The handle http://hdl.handle.net/1887/20270 holds various files of this Leiden University dissertation.

Author: Calero Medina, Clara
Title: Links in science: linking network and bibliometric analyses in the study of research performance
Date: 2012-12-11
1 General Introduction
1.1 Introduction

Knowledge has always been at the core of economic growth and social welfare. The capacity to invent and innovate, to create new knowledge and new ideas that later become part of products, processes and organizations, has always fostered development. Many organizations and institutions have been effective in creating and disseminating knowledge: from the corporations of the Middle Ages to the large firms at the beginning of twentieth century, and from the Cistercian abbeys to the royal academies of science that began to appear in the seventeenth century (David & Foray, 2003).

But even though knowledge has always been important for economic development, the term "knowledge-based economy" is quite recent (OECD, 1996), and thus marks a break and introduces a discontinuity with respect to previous periods. Historical explanations of the abundance (or scarcity) of natural resources have lost much of their effectiveness in explaining disparities in productivity and growth across countries. In contrast, the improved quality of physical equipment and human capital represents a better explanation, as this relates to the creation of “new knowledge and new ideas and incorporate them into the equipment and people” (David & Foray, 2003). Since the beginning of the twentieth century a new characteristic of economic growth has been detected which consists in the growth of the share of intangible capital as compared to tangible capital (Abramovitz and David, 1996). Part of the intangible capital consists of investments in training, education, R&D activities, information and coordination; this means investments devoted to the production of knowledge and human capital.

The knowledge-based economy arises when a group of people produce and exchange new knowledge intensively with the help of information and communication technologies. Therefore, three elements may be distinguished: (1) the production and reproduction of new knowledge is taken up by a significant number of community members; (2) the community creates a "public" space-sharing knowledge movement through new information technologies; and (3) the communication to encode and transmit the new knowledge is intensive.

One of the main issues in a knowledge-based economy is to measure effectiveness in the production, measurement and use of knowledge. Therefore, in this context it is not the mere accumulation of knowledge that is important but the ability to use it in meaningful ways (OECD, 1996). However, knowledge is a concept that is difficult to quantify and/or put a price on. Traditionally, knowledge has been classified as
**General Introduction**

basic or applicable, and depending on how it is stored knowledge can also be classified as codified or tacit. According to van Raan’s (2004) definition, codified knowledge is ‘archived & publicly accessible’, and the non-codified or tacit knowledge is ‘craftsmanship’. Both codified and non-codified knowledge are essential parts of the knowledge-generating processes; codified knowledge helps diffusion and exchange, and non-codified knowledge, located in individuals, is essential to the understanding and use of the former kind.

In knowledge-based economies the science system increases in importance. Public research laboratories and universities are at the core of the science systems, a core extended to government science institutions and research councils, R&D intensive companies, and the supporting infrastructure (OECD, 1996). Consequently, the professional communities that are most engaged in the knowledge-based economy are scientific communities. These are indeed the communities in which, by definition, most members are producers of knowledge to be shared (Dasgupta and David, 1994) and that historically have always been pioneers in the use of new information technologies. Scientific production, however, embedded in the knowledge process, has a complex structure, shaped by technical and social influences (Schmoch, Schubert, Jansen, Heidler and von Gortz, 2010). This complexity is related to the trend towards multidisciplinary and interdisciplinary research, and the increased desire for and necessity of collaboration between researchers. (PREST, 2000). Thus, research is a collective effort combining diverse actors, competences and capabilities, and emphasizing the collective setting, the interface between individual researchers and research institutions (Laredo, 2003).

One of the major forms of scientific output, embedded in this complex system of scientific production, is scientific publishing. Scientific publications represent a specific but immense collection of codified knowledge that can be easily disseminated and absorbed among knowledge users. They can also be easily stored for future use. Publications also provide an important indication of what is leading-edge research, and where it is being performed (Hauser & Katz, 1998). But scientific publications themselves are also an excellent platform for studying how knowledge is shared and disseminated inside the scientific community.

The interconnections between scientific publications (e.g. citations given and received from one paper to another) and inside them (e.g. researchers co-authoring papers) allow us to study the way in which scientists create and share new knowledge by means of network analysis, which may help
to reveal the conditions behind the successful share and transfer of knowledge. Derek de Solla Price (1965) already showed the structure of science as a network of interconnected publications. In the last few decades a diverse group of scientists, including mathematicians, physicists, computer scientists, sociologists, biologists as well as information scientists, have been actively working on network theory in an effort to understand and explain its properties. In network theory terminology the number of citations of a paper is the *in-degree* of a paper, being a local property of the citation network. This quantity gives information about the characteristics of the network around the nodes, but it does not help to uncover the highly clustered structure of the scientific network. In order to understand the complexity behind knowledge production we also need to study the structure of interconnected publications, otherwise we may in fact be missing some important and crucial phenomena. Traditionally, the first approach to analyze the structure underlying a network is to make picture of it. During the last years there has been a rapid development in the field of information science applying different techniques to visualize bibliometric networks. Next to visualization techniques (‘mapping’) the structural characteristics of scientific networks can be studied using measures and metrics developed in network theory through the years. The developments in the last years in network theory is helping to incorporate these measures to the studies of scientific networks with the goal of getting to understand better the process of knowledge creation and sharing.

This thesis originated from the need to identify groups of related nodes within the collaboration and citation networks. In the study of collaboration networks the main goal is to identify research groups, potential research groups or patterns of collaboration. The analysis of citations networks through specific measures and metrics, on the other hand, makes it possible to identify main lines of research through the years. Thus, such analyses improve our understanding of the growth and decline of fields, including phenomena such as paradigm shifts and emerging research themes. Network measures and metrics also allow for the identification of important nodes (e.g., journals, articles) embedded in the citation net.

### 1.2 Network analysis

The aim of this chapter is to show the main lines of the historical developments in network theory, together with a number of representative concepts. The objective is to get a feeling for the kind of properties of networked system that can be measured or modeled and
how these properties are related to practical issues presented in this thesis.

**Historical developments**

In 1736 the mathematician Leonhard Euler took an interest in a mathematical puzzle inspired by an actual situation called the Seven Bridges of Königsberg. The city of Königsberg, Prussia (now Kaliningrad, Russia) on the Pregel River included two large islands which were connected to each other and to the mainland by seven bridges (Figure 1). The popular question at that time was whether it was possible to walk a route that crossed each bridge exactly once, and then returned to the starting point. Euler proved with a graph that this was not possible. A graph is a mathematical object consisting of points (nodes, vertices) connected by lines (links, edges, arcs). In Euler’s graph the four nodes representing the four pieces of land were connected by seven edges representing the seven bridges (Figure 2). As Newman, Barabasi and Watts (2006) explain, the bridge problem can be phrased in mathematical language as the question of whether there exists an Eulerian path in the network. An Eulerian path is a path that traverses each edge exactly once. Euler's proof is considered by many to be the first theorem in a mathematical field called graph theory, which is the main mathematical framework in which properties of networks are described (Harary, 1996).

![Figure 1. Map of Königsberg in Euler's time showing the layout of the seven bridges, highlighting the river and the bridges](image)

The strength of a graph is that the nodes and the edges can be almost anything, since many systems can be simplified to a network structure while maintaining complexity (Rosvall, 2006). The complexity can be retained because a complex system is made up of a large number of
components, or agents, interacting in such a way that their collective behavior is not a simple combination of their individual behavior (Newman, 2002). However, it is important to remark that to be able to abstract a system into its underlying network the units have to be unique, such as for instance humans, proteins, scientific publications, or web pages. A system containing interchangeable units, such as atoms or electrons, cannot be reduced to a network. By abstracting away the particulars of a problem, network theory is capable of describing major topological features with a clarity that would be impossible if all the details were retained. This is why network theory has expanded outside its original domain of pure mathematics (Newman, Barabási, & Watts, 2006).

Figure 2. Left: A simplified representation of the pattern of the river and bridges in the Königsberg bridge problem. Right: the corresponding network of vertices and edges. (Source: Newman, Barabási, and Watts, 2006)

From the 1930s the mathematical language of graph theory has been adopted by social scientists to help them to understand data from ethnographic studies (Borgatti, Mehra, Brass, & Lambiaca, 2009). Since then, social network analysis has emerged as an integrated scientific speciality concerned with the structural analysis of social interaction (Hummon & Carley, 1993). Social network analysis concentrates on the interpretation of the social nature of the nodes, and on the edges between them (Marion, Garfield, Hargens, Lievrouw, White, & Wilson, 2003).

At the beginning of the 1950s mathematicians began to think of graphs as the tool to study the spread of various ‘modes of influence’ – especially information and diseases. The structural properties of networks, particularly their connectedness, became linked with behavioral characteristics such as the expected size of an epidemic or the possibility of a global information transmission. This line of research also included
the notion that graphs should be regarded as stochastic rather than purely deterministic objects, so that graph properties can be thought of in terms of probability distributions – which is the link with the new developments in network theory in recent years (Newman, Barabási, & Watts, 2006).

The novelty of recent developments in network theory is that researchers, mainly physicists, have started to use the principles of statistical mechanics to analyze large networked structures (Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002; Newman, Barabási, & Watts, 2006). This ‘complex network theory’ mainly concentrates on analyzing degree distributions, clustering coefficients, and theoretical mathematical models to explain empirical findings. There have been many discoveries and developments in network theory during the last decade driven by the ever increasing availability of empirical data. Probably the most surprising finding is that many real networks, independent of their age, scope, and function, converge to structures with similar properties (Barabási, 2009).

**Basics of Network Theory**

Often a first step in analyzing the structure of a network is to make a picture of it. A network – ‘graph’ in mathematics- is made up of points, called nodes or vertices, and lines connecting them, usually called edges. Figure 3 shows the example of a network.

![Figure 3. Example of a network](image)

The structure of a network is described by its adjacency matrix $A$, which in the simplest case is a $n \times n$ symmetric matrix, where $n$ is the number of nodes in the network and $A_{ij}$ are the elements.
$A_{ij} = \begin{cases} 
1 & \text{if there is an edge between vertices } i \text{ and } j, \\
0 & \text{otherwise}.
\end{cases}$

$A$ is the adjacency matrix of the network in Figure 3.

$$A = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

This is the case of a symmetric matrix, since there is an edge between $i$ and $j$ and an edge between $j$ and $i$, thus $A_{ij} = A_{ji}$. The symmetric matrix represents an undirected network. If the edges have directions, i.e., if the edge points from $i$ to $j$ or from $j$ to $i$ (but not both) the network is directed. In the case of directed networks $A_{ij} \neq A_{ji}$.

The edges can also be weighted to represent stronger and weaker connections. Thus, the adjacency matrix can be generalized to values other than unity to represent the strength of the connections.

**Network measures**

An example of an important class of network measures is centrality. Centrality is a family of node level properties relating to the structural importance or prominence of a node in a network (Borgatti, Mehra, Brass, & Lambiaca, 2009). Social network analysts have studied and developed measures of centrality for a long time; as a result there is a wide range of concepts and definitions about what it means to be central to a network.

The simplest centrality measure is the degree centrality. The degree of a node in a network is the number of edges attached to it. For instance, in a friendship network between individuals the degree of a person will be the number of friends the individual has within the network.
If the edges are directed a node will have two degree measures: in-degree and out-degree. A well-known measure in a citation network is the in-degree of a paper: the number of its (received) citations. It is the basic standard measure for quantifying impact (De Solla Price, 1965; Egghe & Rousseau, 1990; Garfield, 1972; Lambiotte & Panzarasa, 2009; Wuchty, Jones, & Uzzi, 2007). On the other side the out-degree of a paper is the number of references given by the paper. A high out-degree is found, for instance, in review papers.

The corresponding global description of the network as a whole is the degree distribution, where the tail of the distribution follows power-law function in the case of citation networks (De Solla Price, 1965; Redner, 2005; Clauset, Shalizi & Newman, 2009) and co-authorship networks (Newman, 2001). Actually, many networks contain a small but important number of nodes of an unusually high degree. The effects of these nodes on the performance and behavior of network systems is one of the main lines of research nowadays, since information on these effects can help to avoid the expansion of a disease or of a targeted attack on internet (Newman, 2010).

There is a second version of the degree measure, which is more complicated but based on the same idea, called ‘eigenvector centrality’. This eigenvector centrality acknowledges that not all edges are equally important, and provides each node a centrality depending on the number and the ‘quality’ (weight) of the connections. The eigenvector centrality turns out to be a useful measure in many situations. Actually, already in the 1970s, Pinski & Narin (1976) and Geller (1978) developed a measure of journal impact based on the eigenvector centrality. Their algorithm considered not only the number of citations from one journal to another, but also the prestige of the citing journal based on the average journal impact. Journals that receive many citations from other prestigious (i.e., highly cited) journals are considered highly prestigious themselves. By iteratively passing prestige from one journal to the other a stable solution is reached which reflects the relative prestige of journals (Bollen, 2006).

This way of measuring prestige is also behind the recent PageRank algorithms used to evaluate the status of web pages. The PageRank is calculated by an iterative algorithm which analyzed prestige values from one web page to another and converges to a stable solution (Brin & Page, 1998; Page, Brin, Motwani, & Winograd., 1998; Pillai, Suel, & Cha, 2005). Kleinberg (1999) also worked on an algorithm to increase the effectiveness of web search engines, using the concepts of hubs and authorities. Hubs & authorities are formal notions of structural prominence of vertices in directed graphs (Brandes & Willhalm, 2002).
We will come back to the characteristics of hubs and authorities in next section.

Another way of measuring the central position of a node in a network is called betweenness. A node with a high betweenness centrality is a node that appears very often in the shortest path that connects any two other nodes from the network. Freeman in 1977 developed this centrality measure and is considered as a measure of the influence of a node in the network in terms of information flow.

Another interesting and well-known concept in network theory is called the small-world effect. The geodesic distance between two nodes in a network is defined as the minimum number of nodes (shortest path) one has to pass through to get from one node to another. The small-world effect shows that in most networks the mean geodesic distance between node pairs is surprisingly short compared to the size of the network as a whole. The idea was first explored mathematically by Pool and Kochen during the 1950s (Pool & Kochen, 1978), by Milgram during the 60s (Milgram, 1967), and by Watts and Strogatz during the 90s (Watts & Strogatz, 1998). The mean geodesic distance varies with the type of network, but the basic principle that you can go from an arbitrarily chosen node to any other node in just a small number of steps is well documented in a wide array of systems (Newman, 2008). The small-world effect has important consequences, for instance for the Internet. One of the reasons why Internet functions is because any computer in the network communicates by only a few “hops” over optical or electronic data lines. In practice, data packets sent over the Internet travel typically in the range of about ten to twenty hops long. The performance of the Internet network would be terrible if the packets had to make a thousand hops instead (Newman, 2010).

Another important network concept is the clustering or network transitivity (Watts & Strogatz, 1998; Watts, 1999). A network shows clustering if the probability of two nodes being connected by an edge is higher when these nodes have a common neighbor. Eckmann & Moses (2002) showed there is a close relation between highly clustered regions of a network and the existence of communities. The way a network breaks down into communities can reveal levels and concepts of organization that are not easy to see without network data, and it can help us to understand how a system is structured (Newman, 2010). The development of methods for finding communities within networks is a prosperous sub-area of the network field, with a large number of different techniques under development. However, methods for understanding
what these identified communities really mean are still in the very early
stages of development (Newman, 2008).

1.3 Bibliometric analyses

Nowadays knowledge producers, especially the public research
laboratories and universities, have to deal with different and sometimes
contradictory demands from society. They have to face unpredictable
policies in education and research mainly linked to budget reductions, as
well as an accelerating rate of knowledge growth together with the
internationalization of the knowledge process itself. The picture becomes
even more complicated if we consider that research itself is a complex
and collective effort combining various actors, competences, and
capabilities. It is essential that academics, research managers, and
policymakers stay abreast of the way research works and the impact that
science policy and research management have on research. This is the
reason why methods for the study of research performance – including
bibliometric analyses – should be conceived as an interdisciplinary effort,
aimed at integrating perspectives, insights, and findings from a series of
relevant scientific-scholarly disciplines.

Bibliometrics, the quantitative analysis of bibliographic data, plays an
important role in the study of research performance. The experience
gained by bibliometricians in the analysis of scientific publications, and
the criteria for their usefulness expressed by research management and
policy makers, keeps the bibliometric analyses in the realms of both
theoretical reflections as well as empirical research of an application-
oriented nature. The vast information contained in scientific publications,
the different analyzing techniques available, and the different questions
we want to help answer, require detailed analysis of scientific
communication.

For the purpose of the work presented here we divided the bibliometric
studies and analyses carried out at the Centre for Science and Technology
Studies (CWTS) into three research lines that are interconnected and
complement each other: (1) performance analysis based on direct counts
of citations received by publications; (2) bibliometric mapping of
science; and (3) detailed collaboration and citation analysis using the
network of linkages between publications. Two of these lines have been
at the core of the CWTS research and studies for decades now:
performance analyses primarily based on bibliometric indicators as part
of processes for the assessment of research performance, and bibliometric
mapping of science to unravel the difficult-to-classify science system and to support the assessment of research performance.

The third line, detailed collaboration and citation analysis, is the most recent. This approach originated from a need to identify groups of related nodes inside the collaboration and citation networks. Regarding detailed collaboration analysis, the main goal is to identify research groups, potential research groups or patterns of collaboration. The detailed citations analysis, on the other hand, makes it possible to identify main lines of research through the years and thus improves our understanding of the growth and decline of specific fields, including phenomena such as paradigm shifts and emerging research themes. The detailed citation analysis also allows for the identification of important nodes (e.g., journals, articles) embedded in the network.

**Performance analysis**

Performance analysis, based on publication output and citations received, is used to assess the performance of research communities. The process of citation is a complex one, and certainly does not provide an "ideal" monitor on scientific performance. This is particularly the case for a statistically low aggregation level, for instance, an individual researcher. But the application of citation analysis to the work, the "oeuvre", of a group as a whole over a longer period of time, does yield in many situations a strong indicator of scientific performance, and in particular of scientific quality. An important and absolutely necessary condition is that applied citation analysis is part of an advanced, technically highly developed bibliometric method. Bibliometric indicators are used to assess the research output of countries, universities or research institutions, and departments or research groups (Moed, De Bruin & Van Leeuwen, 1995). The work done by Garfield (1979), Martin & Irvine (1983), Narin (1990), Van Raan (1997), and Schubert, Glänzel & Braun (1989) shows the importance and strength of the performance indicators when it comes to assessing the output of a research unit.

Performance has three central aspects: activity, productivity, and impact. Connecting the scientific output of a research unit to the number of citations received (in-degree in the citation network) provides us with an indicator of impact, influence, or at least visibility (Noyons, 1999). CWTS has been working for many years on improving the bibliometric indicators and adapting them to the specific demands of researchers, research managers, and policy makers (van Leeuwen, 2004). Many studies support the use of the bibliometric methodology developed at CWTS for assessing performance in different fields such as physics.
(Rinia et al., 2001), biology (Nederhof & Visser, 2004), electrical and electronic engineering (Van Leeuwen et al., 2000), chemistry (Van Leeuwen et al., 2003), humanities (Nederhof, 2006; Tijssen et al., 2006), medicine (Tijssen et al., 2002), and social and behavioural sciences (Nederhof, 2006).

**Bibliometric mapping of science**

Each year about a million scientific articles are published. How to keep track of all these developments? Are there specific patterns ‘hidden’ in this mass of published knowledge, at a ‘meta-level’, and if so, how can these patterns be interpreted (Van Raan & Noyons, 2002)? Structuring science is about identifying fields, sub-fields, and research themes and relating them to each other. The mapping of science by means of co-word and co-citation approaches has also been part of bibliometric studies for a long time (Braam, Moed & van Raan, 1991a, 1991b; Callon, Law, & Rip, 1986; Chen, 2003; Garfield, Pudovkin, & Istomin, 2003; Small, 1999; Tijssen & van Raan, 1989). This became necessary because the traditional science classification system is imperfect, especially for highly multidisciplinary environments, and it helps to assess performance.

The data behind science mapping are bibliometric networks and until now this technique has been used mainly with co-occurrence networks (based on keywords in publications) and with co-citation networks (based on citations received and given by publications, authors or journals). Because these maps usually cover many publications, a simple network representation, i.e. a set of nodes and edges, is of no use since the human eye can not catch the information in a big and dense network graph (Newman, 2010). This is why more advanced visualization techniques that allow the representation of the network data in comprehensible maps are used. At CWTS the work carried out through the years (i.e., Noyons, 1999; van Eck & Waltman, 2007, van Eck & Waltman, 2010) shows the importance of this procedure as a research management and science policy tool.

**Collaboration and Citation Analyses**

- **Detailed Collaboration Analysis**

It is often said that in recent decades there has been a sharp increase in the number of papers that involve collaboration among researchers detrimental to papers without collaboration (Hicks and Katz, 1996). Part of the reason for this increase in the proportion of collaborative work, lies in the need for more specialized and concentrated resources, together with an increase in interdisciplinarity (Gibbons et al., 1994). Moreover,
numerous studies have highlighted the positive relationship between research productivity and quality on the one hand, and collaboration between many researchers on the other (e.g., Lawani, 1986, Peters & van Raan, 1994). In addition, the characteristics that influence the intensity of collaboration are various, depending, for instance, on scientific discipline, institutional level, or geographic level (local, national or international) (Katz and Martin, 1997).

Bibliometric analyses play an important role in measuring these tendencies, and a number of studies have been carried out since the 1990s. The co-authorship data were used in many studies to measure collaboration (e.g., Persson & Beckmann, 1995; Martin-Sempere et al., 2002; Melin & Persson, 1996; Bordons & Gomez, 2000; Van Raan, 1998; Seglen & Aksness, 2000). From around 2000 several researchers began the construction of large-scale networks using co-authorship data in mathematics (Barabási et al., 2002); biology, physics and computer science (Newman, 2001); and neuroscience (Barabási et al., 2002).

During the last decade network researchers have been working to reveal the highly clustered nature of scientific production, showing that co-authorships networks are made up of several dense groups of nodes, called ‘communities’ (Lambiotte & Panzara, 2009).

Our daily work with research managers in highly interdisciplinary research centers shows the need for new approaches to help them reorganize their centers, which often are still organized in traditional, disciplinary ways. The novelty of our approach is that we have combined different methods in order to identify communities and functional or potential research groups. Regarding collaboration analysis we used two techniques developed in network theory:

- A technique to identify ‘regions’ between the nodes, called k-core. A k-core is a subgraph in which each node is connected to at least a minimum fixed number (k) of the other nodes in the subgraph (Seiman, 1983). The k-core approach allows actors to join the group if they are connected to k members, regardless of how many other members they may not be connected to (Wasserman & Faust, 1994).

- The Girvan and Newman algorithm to identify the communities and groups based on co-publication networks (Girvan & Newman, 2002; Newman, 2004; Newman & Girvan, 2004). This Girvan-Newman uses the edge betweenness measure as the basis of their algorithm. Based of the same idea of the node betweenness developed by Freeman (see section 1.2), the edge betweenness of an edge measures the times an edge is used in the shortest paths that connect two other
nodes from the network. The edges that connect highly clustered communities have a higher betweenness so cutting these edges should separate communities. The method finds divisions of networks into closely knit groups by looking for the edges that connect groups (Lusseau & Newman, 2004).

The work presented here shows how we can identify communities and groups ('functional research groups', see Seglen & Aksnes, 2000; Calero et al., 2006) by using network analysis techniques to analyze collaboration data and combine the results with other bibliometric techniques (bibliometric mapping of science and performance analysis). This approach may lead to a better understanding of how complex interdisciplinary organizations work and may therefore support research managers to reorganize their organization in a more efficient and practical way.

- **Detailed Citation Analysis**

Citation network analysis began with the study by Garfield, Sher & Torpie (1964) of Asimov’s history of DNA. Isaac Asimov described in his book “The Genetic Code” the major scientific developments that enabled the duplication in a laboratory of the protein synthesis process under control of DNA. Garfield and colleagues created a citation network taking as starting point the papers where the main milestones mentioned by Asimov where published and the citations between these papers as links. They showed that there was “a high degree of coincidence between an historian’s account of events and the citation relationship between these events”. In terms of citations, the representations of fields or areas of specialization are not just ‘formless’ sets of articles. On the contrary, they represent sets of papers with a particular structure that emerges from the citation practices of the researchers active in that field. They emphasize the importance and visibility of certain theoretical and methodological approaches while marginalizing others. We could say that citation practices represent a “knowledge-construction” process that outlines the manner in which we think about and engage with our research.

In all scientific fields there are key concepts that form the basis for theoretical developments through the years. Researchers from the same specialty tend to cite each other in order to position their work in the field on the basis of previous knowledge. Scientific knowledge is assumed to increase over time following a “smooth path”; the papers that introduce important new insights are cited until they are modified or contradicted by new results. The scientific revolutions, i.e., sudden paradigmatic
changes resulting from new insights (Kuhn, 1969), are reflected by abrupt changes in the citation network. In this context, following Small (1978), a cited document stands for a concept. Highly cited documents have a significant content that is shared by a community of scientists.

The citation network has enabled us to analyze the data from two perspectives in terms of time: longitudinal studies and cross-sectional studies. The techniques of longitudinal network analysis show the changes over time in the connectedness of the system. In the evolution of knowledge, phases of consolidation of past results coexist with the exploration of new approaches. One of the techniques is called main path analysis. The second technique is a cross-sectional analysis of the citation network at a well-defined time, using a specific algorithm to identify prominent nodes in the citation network called hyperlink-induced topic search (HITS). These two perspectives are important because they highlight different parts of a citation network.

**Main Path**

The main path analysis makes it possible to unravel the dynamics of convergence and divergence between ‘investigation streams’ (Ramlogan et al., 2007). If knowledge flows through citations, a citation that is a necessary step in many paths between many articles is more important than a citation that hardly plays any role in linking articles (De Nooy, Mrvar & Batagelj, 2005). Among all possible “chains” of citations, from the most recent to the oldest, the network algorithm computes the paths that are most frequently encountered. These paths can be regarded as the backbones of a research tradition (Hummon & Doreian, 1989, 1990; Hummon & Carley, 1993; Batagelj, 2003; De Nooy, Mrvar & Batagelj, 2005). These results identify the path that is most frequently used to ‘walk’ from the present to the past (back in time) in a ‘field’ of papers; this path is called the ‘main path’. It is important to stress that this method does not involve the absolute count of maximum numbers of citations received, but the simultaneous computations of all possible paths through the whole dataset and the choice of the one that is most frequently used through time (Mina et al., 2007).

**HITS - Hubs and authorities**

The concept at the basis of ‘hubs’ and ‘authorities’ in a network can be dated back to Pinski and Narin (1976). They proposed to measure the prominence of scientific journals by taking into account not simply the number of citations that a journal receives, but also the prestige (in terms of citations received) of the journals that cite it. Journals that receive many citations from prestigious journals are considered highly
prestigious themselves and, by iteratively passing prestige from one journal to another, a stable solution is reached which reflects the relative prestige of journals (Bollen et al, 2006). This way of measuring prestige is the basis of the algorithms for evaluating the status of web pages developed by Brin and Page (1998) and Kleinberg (1999). Kleinberg (1999) constructed a ‘centrality’ algorithm to increase the effectiveness of web search engines. This algorithm, called hyperlink-induced topic search (HITS), is based on the idea that there are two types of important nodes in a directed network: hubs and authorities. Hubs and authorities are formal notions of structural prominence of vertices in directed graphs (Brandes & Willhalm, 2002). The algorithm gives each node in a network an authority centrality and a hub centrality. A hub is a node with a large number of links (hub centrality). A node with high authority centrality is one that has many links with hubs, i.e., many other vertices with high hub centrality. The characteristic of a node with high hub centrality is that it points to many nodes with high authority centrality (Newman, 2010).

Authorities are nodes that contain useful information on a topic of interest; hubs are nodes that tell us where the best authorities are to be found. An important scientific paper (in the authority sense) is one that is cited by many important reviews (in the hub sense). On the other hand, an important review is one that cites many important papers. However, “ordinary” papers can also have high hub centrality if they cite many other important papers, and papers can have both high authority and high hub centrality. The reviews, too, may be cited by other hubs and hence have high authority centrality as well as high hub centrality (Newman, 2010).

As an example, Figure 4 below shows the citation network between 10 papers labeled by the publication year (ten different years). The lines (directed edges) show the citation relation between the papers. The direction of the arrow indicates if a paper is cited by (receiving an arrow) the paper on the other side of the line. Notice how the citation flow is related with the year of the publication, this means that for instance the paper from 2005 can not be cited by the paper from 2001. From the citation network the nodes with two highest authority centrality measures and the two highest hub centrality measures have been highlighted. The 2002 paper (blue square) is one of the two papers with the highest authority centrality while the paper from 2008 is one of the two papers with the highest hub centrality measure. The paper from 2005 (the yellow diamond) is the paper that has the other highest hub centrality measure and the other highest authority measure. As the graph shows the paper from 2005 is not only citing to many papers but it is citing also to the one from 2002 that it is an important paper (in the authority sense), and at the
same time the paper from 2005 is not only cited by many papers but one of them (2008) is an important paper (in the hub sense).

![Citation Network between 10 papers](image)

**Figure 4. Citation Network between 10 papers**

For the software Pajek, Batagelj and Mrvar adapted Kleinberg’s hubs/authorities algorithm. The results from the analyses presented in this thesis are based on Pajek.

### 1.4 Linkages in bibliometric studies and thesis outline

The work presented in the next chapters shows how in the past years, due to the increasing need to understand the collaboration and citation process and the developments in network theory, we have developed a third line of research embedded in the bibliometric analyses and combined with the other two lines. In Figure 5 we illustrate this combination of the three lines of bibliometric analysis.

In general terms we could say that performance analysis with bibliometric indicators yields *specific output- and impact-related information* about a specific entity (e.g., country, university, department), whereas science mapping yields *more general structural information* about a research field. Our analyses allow us to link the two approaches. As Noyons, Luwel & Moed (1999) have shown, the mapping procedure can improve the performance analyses so that the performance in a research field may be investigated in more detail. For instance, the position of a research institute on the map. On the other hand, the performance indicators contribute to the validation of the structures
obtained by means of the science maps. Still, there was a gap between the specific and the general analysis. A gap an “in-between” of what happens in the networked system in terms of levels, concepts and ‘natural’ communities within organizations that are not easy to see without analyzing publications using network measures and metrics to understand how the scientific system is structured. The above also explains the difference in bibliometric practice between the more general structural analysis by science mapping, and the more detailed structural analysis by the network approaches.

![Detailed Collaboration and Citation Analysis](image1)

![Bibliometric Mapping of Science](image2)

*Figure 5. The three main CWTS bibliometric analyses*

The research described in this thesis aims to establish the use of detailed collaboration and citation analysis combined with other forms of bibliometric analysis as a tool enabling a better understanding of the organization of scientific communities and the way knowledge is spread inside scientific communities. In this perspective there are three key questions that we address in this thesis:

*Can we identify communities, research groups and potential research groups?*
The answer to this question is crucial for helping research managers and policymakers to organize complex organizations in a more efficient and practical way.

**Can we identify main lines of research through the years, and the articles that linked them into a research tradition that can be considered as the backbone of the field?**

The answer to this question depends on a better understanding of the growth and decline of specific fields, including phenomena such as paradigm shifts and emerging research themes.

**Can we identify important nodes that play a key role in the citation networks?**

The identification of important nodes (e.g., journals, articles) embedded in the network is related to understanding how information flows.

In Chapter 2 and 3 we present two approaches to identify research groups in a particular research field, or inside an organization. Both approaches deal with the complex issue of the position of research groups within a changing structure of scientific research. In particular, in Chapter 2 we identify and classify clusters of authors to represent research groups by means of a combination of bibliometric science mapping techniques and detailed network-based collaboration analysis. We present two types of outcomes: actual research groups and potential research groups. The former enable us to define research groups beyond the formal organizational structure, and the latter can be used to identify potential partners for collaboration.

In Chapter 3 we combine data on bibliometric indicators with detailed collaboration analysis to examine the formal organization of a University Hospital. Allowing the co-publication network itself to identify communities and groups inside this interdisciplinary research centre (in fact a kind of self-organization) may lead to a better understanding of how this complex organization works and how to reorganize research in a more efficient and practical way.

In Chapter 4 we present a study on research cooperation within multinational enterprises (MNE) in the bio-pharmaceutical industry. We use the publications of the MNEs to examine structural factors characterizing research cooperation networks within the industry at the level of major geographical regions (North America, Europe, Pacific-
Asia), with a breakdown into within-MNE and between-MNE network linkages.

In Chapter 5 we present bibliometric characteristics of the world and European universities with the largest scientific output in terms of publications. We compare US universities with European institutions for a number of different aspects, for instance countries with a strong concentration of academic research activities within a core group of universities and countries with a more even distribution of research over institutions. We present a ranking of universities based on indicators calculated for all research fields with a ranking for just one specific field (Oncology). Here we distinguish between general, broad universities and specialized universities. We also present results for rankings based on a single indicator with collaboration maps combining network analysis and a series of indicators.

In Chapter 6 we present a study in which we combine bibliometric science mapping based on co-word network analysis and a specific analysis inside the citation network to investigate the process of knowledge creation and dissemination through scientific publications. We analyze the citations of a very influential paper that introduced a term in a field to identify the articles that influenced the research for some time and to link them to a research tradition that can be considered the ‘backbone’ of the field.

In Chapter 7 we present a method based on the application of network theory to citation networks in order to identify the most important journals related to a given journal, the ‘seed journal’. In just one simple network map we can sketch the relevant citation environment of these seed journal. This approach is of interest to publishers, librarians, scientists, and to science policy makers.

Finally, in Chapter 8 we summarize our conclusions and illustrate the prospects for future research.
References


