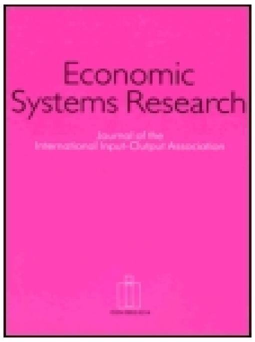

Economic Systems Research

Journal of the International Input-Output Association

Economic Systems Research

Routledge
Taylor & Francis Group

ISSN: 0953-5314 (Print) 1469-5758 (Online) Journal homepage: http://www.tandfonline.com/loi/cesr20

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

önder Nomaler & Bart Verspagen

**To cite this article:** önder Nomaler & Bart Verspagen (2008) Knowledge Flows, Patent Citations and the Impact of Science on Technology, Economic Systems Research, 20:4, 339-366, DOI: 10.1080/09535310802551315

**To link to this article:** https://doi.org/10.1080/09535310802551315

Published online: 05 Dec 2008.

*Economic Systems Research*
*Vol. 20, No. 4, 339–366, December 2008*

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

ÖNDER NOMALER* & BART VERSPAGEN**,†

*Department of Technology Management, Eindhoven University of Technology, The Netherlands;
**Department of Economics, Maastricht University, The Netherlands; †MERIT, United Nations University, Maastricht, The Netherlands

(Received July 2007; revised July 2008)

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 of 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.*

KEY WORDS: Knowledge input–output analysis, knowledge flow matrices, science-to-technology transfer

## 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

*Correspondence Address*: Bart Verspagen, Department of Economics, Maastricht University, PO Box 616, NL-6200 MD Maastricht, The Netherlands. Email b.verspagen@algec.unimaas.nl.

0953-5314 Print/1469-5758 Online/08/040339–28 © 2008 The International Input–Output Association
DOI: 10.1080/09535310802551315

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 *et al.*, 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 *et al.*, 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 *et al.* (2002, p. 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 *et al.*, 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 *et al.* (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. Hicks *et al.*, 2001; Leydesdorff, 2004; Tamada *et al.*, 2006). Meyer (2002, p. 197) 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 (Meyer, 2002, 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, 1982; 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 present the results. Finally, in Section 7, 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 labor, 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 we do not consider the role of capital and labor 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 part;[1] (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 Figure 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 connecting 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 Figure 1 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.

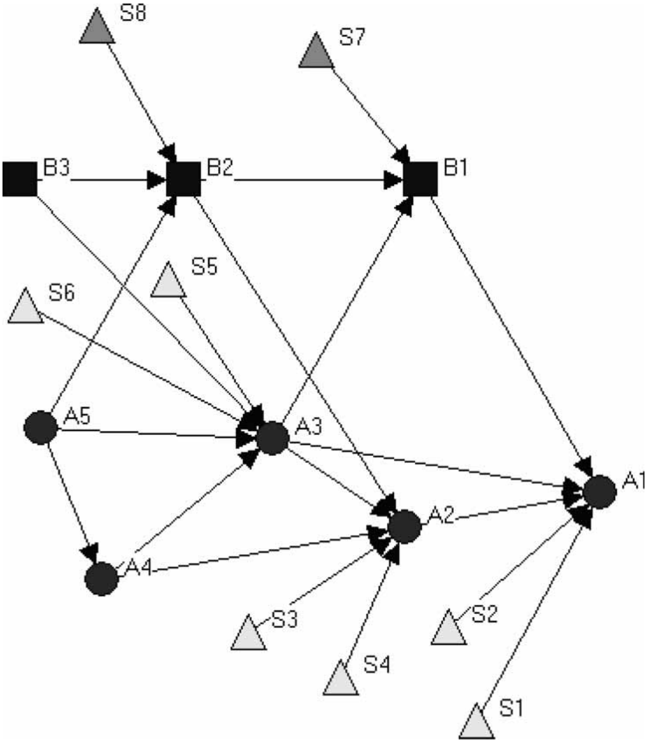

**Figure 1.** Stylized patent-citation network graph

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.

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 *et al.*, 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 four 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 span over the borders of technology sectors, i.e. the science ideas S7 and S8 flow into sector B patents and will eventually also flow into ideas of sector A, and similarly the science ideas S5 and S6 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 Figure 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 Figure 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 Figure 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 (EPO), the US Patent and Trademark Office (USPTO), or other national patent offices). In the example of Figure 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 Figure 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 assuming 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 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 (Figure 1). Although the use of patent citations has by now become quite widespread in the literature (see, for example, 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 systems other 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 with 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 applications[3] to analyze technology flows. Our data are extracted from the Bulletin CDROM issued by the European Patent Office, 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 *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 patent systems other 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

We now set out the procedure by which we construct a knowledge flow table, or citations flow table, which resembles an economic input–output table. This enables us to draw upon a number of established methods from input–output economics to analyze knowledge flows. The analogy that we propose rests on the idea that a citation can be interpreted as a delivery of knowledge (by the cited patent to the citing patent). Furthermore, we distinguish between intermediate 'deliveries' (knowledge flows or citations between industrial sectors), and deliveries from a primary factor (knowledge flows from the science sector).

There is one crucial problem in this analogy between input–output economics and the citations flow system that we propose, relating to the period of observation. In input–output economics, the period in which we observe production flows (a year) typically captures a large part of the chain of intermediate deliveries that make up production. For example, consider the production of a wooden chair in the furniture industry that requires wood and (metal) screws. The production of wood in turn requires inputs from forestry, and the metal screws require steel from the steel industry, which requires iron ore from the mining industry. Moreover, at various stages of production, services from the transport industry are used. The typical situation is now that we observe a large part of this chain of deliveries in the time span of a single year. Inputs that we do not observe being produced within the single year, like hammers and other tools, are treated as primary inputs (capital).

In our citations flow system, however, the typical time lags are much longer than what is observed in a production system (the typical citation lag is several years). In terms of the example, we observe the producer of the wooden chair obtain the wood and metal screws, but we do not observe the delivery of the steel that goes into the screws, nor the trees that make up the wood. These deliveries were made earlier, before we were able to observe them. Note that we do observe some deliveries of steel to the screw-making industry, but these are associated with *future* deliveries of screws to other sectors. Because production scale generally varies over time, interpreting these flows as associated with the *current* amount of screws used in furniture making would be wrong, and lead to a biased view of the intermediate flows.

The way out of this problem is to assume fixed coefficients. Thus, we may use the currently observed input coefficient of steel for the production of screws to derive the amount of steel that was 'indirectly' necessary to produce the wooden chairs that we observe in the current period. In this way, we need to construct a citations flow table, rather than actually observing it directly.

Thus, on the basis of data on citations made by patents of a given year $t$, plus an analytical framework (explained below), we construct an input–output table that describes the gradual knowledge accumulation process that has taken place in the time interval $[-\infty, t]$. Table 2 (later, at the end of this section) fully describes the construction of this table. In a second stage (i.e. the next section) we will describe our various indicators on science intensity and pervasiveness of knowledge flows, all of which are based on the constructed knowledge flow table (input–output table). Some of these indicators (such as the multipliers) are in the standard toolkit of the input–output economics, while some are custom-built by us in order to shed further insights on the intersectoral pervasiveness of knowledge flows.

Before we start, let us make a note on the matrix notation necessary to construct the citations flow table. 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 $\mathbf{X}$, a column vector (of size $n$) by a boldface small letter as in $\mathbf{x}$, and a row vector as $\mathbf{x}'$, where a prime 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 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 $\hat{\mathbf{x}}$ 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. $\mathbf{i}' = [1, 1, 1, \ldots, 1]$).

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 **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

$$c_j^t = cnpl_j^t + cpl_{1j}^t + cpl_{2j}^t + \ldots + cpl_{nj}^t, \tag{1}$$

where $cnpl_j^t$ stands for the number of citations to non-patent literature made by year $t$, industry $j$ patents[4] 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 *et al.* (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.

As in input–output economics, let us express the numbers of citations to various entities in terms of input coefficients. The share of year $t$ citations to the science sector (i.e. the non-patent literature) in all year $t$ citations of industry $j$ is defined as $v_j^t \equiv cnpl_j^t / c_j^t$, and the share of year $t$ citations of industry $j$ patents to industry $i$ patents is $a_{ij}^t \equiv cpl_{ij}^t / c_j^t$. Thus, $v_j^t + \Sigma_i a_{ij}^t = 1$ by definition.

The patents that are cited by year $t$ patents also carry some science content due to their direct citations to the non-patent literature. Considering this indirect effect, it is clear that the science content in year $t$ patents is higher than just $v_j^t \equiv cnpl_j^t / c_j^t$. This indirect effect continues backwards in time *ad infinitum* because every generation of patents cites another older generation of patents, as well as science directly. Under the assumption (used throughout the paper) that the input coefficients ($a$ and $v$) are constant over time, one can calculate the total (i.e. direct plus indirect) science content accumulated in year $t$ patents by backward summation over the time horizon $[-\infty, t]$.

Figure 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 dark grey and light grey. The squares represent the science sector. The knowledge from the science sector is further broken down into two types of science inputs, black and white, based on the sector that they feed into. Thus, scientific knowledge is broken down into industry-specific categories. The citations of the patents of the light grey industry bring in only the white type of scientific knowledge and those of the dark grey industry bring in only the black type of scientific knowledge. 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.

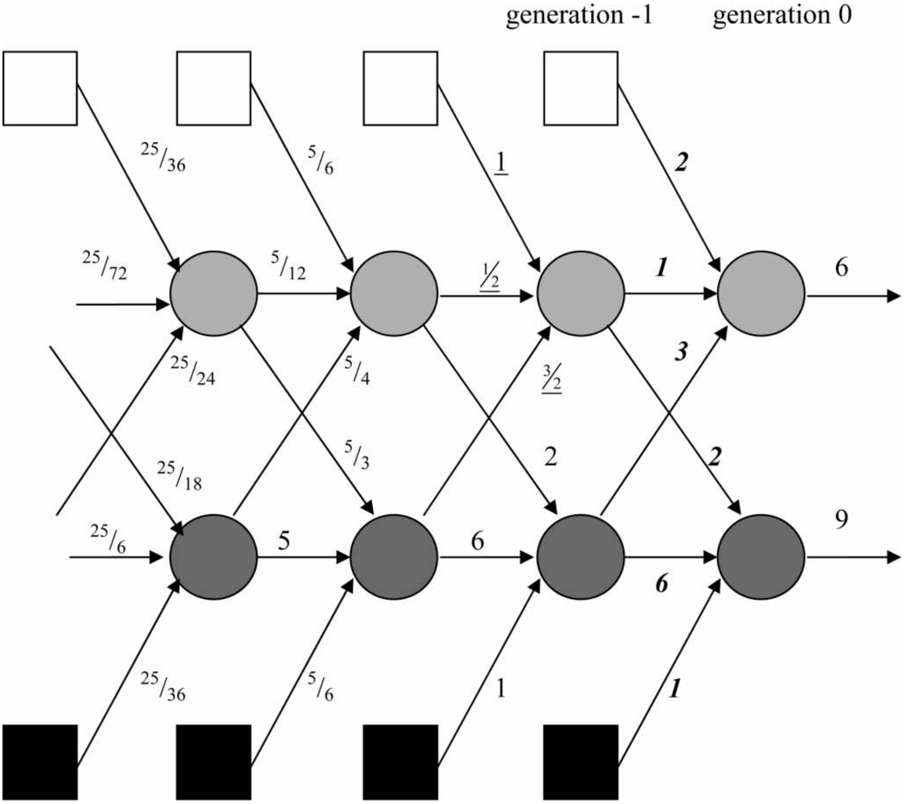

**Figure 2.** Stylized citation network graph used to illustrate formal approach

Assume that we observe a subset of these observations, i.e. the ones made by 'generation 0' patents (this is the rightmost set of patents). These 'actually-observed' citations are indicated by a bold and italic font in the diagram. Thus, we see that for generation $t = 0$, the light grey sector makes a total of six citations (incoming arrows), of which three are to patents of the dark grey sector, one is to patents of the light grey sector itself, and two are to science references. The dark grey sector makes a total of nine citations, of which six are to the dark grey sector itself, two are to patents of the light grey sector, and one is to the science sector. These figures can be used to calculate the input coefficients (the *a*s and *v*s) which set up the specific realizations of **A** and **v**′ as given in Table 1.

**Table 1.** Input coefficients

| | | Light grey | Dark grey |
|---|---|---|---|
| **A** | Light grey | $^1/_6$ | $^2/_9$ |
| | Dark grey | $^3/_6 = {}^1/_2$ | $^6/_9 = {}^2/_3$ |
| **v**′ | Science | $^2/_6 = {}^1/_3$ | $^1/_9$ |

Under the assumption of a fixed **A** and **v** prior to generation 0, and the assumption that total knowledge that flows into each circular node of the diagram is equal to total knowledge that flows out from that node, we can 'estimate' the number of citations that must theoretically be present on the other arrows in the figure (prior to generation 0). These numbers are indicated in the diagram in non-bold and non-italics type. As an example, let us see how this works for citations made by generation –1, i.e. the numbers that are underlined in the diagram. It was already observed that a total of three citations leave the light grey sector at generation –1 (output), and this must be matched by an equal input. Of this total input of three, ($v_{light} = {}^1/_3) \times 3 = 1$ comes from the science sector, ($a_{dark,light} = {}^1/_2) \times 3 = {}^3/_2$ comes from the dark grey sector, and ($a_{light,light} = {}^1/_6) \times 3 = {}^1/_2$ comes from the light grey sector itself ($1 + {}^3/_2 + {}^1/_2 = 3$). In this way, all remaining links in the diagram have been filled in.

The interest of the analysis is in finding the total, i.e. directly and indirectly, accumulated science content (non-patent literature references) embedded in year *t* patents of each other industry *j*. This corresponds to finding the sum of all values on the arrows originating from the black (bottom) and white (top) squares (science sector). Obviously, in order to obtain this total sum, we would have to extend the diagram infinitely to the left. However, using analytical methods simplifies the process, and we will show below that the total science content is equal to the total amount of citations made by patents of generation 0. In the diagram, the sum of direct and indirect science contributions to generation 0 is equal to 15 ( $= 6 + 9$).

To show how this fundamental property of the system arises, define the $n \times n$ matrix **A,** whose elements are the $a^t_{ij}$ 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 that supply these inputs. Let us also construct an $n \times n$ diagonal matrix $\hat{\mathbf{v}}$ with elements $v^t_j$ on the diagonal and zero, otherwise.

Clearly, $\hat{\mathbf{v}}$ gives the fraction of 'direct science content' embedded in a single patent (per industry). The first round of indirect science content can be represented by the matrix product $\hat{\mathbf{v}}\mathbf{A}$, 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 **A**). Similarly, we can envisage a third round of embedded science, represented by $\hat{\mathbf{v}}\mathbf{A}^2$, and a fourth round $\hat{\mathbf{v}}\mathbf{A}^3$, etc. The complete citation chain, for a single patent, is described by the following matrix product:

$$\mathbf{D} = \hat{\mathbf{v}}(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^\infty) \tag{2}$$

where **I** is the ($n \times n$) identity matrix.

The term in brackets 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 **A** are strictly less than one, and all coefficients are non-negative (both of which are naturally satisfied in our matrix **A**). Thus, equation (2) can also be written as:

$$\mathbf{D} = \hat{\mathbf{v}}(\mathbf{I} - \mathbf{A})^{-1} \tag{3}$$

Let us now define the vector $\mathbf{s}' \equiv (\mathbf{Dc})'$, where **c** is the column vector of the $c^t_j$ values, i.e.

**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 **D** all sum to 1.[5] This implies that the sum of the elements of **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, which is what we set out to show. 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.

In terms of the example in Figure 2 (where **A** is given in Table 1, and $\mathbf{c}' = [6{,}9]$), the actual calculation $\mathbf{s}' \equiv (\mathbf{Dc})'$ yields $\mathbf{s}' = [8{,}7]$. Hence, our assumption on the time invariance of the input coefficients implies that, throughout the period $[-\infty,t]$, a total of 8 units of pure (white/top) science (theoretically) must have been introduced into the system by the light grey sector, and 7 units of pure (black/bottom) science by the dark grey sector. In a version of a diagram that would extend infinitely to the left, the sum of flows originating from the bottom (black) science row would be 7, that from the top (white) 8. Note also that, indeed, the sum of elements of **s** is equal to the sum of elements of **c** $(8 + 7 = 6 + 9 = 15)$.

So far, our perspective has been backward, i.e. we asked how many pure science units are embodied in the citations made by patents published in year $t$. For that, we made a decomposition that took account of the citations made by industries at a given point in time. Similarly, by taking account of the citations received by the industries, we may take a forward perspective, asking how much knowledge (in pure science units) is passed on to future generations. Let us introduce the column vector **g** to denote this forward flow of knowledge, where the convention is that the $j$th element of **g** denotes the amount of knowledge passed on by patents of industry $j$ to future generations.

Obviously, we do not observe the citation behavior of future patents (yet), and hence there is no way in which we can observe **g**. However, if we also apply our assumption that at each generation incoming knowledge flows are equal to outgoing knowledge flows to forward streams, we deduce that **c** will be equal to **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. In the case of the example in Figure 2, the quantities of knowledge passed by generation 0 patents onto future generation are thus $\mathbf{g} = [6,\ 9]$ which are indicated by the values on the two rightmost arrows.

The cumulative implications of the knowledge accumulation process (which takes place during the time interval [-∞, t]) as described so far can be put together in tabular form as in Table 2 (the cumulative citations-flow table). Readers who are familiar with the Leontief economic input–output system may recognize that this citations-flow table is, in many regards, similar to the economic input–output table, which breaks national accounts down to the sector level. The matrix **F** is reminiscent of the matrix of intermediate flows

**Table 2.** The construction of the citations-flow table

| | | |
|---|---|---|
| **INTERMEDIATE CITATION FLOWS IN PERIOD $-\infty$ TO *t*:**<br>$(n \times n)$<br>The matrix **F**, with elements $a_{ij} \cdot x_j$ | **TOTAL SCIENCE KNOWLEDGE PROVIDEDTO POST-*t* PATENTS**<br>$(n \times 1)$<br>$\mathbf{g} = \mathbf{c}$<br>**(by assumption): TOTAL CITATIONS MADE IN YEAR *t*** | **CITATIONS RECEIVED IN PERIOD $-\infty$ TO *t*:**<br>$(n \times 1)$<br>$\mathbf{y} = \mathbf{Fi} + \mathbf{g}$<br>$= (\mathbf{I}\text{–}\mathbf{A})^{-1}\mathbf{g}$<br>$= \mathbf{x}$ |
| **TOTAL SCIENCE KNOWLEDGE TAKEN IN PERIOD $-\infty$ TO *t*:**<br>$(1 \times n)$<br>$\mathbf{s}' = (\mathbf{Dg})'$<br>$= [\hat{\mathbf{v}}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{c}]'$ | | |
| **CITATIONS MADE IN PERIOD $-\infty$ TO *t*:**<br>$(1 \times n)$<br>$\mathbf{x}' = \mathbf{s}'\hat{\mathbf{v}}^{-1}$<br>$= [(\mathbf{I} - \mathbf{A})^{-1}\mathbf{c}]'$ | | |

of goods between sectors. A crucial difference, however, between our matrix **F** 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 constructing our matrix **F** for the period - ∞ to *t*, instead of directly observing it in entirety for a fixed period of time. The construction of **F** (equations (2) and (3) and the discussion around them) may seem somewhat tedious, but if one is interested in looking at indirect flows, it is essential that this procedure is followed, rather using the observed matrix **A** directly (as many studies have done, e.g. Johnson and Evenson, 1997; Verspagen, 1997).

Continuing the analogy to input–output economics, science inputs play the role of primary inputs, and $\mathbf{s}'$ 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 future patents (investment).

Scientific knowledge is brought into the system in a cumulative manner by the non-patent literature citations, given by the vector $\mathbf{s}'$, is transmitted forward (up to year *t*) by the patent citations given by the intermediate flow matrix **F**, 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 number of (patent-to-patent and patent-to-non-patent literature) citations made by the sectors are given by the row vector $\mathbf{x}'$, and similarly the total number of citations received by the sectors is given by the column vector **y**. Just similar to the equality of total expenditures to total output in an economic input–output system, in our system $\mathbf{x} = \mathbf{y}$, and similar to the equality of total final demand to total value added, in our system $\mathbf{s}'\mathbf{i} = \mathbf{i}'g$.

Going back to the example of Figure 2 for a final time, the reconstruction framework described in Table 2 yields:

$$\mathbf{F} = \begin{bmatrix} 4 & 14 \\ 12 & 42 \end{bmatrix}, \mathbf{x}' = \mathbf{y}' = [24, 63] \text{ and } \mathbf{s}' = [8, 7]$$

These figures indicate that in the period [-∞, *t*], the patents of the light grey sector must have made (and also received) a total of 24 citations, while the dark grey sector must have made (and also received) a total of 63 citations. Out of these 24 (63) citations made by the light (dark) grey sector, 8 (7) are directed to the science sector. These $8 + 7 = 15$ units of scientific knowledge accumulated in the system during the period [-∞, *t*] have been transmitted to the patents of period *t* through a total of $4 + 12 = 16$ ($14 + 42 = 56$) patent-to-patent citations made by the light (dark) grey sector.

## 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**.[7] 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 **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 *et al.* (2002, pp. 421–424) 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 $\hat{\mathbf{v}}$ upward. We discuss this problem in more detail in a working paper version of this article,[8] 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.

### *5.1. Indicators*

Based on the patent flow table, 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. Before that, note that our indicators analytically approach the transfer of knowledge from two (complementary) directions: The forward and the backward approach. This duality refers to the arrow of time through which knowledge accumulates. The backward approach looks backwards in time and decomposes already accumulated knowledge according to its potential sources. The forward approach, however, decomposes the pure categories of (scientific knowledge) according to where they are, respectively, eventually delivered. In other words, the backward approach considers the year $t$ patents of each industry as another sink of scientific knowledge and pursues a decomposition according to the various sources, while the backward approach considers each industry as a source that brings in scientific knowledge through citations to science made in the period $[-\infty, t]$ and for each source, traces where (i.e. which industries) the associated type of knowledge eventually ends up.

*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 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. $\mathbf{i}'(\mathbf{I}-\mathbf{A})^{-1}$. Similarly defined backward multipliers are also useful indicators in our patent citation flows table, although their interpretation is not quite analogous to input–output economics.

Let the vector $\boldsymbol{\lambda}_t$ denote the counterpart of the backward multiplier vector as applied to the constructed cumulative citation-flows table year $t$. That is, let

$$\boldsymbol{\lambda}_t' = \mathbf{i}'(\mathbf{I}-\mathbf{A}_t)^{-1} \tag{4}$$

Given this specification, for each industry $j$, the backward multiplier $\lambda_{tj}$ 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 Figure 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 a total of one. The backward multiplier measures by how much the value of the arrows between patents (circles) will have to increase to accommodate this.

Note that since $\mathbf{D} = \hat{\mathbf{v}}(\mathbf{I} - \mathbf{A})^{-1}$ as described in equation (3), one can also express the backward multipliers as $\boldsymbol{\lambda}'_t = \mathbf{i}'\hat{\mathbf{v}}^{-1}\mathbf{D}$, or equivalently as $\lambda_j = \Sigma_{i=1}^{N} d_{ij}/v_i$. 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 **A** that we collect from the data does not allow the calculation of output coefficients, once the cumulative patents citations flow table (Table 2) is constructed, one can calculate an output coefficient matrix $\mathbf{B}_t$, where $b_{ij}^t = f_{ij}^t/x_i^t$. Given this matrix, the vector of forward multipliers is calculated as

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

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).

While backward multipliers 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 trough 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. Figure 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 distributes these forward over the patents of a variety of industries.

*Net science multipliers*

The next indicator that we will discuss 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.

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 that is introduced into the patent system by industry $i$ patents during the time interval [-∞, $t$], 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 $\mu_{ti}$, which is the ratio of scientific knowledge introduced by *all* industries that 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 $\mu_t$, which is a vector whose $i$th element is equal to $\mu_j = g_i/s_i$, as follows:

$$\boldsymbol{\mu}_t' = \mathbf{s}'\hat{\mathbf{x}}^{-1}(\mathbf{I} - \mathbf{A})^{-1}\hat{\mathbf{g}}\hat{\mathbf{s}}^{-1} \tag{6}$$

On the basis of Table 2 (i.e. $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{c}$ and $\mathbf{s}' = [\hat{\mathbf{v}}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{c}]'$), 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 $\hat{\mathbf{x}}^{-1}$ is $1/\Sigma_{j=1}^{N} l_{ij}c_j$, whereas the $i$th element of the row vector $\mathbf{s}'$ is $v_i\Sigma_{j=1}^{N} l_{ij}c_j$, which implies that the term $\mathbf{s}'\hat{\mathbf{x}}^{-1} = \mathbf{v}'$. Consequently, given $\mathbf{v}'(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{i}'$, the calculation of $\boldsymbol{\mu}_t$ reduces to

$$\boldsymbol{\mu}_t' = \mathbf{g}'\hat{\mathbf{s}}^{-1} \tag{7}$$

*Self-reliance of sectors*

The matrix $\mathbf{D} = \hat{\mathbf{v}}(\mathbf{I} - \mathbf{A})^{-1}$ 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 matrix $\mathbf{B}_t$ as well as the diagonal matrix $\hat{\mathbf{r}}$, which gives the ratio $g_j/x_j$ (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

$$\mathbf{K} = (\mathbf{I} - \mathbf{B})^{-1}\hat{\mathbf{r}} \tag{8}$$

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 ($\mathbf{Ki} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{r} = \mathbf{i}$). Using $\mathbf{K}$, we introduce a similar measure to $\delta$, 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 versus 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 $\mathbf{D} = \hat{\mathbf{v}}(\mathbf{I} - \mathbf{A})^{-1}$ 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:

$$h_j^{src} = \frac{1}{\sum_{i=1}^{N,i\neq j} \left(d_{ij}/1 - d_{jj}\right)^2} \tag{9}$$

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 $h_j^{src}$ – 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:

$$h_j^{snk} = \frac{1}{\sum_{i=1}^{N,i\neq j} \left(k_{ij}/1 - k_{jj}\right)^2} \tag{10}$$

$h_i^{snk}$ 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 industries, and is a measure of how pervasively industry *i* influences the other industries in the system.

## 6. 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. Owing 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 are representative of the other years between 1985 and 2004.[9]

Table 3 documents the scores of the sectors on the indicators that we presented in the previous section. The leftmost column gives the 'direct' science intensity of the sectors (number of non-patent citations per patent), which corresponds to the indicator resulting from an analysis that does not consider intersectoral flows. Given our interest in the value added of the intersectoral approach, we start our empirical analysis by looking at the relationship between this indicator and the other indicators, which are specific to the intersectoral nature of the analysis. The correlation coefficients between the indicators, documented in Table 4, are useful in making this comparison.

Science citations per patent indicator correlate strongly negatively with backward and forward multipliers. This is intuitive because the multipliers measure the amount of patent-to-patent citations that are necessary to pull one additional unit of science knowledge from the past (backward multiplier) or carry forward one additional unit of science knowledge into the future (forward multiplier). Hence, the focus is on indirect effects through patent-to-patent citations. The science intensive sectors (those that have a high number of direct science citations per patent) by definition have a lower reliance on indirect science citations.

On the other hand, the correlation with the net multiplier is strong and positive. This captures to what extent a sector is a net supplier or user of knowledge. The positive correlation suggests that highly science intensive sectors are generally responsible for a larger fraction of the knowledge introduced into the system than the fraction of knowledge used. In other words, some the knowledge that these sectors introduce into the system is diffused to other sectors. But high science intensity also correlates positively with high self-dependence, especially so in terms of using knowledge. This 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.

The correlation between science intensity and the two pervasiveness indicators is negative, indicating that science intensive sectors generally tend to be less pervasive. But these correlations are weaker than those between science intensity and the other indicators. This indicates that one particularly interesting finding of our intersectoral perspective relates to pervasiveness. Figure 3 plots the two pervasiveness indicators against each other. 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

**Table 3.** Indicator scores of the sectors (definitions are given in the text, all data refer to calculations made with 1992 data)

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

**Table 4.** Correlation coefficients between the indicators

| | Science citations per pat. | Backward multiplier | Forward multiplier | Net science multiplier | Self-use | Self-supply | Use pervasiveness | Supply pervasiveness |
|---|---|---|---|---|---|---|---|---|
| Science citations per pat. | 1.00 | | | | | | | |
| Backward Multiplier | −0.88 | 1.00 | | | | | | |
| Forward Multiplier | −0.85 | 0.98 | 1.00 | | | | | |
| Net science multiplier | 0.85 | −0.89 | −0.80 | 1.00 | | | | |
| Self-use | 0.86 | −0.88 | −0.88 | 0.87 | 1.00 | | | |
| Self-supply | 0.54 | −0.53 | −0.60 | 0.46 | 0.82 | 1.00 | | |
| Use pervasiveness | −0.69 | 0.70 | 0.71 | −0.46 | −0.61 | −0.52 | 1.00 | |
| Supply pervasiveness | −0.46 | 0.22 | 0.27 | −0.10 | −0.29 | −0.32 | 0.61 | 1.00 |

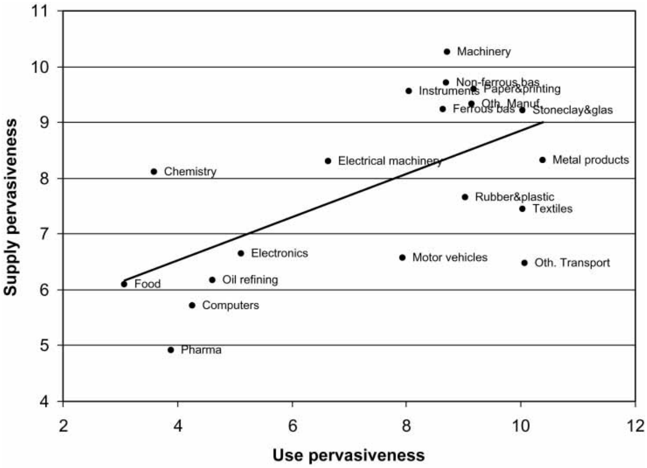

**Figure 3.** Use pervasiveness versus supply pervasiveness, solid line is linear regression line

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). As could be expected from the correlation table, 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. Oil refining is the only sector that is clearly in the low pervasiveness group *and* has a low science intensity and net multiplier. All other sectors with low pervasiveness have net multipliers very close to or larger than one (and correspondingly, high science intensities, $>1/2$ science citation per patent).

On the other hand, and pointing to the imperfect correlation, 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. These sectors 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.

The second way in which we can classify the sectors in Figure 3 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 relatively 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.

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 system, which exchange a lot of knowledge within them, but not so much (relatively) to the rest of the system. 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**.

The MDS results are in Figure 4, 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 shows 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 clusters 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.

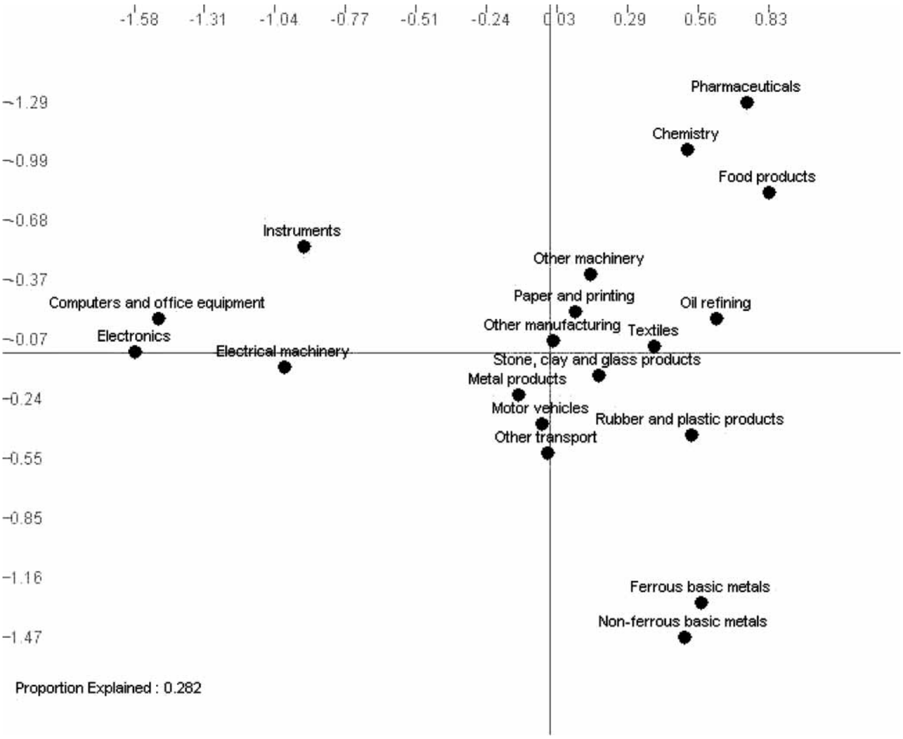

**Figure 4.** MDS map of mutual knowledge dependencies among the sectors

The metallurgy cluster is very different. It 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).

## 7. 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 great 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 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.

## Acknowledgment

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

## Notes

[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, for example, 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.

[2] A further complication results from the fact that 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 later.

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

[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.

[5] The vector of the column sums of matrix $\mathbf{D}$ is $\mathbf{i}'\hat{\mathbf{v}}(\mathbf{I}-\mathbf{A})^{-1}$. Let us call this vector of column sums $\boldsymbol{\varepsilon}'$. Then $\mathbf{i}'\hat{\mathbf{v}} = \mathbf{v}' = \boldsymbol{\varepsilon}(\mathbf{I}-\mathbf{A})$. The identity $v_j^t + \Sigma_i a_{ij}^t = 1$ yields $\mathbf{v}' = \mathbf{i}' - \mathbf{i}'\mathbf{A} = \mathbf{i}'(\mathbf{I}-\mathbf{A})$. Hence, $\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{A}) = \mathbf{v}' = \mathbf{i}'(\mathbf{I}-\mathbf{A})$. Postmultiplying by $(\mathbf{I}-\mathbf{A})^{-1}$ gives $\boldsymbol{\varepsilon} \ = \mathbf{i}$.

[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.

[7] 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.

[8] The working paper version is available, among others, at http://www.merit.unu.edu as Working Paper 2007–22.

[9] The working paper version of this paper includes a section that extensively documents the time-invariance.

## References

Cohen, W.M., Nelson, R.R. and Walsh, J.M. (2002) Links and impacts: the influence of public research on industrial R&D, *Management Science*, 48, pp. 1–23.

Dietzenbacher, E. (2005) More on multipliers, *Journal of Regional Science*, 45, pp. 421–426.

Dosi, G., Llerena, P. and Sylos Labini, M. (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, pp. 1450–1464.

Ghosh, A. (1958) Input–output approach in an allocation system, *Economica*, 25, pp. 58–64.

Hall, B.H., Jaffe, A.B. and Trajtenberg, M. (2002) The NBER patent-citations data file: lessons, insights and methodological tools, in: A.B. Jaffe and M. Trajtenberg (Eds) *Patents, Citations & Innovations. A Window on the Knowledge Economy*, pp. 403–460 (Cambridge, MA: MIT Press).

Hicks, D., Breitzman, T., Olivastro, D. and Hamilton, K. (2001) The changing composition of innovative activity in the US – a portrait based on patent analysis, *Research Policy*, 30, pp. 681–703.

Jaffe, A.B., Trajtenberg, M. and Henderson, R. (1993) Geographic localization of knowledge spillovers as evidenced by patent citations, *Quarterly Journal of Economics*, 108, pp. 577–598.

Jaffe, A.B. and Trajtenberg, M. (Eds) (2002) *Patents, Citations & Innovations. A Window on the Knowledge Economy* (Cambridge, MA: MIT Press).

Johnson, D. and Evenson, R.E. (1997) Innovation and invention in Canada, *Economic Systems Research*, 9, pp. 177–192.

Kline, S.J. and Rosenberg, N. (1986) An overview of innovation, in: R. Landau and N. Rosenberg (Eds) *The Positive Sum Strategy: Harnessing Technology for Economic Growth* (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, pp. 991–1001.

Los, B. (1999) The impact of research and development on economic growth and structural change. PhD thesis, University of Twente, Enschede, The Netherlands.

Maurseth, P.-B. and Verspagen, B. (2002) Knowledge spillovers in Europe. A patent citation analysis, *Scandinavian Journal of Economics*, 104, pp. 531–545.

Meyer, M. (2002) Tracing knowledge flows in innovation systems, *Scientometrics*, 54, pp. 193–212.

Michel, J. and Bettels, B. (2001) Patent citation analysis. A closer look at the basic input data from patent search reports, *Scientometrics*, 51, pp. 185–201.

Miller, R.E. and Blair, P.D. (1985) *Input–Output Analysis. Foundations and Extensions* (Englewood Cliffs, NJ: Prentice Hall).

Narin, F., Hamilton, K.S. and Olivastro, D. (1997) The increasing linkage between U.S. technology and public science, *Research Policy*, 26, pp. 317–330.

Oosterhaven, J. and Stelder, D. (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 Duarte, R. (2003), Analysing pollution by way of vertically integrated coefficients, with an application to the water sector in Aragon, *Cambridge Journal of Economics*, 27, pp. 433–448.

Scherer, F.M. (1982) Inter-industry technology flows and productivity measurement, *Review of Economics and Statistics*, 64, pp. 627–634.

Tamada, S., Naito, Y., Kodama, F., Gemba, K. and Suzuki, J. (2006) Significant difference of dependence upon scientific knowledge among different technologies, *Scientometrics*, 68, pp. 289–302.

Van Moergastel, T., Slabbers, M. and Verspagen, B. (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*, 9, pp. 47–65.

Webb, C., Dernis, H., Harhoff, D. and Hoisl, K. (2004) A first set of EPO patent database building blocks for analysing European and international patent citations. OECD mimeo, OECD, Paris.

## Appendix. Indicators Used for Graphing Technological Clusters

The first indicator is based on the matrix **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 **D**, we define a proximity matrix that captures the pair-wise dependency of the sectors. We call this symmetrical square matrix **SimD**, and the elements of this are defined as $SimD_{jj} = 1$ for each sector $j$ and for $i \neq j$as

$$SimD_{ij} = \left(\frac{d_{ij}}{1 - d_{ii}} \cdot \frac{d_{ji}}{1 - d_{jj}}\right)^{1/2}$$

Since $\Sigma_{i=1}^{19} d_{ij} = 1$ 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.

Analogously, we can build an alternative matrix **SimK** of the mutual knowledge dependencies among the sectors on the basis of the matrix **K**. Since **K** is based on the output coefficient matrix **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 **SimK** and **SimD** simultaneously. This is the matrix **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.