



UNITED NATIONS
UNIVERSITY

UNU-MERIT

Working Paper Series

#2007-022

**Knowledge Flows, Patent Citations and the Impact of
Science on Technology**

Önder Nomaler & Bart Verspagen

Knowledge Flows, Patent Citations and the Impact of Science on Technology

Önder Nomaler & Bart Verspagen

Eindhoven University of Technology
o.nomaler@tm.tue.nl, b.verspagen@tm.tue.nl

June 2007

Abstract

Technological innovation depends on knowledge developed by scientific research. The number of citations made in patents to the scientific literature has been suggested as an indicator of this process of transfer of knowledge from science to technology. We provide an intersectoral insight into this indicator, by breaking down patent citations into a sector-to-sector matrix of knowledge flows. We then propose a method to analyze this matrix and construct various indicators of science intensity of sectors, and the pervasiveness of knowledge flows. Our results indicate that the traditional measure of the number of citations to science literature per patent captures important aspects intersectoral knowledge flows, but that other aspects are not captured. In particular, we show that high science intensity implies that sectors are net suppliers of knowledge in the economic sector, but that science intensity does not say much about pervasiveness of either knowledge use or knowledge supply by sectors. We argue that these results are related to the specific and specialized nature of knowledge.

Keywords: knowledge input-output analysis, knowledge flow matrices, science-to-technology transfer

This research has benefited from support under DIME, which is a project in the European Union 6th Framework Programme.

UNU-MERIT Working Papers
ISSN 1871-9872

Maastricht Economic and social Research and training centre on Innovation and Technology,
UNU-MERIT

UNU-MERIT Working Papers intend to disseminate preliminary results of research carried out at the Centre to stimulate discussion on the issues raised.

1. Introduction

Technological change results in largest part from investments, such as R&D, by commercial firms. Among the inputs that firms use to produce technological knowledge is knowledge generated in the ‘science’ sector, e.g., from universities and public research organizations. The interaction between the (public) science sector and the (private) firm sector is seen as an important determinant of the technological competitiveness of firms and, at a higher aggregation level, regions and countries. For example, it is an often-held policy view that an important reason why Europe lags behind the United States in terms of technological performance, is that the interaction between the science and technology spheres is less developed in Europe than in the United States (Dosi, Llerena and Sylos Labini, 2006, summarize this argument and discuss it critically).

Knowledge flows, or interaction in a more general sense, between science and technology takes many different forms, each associated with specific channels and types of knowledge (Cohen, Nelson and Walsh, 2002). For example, knowledge may be transferred by means of personal contacts at conferences and workshops, or by mobility (change of jobs) of researchers, by (graduate) students, by joint research projects, or by publication channels such as scientific articles and patents. With regard to the relative importance of these channels or sources, Cohen, Nelson and Walsh (2002: 14), reporting on the outcome of a survey among R&D managers in US firms, conclude that “publications/reports are the dominant channel, with 41% of respondents rating them as at least moderately important”.

One way of quantifying the impact of the ‘publication channel’ on technology development is through the use of citations by patents to scientific publications (Narin, Hamilton and Olivastro, 1997). This makes use of the need for patents to cite the ‘state-of-the-art’ with regard to the invention described in the patent. An important part of this state-of-the-art is provided by means of citations, either to other patents or to so-called non-patent literature. The latter often are citations to scientific articles or handbooks. Narin, Hamilton and Olivastro (1997) count the frequency of such non-patent literature citations, and trace the nature and geographical origin of the cited works. They conclude that the ‘science intensity’ of patents has increased over time, as evidenced by a rise in the average number of citations to science in a single patent, that the nature of the citation links is often geographically biased (patents tend to cite science from the same country), and that there are substantial differences between technology fields with regard to science intensity.

The number of ‘science references’ per patent has now become a standard way of quantifying the impact of science on technology (e.g., Tamada et al., 2006, Leydesdorff, 2004, Hicks et al., 2001). Meyer (2002) classifies citation analysis as one of the three available methods for quantifying the science – technology link (the other two methods are looking at industrial science publications and university patenting), and argues that it is the most widely used of the three (p. 197).

The use of citation analysis to measure the science – technology linkage is, to our knowledge, limited to the use of citations in patents to non-patent literature. Citations in patents to other patents are sometimes used as a frame of reference (e.g., benchmark the average number of citations to non-patent literature against the average number of citations to patents), but what is usually disregarded is the second- and higher-order effects that may occur when citations to non-patent literature propagate forward when the patent that makes the citation to science is cited by other patents. It is our aim in this paper to provide a method of analyzing this citation process, taking account of ‘direct’ citations, as well as the ‘indirect’ effects that occur as a result of the forward propagation described above. In other words, our aim is to provide a method that provides a more complete impression of the science – technology linkages than is traditionally obtained by only looking at ‘direct’ citations.

Our proposed method draws inspiration from existing literature that uses patent citations to study technological interdependence between economic sectors. This issue has been addressed in the economic literature by studying so-called technology flow matrices (e.g., Scherer, 1984, Johnson and Evenson, 1997, Verspagen, 1997, Los, 1999). Usually, these are used to construct measures of so-called ‘indirect R&D’ (i.e., technology spillovers), and to relate this to productivity growth. We introduce citations to non-patent literature in such technology-flow matrices, and utilize a number of methods that are broadly known in input-output analysis (Miller and Blair, 1985) in order to quantify the influence of science on different economic sectors.

We start our discussion, in Section 2, with a general conceptual framework of how technology flows operate. This sets the general context of our theoretical approach, and links it to an observable database (i.e., patent citations). Section 3 discusses the general nature of our database, and the way in which patent citations can be interpreted as indicators of technology flows. In Section 4, we provide a formal theoretical framework for assessing the science – technology linkages at the sectoral level. Our approach is based on an aggregation of the citations data and an analytical abstraction that draws inspiration from input-output economics. Section 5 presents the empirical indicators that we derive from the methodology. In Section 6, we review trends in the database, and investigate to which extent the assumptions we made in the formal theory are justified by the data. Section 7 presents the results. Finally, in Section 8, we summarize the argument and conclude.

2. A graph-theoretic view of technological change

The aim of our analysis is to build a theoretical model of the flows of ideas in the inventive process. Invention, or innovation more broadly, can be seen as a process that takes labour, capital and prior knowledge as inputs, and produces new knowledge. We limit ourselves here to the part of this process in which prior knowledge contributes to the development of new knowledge, and hence do not consider the role of capital and labour in the inventive process. Our perspective is based on the idea of a network graph, in which new ideas (patents) are drawn from previous ideas (patents or pieces of scientific knowledge). We will use patent citations to indicate the relationships between ideas.

The view of patent citations as a network graph rests on five assumptions: (i) the complete knowledge domain can be divided into two broad categories, *science* and *technology*, (ii) within each of these two broad categories, knowledge can further be distinguished into different types or *fields* (e.g., electronics and mechanical engineering in technology, and physics and biology in science), (iii) we can usefully analyze the technology part of the system without considering its inputs back into the science part¹, (iv) knowledge is cumulative: prior (accumulated) knowledge embedded in a set of patents, is propagated forward if this patent is cited, (v) on average, the magnitude of the knowledge transmitted forward by a single citation is constant across citations.

Our notion of a network of ideas is illustrated in a stylized example in Diagram 1, which displays a network of knowledge flows (patent citations). The nodes in this network (the squares, circles and triangles) represent pieces of knowledge (ideas), and the arrows connect-

¹ The latter assumption is perhaps the most controversial, as it seems to suggest a ‘linear’ view in which science impacts on technology, but the reverse impact (from technology to science) is absent (see, e.g., Kline and Rosenberg, 1986 for a critique of such a view). Although we are sympathetic to a more interactive view of the relationship between science and technology, we are willing to accept the assumption. The main reason is that our data, which are patent citations, allow us to observe the inputs of science (the scientific literature) into technology (patents), but not the reverse.

ing them illustrate the cumulative relations between the ideas. Thus, for example, idea A4 is an input into idea A2, which in turn is an input into idea A1. The different types of knowledge are represented by different symbols (Type A by a circle, Type B by a square, and Type S by a triangle). Type S represents ideas that are in the realm of science, and, according to assumption (iii) do not have any inputs from the other two types of ideas. These types, A and B, represent different fields in technology.

Relations between fields are selective and specific, i.e., some types of relations are more frequent than others. For example, in the diagram the composition of inputs into Type A ideas is different from those into type B ideas. Idea A1 takes as input two ideas of Type A (i.e., from its own field), one idea of Type B, and two ideas from science. Idea B1 takes as inputs one idea of Type A, one of Type B and one from science.

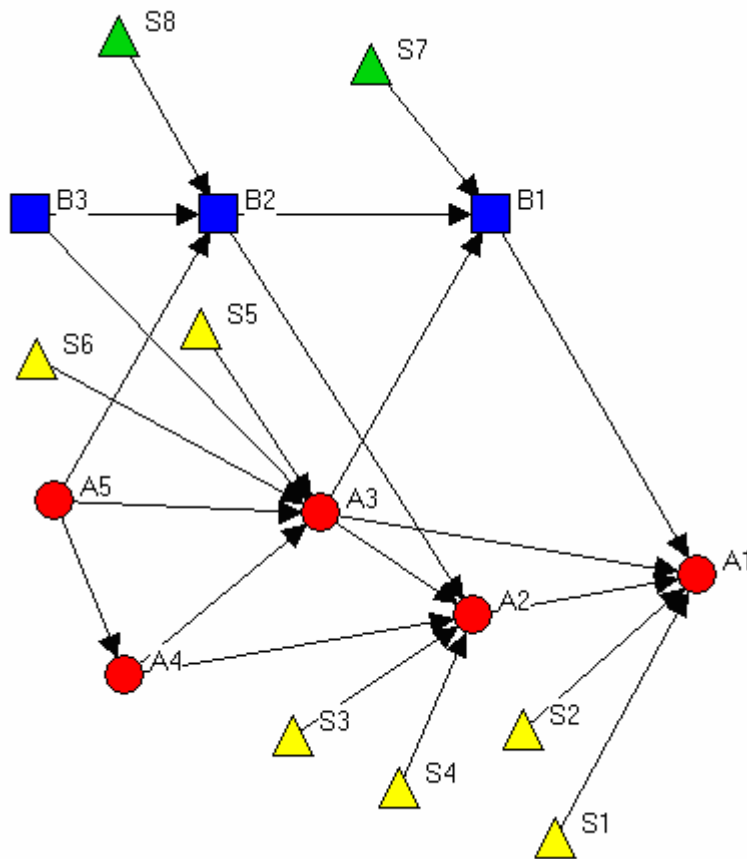


Diagram 1. Stylized patent-citation network graph

The issue that interests us is whether some types of technology have a higher dependence on, or input from science, than other types of technology. Traditionally, (e.g., Narin, Hamilton and Olivastro, 1997) this is measured by the number of citations to science, per patent. In terms of the network, this means that the number of incoming links from triangles is counted. But the diagram clearly brings out that there are also indirect inputs from science. For example, the impact of idea S5 is immediate in the development of idea A3, but A3 in turn leads to A1, A2 and B1. Hence S5 is ‘embodied’ somehow in 4 ideas in the technology realm, something that its shares with idea S6. Conversely, when we look at idea A1, it builds directly on S1 and S2, and indirectly on all other science ideas (S3-S8). The ideas of Type B generally show a much lower number of science ideas embodied in them. Note also that these indirect relations pan over the borders of technology sectors, i.e., the green science ideas that flow

into sector B patents will eventually also flow into ideas of sector A, and similarly the yellow science ideas also reach both sectors.

In order to assess the ‘science content’ of technology fields, or sectors, we need a map of the actual network. As we will explain in detail below, our approach will be to construct such a stylized map on the basis of patent citation data. Our empirical implementation assumes that citations between patents capture the links between ideas of type A and B (technology) in Diagram 1, and that citations in patents to non-patent literature capture inputs from the science realm to the technology realm. We assume that knowledge flows from the cited patent (or science reference) to the citing patent. Section 3 will provide a more detailed discussion of the nature of patent citations and how we capture them.

Although patent databases are large (e.g., the main database that we will use has approximately 1.6 million patents), modern computers allow us to analyze the actual citation network graph that results from this, and hence we would be able to make direct observations based on such a micro-account. However, this suffers from particular limitations in the available data, related to the fact that the actual databases that we have are truncated in various ways, and hence that we can only observe particular sub-parts of the whole knowledge flows network. Two types of truncation are relevant: in time and between patent systems.

With regard to time-truncation, we have left- and right-truncation. With right-truncation, the problem is that we can only observe patent citations up to the most recently published patent. For this patent, we know where its (direct) inputs came from (what it cited), but we do not know into which other patents (ideas) it will become an input. In Diagram 1, ideas A1 and B1 are examples of right truncation. Left-truncation in time results from the fact that patents and patent citations were not recorded from the beginning. For example, our patent data start in 1979, and no citations to patents prior to this year are available. In Diagram 1, ideas B3, A4 and A5 are examples of this. We observe the patents that cite these patents, but we do not observe what they cite themselves, and we cannot be sure that there are no other patents (published before the left truncation) that cite these patents.

An additional truncation problem occurs because patents can be filed under different national or international patent systems (associated with different patent offices, e.g., the European Patent Office, the US Patent and Trademark Office, USPTO, or other national patent offices). In the example of Diagram 1, it may be the case that patents A1 and B2 are filed with the USPTO, and the other patents are filed with the EPO. In reality, such citations between the different patent systems are frequent.² The truncation problem arises because we have only information on patent characteristics of patents of a single patent system (the EPO), and we lack complete information on patents in the other systems. Thus, in terms of Diagram 1, if patents A1 and B2 are indeed outside the EPO system, we do not have specific information on them (e.g., we do not know their field of origin). Obviously, this distorts our picture since, for example, we cannot observe where one of the inputs into A2 comes from, or where A2 sends one of its outputs.

In order to avoid these truncation problems, we implement a more aggregate (sector-level) approach to mapping the knowledge flows network. This essentially consists of constructing for a single point in time a set of probabilities that a knowledge flow emerges between two sectors, and assume that these probabilities are constant in time, so that we can extrapolate them (we will test for the assumption of constant probabilities). This approach is based on the methods developed in input-output economics. In short, we avoid the truncation problems by sampling the data rather than summing it up. The sample is based on a particular generation

² A further complication results from the fact often one idea is filed under different patent systems, resulting in two ‘varieties’ (e.g., a USPTO and an EPO) of the same patent. The procedure we use to construct our patent citations dataset takes this into account, and standardizes such cases to the single EPO variety of the patent. Details are given below.

of patents and their backward linkages (citations). We define such a generation of patents as all patents belonging to a particular year. The network representation that we build (described in Section 4) is based only on the direct citation inputs into this generation of patents, but it assumes that the indirect inputs (i.e., the citations made by the cited patents) can usefully be described by the same probabilities as observed in the single generation.

3. Patent citations: measurement and interpretation

We have already briefly discussed how we will use patent citations as a representation of the flows in our technology network (Diagram 1). Although the use of patent citations has by now become quite widespread in the literature (see, e.g., the overview of contributions in Jaffe and Trajtenberg, 2002), there are certain problems with this particular interpretation. Before we actually proceed to develop a theoretical framework and use it for empirical analysis, we briefly discuss these issues here.

Central in our approach is the notion of a patent citation. But, of course, patent citations were not introduced to facilitate the economic analysis of science and technology. Instead, the (legal) purpose of the patent citations is to indicate which parts of the described knowledge are claimed in the patent, and which parts other patents have claimed earlier. From an economic point of view, however, the assumption is that a reference to a previous patent indicates that the knowledge in the latter patent was in some way related to the new knowledge described in the citing patent.

Authors like Jaffe et al. (1993) and Maurseth and Verspagen (2002) have argued that the citation link can be interpreted as a knowledge spillover, i.e., an externality for the citing party. However, we are not specifically interested in the notion of knowledge spillovers, but instead in the broader notion of technology flows (i.e., flows irrespective of whether they represent and externality in the economic sense), and hence accept patent citations as a broad indicator of knowledge relatedness and flows.

We will use only citations between European patents (including international patents under the PCT system filed through EPO), i.e., we will only consider patent citations where both the citing and cited patent are applied for at the EPO. Besides a practical reason (we do not have information on patents in other systems than the US and EPO systems), there is also a more fundamental reason to limit our citations information to the EPO patents. This is the fact that there are major differences between citation practices at the two patent offices. In the USPTO system the applicant, when filing a patent application, is requested to supply a complete list of references to patents and non-patent documents that describe the state-of-the-art of knowledge in the field. In the EPO system, the applicant may optionally supply such a list. In other words, while in the US there is a legal requirement and non-compliance by the patent applicant can lead to subsequent revocation of the patent, in Europe it is not obligatory. As a result, applicants to the USPTO “rather than running the risk of filing an incomplete list of references, tend to quote each and every reference even if it is only remotely related to what is to be patented. Since most US examiners apparently do not bother to limit the applicants’ initial citations to those references which are really relevant in respect of patentability, this initial list tends to appear in unmodified form on the front page of most US patents.” (Michel and Bettels, 2001, p. 192). This tendency is confirmed by the number of citations that on average appear on USPTO patents. Michel and Bettels report that US patents cite about three times as many patent references and three and a half times as many non-patent references compared to European patents. Thus, our strategy of using only EPO citations implies that we take a more conservative view of knowledge flows.

In more specific terms, we use data from European patent applications³ to analyze technology flows. Our data are extracted from the Bulletin cdrom issued by the European Patent Office (EPO), and from the REFI-dataset supplied to us on dvd by the EPO. The Bulletin dataset supplies us with the date of each individual patent, countries of residence of its inventors, and the technology class (International Patent Class, IPC) assigned to it by the patent examiner. We use the priority date of the patent (which is the date at which the knowledge in the patent was first patented, worldwide) to assign it to a year (when priority date is missing, we assume the patent was first applied at the EPO, and hence use the EPO application date).

We also utilize a database supplied by the OECD covering the phenomenon of international patent families. In this context, the term patent family is used to describe a set of patents filed under different patent systems (e.g., EPO, USPTO), but covering the same invention. The OECD database that we use (Webb, Dernis et al., 2004) provides a list of so-called equivalent patent numbers (e.g, EPO patent 1234567 is equivalent to USPTO patent 7654321). This database is updated using data from the Espacenet webserver, which uses the same raw database as was used to construct the OECD database.

The REFI-dataset that is the source for our citations data also contains citations made in other patent systems than the EPO. The start of our citations database is a list of citing and cited patents (a so-called citation pair), covering a range of patent systems including the EPO. From this list, we identify the citation pairs in which the citing patent is either an EPO patent, or where the citing patent is found by our patent families database to be equivalent to an EPO patent. In the latter case, we substitute the original (non-EPO) citing patent by the equivalent EPO patent. Thus, we have, as an intermediate result, a list of citation pairs where all citing patents are EPO patents. We then select the subset of this list where also the cited patent is an EPO patent, or where the cited patent has an EPO equivalent. The final citation database, used in the analysis below, is then a list of approximately 1.64 million citation pairs, involving approximately the same number of EPO patents.

Obviously, related to the inter-industry point of view that we take, the assignment of a patent to an (economic) industry (sector) is crucial. We use the Merit IPC-Isic concordance table (Van Moergastel et al., 1994) to make this assignment. This concordance table is based on a detailed comparison of the content of the IPC and Isic (rev. 2) classification schemes, and a matching of the activities described in both. The principle of the matching is that the patent is assigned to its most likely industry of origin (e.g., a textiles machine is assigned to the machinery sector, not the textiles sector). The concordance is done at the 4-digit IPC level, and a mixture of 2-, 3- and 4-digit Isic industries (these will be introduced below when we discuss the data). We use only the manufacturing sectors in the concordance, and opt to aggregate the 22 sectors found in the concordance to 19. The concordance allows the assignment of a single IPC class to multiple Isic industries, based on a weighting scheme. This implies that patents are assigned fractionally, i.e., we do not necessarily have an integer number of patents in each industry.

4. Approximating the knowledge flow network by input-output methods

Before we start, let us make a note on our matrix notation for consistency. A square matrix (of size $n \times n$, where n is the number of industries) will be indicated by a boldface and capitalized letter as in \mathbf{X} , while a column vector (of size n) by a boldface small letter as in \mathbf{x} , and a row vector as in \mathbf{x}' , where ' stands for transposition. We refer to individual elements of matrices by small letters with two subscripts (i.e., x_{ij} stands for the i^{th} row j^{th} column element of

³ Below, we will use the term 'patents' to refer to patent applications, and we consider these applications whether or not they are granted.

matrix \mathbf{X}), while elements of vectors will be referred to by small letters with a single subscript (x_i stands for the i^{th} element of the vector \mathbf{x}). The format $\langle \mathbf{x} \rangle$ will be used to indicate a diagonal matrix (of size $n \times n$) constructed from the vector \mathbf{x} , which has x_i at its i^{th} diagonal and zeros elsewhere. Finally, \mathbf{i} (\mathbf{i}') refers to the column (row) summation vector (i.e., $\mathbf{i}' = [1, 1, 1, \dots, 1]$).

In constructing a knowledge flow table in raw form, we start at the patent citation level. For each of our citation pairs, we have information on the industries of the citing and cited patents. Furthermore, patents are also classified according to the year of the priority date. We follow the usual approach in the literature by constructing a citation flow matrix \mathbf{CPL} (we will omit time superscripts in our matrix notation, but all matrices refer to a specific year, unless otherwise indicated in the text) for year t , where t refers to the priority (invention) year of the citing patent (the patent that receives the knowledge flow). The rows and columns in the citation matrix represent industries of origin of the cited (row) and citing (column) patent. A column of this matrix will break down the citations made by industry j patents of year t (c_j^t) into $n+1$ (where n is the number of industries) categories such that

$$c_j^t = cnpl_j^t + cpl_{1j}^t + cpl_{2j}^t + \dots + cpl_{nj}^t, \quad \mathbf{1}$$

where $cnpl_j^t$ stands for the number of citations to non patent literature made by year t , industry j patents⁴ and cpl_{ij}^t for the number of citations to patents originating in industry i , made by year t , industry j patents. The number $cnpl_j^t$, usually scaled by the number of patents in industry j , year t is what Narin, Oliveira and Hamilton (1997) used as an indicator of the science intensity of industry j patents. However, from an input-output perspective, this is hardly a satisfactory measure, since it only captures the direct citations to science that industry j makes.

Let us define $v_j^t \equiv cnpl_j^t / c_j^t$ and $a_{ij}^t \equiv cpl_{ij}^t / c_j^t$ (note that by definition, $v_j^t + \sum_i a_{ij}^t = 1$). If the a 's and v 's are assumed to be constant over time, we can calculate the indirect science intensity of industry j patents, by accumulating the $cnpl_i^t$ that are made in the patents that industry j cites (cpl_{ij}^t). But these patents in turn cite other patents, which cite non patent literature as well as patents, and the process continues *ad infinitum*. To represent this process, which is not un-akin to a Leontief multiplier process, we define the $n \times n$ matrix \mathbf{A} , whose elements are the a_{ij}^t values. In terms of its interpretation, this matrix is clearly analogous to the input-coefficient matrix of input-output economics, which decomposes the input requirements of a number of economic sectors over the sectors which supply these inputs. Let us also construct an $n \times n$ diagonal matrix $\langle \mathbf{v} \rangle$ with elements v_j^t on the diagonal and

Now consider the following line of reasoning aimed at finding the total, i.e., directly and indirectly, accumulated science content (non patent literature references) embedded in patents of industry j , year t . Clearly, $\langle \mathbf{v} \rangle$ gives the fraction of 'direct science content' embedded in a single patent (per industry). But a 'second round' of science content flows from the patent literature citations and the non patent literature that they (directly) cited in the past. This can

⁴ Note that we do not make any direct observation about the nature of the cited science knowledge. For example, if the chemistry sector cites a paper in electrical engineering, it will be recorded in the chemistry column, not in the electrical machinery or electronics column. Also note that we cannot observe whether a particular scientific paper (or other non-patent-literature) is cited more than once. Hence we treat each non-patent-literature reference as if it were unique.

be represented by the matrix product $\langle \mathbf{v} \rangle \times \mathbf{A}$, which, for industry j , captures both science inputs that entered the system in industry j itself, and science inputs that entered the system in other industries (depending on \mathbf{A}). Similarly, we can envisage a third round of embedded science, represented by $\langle \mathbf{v} \rangle \times \mathbf{A} \times \mathbf{A}$, and a fourth round $\langle \mathbf{v} \rangle \times \mathbf{A} \times \mathbf{A} \times \mathbf{A}$, etc. The complete citation chain, for a single patent, is described by the following matrix product:

$$\mathbf{D} = \langle \mathbf{v} \rangle \times [\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \dots + \mathbf{A}^\infty], \quad \mathbf{2}$$

where \mathbf{I} is the $(n \times n)$ identity matrix.

The term in the brackets is quite familiar from input-output economics. It is the power series expansion of the Leontief inverse $(\mathbf{I} - \mathbf{A})^{-1}$, which is convergent if all column sums of the elements of matrix \mathbf{A} are strictly less than one, and all coefficients are non negative (both of which are naturally satisfied in our matrix \mathbf{A}). Thus, equation (2) can also be written as

$$\mathbf{D} = \langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1}. \quad \mathbf{3}$$

Let us now define the vector $\mathbf{s}' \equiv (\mathbf{D} \times \mathbf{c})'$, where \mathbf{c} is the column vector of the c'_j values, i.e., \mathbf{c} is simply the total citations (to patents and non patent literature) made by the industries. \mathbf{s}' represents the total (direct and indirect) science input in citations that has flown into the system through the various industries, either in year t or prior to that, and transmitted forward in time through (a long chain of) patent citations. The j^{th} element of \mathbf{s}' represents total science inputs that was introduced into the system by patents of industry j .

The columns of the matrix \mathbf{D} all sum to 1⁵. This implies that the sum of the elements of \mathbf{c} is equal to the sum of the elements of \mathbf{s}' . Because \mathbf{s}' represents the distribution of the 'production' of science inputs (non patent citations), we can conclude that in the formal system described so far, the total number of citations made by patents in year t is equal to the total number of science references embodied in these citations. In other words, if we define an average 'composite' citation by the fractions a_{ij} and v_j (i.e., a composite citation made by industry j cites v science references and a_{ij} patents of industry i), this embodies exactly a single unit of science input.⁶ Thus, a unit of 'pure' science references is a natural measurement of knowledge in our system.

Diagram 2 represents a stylized example of a chain of citations that is described by the system as introduced so far. The circles represent two industries, distinguished by the colors red and yellow. The squares represent the science sector. We also define two types of science inputs, blue and green, based on the sector that they feed into. The arrows represent citations between the different units (industries and the science sector), and the numbers indicate the number of citations made on a particular link. Thus, we see that for generation $t = 0$, the yellow sector makes a total of 6 citations, of which 3 are to patents of the red sector, 1 is to pat-

⁵ The vector of the column sums of matrix \mathbf{D} is $\mathbf{i}' \times \langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1}$. Let us call this vector of column sums $\boldsymbol{\varepsilon}$. Then $\mathbf{i}' \times \langle \mathbf{v} \rangle = \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon} \times \mathbf{A}$, which, due to the identity $v'_j + \sum_i a'_{ij} = 1$ implies the equation system

$$\boldsymbol{\varepsilon}_j - \sum_{i=1}^n \boldsymbol{\varepsilon}_i a_{ij} = v_j = 1 - \sum_{i=1}^n a_{ij} \quad \text{for all } j=1,2,\dots,n.$$

If we define the instrumental vector $\boldsymbol{\delta}$, where $\boldsymbol{\delta}_j = 1 - \boldsymbol{\varepsilon}_j$, this equation system is equivalent to $\boldsymbol{\delta} = \boldsymbol{\delta} \times \mathbf{A}$. Clearly, unless each column sum of matrix \mathbf{A} adds up to 1 (which is never the case for any \mathbf{A} matrix in IO systems), $\boldsymbol{\delta} = \boldsymbol{\delta} \times \mathbf{A}$ solves for $\boldsymbol{\delta} = \mathbf{0}'$, which implies $\boldsymbol{\varepsilon} = \mathbf{i}'$.

⁶ This characteristic of the system is partly the result of the implicit assumption that in the citation system total inputs in a sector are equal to total outputs. We will relax this assumption below.

ents of the yellow sector itself, and 2 are to science references. The red sector makes a total of 9 citations, of which 6 are to the red sector itself, 2 are to patents of the yellow sector, and 1 is to the science sector. These numbers can be used to set up the following specific realizations of \mathbf{A} and \mathbf{v}' :

		Yellow	Red
\mathbf{A}	Yellow	1/6	2/9
	Red	3/6=1/2	6/9=2/3
\mathbf{v}'	Science	2/6=1/3	1/9

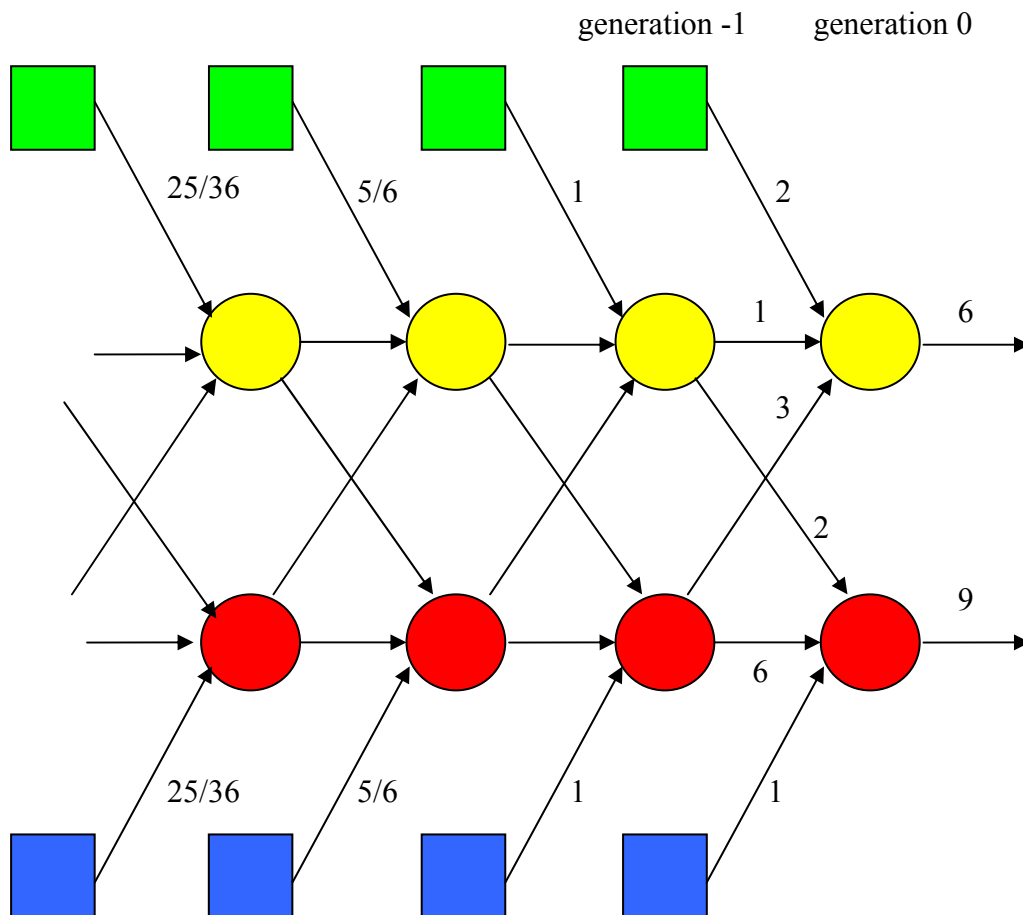


Diagram 2. Stylized citation network graph used to illustrate formal approach

Under the assumption of a fixed \mathbf{A} and \mathbf{v} prior to generation 0 (this is the rightmost set of patents), we can derive the number of citations that must be present on every one of the arrows that points to generations of patents prior to generation 0. For example, at generation 1, a total of 3 citations leave the yellow sector (output), and this must be matched by an equal input. $v_{\text{yellow}} = 1/3$ of this (i.e., a total of 1) comes from the science sector, $a_{\text{red,yellow}} = 1/2$ of

this (a total of $3/2$) comes from the red sector, and $a_{yellow,yellow} = 1/6$ of this (a total of $1/2$) comes from the yellow sector itself ($1 + 3/2 + 1/2 = 3$).

In this way, all the links originating directly from science have been filled in in the figure. Obviously, in order to obtain a complete picture, we would have to extend the diagram infinitely to the left. In that (imaginary) case, the sum of all values on the arrows originating from the blue and green squares (science sector) would be equal to 15 ($6 + 9$), which is the total amount of citations made by patents of generation 0 (i.e., the sum of elements of vector \mathbf{c}). \mathbf{c} is represented in the diagram by the two horizontal arrows on the right hand side.

The vector \mathbf{s}' may be obtained by summing the top (green) and bottom (blue) rows of science inputs. Making the actual calculation $\mathbf{s}' = (\mathbf{D} \times \mathbf{c})'$, we obtain $\mathbf{s}' = [8, 7]$. Hence, a total of 8 units of pure (green) science have been introduced into the system by the yellow sector, and 7 units of pure (blue) science by the red sector. In a version of a diagram that would extend infinitely to the left, the sum of flows originating from the bottom (blue) science row would be 7, that from the top (green) 8. Note that, indeed, the sum of elements of \mathbf{s} is equal to the sum of elements of \mathbf{c} ($8 + 7 = 15 = 6 + 9$).

So far, our perspective has been backward, i.e., we have asked how many pure science units are embodied in the citations made by patents published in year t . We may also take a forward perspective, which asks how much knowledge (in pure science units) is passed on to future generations. Let us introduce the column vector \mathbf{g} to denote this forward flow of knowledge, where the convention is that the j^{th} element of \mathbf{g} denotes the amount of knowledge passed on by patents of industry j to future generations.

Obviously, we do not observe the citation behaviour of future patents (yet), and hence there is no way in which we can observe \mathbf{g} . However, if we apply our assumption that at each generation incoming knowledge flows are equal to outgoing knowledge flows also to forward streams, we deduce that \mathbf{c} will be equal to \mathbf{g} , or, in words, that the total number of citations that a generation of patents makes (by sector) is also equal to the amount of knowledge it passes on to future generations of patents. The intuition behind this is that each generation of patents simply passes on the knowledge it received from previous generations of patents (by means of patent-to-patent citations), plus the knowledge it took directly from the science sector.

The system as described so far can also be summarized in tabular form as in Table 1. A narrative of constructing this patent flow table starts with the calculation of the (row) vector \mathbf{s}' . This vector results from the empirically observed citation flow matrix \mathbf{CPL} (equation 1 above), and the matrices \mathbf{A} and $\langle \mathbf{v} \rangle$ and vector \mathbf{c} that are all calculated from \mathbf{CPL} . Given $\langle \mathbf{v} \rangle$, we calculate the total amount of citations that was necessary to bring the amount of knowledge represented by \mathbf{s}' to generation t patents. This is represented by the row vector \mathbf{tcm}' . In terms of the above diagram, this is the sum of all arrows in the diagram, except the two rightmost arrows leaving generation 0 patents.

In a similar way, and using the elements of \mathbf{A} , we may fill in the top-left block of the table, which represents the patent-to-patent citations part of \mathbf{tcm} . Row-wise, the elements in this block of the table represent the patent-to-patent citations that leave patents of a specific sector, or, in other words, knowledge that is transmitted from this sector to other sectors. The column vector \mathbf{g} is put to the right of this, so that we capture in every row also the knowledge that is transmitted to future generations of patents. As we have already explained above, the assumption $\mathbf{g} = \mathbf{c}$ is a reflection of our assumption that each generation of patents simply passes on the knowledge it received from previous generations of patents plus the knowledge it took directly from the science sector. Finally, the rows will sum to the same total as the columns sum to, and hence we put \mathbf{tcm}' as the row totals on the far right of the table.

<p>INTERMEDIATE CITATION FLOWS IN PERIOD – ∞ TO t: (n × n)</p> <p>The matrix ICF, with elements $a_{ij} \cdot tcm_j$</p>	<p>TOTAL SCIENCE KNOWLEDGE PROVIDED TO POST-t PATENTS</p> <p>(n × 1)</p> <p>g = c (by assumption): TOTAL CITATIONS MADE IN YEAR t:</p>	<p>CITATIONS RECEIVED IN PERIOD – ∞ TO t:</p> <p>(n × 1)</p> <p>tcr = (ICF × i) + g = (I - A)⁻¹ × g = tcm</p>
<p>TOTAL SCIENCE KNOWLEDGE TAKEN IN IN PERIOD – ∞ TO t: (1 × n)</p> <p>s' = [D × g]' = [<v> × (I - A)⁻¹ × c]'</p>		
<p>CITATIONS MADE IN PERIOD – ∞ TO t: (1 × 19)</p> <p>tcm' = s' × (<v>)⁻¹ = [(I - A)⁻¹ × c]'</p>		

Table1. The construction of the patent citation flow table

Readers who are familiar with the Leontief economic input-output system may recognize that our constructed patent-flow table is in many regards similar to this.⁷ The matrix **ICF** is reminiscent of the matrix of intermediate flows of goods between sectors. A crucial difference, however, between our matrix **ICF** and a matrix of intermediate deliveries lies in the time horizon applied to construct these matrices. In national accounts, the complete chain of intermediate goods flows that lead to a final product are typically observed within a year. On the other hand, the average lag between a cited and citing patent is typically several years, implying that a chain of several patent citations can quickly run over a period of decades. This is why we have to resort to indirect observations based on the assumption of constant $\langle v \rangle$ and **A** to construct our matrix **ICF** for the period – ∞ to t, instead of directly observing it in entirety for a fixed period of time.

Continuing the analogy to input-output economics, science inputs play the role of primary inputs, and **s'** is similar to a value added vector. Finally, **g** is similar to the investment part of final demand. Because we have no equivalent of final consumers in our system, all ‘final demand’ is necessarily carried on to the future (investment).

Scientific knowledge is brought into the system in a cumulative manner by the non patent literature citations, given by the vector **s'**, is transmitted forward (up to year t) by the patent citations given by the intermediate flow matrix **ICF**, and is finally transmitted to the future, (i.e. post year-t patents) by the final set of **g** patent citations. Throughout this process the total

⁷ Note that in our own view, an analogy between the economic input-output table and our patent citation flow table is natural, but there is no need to subscribe to this analogy in order to appreciate the calculations that we will provide in the remainder of this paper.

number of (patent-to-patent and patent-to-non-patent literature) citations made by the sectors are given by the row vector \mathbf{tcm}' , and similarly the total number of citations received by the sectors is given by the column vector \mathbf{tcr} . Just similar to the equality of total expenditures to total output in an economic input-output system, in our system $\mathbf{tcm} = \mathbf{tcr}$, and similar to the equality of total final demand to total value added, in our system $\mathbf{s}'\mathbf{x} = \mathbf{i}'\mathbf{g}$.

5. The empirical implementation and proposed indicators

In actually constructing the knowledge flow table in its raw form (as represented in equation 1), we start at the patent citation level. We represent patent citations as pairs of citing and cited patents. For each of those pairs, we have information on the sectors of the citing and cited patents. Denoting the sector of the cited patent by i , and the sector of the citing patent by j , we record this particular citation as a knowledge flow of value 1, in the cell (i,j) of the matrix called **CPL**.⁸ We have yearly versions of the **CPL** matrix for 1979 – 2005, where the year refers to the priority date of the citing patent. Citations to non-patent-literature are entered into the row-vector \mathbf{cpnl}' , in the column i , where i is the sector of the citing patent. Because we look at direct citations only, the problem of right time-truncation is solved: we always observe all incoming citation links in a matrix for a particular year.

But this is obviously not the case for left-truncation: any citation to a patent that falls before the first date for which we have patent information (e.g., sector/IPC class) is not recorded. Hall, Jaffe & Trajtenberg, (2002, p. 421-4) show that, in the USPTO database, about half of all citations to a particular (average) patent are made within a period of 10 years. For the most recent year for which we can construct a reliable matrix (1999 or 2000), we can broadly cover cited patents over a 15-20 year time lag. Hence, we can be confident that the large majority of incoming citations are covered.

The problem of truncation between patent systems is somewhat harder to deal with in a completely adequate way, although we did correct partially for this by using the (updated) OECD database of patents equivalents (see above). But we still have the problem that we do not capture all patent-to-patent citations, while we capture all (direct) citations to the non-patent literature. This will bias the value of elements of matrix $\langle \mathbf{v} \rangle$ upward. We discuss this problem in more detail in the appendix, where we conclude that an imperfect correction for this truncation problem is feasible. We have used this correction method on our data, but choose to report results based on uncorrected data. The impact of the correction is not large and corrected results are available on request.

Indicators

Based on the input-output tables discussed so far, we construct a number of indicators that are aimed at scoring the sectors in terms of their role in the knowledge transfer system. We discuss these indicators one by one.

Backward multipliers: In input-output economics, backward multipliers capture the general idea that due to the derived demand for intermediate goods, an increase in demand for one sector will increase the total gross production of the economy by more than the original increase in demand. Furthermore, the resulting output increase is not confined to the sector

⁸ As explained above, our IPC-Isic concordance sometimes assigns a patent to multiple Isic sectors, with particular weights assigned to each of the sectors. We apply a fractional counting method for these cases. The value assigned to cell (i,j) of the matrix is equal to the product of the weights of the sectors i and j . Because the set of weights sums to one for both the citing and cited patent, each citation will count for one after it has been divided over all possible combinations of citing and cited sectors.

where the original increase in demand takes place. Backward multipliers are generally calculated as the column j sum of the so-called Leontief inverse, i.e., $\mathbf{i}' \times (\mathbf{I} - \mathbf{A})^{-1}$. Similarly defined backward multipliers are also useful indicators in our patent citation flows table, although their interpretation is not completely analogously to input-output economics.

Let the vector λ_t denote the counterpart of the backward multiplier vector as applied to the constructed cumulative citation-flows table year t . That is, let

$$\lambda_t' = \mathbf{i}' \times (\mathbf{I} - \mathbf{A}_t)^{-1}. \quad 4$$

Given this specification, for each industry j , the backward multiplier λ_{jt} indicates the total number of patent-to-patent citations that is necessary to make one (extra) unit of composite scientific knowledge available in year t patents of industry j , which can, in turn, be transmitted forward to post-year t patents (of various industries). In terms of Diagram 2, the backward multiplier measures what happens, under constant input-coefficients, if the rightmost arrow in one of the sectors is increased by one. For this to happen, the value of science inputs (arrows originating from squares) in the diagram will also have to increase by one. The backward multiplier measures by how much the value of the arrows between patents (circles) will have to increase to accommodate this increase.

Because there is a strict generational separation between the citations in the diagram, and the input coefficients are fixed, an increase of the value on the patent-to-patent arrows in the diagram corresponds to a citation chain that stretches further and deeper to the left (i.e., backward in time). Hence the backward multiplier is also an indicator of the average time-lag for scientific knowledge to become embodied in the present generation of patents (and available for future citation), and hence an indicator for the relative age of knowledge accumulated in the patents of the industry.

Let us give a single-sector example to clarify the interpretation of these backward multipliers. Assume that out of all citations, $100x$ ($0 < x < 1$) percent goes to NPL and $100(1-x)$ percent goes to other patents. Note that with smaller x , less science knowledge flows in at the present generation of patents, or, in other words, more science knowledge flows in at earlier generations. With constant x , this naturally holds at all generations, which implies that with smaller x , the average age of the knowledge embodied in current generations is higher.

Our notation implies that the 1×1 input coefficient matrix $\mathbf{A} = 1 - x$, and $\mathbf{v} = x$. Thus, x of a 1 unit increase of knowledge accumulated in year t patents comes from the x immediate NPL citations. The rest of the knowledge comes from the $1 - x$ citations that are directed to older patents, which pass on 'second hand' scientific knowledge to generation t . The second hand knowledge transmitted by these $1 - x$ citations had been supplied by the $x(1 - x)$ NPL citations of the previous generation patents, plus $(1 - x)^2$ patent citations directed to the earlier generation of patents. Similarly, the knowledge transmitted by these $(1 - x)^2$ patent citations had been supplied by the $x(1 - x)^2$ NPL citations of the previous generation patents, plus $(1 - x)^3$ patent citations directed to the (even) earlier generation of patents. Keeping on compounding backwards through the history of citations, it is confirmed that each unit of knowledge accumulated in patents of t , was introduced into the system by a total of $x + x(1 - x) + x(1 - x)^2 + \dots + x(1 - x)^\infty = 1$ NPL citations, and this was transmitted to patents of t through a total of $(1 - x) + (1 - x)^2 + \dots + (1 - x)^\infty$ patent-to-patent citations. Adding these two numbers up, we arrive at the total number of citations (to patent and non-patent literature) that is responsible for the extra unit of knowledge accumulated in patents of year t , which is equal to $1 + (1 - x) + (1 - x)^2 + \dots + (1 - x)^\infty = 1/x$. Thus, the backward multiplier λ is equal to $1/x$, and since x is related to average age of embodied knowledge, the multiplier serves as an indicator of age. More precisely, since x is inversely related to average age of embodied knowledge, the multiplier positively related to this.

Note that since $\mathbf{D} = \langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1}$ as described in equation (3), one can also express the backward multipliers as $\lambda'_j = \mathbf{i}' \times \langle \mathbf{v} \rangle^{-1} \times \mathbf{D}$, or equivalently as $\lambda_j = \sum_{i=1}^N \left(\frac{1}{v_i} \cdot D_{ij} \right)$. This implies

that the backward multipliers are actually the weighted average of the (inverse of the) share of non patent literature citations of all sectors, where the set of weights for each sector j is given by the shares of each sector-specific type of scientific knowledge embedded in the total knowledge stock of sector j patents.

This illustrates that the backward multipliers are structural indicators that reflect the idea of vertically integrated sectors, as in Pasinetti (1981), or Sánchez-Chóliz and Duarte (2003). The citation network structure captured by our patent citations flow table indicates that, at the industrial aggregate level, only a part of the scientific knowledge that is eventually transmitted to industry j patents of year t comes from the immediate NPL citations of this industry itself. Another good deal comes from citations to older patents, including some of other industries, and this goes back in time *ad infinitum*. Thus, the total knowledge embodied in current generation patents is generally a mixture of bits and pieces of various types of industry-specific knowledge, and the backward multipliers capture this.

Forward Multipliers: These multipliers are technically similar to the backward multipliers but they are based on output coefficients, not on input coefficients. Accordingly, in economics, these multipliers capture supply-push effects rather than demand-pull effects: The forward multiplier of industry j indicates the increase in total expenditures of the economy that would be caused by a unit increase in sector j value added. Since the idea of a supply-driven model is originally introduced by Ghosh (1958), these multipliers are also referred to as the Ghosh multipliers.

Although the direct citation matrix \mathbf{A} that we collect from the data does not allow the calculation of output coefficients, once the cumulative patents citations flow table (Table 1) is constructed, one can calculate an output coefficient matrix \mathbf{B}_t , where $b_{t,ij} = icf_{t,ij}/tcr_{t,j}$. Given this matrix, the vector of forward multipliers is calculated as

$$\boldsymbol{\gamma}_t = (\mathbf{I} - \mathbf{B}_t)^{-1} \times \mathbf{i}. \quad 5$$

In the context of the patent citation network, these forward multipliers have the following interpretation. For each industry j , the forward multiplier $\gamma_{t,j}$ indicates the total number of citations that is necessary to transmit 1 (extra) unit of industry j -specific scientific knowledge, through the citation network, to patents of year t (of potentially a number of different industries). In line with the interpretation of the backward multiplier, the forward multiplier of industry j is a relative indicator of the average age of industry j -specific scientific knowledge (i.e., knowledge introduced by industry j patents), as embodied in the current generation of patent citations. The intuition in terms of age of knowledge behind these forward multipliers is quite similar to that of the backward multipliers (in a single sector world, e.g., forward and backward multipliers are equal).

Still, forward and backward multipliers are quite different in terms of their interpretation. While backward multiplier are about the process of accumulation of the pool of different types of knowledge embodied in patents of a given industry, the forward multipliers are about the process through which a given type of knowledge is transmitted forward and distributed over the patents of all industries. This highlights the importance of a conceptual distinction between the two alternative temporal directions in which one can look at our cumulative citation flows system (cf. Diagram 2). A backward looking approach perceives the patents of different industries as different sinks. A different composition of a variety of different types of

industry-specific scientific knowledge of different industries eventually accumulates in each sink, and the backward looking approach considers the composition of knowledge found in each sink. On the other hand, the forward looking approach looks at sources, each of which introduces a different type of industry-specific scientific knowledge and distribute these forward over the patents of a variety of industries.

The next indicator that we will discuss below aims at analyzing the relative strength of the patents of different industries in terms of their double role in performing as sources and sinks at the same time. As we will argue, some industries are relatively more active in their role to perform as sources than they are in performing as sinks and vice versa.

Net science multipliers: The magnitude $g_i (= c_i)$ is the amount of *composite* scientific knowledge that is accumulated in industry i patents of year t , which is made available to future (i.e., post-year t) patents. On the other hand, s_i is the amount of *pure* industry i -specific scientific knowledge which is introduced into the patent system by industry i patents during the time interval $[-\infty, t]$, and distributed over patents of various industries. $g_i > s_i$ would indicate that industry i patents are more important as sinks of knowledge than as sources, and vice versa for $s_i > g_i$. Therefore, we define a ratio μ_{ti} , which is the ratio of scientific knowledge introduced by *all* industries which eventually ends up in industry i patents of year t , and the scientific knowledge introduced by industry i that ends up in all patents of year t . Thus $\mu_{ti} > 1$ (< 1) would indicate that industry i is a net knowledge supplier (user).

We note that this idea is quite similar to what Oosterhaven and Stelder (2002) call net multipliers in an economic input-output context. Dietzenbacher (2005) shows that such net (value added) multipliers give the ratio of value added to final demand. This is obviously very similar to our source/sink interpretation of knowledge flows. This is why we call this indicator the net science multiplier indicator. Following Oosterhaven and Stelder (2003) and Dietzenbacher (2005), we calculate the row vector μ_t , which is a vector whose i^{th} element is equal to $\mu_{ij} = g_i/s_i$, as follows:

$$\mu_t' = \mathbf{s}' \times \langle \mathbf{tcm} \rangle^{-1} \times (\mathbf{I} - \mathbf{A})^{-1} \times \langle \mathbf{g} \rangle \times \langle \mathbf{s} \rangle^{-1}. \quad 6$$

On the basis of Table 1, (i.e., $\mathbf{tcm} = (\mathbf{I} - \mathbf{A})^{-1} \times \mathbf{c}$ and $\mathbf{s}' = [\langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1} \times \mathbf{c}]'$), and also letting \mathbf{L} denote the Leontief inverse $(\mathbf{I} - \mathbf{A})^{-1}$, it is clear that the element in the j^{th} diagonal of the inverse matrix $\langle \mathbf{tcm} \rangle^{-1}$ is $1 / \sum_{j=1}^N L_{ij} C_j$, whereas the i^{th} element of the row vector \mathbf{s}' is $v_i \sum_{j=1}^N L_{ij} C_j$, which implies that the term $\mathbf{s}' \times \langle \mathbf{tcm} \rangle^{-1} = \mathbf{i}' \times \langle \mathbf{v} \rangle$. Consequently, given $\mathbf{i}' \times \langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1} = \mathbf{i}'$, the calculation of μ_t reduces to

$$\mu_t' = \mathbf{i}' \times \langle \mathbf{g} \rangle \times \langle \mathbf{s} \rangle^{-1}. \quad 7$$

Self-reliance of sectors: The matrix $\mathbf{D} = \langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1}$ breaks the embedded knowledge in the patents of each sector down to its sector-specific components. The extent to which a sector relies on scientific knowledge introduced by itself can be assessed by looking at the weight on the diagonal of this matrix. The most straightforward way of measuring this is simply

$$\delta_j = D_{jj},$$

which we refer to as the self-use of knowledge indicator. It is the share of sector j -specific knowledge in the composite knowledge mix of sector j patents.

Similar to \mathbf{D} , we construct a matrix \mathbf{K} that decomposes the knowledge supplied by an industry i in terms of the industries that use its knowledge. This uses the output coefficient ma-

trix \mathbf{B}_t as well as the diagonal matrix $\langle \mathbf{f} \rangle$, which has the ratio g_j/tcr_j (i.e., the share of citations received by industry j patents of year t in all citations received by industry j patents of $[-\infty, t]$) on its j^{th} diagonal. The matrix

$$\mathbf{K} = (\mathbf{I} - \mathbf{B})^{-1} \times \langle \mathbf{f} \rangle \quad 8$$

is then quite similar to \mathbf{D} : k_{ij} gives the share of the industry i -specific knowledge (introduced and transmitted by industry i patents in $[-\infty, t]$) which is eventually transmitted to industry j patents of year t . Since these are shares, all row sums of \mathbf{K} are equal to 1 ($\mathbf{K} \times \mathbf{i} = (\mathbf{I} - \mathbf{B})^{-1} \times \langle \mathbf{f} \rangle \times \mathbf{i} = \mathbf{i}$). Using \mathbf{K} , we introduce a similar measure to δ , but which focuses on the knowledge supply of the sectors. This is

$$\kappa_j = K_{jj},$$

which is the share of sector i -specific scientific knowledge transferred by sector i to itself, or the self-supply indicator. It is an indicator of how much a sector generates internal knowledge vs. knowledge that is used by other sectors.

Pervasiveness of knowledge suppliers and users: Independently of the amount of knowledge that an industry supplies to the aggregate system, or the amount that it uses, the distribution of its knowledge supply or demand over the range of industries is an important indicator. For example, the distribution of knowledge inputs of a certain industry over all industries in the system indicates to which extent that industry is dependent on a small range of industries for its (ultimate) knowledge inputs. Similarly, the distribution of knowledge supply of an industry over all industries is an indication of how pervasive knowledge of a particular industry is.

We may again use the matrix $\mathbf{D} = \langle \mathbf{v} \rangle \times (\mathbf{I} - \mathbf{A})^{-1}$ to construct an indicator for this. We start by looking at the off-diagonal elements of this matrix (the diagonal elements are captured in the previous indicator), and consider the column for every sector. For this column, we calculate an inverse Herfindahl index for industry j as

$$h_j^{sc} = \frac{1}{\sum_{i=1}^{N, i \neq j} \left(\frac{d_{ij}}{1 - d_{ij}} \right)^2}. \quad 9$$

This gives the Herfindahl equivalent number of industries that supply industry j with various types of industry-specific scientific knowledge, and is an inverse index of concentration of the knowledge sources of industry j . Clearly h_j^{sc} , which is our indicator of knowledge-use pervasiveness, indicates the variety in inter-industrial backward citation linkages between industry j and all other industries.

Similarly, we construct an inverse Herfindahl indicator for knowledge supply, using the row-wise, off-diagonal elements of \mathbf{K} as follows:

$$h_j^{sk} = \frac{1}{\sum_{i=1}^{N, i \neq j} \left(\frac{k_{ij}}{1 - k_{ij}} \right)^2}. \quad 10$$

h_i^{sk} gives the Herfindahl equivalent number of industries that eventually embed the industry i -specific knowledge that is introduced and transmitted by industry i patents. It indicates the variety in inter-industrial forwards citation linkages between industry i and all other indus-

tries, and is a measure of how pervasively industry i influences the other industries in the system.

6. Some properties of the citation network

In this section, we provide an overview of some of the basic trends in our database. Our guiding question is how citation behaviour of patents differs between sectors and over time. Note that our research questions and methodology start from the basic notion that there are important sectoral differences in terms of spillovers and the associated citations. This section will document such differences at a basic level, thereby illustrating the interest of the analysis on a general level. Moreover, the specific methodology that we adopted (the construction of the citation flow table) adopts a number of specific assumptions, among which the time invariance of the coefficients is perhaps the most important one. Therefore we investigate to which extent this is indeed a valid assumption given the trends in the data.

Figure 1 shows the trends in the raw data on patents and citations. The EPO was established in 1979, and we see that since then the number of patents introduced per year has been increasing. During 1980 – 1989, the number of patents increases more or less linearly, after which the growth rate decreases until 1997. Despite the rather smooth behavior of the motion of the number of patents, during the 3 years that follow 1997, we observe a sudden peak. This is due to the fact that in our database, we also track patents that were filed under the so-called Patent Cooperation Treaty (PCT), which is a global patent system operating in much the same way as the EPO operates for Europe. This means that applicants may gain multiple national patents as a result of just one filing. The EPO is one of the regional patent offices at which patents can be filed for global protection under the PCT. The peak around 1999 is associated with a rise in the number of such PCT filings. The eventual fall in the number of patents since 2001 is a result of the (right) truncation of our database.

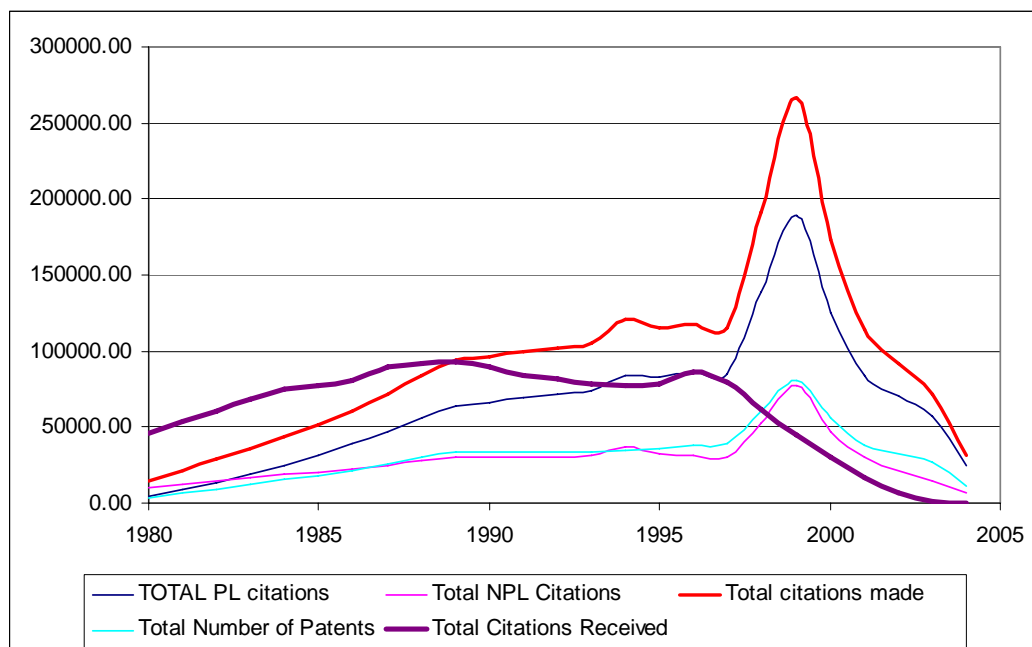


Figure 1. Number of citations made and number of patent applications

The number of citations to patents or to non patent literature rises in close connection with the number of patents itself. Due to the right truncation problem, the number of citations re-

ceived by the patents in our database shows first a gradual, then sharp decrease after 1989. The peak around 1999 associated with the PCT filings is magnified in the citations data.

Figure 2 shows the number of citations made or received per patent. After 1985, the number of citations made per patent remains rather stable, around 3 citations to patents, and slightly below one to non patent literature. Also note that the PCT-related peak in Figure 1 does not translate to Figure 2. Despite the stable pattern, however, citation data are still affected by the truncation problem towards the end of the period, because patent citations are published in the so-called search report (issued after the first phase of examination), which trails the patent application by a period of up to several years.

Unlike the average number of citations made per patent, the average number of citations received per patent shows a strong decreasing trend over the 25 years. This is clearly an inevitable outcome of the right truncation of the data. Simply the size of the set of patents that can cite a patent is a decreasing function the age of the patent in question, and given the citation lags, 25 years is not a sufficiently long period for this ratio to exhibit steady state behavior.

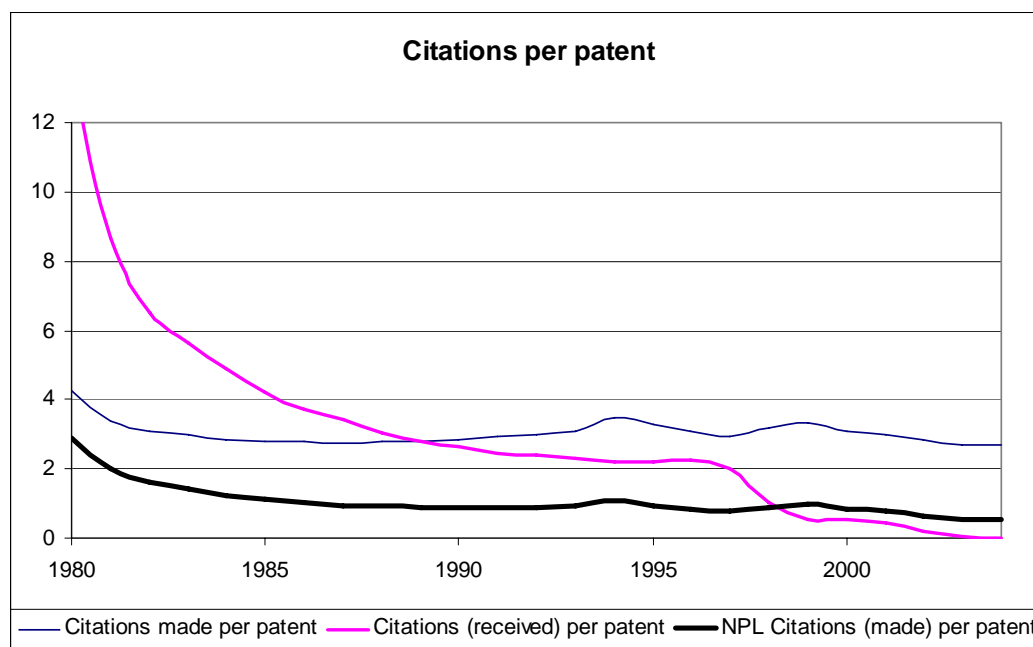


Figure 2. Number of citations per patent

For our input-output framework, the distribution of all citations over the sectors that make or receive the citations is also an important issue to consider. At the sectoral level, the data exhibit two key properties. First, one observes strong heterogeneity over the sectors in terms of citation rates (i.e., citations per patent, made or received) and also in terms of the distributions of citations over the entities that receive the citations (i.e., patents or non patent literature). Second, the pattern of such cross-sectional heterogeneity is rather invariant over time. These two properties are crucial for the plausibility of some of the assumptions of our methodology, and therefore deserve a detailed illustration.

Table 2 introduces our sectors and also gives a number of indicators regarding the sector-level heterogeneity in citation behavior for patents of 1992. As we will argue below, the 1992 values are representative of the period until the early 2000s. The first column of the table indicates the average number of citations per patent for each of the 19 sectors. The following two columns break these averages down to the average number of non patent literature citations and the average number of patent (i.e., patent literature) citations (the first column is the

sum of the second and the third columns). The fourth column gives the share of non patent literature citations in all citations (i.e., the fourth column is the ratio of the second column to the first column). The sectors are sorted in a descending order according to the second column.

For all 4 variables, of the 19 sectors, 12 have a value below the (unweighted) average, 5 sectors have a value between the average and one standard deviation above it, and 2 sectors show a value above the average plus one standard deviation. The sectors that are above average are almost all usually considered as high- or medium-tech (food products and non-ferrous basic metals are the exceptions). Two sectors stand out from the sample: pharmaceuticals and food products. Both sectors, or at least part of their cutting-edge technologies as reflected in our database, have a stronghold in biotechnology. The pharmaceuticals sector is obviously known as a strongly science-based sector, and the high citation rate is in line with this impression (Narin, Hamilton and Olivastro, 1997). ICT-related sectors (computers and office machinery, and electronics) rank as above average, and so does chemistry (excl. pharmaceuticals). Non-ferrous basic metals is also somewhat of a surprise in the top-list of citation rates.

Among all these four indicators the extent of heterogeneity is the highest in the number of NPL citations per patent (the coefficient of variation is around 120%) and lowest in the number of patent citations per patent (the coefficient of variation is around 12%). The first, second and the fourth columns are highly correlated with each other (correlation coefficients around 95%), while the correlation between any of these three with the third column is positive yet lower (around 50-60%).

Sector	Citations per patent	NPL citations per patent	PL citations per patent	Share of NPL citations in all citations
Pharmaceuticals	4.347	1.907	2.440	0.439
Food products	3.672	1.435	2.237	0.391
Unweighted mean+standard deviation	3.283	1.058	2.303	0.308
Electronics	3.124	1.052	2.072	0.337
Computers and office equipment	3.231	0.983	2.248	0.304
Non-ferrous basic metals	2.922	0.826	2.096	0.283
Chemistry	3.048	0.622	2.426	0.204
Instruments	2.811	0.613	2.197	0.218
Unweighted mean	2.645	0.568	2.077	0.191
Ferrous basic metals	2.485	0.564	1.922	0.227
Stone, clay and glass products	2.576	0.532	2.044	0.206
Electrical machinery	2.359	0.476	1.883	0.202
Paper and printing	2.389	0.360	2.029	0.151
Other manufacturing	2.235	0.238	1.997	0.106
Textiles	2.235	0.224	2.011	0.100
Metal products	1.966	0.210	1.756	0.107
Other transport	1.923	0.186	1.737	0.097
Other machinery	2.086	0.177	1.909	0.085
Oil refining	2.745	0.176	2.569	0.064
Motor vehicles	2.054	0.128	1.926	0.062
Rubber and plastic products	2.040	0.082	1.958	0.040

Table 2. Patent-to-patent (PL) and non patent literature (NPL) citations per patent, 1992

These properties can be summarized as follows. The citation rates exhibit strong heterogeneity across the 19 sectors. Although some part of this heterogeneity lies in the heterogeneity in citations to patents, the dominant reason is the strong heterogeneity in citations to the non patent literature. The sectors that are often considered as high-tech make relatively more citations to other patents. However, their outstanding position in terms of their higher tendency to cite non patent literature is significantly more pronounced than their stronger tendencies to cite patents.

In order to demonstrate the invariance of heterogeneity in citation behavior across the sectors, we calculated the first column of Table 2 (the vector of citation rates) for each individual year between 1980 and 2004. For each possible pair of years, we calculated the correlation between the two corresponding vectors of citation rates. The matrix of these pair-wise correlations are given in Figure 4 in the form of a gray-scale grid where darkness is a decreasing function of the correlation (i.e., white areas correspond to high correlations).

The temporal correlation structure in Figure 3 suggests some kind of a structural break around 1986 (but not around 1999, which is the time of the PCT related peak in Figure 1). The set of citation rate vectors that belong to years in 1986-2004 are highly correlated. Although the pair wise correlations between the years within this interval are generally decreasing with the length of the time interval between the two years considered, they generally show very high correlations (the lowest coefficient is around 90%). This situation is similar if one considers the interval 1980-1985, however, between the two intervals (i.e., correlations between year involving both intervals) the correlation is much lower, and can go down to levels around 40%.

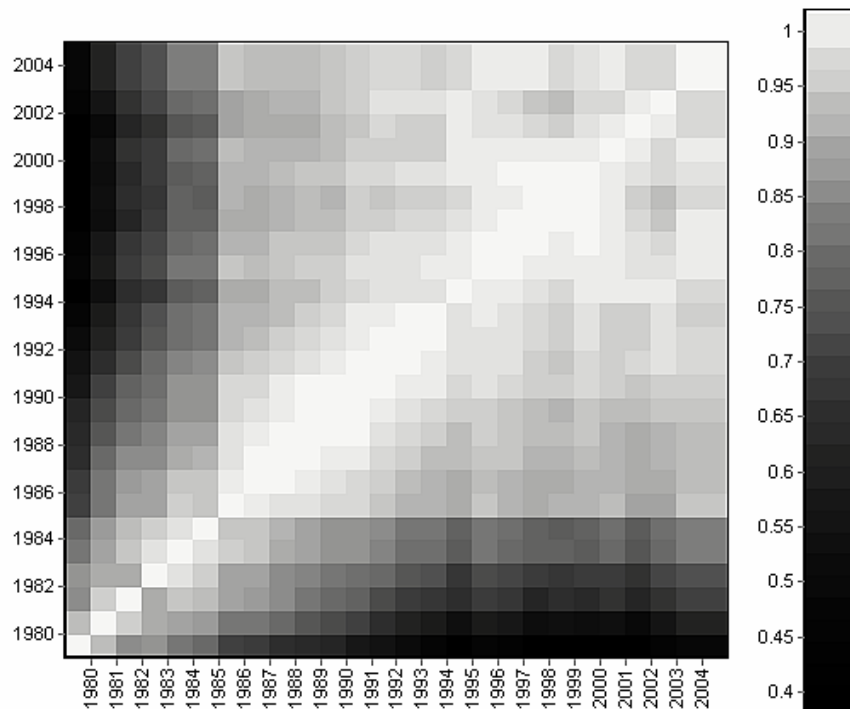


Figure 3. Intertemporal correlations between the citation rates of industries

We also made similar analyses of the distribution of total patents and the total number of citations for the individual 19 sectors (not documented in detail). As shown by Figure 1, both of these tend to grow over time, yet their distribution over the sectors exhibits a rather time-invariant pattern especially after 1986. In this case, the correlations do not fall between 60%.

It seems that the distributions of total patents and total citations over the sectors are even more time-invariant than the citation rates.

There is yet another property of which the time-invariance is utterly important in terms of the empirical justification of the key assumptions of our methodology: the heterogeneity of the share of non patent literature citations in all citations made, and the distribution of the patent citations over the sectors that receive these citations. The patent citation flow table that we constructed above depends crucially on the citation structure between sectors being stable over time, because we use citation data on patents of a single year and assume fixed coefficients in backcasting the flow of citations that feeds into these patents. We assess the time invariance of these fixed coefficients by looking at the input coefficient matrices \mathbf{A}_t and the vector \mathbf{v} which we both calculate individually for each year t (on the basis of the distribution of all citations made in the given year t). The intertemporal correlation of the elements of the \mathbf{A} matrix and the \mathbf{v} vector, pair wise between years, is depicted in Figure 4 in the form of a gray-scale grid.⁹ This shows an even stronger time invariance than observed in Figure 3. In the period 1986-2004, the correlation between the input coefficient matrices of any given two years is never less than 95%. Hence we can be confident that the assumption of time-invariance of the \mathbf{A} matrix is reasonable, and this justifies the power series expansion interpretation of the Leontief inverse, which lies at the very core of our knowledge decomposition/accumulation exercise.

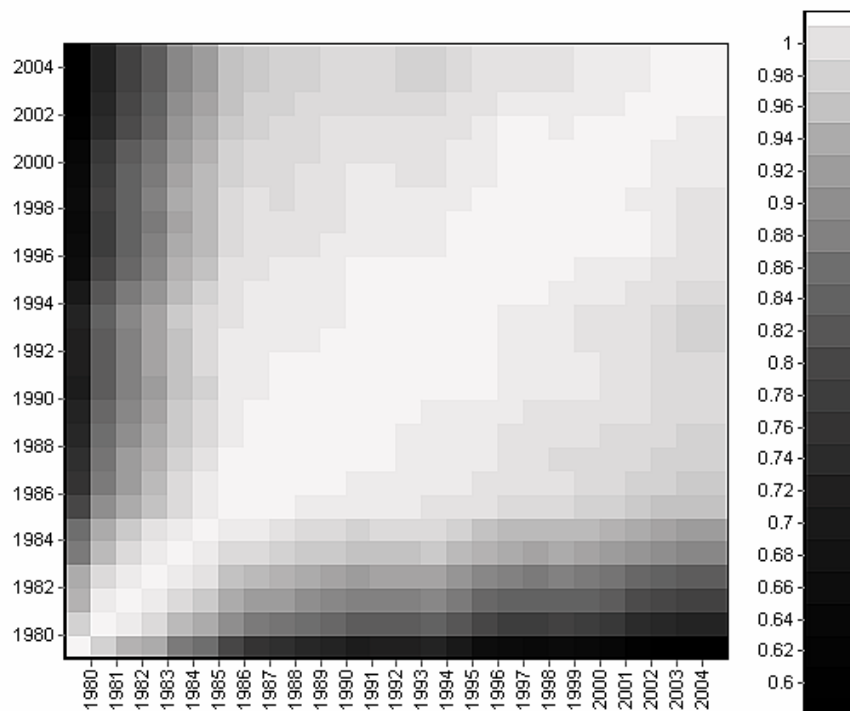


Figure 4. Intertemporal correlations between the input coefficient matrices

Finally, we must address a last assumption of our methodological framework. The internal logic of our knowledge decomposition exercise demands that knowledge is preserved through consecutive generations of patents. Each step of the knowledge transmission process decomposes the knowledge absorbed by generation t patents (from generation $t-1$ patents as well as cited non patent literature) according to the citation structure of generation $t-1$ patents. In

⁹ Each of the correlations in the figure is based on 19x20 observations (19x19 sectoral links in \mathbf{A} , and 1x19 observations in \mathbf{v}). We have done a similar exercise separately for \mathbf{A} and \mathbf{v} , this yields the same qualitative outcomes.

terms of our convention that one citation contains one unit of knowledge, this means that ultimately, the number of citations received by patents of a specific generation and a specific sector must be equal to the number of citations made. But since the number of citations received is not observable due to right truncation of the dataset, we cannot test this directly. Therefore, we consider the correlation, over sectors and for a given year, between the number of citations made and the number of citations received. We will consider a positive correlation as support for our assumption of equal knowledge input and output.

Figure 5 shows the correlations between the total number of citations received by patents of a given year, across the sectors, and several indicators of citations made. The highest correlation is with the total number of citations made to the patent literature. With the exception of the very early and very late years, this is always above 95%. Thus, we find indeed that the sectors in which the patents make relatively more citations are also cited relatively more.

This may be caused by self citations, at the sectoral level, which would make the outcome somewhat trivial. As also shown by Figure 5, if we exclude such sectoral self citations, the correlation is lower, but remains high, reaching approximately 90% in the late 1990s. Thus, self citations are part of, but not the only or the most dominant reason for the strong correlation between citations made and citations received. Thus, the sectors that make relatively more citations to patents of other sectors are also cited relatively more by the patents of these other sectors. Finally, Figure 5 also suggests that the sectors which have a higher share of NPL citations in the composition of their citations made, are cited relatively more often. This latter relation is not as strong as the former ones, but it is positive and statistically significant at conventional levels. This finding confirms that a relation observed in other studies (Fleming and Sorensen, 2004) at the individual patent level also holds at the sectoral level in our dataset.

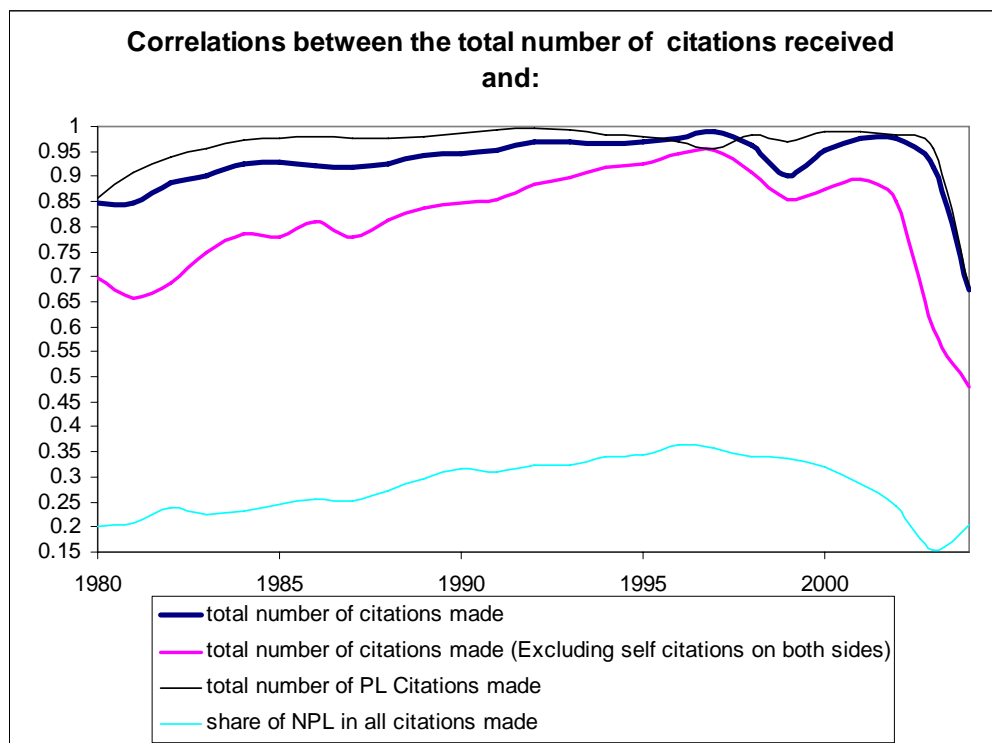


Figure 5. Intertemporal correlations between the total citations of industries

7. Empirical results

Our research question is about the impact of science on technology, and about the insights that can be gained by our proposed intersectoral approach in comparison to an approach based on the single indicator of the number of patent-to-science citations per patent. Due to the strong time-invariance of our annual data, the indicators based on the intersectoral perspective are also invariant over time. Thus we present here the results for 1992 which is representative of the other years between 1985 and 2004.

Given our interest in the value added of the intersectoral approach, we start our empirical analysis by a factor analysis of the various indicators. This enables us to reduce the number of dimensions in which we score the sectors, and to assess the role of the traditional indicator (patent-to science citations) in comparison to the other indicators. In addition to this indicator, we include the forward and backward multipliers, the net multipliers, the two diagonal share indicators, and the two pervasiveness indicators. All of these indicators have been introduced above, their numerical values are documented in the appendix. The results of the factor analysis (principal components) are documented in Table 3.

Table 3. Factor analysis on the indicators

Indicator	Factor 1 loadings (science intensity)	Factor 2 loadings (pervasiveness)
Backward multiplier	-0.953	0.195
Forward multiplier	-0.917	0.263
Net multiplier	0.942	0.003
Science citations per patent	0.846	-0.400
Self-use	0.929	-0.262
Self-supply	0.605	-0.296
Use pervasiveness	-0.528	0.723
Supply pervasiveness	-0.036	0.953
% variance explained (cumulative)	70.1	14.7 (84.1)

Factors rotated using varimax algorithm.

The results show very clearly that the science citations per patent indicator only capture one dimension in the data. This is the first factor, and hence the one that explains that largest proportion of the variance (almost 70%). This factor loads high on the number of science citations per patent, on the net multiplier, on self-use, and, to a lower extent, self-supply. It loads strongly negative on backward and forward multipliers, and, to a lesser extent, on use pervasiveness. We label this factor ‘science intensity’, although this does not cover the complete content of the factor. In particular, the loadings suggest that science intensity implies low backward and forward multipliers, and high self-dependence of the sector.

The results for the backward and forward multipliers are in line with our expectations, based on the interpretation of the multipliers as the average age of the scientific knowledge embodied in patents. Sectors that are highly science intensive have a rapid turnaround time of scientific knowledge, and hence score low on both types of multipliers. Interestingly, the normative interpretation of these multipliers (interpreted in terms of age of knowledge) runs against the interpretation in economic input output accounts, where high multipliers usually have a positive normative interpretation.

Note that the result is quite opposite for the net multiplier, which captures to what extent a sector is a net supplier or user of knowledge. Here we observe a positive correlation between the science intensity and the net multiplier, i.e., highly science intensive sectors are generally

responsible for a larger fraction of the knowledge introduced into the system than the fraction of knowledge used.

The first factor also shows that high science intensity comes with high self dependence, especially so in terms of using knowledge. Note that this is, contrary to the result for the backward and forward multipliers, not related to any specific mechanism embedded in the definition of these indicators. However, it is intuitively plausible if we accept that scientific knowledge is specific to the sectors in our analysis: if a sector depends highly on input of scientific knowledge, and if this knowledge is specific, it will have to rely to an important extent on sourcing this knowledge itself.

More interesting from the point of view of our intersectoral perspective is the second factor, which explains around 15% of the variance. This is the factor that has a factor loading for the number of science citations per patent that is relatively low, even in terms of its absolute value (e.g., the absolute value of 0.4 is below a threshold of $\frac{1}{2}$ that is often used to indicate important factor loadings). This means that the second factor captures an element in the data that is not strongly related to the direct science intensity of a sector. Hence this aspect is specific to the intersectoral approach that we used, and would not be apparent in the usual approach of only looking at the number of science citations per patent.

The second factor only loads high on the two pervasiveness indicators. High values of these indicators (inverse Herfindahl) indicate that a sector caters its knowledge to a broad range of other sectors (supply), or sources its knowledge from a broad range of other sectors (use). The factor loadings indicate that these two processes (pervasive use and supply) tend to go together, but they are largely uncorrelated to the first dimension of science intensity.

Hence we conclude that science intensity and pervasiveness of sectors in terms of knowledge flows are two separate dimensions. The first of these, science intensity, may well be captured by the normal approach of measuring science citations per patent, but it says very little about pervasiveness of knowledge flows from a sector. This suggests that the value added of a multisectoral approach lies mainly in being able to provide a better picture of the pervasiveness of knowledge flows.

We have already seen that science intensity, as measured by the number of science citations per patent, tends to be related to the net multiplier, i.e., whether sectors are net suppliers or users of knowledge. Figure 6 plots these two indicators against each other. The relationship occurs as somewhat nonlinear, with the net multiplier flattening off for high values of science intensity. The figure also shows that the observations (roughly) fall into two quadrants: sectors with more than half a science citation per patent are the ones that have a net multiplier of one or higher.

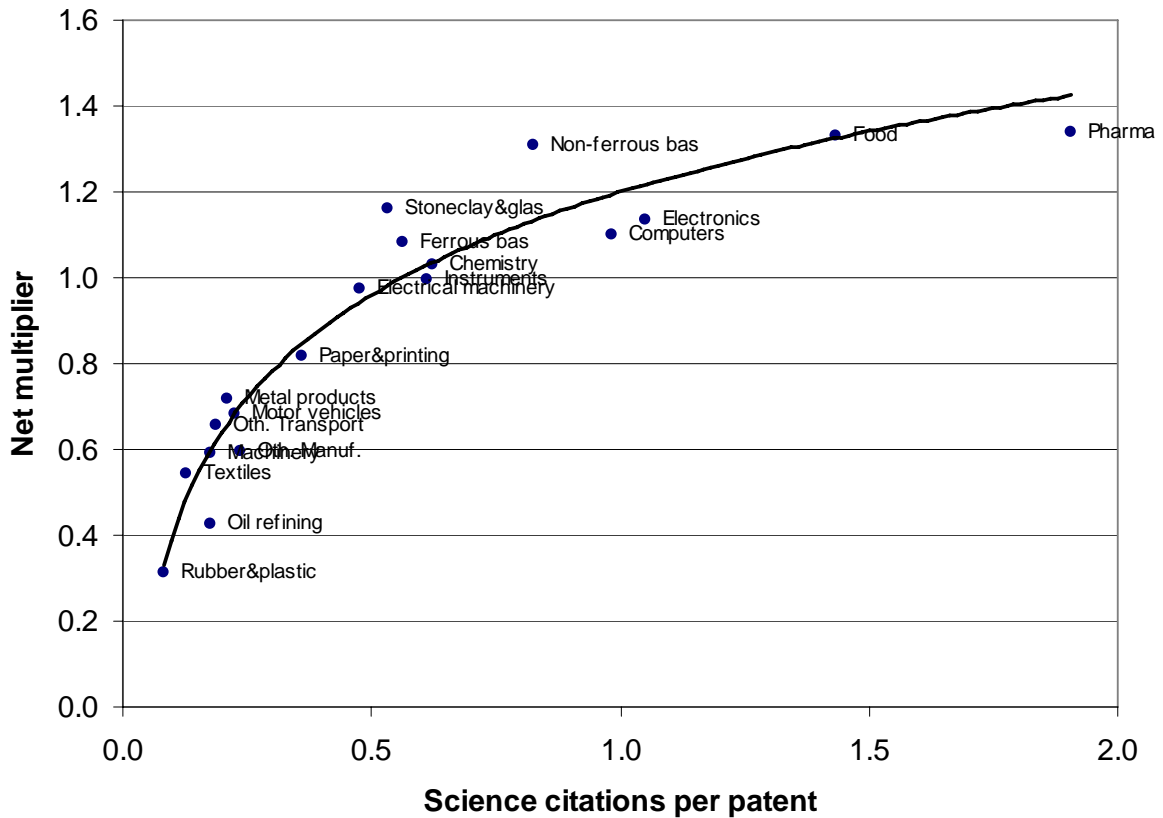


Figure 6. Science intensity and net science multipliers (λ and γ), solid line is logarithmic regression line

Figure 7 plots the two indicators that provide additional information relative to the dimension of science intensity, i.e., the two pervasiveness indicators. There is a broad positive relation between them, but it is far from perfect. The figure allows us to broadly classify the sectors in two ways. First, we can observe that there is a dichotomy between the sectors in terms of their general level of pervasiveness (both use and supply): we have a group of sectors on the left bottom of the figure, and one on the right top, with only a very limited number observations in between (motor vehicles and electrical machinery). Interestingly, among the sectors in the low pervasiveness part of the graph, we predominantly find sectors that also have high science intensity and a high net multiplier (cf. Figure 6). Oil refining is the only sector that is clearly in the low pervasiveness group *and* in has a low science intensity and net multiplier. All the other sectors with low pervasiveness have net multipliers very close to or larger than one (and correspondingly, high science intensities, $> \frac{1}{2}$ science citation per patent). On the other hand, some of the highly science intensive sectors, those related to materials (basic metals, stone clay glass, rubber and plastic), are among the highly pervasive ones, together with paper and printing, and machinery.

Thus, there are a few sectors that are exceptions to the general tendency that science intensive sectors have low pervasiveness. Instruments is the strongest exception, but also the two basic metals sectors and stone-clay-glass. These are all sectors that have high science intensity and corresponding high multipliers, but also high pervasiveness. Note that instruments is the only sector in this list of exceptions that is generally considered as high-tech.

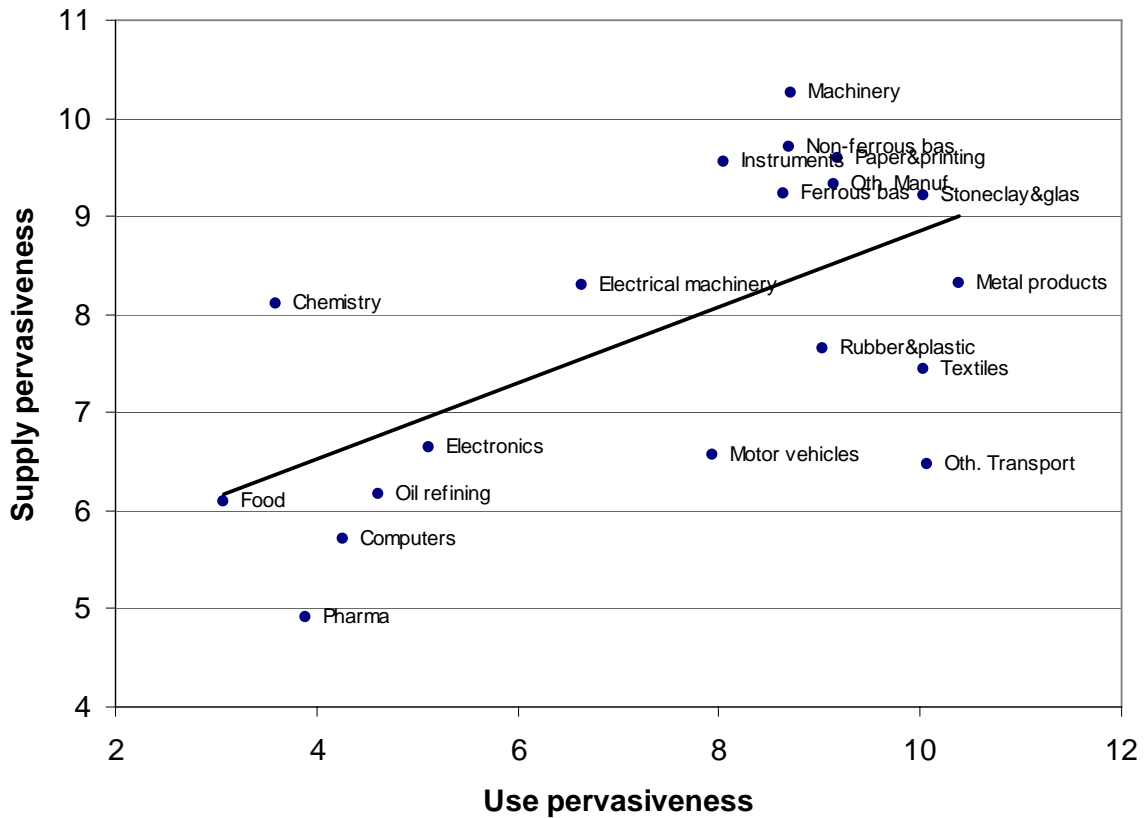


Figure 7. Use pervasiveness vs. supply pervasiveness, solid line is linear regression line

The second way in which we can classify the sectors in Figure 7 is by whether they are particularly pervasive in terms of supply or use of knowledge. This can be evaluated on the basis of whether sectors are above or below the regression line. Sectors that are above (below) the line are particularly pervasive with regard to their knowledge supply (use). In the group of sectors with low pervasiveness, most are relative more pervasive in terms of their use than in terms of their supply. Chemistry is the only exception, it is relatively much more pervasive with regard to its supply than with regard to its use. In the group of highly pervasive sectors, the sectors are more evenly distributed above or below the regression line. Interestingly, the sectors with high science intensity (science citations per patent) tend to be above the line, the lowly science intensive sectors below the line.¹⁰

We are thus left with a somewhat paradoxical situation. The science intensive (and often high-tech) sectors appear to be the largest net-suppliers of knowledge (according to our net science multiplier measure, $\mu_j > 1$), but only in a limited number of cases does this come together with a pervasive influence on a large range of other sectors. Especially in the science intensive sectors that are also high-tech, we find a strong concentration of knowledge flows to and from a rather limited number of other sectors. The highly science intensive sectors that are exceptions to this rule tend to be the ones that are science intensive but not generally considered as high-tech (instruments is the odd case).

This paradoxical relationship between science intensity and pervasiveness seems to point to the existence of clusters of strongly technologically related sectors in the knowledge network, which exchange a lot of knowledge within them, but not so much (relatively) to the rest

¹⁰ The positive correlation between science intensity and supply / use pervasiveness is confirmed by a plot of science citations per patent vs. supply/use pervasiveness (not documented but available on request).

of the network. In other words, these are relatively self-sufficient subsets of the knowledge system, due to the specialized and specific nature of knowledge. In order to observe whether these clusters indeed exist, and how they relate to the specific results on science intensity and pervasiveness that we obtain, we present a Multi-Dimensional Scaling (MDS) analysis. MDS is often used for visualization of multi-dimensional data. The underlying logic of this dimension reduction method is as follows. Given a matrix that gives the similarities between pairs of entities (sectors), we aim to find a two-dimensional map, in which the distances between the entities is consistent with the ranking of the inverse similarities in the original matrix. A heuristic algorithm is used to find such a map in an iterative way. Note that the (horizontal and vertical) dimensions on this map have no other function than to provide a number of degrees of freedom for the mapping (a 3D mapping would provide better results, but the 2D map we use is easier to interpret), and they have no particular *a priori* interpretation. The MDS analysis is based on an indicator of mutual dependency between the sectors, which we explain in detail in the appendix. This indicator is based on our matrices **D** and **K**.

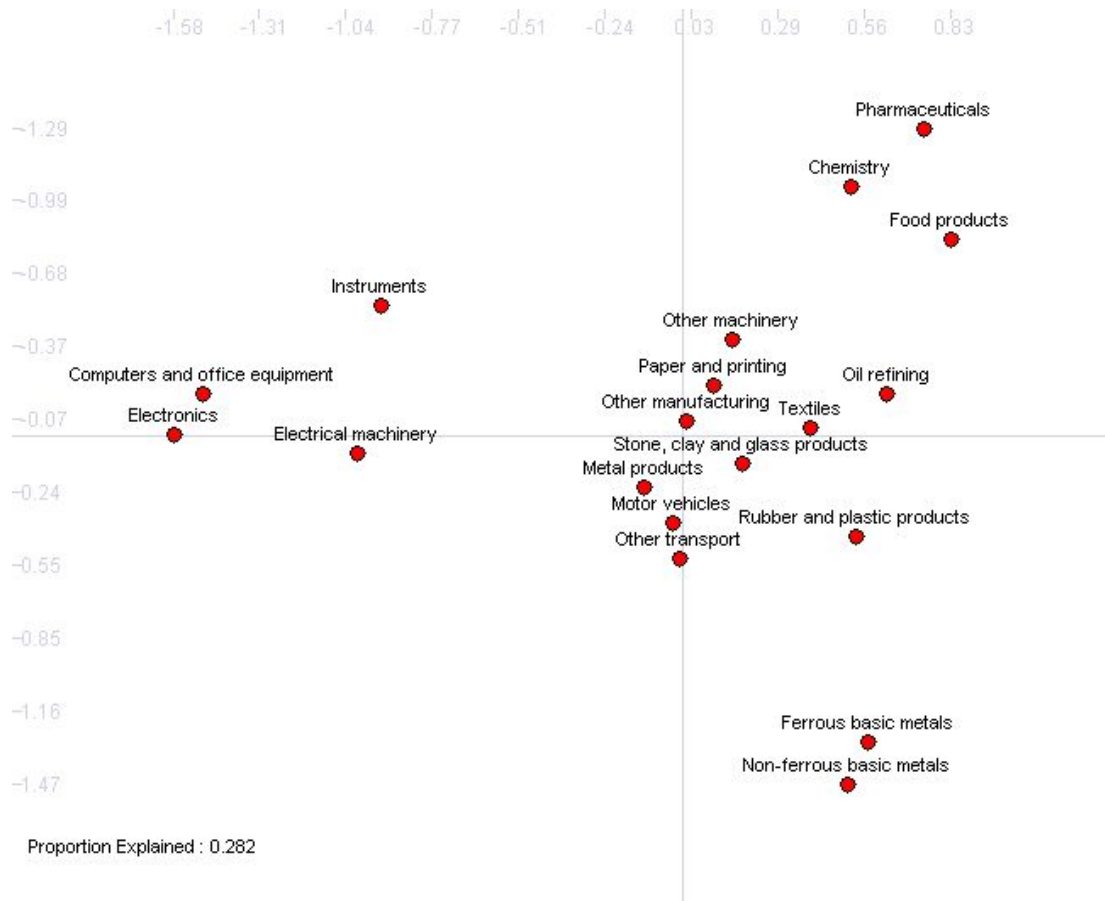


Figure 8. MDS map of mutual knowledge dependencies among the sectors.

The MDS results are in Figure 8, where we observe three distinct clusters, plus a large group of sectors in the center of the graph. Note that the fact that two sectors are near to each other in this map is an indication of the fact that they have intensive knowledge exchange relationships. The three (peripheral) clusters indeed correspond closely to the general intuition about technological relatedness. The first cluster, on the left of the figure, includes sectors that are strongly related to ICTs (electronics, electrical Machinery, computers and office equipment, and finally instruments somewhat closer to the center). The sectors of the second

cluster (pharmaceuticals, chemistry and food products) share an agri-bio focus. The third cluster (ferrous and non-ferrous basic metals) is metallurgy-based. As the pervasiveness indicators suggested, the large group of sectors in the middle show a much weaker mutual interdependence structure, and the dependence of each sector in the center is rather distributed. These sectors also have relations with each of the three other clusters.

This map of technological relatedness suggests that the dual structure of the relationship between science intensity and pervasiveness of knowledge flows is rooted in the existence of the three clusters. Both the agri-bio and ICT cluster both have strong internal knowledge flows, and this leads to the emergence of their central sectors (e.g., pharmaceuticals in the agri-bio cluster, electronics and computers in the ICT cluster) as net knowledge suppliers. However, the influence of these central sectors is much stronger within the cluster than in the total system, leading to their limited pervasiveness.

This is much different for the metallurgy cluster. This is made up of only two sectors, which are strongly related, but are also small, and have a more pervasive linkage to the rest of the system. In other words, they lack the relatively strong internal dynamics found in the other clusters, and hence appear as much more pervasive.

Summarizing, the analysis suggests a dichotomy within the group sectors that are net suppliers of knowledge. On the one hand, there is a group of sectors that is a net supplier of knowledge mainly as a result of very intensive connections within a specialized cluster of technologically related sectors. This group includes two broad clusters, i.e., ICT-related and bio-food related. On the other hand, there is a group of sectors that is both science intensive and has intensive knowledge flows to and from a relatively large set of other sectors. This includes a set of sectors that are traditionally seen as low-tech, but appear as science intensive in our data (in particular the basic metals sectors and other materials related sectors).

8. Summary and conclusions

We have proposed a methodology, broadly related to economic input-output analysis, that can be used to analyze intersectoral knowledge flows. The focal point of the analysis is the extent to which economic sectors use knowledge from the scientific domain, and transfer this knowledge through the system of interrelated economic sectors. Previously, the average amount of citations made to the scientific literature in a patent produced in a particular sector has been used as a proxy of the science intensity of R&D activities in the sector. Our analysis provides an intersectoral insight into this indicator.

We propose measures that assess the net supply of knowledge to other sectors (i.e., the amount of scientific knowledge supplied to other sectors minus the amount used from other sectors, the so-called net multiplier), the knowledge self-reliance of the sectors (i.e., the relative amount of knowledge that the sector uses from or supplies to itself), and the pervasiveness of knowledge use and supply (i.e., the extent to which a broad range of other sectors is served by knowledge flows from a sector, or the extent to which knowledge is sourced from a broad range of other sectors).

We arrive at the following empirical results. First, we show that high science intensity generally also implies net supply of knowledge to other sectors. Thus, it seems to be the case that high science intensity (a high amount of science citations per patent) is indeed an indication of the potential of a sector to diffuse scientific knowledge into the economic system. Second, we show that science intensive sectors also rely to an important extent on their own knowledge imports. The diagonal elements of our sector-by-sector knowledge flow matrix carry relatively much weight in the science intensive sectors. This is a first indication of the fact that scientific knowledge is highly specialized and specific. Finally, we show that the number of science citations per patent is not a good indicator for the knowledge pervasiveness.

ness of sectors. Such pervasiveness, both in terms of knowledge use and supply, appears in our analysis as a dimension that is quite separate from science intensity as measured in the traditional way. In particular, the traditional high-tech sectors, which are only a subset of the science intensive sectors, are particularly non-pervasive, especially so in terms of their knowledge supply to other sectors. They tend to cater knowledge to and from a subset of sectors, including themselves. We show that these subsets of sectors form closely interacting knowledge clusters (in particular, ICT and bio-food). Other highly science intensive sectors (materials and machinery) are more pervasive than these high-tech sectors.

In general terms, our results thus stress two broad conclusions with regard to the economic nature of scientific knowledge. First, sectors differ to an important degree with regard to the degree they are capable of transferring knowledge from the scientific to the economic domain. The economy relies on a relatively small range of sectors to achieve this transfer. At the same time, our second conclusion argues that often, scientific knowledge is highly specialized and specific to the sectoral context in which it is introduced in the economy. When this is the case, once knowledge is introduced by the science intensive sectors, it tends to stay within a limited cluster of technologically related sectors.

We can speculate that the latter phenomenon is related to the specific nature of knowledge and production relations in which knowledge is applied. How it comes about is something that our analysis cannot explain. But it is clear from our results that the specific and specialized nature of knowledge provides a limit to the intersectoral diffusion of it, and this constitutes a potential source of unbalanced growth and development between economic sectors, much in line with what is observed and analyzed in intersectoral economic analysis.

References

- Cohen, W. M., R. R. Nelson, and J.M. Walsh (2002). "Links and impacts: the influence of public research on industrial R&D." *Management Science* **48**(1): 1-23.
- Dietzenbacher, E. (2005), 'More on Multipliers', *Journal of Regional Science*, **45**, pp. 421–426.
- Dosi, G., P. Llerena, and M. Sylos Labini (2006). "The relationships between science, technologies and their industrial exploitation: An illustration through the myths and realities of the so-called 'European Paradox'." *Research Policy* **35**: 1450-1464.
- Fleming, L. and O. Sorenson (2004): Science and diffusion of knowledge, *Research Policy*, **33**, 1615-1634
- Ghosh, A. (1958), 'Input-Output Approach in an Allocation System', *Economica*, **25**, pp. 58-64.
- Hall, B.H., Jaffe, A.B. and M. Trajtenberg (2002), 'The NBER Patent-citations data file: lessons, insights and methodological tools', in Jaffe & Trajtenberg (2002), pp. 403-460.
- Hicks, D., T. Breitzman, et al. (2001). 'The changing composition of innovative activity in the US — a portrait based on patent analysis', *Research Policy* **31**: 681–703.
- Jaffe, A.B., M. Trajtenberg and R. Henderson (1993). Geographic localization of knowledge Spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, **108**, 577-598.
- Jaffe, A. B. and M. Trajtenberg (2002), *Patents, Citations & Innovations. A Window on the Knowledge Economy*, Cambridge MA: MIT Press.
- Johnson, D. and R.E. Evenson (1997), "Innovation and Invention in Canada", *Economic Systems Research*, vol. 9, pp. 177-192.
- Kline, S. J. and N. Rosenberg (1986). 'An overview of innovation', in: *The positive sum strategy: harnessing technology for economic growth*, edited by R. Landau and N. Rosenberg. Washington DC, National Academic Press.

- Leydesdorff, L. (2004). "The University–Industry Knowledge Relationship: Analyzing Patents and the Science Base of Technologies." *Journal of the American Society for Information Science and Technology* **55**: 991–1001.
- Los, B. (1999), *The impact of research and development on economic growth and structural change*, PhD thesis University of Twente.
- Maurseth, P.-B. and B. Verspagen (2002), "Knowledge Spillovers in Europe. A Patent Citation Analysis", *Scandinavian Journal of Economics*, vol. 104, pp. 531-545.
- Meyer, M. (2002). "Tracing knowledge flows in innovation systems." *Scientometrics* **54**: 193–212.
- Michel, J. and B. Bettles (2001). Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics*, **51**, 185–201.
- Miller, R.E. and Blair, P.D. (1985). *Input-output analysis. Foundations and extensions*, Englewood Cliffs: Prentice Hall.
- Narin, F., K.S. Hamilton and D. Olivastro (1997), "The Increasing Linkage between U.S. Technology and Public Science", *Research Policy*, vol. 26, pp. 317-330.
- Oosterhaven and Stelder (2002), 'Net multipliers avoid exaggerating impacts: with a bi-regional illustration for the Dutch transportation sector', *Journal of Regional Science*, 42, pp. 533-543.
- Pasinetti, L. L. (1981). *Structural Change and Economic Growth. A Theoretical Essay on the Dynamics of the Wealth of Nations*. Cambridge, Cambridge University Press.
- Sánchez-Chóliz, J. and R. Duarte, (2003), Analysing pollution by way of vertically integrated coefficients, with an application to the water sector in Aragon, *Cambridge Journal of Economics*, Vol. 27, pp. 433–448.
- Scherer, F.M. (1982), "Inter-Industry Technology Flows and Productivity Measurement", *Review of Economics and Statistics*, vol. 64, pp. 627-634.
- Tamada, S., Y. Naito, et al. (2006). "Significant difference of dependence upon scientific knowledge among different technologies." *Scientometrics* **68** (289-302).
- Van Moergastel, T., Slabbers, M. and B. Verspagen (1994), *MERIT concordance table: IPC - ISIC (rev. 2)*, MERIT Research Memorandum 2/94-004, University of Maastricht.
- Verspagen, B. (1997), "Measuring Inter-Sectoral Technology Spillovers: Estimates from the European and US Patent Office Databases", *Economic Systems Research*, vol. 9, pp. 47-65.
- Webb, C., H. Dernis, D. Harhoff and K. Hoisl (2004). A first set of EPO patent database building blocks for analysing European and international patent citations. *OECD mimeo*.

Appendix 1 – Correcting for the patent system truncation problem

As explained in the main text, the construction of our ('raw') matrix of technology flows is based on a count of citations made by a single generation (year) of EPO patents. Because we need information on the sector of origin of the cited patent, we must limit this count to those cited patents that are inside the EPO-system. Thus, if, for example, a patent cites 5 other patents, and 2 of those cited patents are EPO patents and the other 3 are non-EPO patents, we must construct our matrix on the basis of the 2 cited EPO patents only.

We can (must) assume that, in the aggregate statistical picture that emerges out of our counts, the sectoral distribution of the 3 non-EPO patents is identical to that of the 2 EPO-patents. Under this assumption, we can 'scale up' the actual number of citations made to EPO-patents by the factor $(NE+E)/E$, where NE is the number of citations made to non-EPO patents, and E is the number of citations made to EPO-patents. In the example, $NE=3$ and $E=2$, and hence the scaling factor becomes $5/2$. Multiplying the total of 2 citations (to EPO-patents) by this factor yields 5 citations.

Such an up-scaling of patent citations is necessary because in our analytical model, the share of $cnpl_i$ in the column-sum c_i plays an important role. As a result of the patent system truncation, we observe, at the patent level, all science citations, but not all patent-citations. Hence we must 'correct' using the up-scaling factor.

In practical terms, we can compute the up-scaling factor at the sector level, and apply it column-wise to the matrix system (1). To be precise, we calculate an up-scaling factor for every sector i by dividing the total number of patent citations made by patents in that sector, by citations to EPO patents made by the same set of patents. We multiply each value c_{ji} by the up-scaling factor for i .

A complication in our database is caused by the fact that a single idea (invention) may be patented under different patent systems (this is the notion of international patent families discussed in the main text). Thus, in terms of the example in Diagram 1 in the main text, it may be the case that patent B2 (filed under the USPTO), and patent B3 (filed under the EPO) in fact refer to the same invention. In that case, we should consider them as a single node, rather than two separate ones.

As described also in the main text, we have partly solved this problem, by using the (updated) OECD database on patent equivalents. But the OECD equivalents database provides only equivalents for EPO patents, and hence does not account for the fact that some of the non-EPO patents that are left in the list of patents cited by EPO patents, are in fact equivalent. For example, if we find that our set of EPO patents cite both USPTO patent 12345 and Australian patent 54321, these two patents may either be equivalent, or not.

Thus, some of the citations to non-EPO-patents are in fact redundant, because the non-EPO patent may be equivalent to a particular EPO-patent that is also cited. The up-scaling factor that we introduced above does not take this into account. Without accounting for the equivalence relations explicitly, we cannot solve this problem. Hence, all we can do is provide results on the basis of non-up-scaled data (representing the case that all non-EPO-patents have an EP-equivalent, and that the corresponding citation is recorded) as well as up-scaled data (representing the case that none of the non-EPO-patents have an EP-equivalent). Our default will be to present results based on the non-up-scaled data, and discuss briefly how up-scaling changes the broad conclusions. More precise results are available on request.

Appendix 2. Indicators used for graphing technological clusters

The first indicator is based on the matrix \mathbf{D} , of which each element d_{ij} indicates the share of industry i -specific scientific knowledge in the total stock of scientific knowledge that is accumulated and eventually transmitted to industry j patents of year t , (based on direct and indirect knowledge flows). Thus, each column j of this matrix (which adds up to 1) decomposes each unit of the accumulated knowledge in sector j patents into its sector-specific knowledge components.

Based on matrix \mathbf{D} we define a proximity matrix which captures the pair-wise dependency of the sectors. We call this symmetrical square matrix \mathbf{SimD} , and the elements of this are defined as

$$SimD_{ij} = \left(\frac{D_{ij}}{1-D_{ii}} \cdot \frac{D_{ji}}{1-D_{jj}} \right)^{\frac{1}{2}} \text{ for } i \neq j,$$

$$SimD_{jj}=1 \text{ for each sector } j.$$

Since $\sum_{i=1}^{19} D_{ij} = 1$ for each sector j , $1-D_{jj}$ is the share of all non-sector j -specific types of knowledge embedded in sector j patents, and therefore $D_{ij}/1-D_{ii}$ is the share of sector i -specific knowledge in all non-sector j -specific knowledge embedded in sector j patents of year t . The presence of $1-D_{jj}$ and $1-D_{ii}$ in the denominator (instead of 1) filters the effects of the heterogeneity in self citation rates from these pair-wise mutual dependence indicators, and the multiplicative nature of the indicator emphasizes the mutuality by the simultaneous consideration of the importance of sector i -specific knowledge in sector j patents and the importance of sector j -specific knowledge in sector i patents. Thus, $SimD_{ij}$ measures the extent to which sectors i and j are mutually dependent in terms of the sector-specific knowledge they *use* from each other.

We can also build an alternative matrix \mathbf{SimK} of the mutual knowledge dependencies among the sectors on the basis of the matrix \mathbf{K} :

$$SimK_{ij} = \left(\frac{K_{ij}}{1-K_{ii}} \cdot \frac{K_{ji}}{1-K_{jj}} \right)^{\frac{1}{2}} \text{ for } i \neq j,$$

$$SimD_{jj}=1 \text{ for each sector } j.$$

Since \mathbf{K} is based on the output coefficient matrix \mathbf{B} , $SimK_{ij}$ measures the extent to which sectors i and j are mutually dependent in terms of the sector-specific knowledge they *supply* to each other.

Finally, we combine the supply and use similarity measures into a single metric, so that we cover \mathbf{SimK} and \mathbf{SimD} simultaneously. This is the matrix \mathbf{SimDK} that is defined by $SimDK_{ij} = SimD_{ij} \cdot SimK_{ij}$. This is the matrix that is used as an input into the MDS analysis.

Appendix 3. Indicator scores of the sectors

Definitions are given in the text, all data refer to calculations made with 1992 data.

Sector	Backward Multiplier	Forward Multiplier	Net science multiplier	Science citations per patent	Self-use	Self-supply	Use pervasiveness	Supply pervasiveness
Electrical machinery	4.91	5.05	0.97	0.48	0.55	0.56	6.64	8.30
Electronics	3.39	3.90	1.14	1.05	0.78	0.69	5.11	6.65
Chemistry	4.67	5.17	1.03	0.62	0.52	0.50	3.60	8.10
Pharmaceuticals	2.84	3.98	1.34	1.91	0.81	0.61	3.89	4.92
Oil refining	6.76	6.01	0.43	0.18	0.18	0.43	4.62	6.17
Motor vehicles	9.11	7.38	0.54	0.13	0.29	0.53	10.04	7.44
Other transport	7.71	6.60	0.66	0.19	0.29	0.44	10.07	6.47
Ferrous basic metals	5.00	5.21	1.08	0.56	0.47	0.44	8.64	9.24
Non-ferrous basic metals	4.41	5.11	1.31	0.83	0.55	0.42	8.70	9.71
Metal products	6.76	6.08	0.72	0.21	0.29	0.40	10.39	8.32
Instruments	4.74	4.87	1.00	0.61	0.63	0.63	8.05	9.55
Computers and office equipment	3.61	4.00	1.10	0.98	0.74	0.68	4.26	5.71
Other machinery	6.95	6.14	0.59	0.18	0.28	0.47	8.71	10.26
Food products	3.25	4.34	1.33	1.43	0.59	0.44	3.07	6.10
Textiles	6.75	6.09	0.68	0.22	0.21	0.30	7.94	6.57
Rubber and plastic products	7.80	6.39	0.31	0.08	0.09	0.29	9.04	7.65
Stone, clay and glass products	5.42	5.65	1.16	0.53	0.40	0.35	10.03	9.21
Paper and printing	5.79	5.51	0.82	0.36	0.30	0.37	9.18	9.59
Other manufacturing	6.30	5.62	0.60	0.24	0.21	0.35	9.15	9.33

The UNU-MERIT WORKING Paper Series

- 2007-01 *Developing science, technology and innovation indicators: what we can learn from the past* by Christopher Freeman & Luc Soete
- 2007-02 *The impact of innovation activities on productivity and firm growth: evidence from Brazil* by Micheline Goedhuys
- 2007-03 *Estimations of US debt dynamics: Growth cum debt and the savings glut in Kouri's model* by Thomas Ziesemer
- 2007-04 *States and Firms on the Periphery: The Challenges of a Globalising World* by Gabriel R.G. Benito & Rajneesh Narula
- 2007-05 *How Do Consumers Make Choices? A Summary of Evidence from Marketing and Psychology* by Zakaria Babutsidze
- 2007-06 *Inter-firm Technology Transfer: Partnership-embedded Licensing or Standard Licensing Agreements?* by John Hagedoorn, Stefanie Lorenz-Orlean & Hans Kranenburg
- 2007-07 *The Behavior of the Maximum Likelihood Estimator of Dynamic Panel Data Sample Selection Models* by Wladimir Raymond, Pierre Mohnen, Franz Palm & Sybrand Schim van der Loeff
- 2007-08 *Location and R&D alliances in the European ICT industry* by Rajneesh Narula & Grazia D. Santangelo
- 2007-09 *How do social capital and government support affect innovation and growth? Evidence from the EU regional support programmes* by Semih Akcomak & Bas ter Weel
- 2007-10 *The Micro-Dynamics of Catch Up in Indonesian Paper Manufacturing: An International Comparison of Plant-Level Performance* by Michiel van Dijk & Adam Szirmai
- 2007-11 *Financial Constraint and R&D Investment: Evidence from CIS* by Amaresh K Tiwari, Pierre Mohnen, Franz C. Palm & Sybrand Schim van der Loeff
- 2007-12 *The Spatial Hierarchy of Technological Change and Economic Development in Europe* by Bart Verspagen
- 2007-13 *The origins and implications of using innovation systems perspectives in the design and implementation of agricultural research projects: Some personal observations* by Andy Hall
- 2007-14 *Technology supply chain or innovation capacity?: Contrasting experiences of promoting small scale irrigation technology in South Asia* by Andy Hall, Norman Clark and Guru Naik

- 2007-15 *Are firms that received R&D subsidies more innovative?* by Charles Bérubé & Pierre Mohnen
- 2007-16 *Foreign direct investment and firm level productivity. A panel data analysis* by Geoffrey Gachino
- 2007-17 *Technological spillovers from multinational presence towards a conceptual framework* by Geoffrey Gachino
- 2007-18 *Technological capability building through networking strategies within high-tech industries* by Wim Vanhaverbeke, Bonnie Beerkens and Geert Duysters
- 2007-19 External technology sourcing: the effect of uncertainty on governance mode choice by Vareska van de Vrande, Wim Vanhaverbeke & Geert Duysters
- 2007-20 *Exploration and Exploitation in Technology-based Alliance Networks* by Wim Vanhaverbeke, Victor Gilsing, Bonnie Beerkens, Geert Duysters
- 2007-21 *ICT Externalities: Evidence from cross country data* by Huub Meijers
- 2007-22 *Knowledge Flows, Patent Citations and the Impact of Science on Technology* by Önder Nomaler & Bart Verspagen