r-project.org](http://www.r-project.org)
  - R-EXTENSIONS
    - [http://www.r-project.org/other-projects.html](http://www.r-project.org/other-projects.html)
  - R-WEKA. [Hornik-2009]
    - [http://cran.r-project.org/web/packages/RWeka/index.html](http://cran.r-project.org/web/packages/RWeka/index.html)
    - To make the different sets of tools from both environments available in a single unified system, an R package RWeka is suggested which interfaces Weka’s functionality to R.

- Augustus [Chaves-2006]
  - PMML 4-compliant scoring engine that works with segmented models. Open Data developed and continues to maintain a suite of analytic tools. Designed for use with statistical and data mining models. Written in python. Augustus is part of a comprehensive suite of tools focused on change detection.

- KNIME (Konstanz Information Miner)[Berthold-2009]
  - [http://www.knime.org](http://www.knime.org)
  - It is a modular environment, which enables easy visual assembly and interactive execution of a data pipeline. The user can model workflows, which consist of nodes that process data, transported via connections between those nodes
  - KNIME-EXTENSIONS
    - [http://labs.knime.org](http://labs.knime.org)

- RapidMiner [Mierswa-2006][RAPIDMINER-2009]
  - [http://rapid-i.com/content/view/181/190](http://rapid-i.com/content/view/181/190)
  - It is an environment for machine learning and data mining processes. A modular operator concept allows the design of complex nested operator chains for a huge number of learning problems. The data handling is transparent to the operators.

- Rattle [Williams-2009]
  - [http://rattle.togaware.com](http://rattle.togaware.com)
  - A Data Mining GUI for R. The Rattle package provides a graphical user interface specifically for data mining using R. It also provides a stepping stone toward using R as a programming language for data analysis
• WEKA [Hall-2009][Bouckaert-2010]
  o http://www.cs.waikato.ac.nz/ml/weka/index.html
  o The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike. It allows users to quickly try out and compare different machine learning methods on new data sets. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided. Extending the toolkit is easy thanks to a simple API, plugin mechanisms and facilities that automate the integration of new learning algorithms with WEKA's graphical user interfaces Tool for Data Mining
  o Comparison R vs. WEKA [PENTAHO-2007]
  o WEKA-EXTENSIONS
    • http://weka.wikispaces.com/Related+Projects
  o MOA [Bifet-2010]
    • http://moa.cs.waikato.ac.nz
    • Data Stream Mining
• Pentaho (based on weka)
  o http://www.pentaho.com
  o Business Intelligence Suite (Reporting, Analysis, Dashboards, Data Integration (ETL), Data Mining)
• XELOPES [PRUSYS-2009]
  o http://www.prudsys.de/en/technology/xelopes
  o Business Intelligence library. Implementations for Java, C++, and C#, CORBA and web service interfaces are currently available
• Orange [Demsar-2004]
  o http://orange.biolab.si
  o Open source data visualization and analysis. Data mining through visual programming or Python scripting
• METAL [Berrrer-2000]
  o http://www.metal-kdd.org
  o A Meta-Learning Assistant for Providing User Support in Machine Learning and Data. This research is supported by the European Union (ESPRIT Reactive LTR 26357).
• Shogun [Sonnenburg-2010]
  o http://www.shogun-toolbox.org
  o Machine Learning Toolbox. Implemented in C++ and interfaces to MATLAB, R, Octave, Python
• PyBrain [Schaul-2010]
  o http://www.pybrain.org
  o Machine learning library for Python
• Dlib-ml [King-2009]
  o [http://dlib.net/ml.html][http://dlib.net/ml.html]
  o Machine learning library for C++
• Java-ML [Aabel-2009]
In the chapter 65 of [Maimon-2011] it is a collection of modern SW. We have updated and completed that collection. Most of the tools are placed in the context of Business Intelligence and big companies in the field of DBMS usually incorporate some extension or plugin to support data mining tasks.

- **Commercial (DM, KDD and BI – business intelligence --)**

  - **Machine learning library for Java**
    - JHepWork [Chekanov-2010]
      - [http://jwork.org/jhepwork](http://jwork.org/jhepwork)
      - Machine learning library for Jython.
    - LibDAI [Mooij-2010]
      - [http://people.kyb.tuebingen.mpg.de/jorism/libDAI](http://people.kyb.tuebingen.mpg.de/jorism/libDAI)
      - Library for discrete approximate inference in graphical models. C++
    - BIRT (Business Intelligence and Reporting Tools)
      - [http://www.eclipse.org/birt](http://www.eclipse.org/birt)
      - Reporting and business intelligence capabilities for rich client and web applications, especially those based on Java and Java EE (ECLIPSE)
    - JMLR MLOSS (Machine Learning Open Source Software) from JMLR
      - [http://jmlr.csail.mit.edu/mloss](http://jmlr.csail.mit.edu/mloss)
      - Implementations of non-trivial machine learning algorithms, toolboxes, etc.

  - **Microsoft SQL Server**
    - BOOK: Smart Business Intelligence Solutions with Microsoft SQL Server 2008 ([Safari](http://www.safari.com))
    - BOOK: Microsoft SQL Server 2008 Bible ([Safari](http://www.safari.com))
  - **IBM SPSS Modeler**
    - [http://www.spss.com/software/modeler](http://www.spss.com/software/modeler)
  - **SAS Enterprise Miner [SAS-2003]**
  - **Oracle Data Miner [ORACLE-2011]**
  - **KXEN**
    - [http://www.kxen.com](http://www.kxen.com)
  - **SPM (Salford Predictive Modeling Suite)**
    - [http://salford-systems.com](http://salford-systems.com)
  - **SAND InDatabase-Analytics**
    - [http://www.sand.com/analytics/indatabase-analytics](http://www.sand.com/analytics/indatabase-analytics)
  - **Teradata Warehouse Miner**
  - **Predixion Insight**
    - [https://www.predixionsoftware.com/predixion/Products.aspx](https://www.predixionsoftware.com/predixion/Products.aspx)
  - **Megaputer PolyAnalyst**
- Inforsense
  - http://www.inforsense.com
- MicroStrategy
  - http://www.microstrategy.com
- TIBCO Spotfire Miner
- Pervasive DataRush (Apache Hadoop cluster)
- Zementis ADAPA [Guazelli-2009a]
  - http://www.zementis.com
  - Scoring engine in the CLOUD.
  - Software as a Service on the Amazon Cloud (NOT FREE)
  - ADAPA (Adaptive Decision And Predictive Analytics) Add-in for Microsoft Office Excel

**Other software relevant to SISOB (Web Reviews)**

- Software Suites for Data Mining, Analytics, and Knowledge Discovery
- Social network analysis, Link Analysis, and Visualization
- Visualization and Data Mining Software
- Network text analysis, Network dynamics and visualization:
  - http://www.textrend.org/