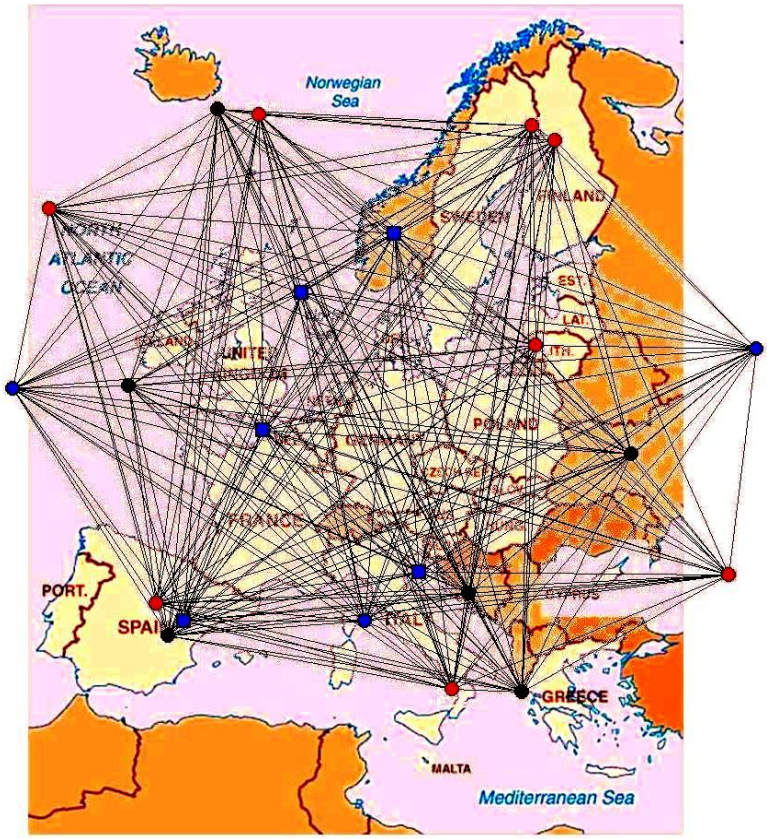
CESPRI - BOCCONI UNIVERSITY



Highly Cited Patents, Highly Cited Publications, and Research Networks

December 2006

Stefano Breschi
Gianluca Tarasconi
Christian Catalini
Lorenzo Novella
Paolo Guatta
Hrannar Johnson

Submitted to DG Research
Directorate M: 'Investment in Research and Links with other Policies'
European Commission

This Report reflects the results of research and analysis conducted by CESPRI.
The results do not necessarily reflect the view of the European Commission.

Preface

This report presents the main results of a study, ***Highly Cited Patents, Highly Cited Publications, and Research Networks***, conducted for the European Commission by the *Centre for Research on Innovation and Internationalization* (CESPRI) of Università Commerciale Luigi Bocconi, Milan. The study purported to appraise the existence and the importance of social network linkages between the authors of scientific publications cited in patents (i.e. scientists) and the generators of patented inventions (i.e. inventors). The study focused upon five technology fields, characterised by high degrees of science intensity as measured by the average number of citations to scientific publications per patent, and by high growth rates in the number of patent applications.

The study developed and applied a quantitative methodological framework for high-quality analysis of social networks linking authors of scientific publications and inventors of patents. In particular, the study built a large scale dataset relating patent and patent citations data to scientific publications cited in patents.

The study was conducted between January 2005 and December 2006 under the direction of Vincent Duchene, Policy Analyst - S&T Indicators and Economic Analysis, Directorate M: 'Investment in Research and Links with other Policies', DG Research, European Commission.

The project team included the following CESPRI affiliates: Stefano Breschi (Principal investigator), Gianluca Tarasconi (CESPRI database administrator), Christian Catalini, Lorenzo Novella, Paolo Guatta and Hrannar Johnson.

Methodological issues and preliminary results were discussed at various workshops and seminars, including the EIASM Workshop on *Management and Complexity* (University of Oxford, June 2006), and the EPO workshop on the *Economics and Management of Patents* (University of St. Gallen, September 2006).

This report reflects the results of research and analysis conducted by CESPRI. The results do not necessarily reflect the view of the European Commission, or any of the experts consulted during the course of the project. Comments on this Report can be sent to Stefano Breschi, CESPRI, Università Bocconi, via Sarfatti 25, 20136 Milan, Italy; stefano.breschi@unibocconi.it.

Executive Summary

This report offers a large-scale empirical appraisal of the social connections linking academic scientists and industrial researchers in five science intensive technology fields, namely transmission of digital information (telecommunications), speech analysis (ICT), semiconductors, lasers, and biotechnology (measuring, testing, diagnostics). It shows that, in spite of different objectives and systems of incentives, the two communities of researchers are socially connected to a much larger extent than one would normally presume. A key role in connecting the two communities is played by specific individuals, i.e. authors-inventors, that act as gate-keepers bridging the boundaries across the two domains. A further important result emerging from our study is that social networks of collaboration among scientists and inventors work as effective conduits of knowledge flows from the realm of science to that of technology. In this respect, our analysis shows that social proximity to authors of scientific publications is a much more fundamental factor affecting the knowledge transfer from scientific research to technological applications than geographical proximity.

The increasing inter-dependency between science and technology has made the theme of university–industry knowledge transfer a key research issue both in economics and management studies, as well as a top entry in the science and technology policy agenda of many countries. In the context of Europe, a general and widespread belief is that the mechanisms leading to the transfer of scientific knowledge into technological applications are somehow impaired and less effective than in other areas of the world, notably the United States. This conjecture has led to interpreting the European lag in some key high tech sectors, such as electronics and biotechnology, as a consequence of its inability to convert its scientific strength into economic profitable innovations. This phenomenon has also deserved the name of "European Paradox" to stress the fact that European strength in the production of high quality scientific output is not matched by the ability of European private companies to benefit from such output.

The existence and the extent of a European weakness in the transfer of knowledge from the domain of scientific research to technological applications is normally predicated on the basis of bibliometric indicators on the quantity and quality of scientific output. To date, however, very few studies have attempted to investigate in depth the actual mechanisms through which knowledge produced within the boundaries of academic organisations gets transferred and translated into technological developments. This study contributes to filling this gap by proposing a large scale quantitative analysis of the social connections linking academic scientists and industrial researchers in five science intensive technology fields.

To this purpose, the study exploits a complex, relational dataset reporting full bibliographical information on patent applications and scientific publications cited in those patent documents. The key analytical tool used to investigate the linkages connecting academic scientists and industrial researchers is represented by social network analysis. Specifically, information on co-authorship and on co-invention is exploited to assess the extent of connectedness among the two social communities of researchers. Likewise, citations from patent documents to scientific publications are used as proxy for the knowledge flows from the realm of science to that of technology.

Main findings

The main findings of the study may be summarised as follows.

High quality scientific publications find their way into a large number of technological developments. Publications that are (highly) cited in patents are not only cited in the realm of technology, but they are also heavily cited by other scientific publications. Besides validating the methodological choice of using patent citations to scientific publications as proxy of knowledge flows from science to technology, this finding suggests that there is not necessarily a conflicting logic between scientific and industrial communities. In this respect, however, it should be also noted that European scientific publications cited in patents receive a lower average number of citations in scientific literature than the corresponding articles published by US authors. This evidence seems to suggest that high quality European publications face more obstacles in translating into technological applications than comparable scientific output in the US.

European science is relatively under-represented among publications that provide key contributions to technological developments. The share of European organisations among scientific publications that are *highly cited* in patents is systematically lower than the its share of all cited publications. This gap is particularly evident in fields such as lasers, semiconductors and biotechnology. This result suggests that European scientific output translates into a lower number of technological developments, thereby providing further support to the conjecture about the existence of weaknesses in the process of knowledge transfer from science to technology.

Private companies account for a large share of scientific publications highly cited in patents. The role played by different types of institutions in the production of scientific publications highly cited in patents varies across technology fields, with universities accounting for a large share particularly in biotechnology. However, a key result emerging from our analysis is that private companies account for a quite large fraction of highly cited publications in all technology fields. In particular, the share of highly cited publications held by private companies is remarkably larger than their share of all scientific publications, which according to other studies may be estimated around 5-10%. This result suggests that corporate labs contribute to a large extent to the scientific research that is incorporated into technological applications.

The European private companies' contribution to the production of scientific publications highly cited in patents is significantly lower than the contribution of private companies located in other areas, notably the United States. A major weakness of the European systems of research, as compared to other geographical areas, especially the United States, is related to the low degree of involvement of private companies in the conduct of research leading to scientific publications cited in patents. Whereas the contribution of the public system of scientific research, i.e. universities and public research organisations, is generally comparable to, and often larger than the contribution of the corresponding system in the US, the fraction of scientific publications accounted for by the private system of research is considerably lower. To the extent that the ability of private companies to profit from scientific output generated in the sphere of science depends on the possession of absorptive capabilities and especially on the existence of boundary-spanning individuals, we believe this characteristic represents one of the major obstacles to the effective diffusion of knowledge from the realm of science to that of technology.

The propensity of European technology to build upon US scientific publications is generally higher than the propensity of US technology to rely upon European science. An analysis of the knowledge flows across geographical areas by origin of citing patents and origin of cited publications reveals that European patents tend

to cite US scientific publications to a larger extent than US patents tend to cite European scientific papers. In other terms, the empirical evidence shows the existence of an asymmetry in knowledge flows between Europe and the US, with a larger amount of knowledge flowing from the US to Europe than vice versa. Likewise, we observed that the propensity of US inventors to rely upon the domestic science base is significantly greater than the propensity of European inventors to exploit their domestic science base.

The two communities of academic scientists and industrial researchers are highly connected to each other. The social network analysis shows that the network of co-inventors is highly disconnected with many components of small size. This means that most collaborators of each inventor come from the same organisation and that few connections exist among teams of industrial researchers. However, when one looks at co-invention and co-authorship relations simultaneously, the key result is that the two communities of researchers are significantly more socially connected that one would probably expect. In three crucial technology fields, such as semiconductors, lasers and biotechnology, 35%, 51% and 53%, respectively, of *all* authors and inventors are either directly or indirectly connected, via co-invention or co-authorship, to each other in a large connected component. In addition to that, 24%, 40% and 32% of *all* inventors are either directly or indirectly connected (i.e. reachable) to each other. Besides indicating that academic scientists and industrial researchers are highly connected, these results suggest that the community of inventors itself is much more connected than data on co-invention only would lead us to presume. Although not directly connected to each other, inventors are indirectly connected through scientific authors and through authors-inventors, i.e. individuals that participate in teams of inventors *and* in teams of scientists.

Authors-inventors play a key role in connecting the communities of scientists and inventors and act as gatekeepers across the two realms. A crucial role in ensuring high degrees of connectivity between the two communities of researchers is played by a specific type of individuals that we have labelled as authors-inventors. They are researchers that do publish scientific articles and patent new inventions, thereby participating into both communities. Social network analysis reveals that such individuals are characterised by a higher degree centrality, i.e. they tend to collaborate on average with a significantly larger number of other inventors *and* authors, than do simple inventors and authors, and by a higher betweenness centrality, i.e. they play a crucial function of knowledge brokers in the network that makes them more in-between than simple inventors and authors, and ensures a rapid diffusion of knowledge and ideas from one domain to the other.

Europe is characterised by a relatively low number and share of science-technology gatekeepers, i.e. authors-inventors. Given the key role played by authors-inventors in bridging the realms of science and technology, we believe that a key finding of our study is that the share of European inventors playing this specific function is lower than its share of simple inventors. To the extent that authors-inventors act as brokers of knowledge from the domain of science to that of technology, we believe this finding has very important implications for our understanding of the gap between Europe and the US in the effectiveness to translate the results of scientific research into commercially useful applications. Proximity to such individuals, and more generally proximity among authors of scientific publications and inventors of patented inventions is in fact a fundamental factor affecting the effective diffusion of scientific knowledge (see below). In addition to this, we do also believe that this result is consistent with our finding that a major European weakness is related to the feeble commitment of private com-

panies in the production of scientific publications relevant for technological developments, given that authors-inventors are most likely to come from such organisations.

The network of academic scientists and industrial researchers has the properties of a "small world". The social network of academic scientists and industrial researchers is characterised by topological properties typical of "small world" graphs. On the one hand, it presents high degrees of *local cliquishness*, i.e. an individual's collaborators tend to collaborate with each other; on other hand, it also presents a low average distance among individuals, i.e. any random pair of individuals is separated by a low number of steps. This means that the network of authors and inventors works at least potentially as an effective means of knowledge transmission and diffusion.

Social proximity among (academic) scientists and industrial researchers is the most important factor affecting the probability that a patented invention will build upon a scientific publication. In this study, we estimated an econometric model for the probability that a patent-paper pair is linked by a citation tie. Our findings reveal that such a probability is apparently affected in a negative way by the geographical distance that separate patent inventors and paper authors. Yet, the effect of spatial distance vanishes once we control for the social distance among them. Inventors that are socially *closer* to authors of scientific publications are much more likely to build upon such publications than are inventors located at a larger social distance. In other terms, knowledge transfer from science to technology takes place mostly through social networks of collaboration among scientists and inventors.

Policy recommendations

Results of this study provide further empirical support to the conjecture that the mechanisms driving the transfer of scientific outputs into technological applications in Europe are somehow impaired and less effective than in other areas of the world, notably the United States. At the same time, they also point out that social networks of (academic) scientists and industrial researchers account for much of the observed patterns of knowledge diffusion from science to technology. In particular, the study has shown that a crucial role in connecting the two communities of researchers is played by a specific type of individuals, i.e. authors-inventors, that act as gatekeepers and channel information and knowledge between groups with different objectives and incentives. In this respect, a major European weakness is related to the comparatively lower involvement of private companies in the conduct of basic and applied research leading to scientific publications and to the consequently lack of authors-inventors that are able to bridge and connect the realms of science and technology. In other words, it is possible that part of the European backwardness in this field is due to a less connected research area. We do believe that increasing such a connectivity should feature prominently in a policy agenda aiming to spur the rate of knowledge transfer from science to technology. In this respect, the mobility of inventors (i.e. industrial researchers) and academic scientists across regions, countries, and organisations represents, in our view, a major policy objective in order to achieve higher degrees of social connectivity among the two communities of research.

Contents

1. OUTLINE OF THE REPORT

1.1. STUDY OBJECTIVES AND FOCUS

This document reports the main results of a large-scale quantitative analysis of the research networks linking inventors of patents and authors of scientific publications. Conducted between January 2005 and December 2006, the study focused upon five technology fields, which were selected from International Patent Classification (IPC) codes used to classify patent documents. The fields examined in this study are characterised by high degrees of science intensity, as measured by the number of citations from patents to scientific publications, high growth rates in the number of patent applications, and high degrees of turbulence, as measured by the weight of new innovative entrants on the total number of patent applications. In brief, the five technology fields considered for the analysis are science-based, highly dynamic and fluid domains of research, in which the degree of interaction between scientists and inventors is expected to be relevant.

The five technology fields included in the analysis are:

- Transmission of digital information
- Speech analysis
- Semiconductors
- Lasers
- Biotechnology (measuring, testing, diagnostics)

Starting from patent applications in the five technology fields over the period 1990-2003, the study has built a complex relational database that includes full bibliographical information on patents and scientific publications cited by those patents. This dataset has been used to address a set of issues relevant for our understanding of the linkages between science and technology. In particular, the focus of the study has been on assessing:

- Whether and to what extent scientific publications cited in patents are cited only within the realm of technology or, conversely, they are also cited by other scientific publications
- What organisations (i.e. companies, universities and research organisations) are responsible for the production of scientific publications highly cited in patents

- The European capability to produce scientific publications that are highly cited in patents and its position with respect to other areas, especially the United States and Japan
- The social network relationships linking scientists and inventors, and the structural properties of the network of patent inventors and paper authors
- The impact of social proximity and spatial proximity in affecting the likelihood that a patent builds upon knowledge produced in the domain of scientific research.

To address these issues, the study has developed and applied a quantitative methodological framework involving the use of statistical analysis, regression analysis and social network analysis tools. The study purports to demonstrate the applicability of social network concepts and analytical tools in appraising the mechanisms that govern the transfer of knowledge from the realm of open science to that of private technology.

1.2. STUDY CONTEXT

University–industry knowledge transfer is nowadays a key research subject both in economics and management studies, as well as a top entry in the science and technology policy agenda of a number of developed and developing countries. Awareness of the problem has been certainly spurred by the increasing interdependency between science and technology, as shown by several studies (Narin, Hamilton, Olivastro, 1997; Verbeek et al., 2003). The ability to have timely access to advanced scientific knowledge represents nowadays a fundamental factor that can explain performance differentials among firms and regions (Cockburn and Henderson, 1998; Zucker et al., 1998).

In the context of Europe, a general and widespread belief is that the mechanisms leading to the transfer of scientific knowledge into technological applications are somehow impaired and less effective than in other areas of the world, notably the United States. This conjecture has led to interpreting the European lag in some key high tech sectors, such as electronics and biotechnology, as a consequence of its inability to convert its scientific strength into economic profitable innovations (Dosi, Llerena, Sylos-Labini, 2005). This phenomenon has also deserved the name of "European Paradox" to stress the fact that European strength in the production of high quality scientific output is not matched by the ability of European private companies to benefit from such output.

Although the very existence of a European Paradox has not gone unchallenged, the aim of this study was not on testing the extent of knowledge transfer from science to technology, but rather to investigate the mechanisms underlying such a transfer. In this respect, "distance" between the two realms of academic and industrial research has been increasingly called in to explain whether the former may, or may not, benefit the latter. Two concepts have attracted most of the attention in the recent literature: *geographical* and *social* distance.

The geographical distance hypothesis suggests that both scientific and technical knowledge are largely "tacit" and "non-codifiable", and require distance-sensitive transmission means such as frequent face-to-face contacts clarifying discussions and on-site demonstrations (Feldman, 1999). According to this hypothesis, therefore, scientific knowledge should benefit mostly those companies located nearby the source of its production (Jaffe, 1989). Although a large empirical literature has developed around this conjecture, one should also point out that most of the evidence produced to support it is just indirect, i.e. it does not measure knowledge flows in a direct way (Breschi and Lissoni, 2001). Moreover, to the extent that spatial proximity is the fundamental mechanism explaining the effectiveness of knowledge transfer, this hypothesis is not particularly helpful for our understanding of the European problems.

An alternative and competing hypothesis is that the exchange of tacit knowledge between university and corporate researchers requires the two social communities to share some acquaintances and/or a few codes of behaviour in terms of reciprocity and fairness (both in case of market transactions and in case of free sharing). Similarly, academic researchers' mobility to and from industrial labs (either in the position of employees or entrepreneurs) requires a web of personal contacts for exchanging information on job and financing opportunities, and again some codes of behaviour that do not punish such mobility by portraying it as free-riding (Balconi et a., 2004). More generally, this perspective emphasises that the creation and the diffusion of knowledge cannot be separated from the social network underpinning it, so that *social proximity*, rather than just spatial proximity, is the key driver of knowledge diffusion from the realm of science to that of technology. In this respect, the literature has focused upon the conflicting incentives and norms of behaviour that characterise the worlds of "Proprietary Technology" and "Open Science" (Dasgupta and David, 1994).

The former is approximately identified with the results of privately sponsored industrial research. Intermediate research results, instruments, and methods are normally shared with other researchers, in order to get feedbacks and gain credit for future help, but not outside some organizational boundaries defined by the research sponsors. Communication with researchers from rival companies is monitored and restricted, and codification efforts (such as those leading to the publication of research papers) are delayed as long as possible. By contrast, the incentive structure of "Open Science" is modelled upon Merton's (1957) sociological account of the function of disclosure norms and publications in forging the career path of academic scientists. The New Economics of Science depicts the community of scientists as composed by many small groups, linked both by career schemes requiring scientists to move across groups, albeit occasionally, and by some degree of across-group legitimization mechanism for individuals' research contributions. Each group of academic scientists (or each set of tightly connected groups) belongs to a wide community of researchers of the same science field (an "epistemic community", as defined by Cowan et al. (2000) and Steinmueller (2000) and contributes to expanding, codifying and securing the reliability of scientific knowledge by establishing mutually recognized research and test procedures, as well as communication codes for both written and oral exchanges. Within each community, codified knowledge is a public good. In turn, links among different groups are as many as it is necessary to spread information on the reputation of individual researchers, both in terms of capabilities and adherence to the behavioural codes of "Open Science".

Do the "Open Science" and the "Proprietary Technology" realms ever get in touch? How do they reconcile their different systems of incentives and social structures? One key mechanism is advanced education: doctoral students trade their willingness to provide free or cheap research assistantship for learning, and most of them will then pursue a career as industrial researchers. In addition, academic scientists can occasionally turn into industrial researchers, and *vice versa*, depending upon the origin (public versus private) of the research funds, and the possibility (for industrial researchers) to spend some time working in close contact with a university or a public research centre. Mobility of researchers to and from universities, public labs, and corporate labs can produce similar contamination effects.

The most recent literature has offered various insights on the key mechanisms through which academic scientists and industrial researchers get in touch and exchange knowledge. On the one hand, a

few studies have analysed the incentives for firms to publish research results in scientific journals, by acknowledging the central importance of forming ties with the academic community, via boundary-spanning "gatekeepers", in order to access socially embedded knowledge (Hicks, 1995). In this vein, Stern (1999) has argued that pure scientific research is costly to the firm, but necessary to attract and recruit research-oriented scientists, who have a taste for publishing and whose skills are needed for transforming research into product development. In such cases, one also expects industrial researchers to adjust their behaviour to the incentive structure of the contingent research program (both in terms of adherence to the research objectives and publication rules). For example, industrial researchers for large corporate labs, more often involved in basic research along with universities and public labs, will find it easier to publish; they will also find it more rewarding, since they entertain hopes of further co-operation in the future. On the other hand, the academic community has shown an increasing interest in the exploitation of scientific research for industrial purposes. The evidence produced by Zucker et al., (1998) shows quite convincingly that star scientists from disciplinary fields prone to commercial exploitation trade their knowledge assets on a market basis, either through founding new companies or other forms of contractual arrangements.

In brief, one can say that, especially in high tech and science intensive industries, academic and industrial researchers tend to form communities and social groups that are increasingly interrelated and exchange knowledge through market and non-market transactions, with different objectives and incentive schemes. However, while case studies on the theme of social distance among scientists and inventors abound, large-scale quantitative research on the same subject is more of a rare breed, limited as it is by highly demanding data requirements. In order to map social groups in Science and Technology we need data on information exchanges between researchers, both within individual companies and academic research groups, and across them.

As long as we regard team-working experiences as a key mean for knowledge exchange, co-authorship of scientific papers is the ideal quantitative indicator to investigate social networks of academic scientists, and indeed there is a long tradition of exploiting them to that purpose (e.g. Melin and Persson, 1996). The most recent research efforts within this line of enquiry draw extensively from graph theory, as it may be applied to social network analysis. They describe the social structure created by Open Science rules as a "small world", i.e. a "distinctive combination of high clustering with short

characteristic path length" (Watts and Strogatz, 1998). Each researcher has a number of links which suffice to involve him deeply in a local network of collaboration (his research group), and a few researchers have as many links with members of other research groups as it is necessary to connect most, if not all, of the epistemic community. News spread fast, as well as chances to engage in research partnerships apt to allow for knowledge exchanges.

What about Proprietary Technology? How to measure social networks there? And which properties will those networks exhibit, especially at the boundary with "Open Science", i.e. when industry–university cooperation occurs? Recent studies have convincingly shown that an extremely useful empirical tool is represented by another traditional indicator, namely patent applications, albeit in a way which mimics closely the use of co-authorship data from scientific publications. More precisely, patent data can be extremely useful in measuring social distance among inventors, as many inventions are the outcome of teamwork, so that the related patent documents list more than one inventor. One can therefore reasonable assume that inventors listed on the same patent know each other, and have possibly exchanged crucial scientific or technical information (Balconi et al., 2004; Breschi and Lissoni, 2006; Singh, 2003).

To date, however, no one has tried to combine co-authorship data with co-invention data and investigate the extent to which the two communities of academic scientists and industrial researchers are linked to each other through collaboration. This is precisely the objective of this study. The aim is to assess whether and to what extent the two communities are connected and whether the "social" connection among them may explain the knowledge flows taking place across the borders of the two communities.

To this purpose, the study has built a large, complex relational database that combines information on inventors reported in patent documents and authors of scientific publications cited by those patents. Social network analysis of co-authorship and co-invention is used to test the degree of connectedness between the two communities, whereas patent citations to scientific publications are used as a proxy of the knowledge flows from the realm of science to that of technology.

1.3. ORGANISATION OF THE REPORT

The report is organised as follows. Section 2 provides a detailed discussion of the technical procedures and the methodological issues related to the construction of the dataset on citing inventors and cited authors. It also provides a discussion of the main data sources used in the study. Section 3 reports the main findings of the study and is divided into two main parts. The first part concerns a statistical analysis of the dataset on citing patents (inventors) and cited publications (authors), which provides a broad set of statistics on the knowledge flows from science to technology. The second part reports the main results emerging from the social network analysis of the co-authorship and co-invention relationships linking academic scientists and industrial researchers. Section 4 summarizes the results and offers policy recommendations.

2. METHODOLOGICAL FRAMEWORK

This section is devoted to discussing the technical procedures and the methodological issues related to the construction of the dataset, which has been used for the analysis of the network linkages between scientists and inventors. Before turning to this, we briefly describe the sources of data that have been used in the course of the study.

2.1 Data sources

All the data sources have been taken in raw format, after which a process of reading, parsing and standardisation has taken place. All the work of database construction has been carried out at CESPRI. The following data sources have been used in this study:

1) EP-CESPRI dataset

The EP-CESPRI dataset, owned and maintained by CESPRI, includes all patent applications to the European Patent Office (EPO), from June 1st 1978 (starting date of the EPO) to June 1st 2004. The data set includes the full set of bibliographic variables concerning each patent application:

- Priority, application, and publication number
- Priority, application and grant dates
- Title and abstract
- Designated states for protection
- Status of the application
- Main and secondary *International Patent Classification* (IPC) codes
- Applicant's name and address
- Inventors' names and addresses
- Reference to other EPO patents

In the construction of this dataset, CESPRI has gone through a thorough process of cleaning and standardisation of applicant names in a major effort to correctly identify the company/institution, which applied for each specific patent. In fact, a major problem with using patent data at the level of individual organizations is that they register their patents under different names. This work has been conducted in collaboration with the *SAS Institute* (Italy), using the SAS/Data Quality Solution® software. In its current version, the EP-CESPRI data set contains 118,396 unique organizations, which

are constructed by combining 154,366 different original applicant names. Each patent applicant in the data set is identified by an internal numerical code (company code) and by its standardised name. Address fields have been also cleaned and harmonised to allow geographical analyses of patenting activities.

A similar work has been undertaken for inventors reported in patent documents. These have been identified first by assigning a unique code to all inventors with the same name, surname and address; and then by running Massacrator©, a programme that assigns scores to any pair of inventors with the same name+surname but different address, on the basis of information suggesting the two inventors may be the same person (such as the technological class of their patents, the identity of their patent applicants, their location in space and the identity of their co-inventors).

Finally, the EP-CESPRI dataset also reports for each patent document all citations made to all prior patents cited by the document itself. To this purpose, the so-called REFI tape citation data set has been purchased, processed and linked to the EP-CESPRI data set by publication number of patents. The REFI database contains, in addition to the patents cited, also the references to non-patent documents. We used this dataset to retrieve and process all citation made by EPO patents to non-patent literature and to identify among them the citations corresponding to scientific articles.

2) United States Patent and Trademark Office (USPTO) database

Beside EPO data, we purchased and processed patent data from the United States Patent and Trademark Office (USPTO). The dataset contains information on all patents granted by the USPTO from January 1st 1975 to December 31st 2003. The dataset includes the following set of bibliographic variables concerning each patent application:

- Application number
- Application and grant dates
- Main United States Patent Classification (USPC) codes
- Applicant's name
- Inventors' names and addresses
- References to other patents
- References to non-patent documents

A few major differences between EPO and USPTO data have to be noted. In the first place, the USPTO dataset does not report any information on the address of patent applicants. To carry out analysis of patenting activity at the geographical level, one has to rely therefore on information concerning the location of inventors. Second, the USPTO classifies patent documents according to both the USPC and the IPC nomenclature. However, a major problem with the USPTO patent data is that the International Patent Classification (IPC) system is not fully reliable, given the little familiarity of US patent examiners with the IPC system. Finally, USPTO raw data share the same problems of EPO raw data with respect to the fact that applicants' and inventors' names are not standardised and therefore may be spelled in quite different ways. Contrary to the EPO, however, we have not performed any cleaning and standardisation of such names, as this was beyond the scope of the study.

3) Science Citation Index ISI-Thomson

In order to identify patent citations to scientific literature, we have used the Science Citation Index (SCI) produced by the Institute for Scientific Information, ISI-Thomson, in Philadelphia, USA. For this study, we used the online version of this dataset available through the Web of Science interface. Raw data were extracted and further cleaned, processed and standardised for the purposes of this study.

The SCI is a multidisciplinary database that covers the most important journals in the natural and life sciences, by providing information from more than 5,700 peer-reviewed international journals across 178 subject fields. The following data are available for each paper covered by the SCI:

- Title
- Names of all authors
- Institutional affiliations and addresses listed by those authors
- Number of citations made to other scientific publications
- Number of citations received by other scientific publications
- Journal title
- Publication year
- Subject field of the journal[1]

[1] Each journal is assigned by ISI one or more Journal Subject Categories, i.e. internally coherent journal sets which represent 'scientific subfields'. In this respect, it has to be noted that several (important) journals, such as Nature and Science, are included in a residual 'multidisciplinary' category, as they span across a variety of scientific domains.

A few comments are needed to explain the nature of the available data and the methodological problems they present.

In the first place, it is important to point out that the SCI only reports the surname and the first letter of the name of each author. This creates major problems for analyses conducted at the level of individuals. Two issues are important in this respect. On the one hand, given that the same [surname+first letter of name] may appear in different scientific papers, it may be quite difficult to determine whether the same individual is responsible for all the papers or whether they have been authored by different individuals, who happen to be homonyms. If a careful, manual checking may help to sort out cases of homonymous individuals, this is only possible for relatively small scale studies and not for the analysis of hundreds or thousands individuals. On the other hand, a similar problem arises when trying to link information on inventors' names and information on authors' names, in order to identify those individuals that have produced both patented inventions and scientific publications. As patent dataset report name and surname of each inventor, whereas the SCI dataset reports only the first letter of name and the surname of each author, the risk in performing a simple matching by surname and first letter of name is that different individuals are identified as the same person, thereby leading to an overestimation of the number of inventors-authors. In what follows, we will discuss how we attempted to overcome such problems.

A second source of problems is related to the identification of the institutional affiliation of authors. The SCI dataset reports for each paper the institutional affiliations and addresses of the authors listed in the paper. However, it does not provide any correspondence between *each* author and her institutional affiliation. In the case of multiple authors and multiple affiliations, it becomes therefore quite difficult to assign each author to her correct affiliation and geographical address. Whereas one can make the hypothesis that the order of the authors in the paper corresponds to the order of the affiliations, so that the first author is paired to the first affiliation reported in the paper and so on, such a hypothesis does not help to solve the problem in those cases where the number of affiliations is greater than the number of authors. To the best of our knowledge, no systematic effort has been so far devoted to solving this problem. In this study, we have attempted to overcome such limitations in a rather ad-hoc way, as we will discuss below. Yet, the lack of cleaning and standardisation algorithms

of authors' names, affiliations and addresses is still a major obstacle for a fuller and more systematic use of data on scientific publications.

Finally, it has to be pointed out that although large in terms of volume and scope, the content of the SCI database is not necessarily a good reflection of all worldwide scientific publication activity. The databases are biased in favour of English-language journals. Research publications from English speaking nations (the US in particular) therefore dominate the databases. There is also a rather strong focus on fundamental research, especially in the natural and life sciences. Nonetheless, one may assume that the international journal publications in these databases provide a satisfactorily representation of internationally accepted high-quality 'mainstream' basic research. The lion's share of the publications therefore originates from universities and other public research institutions. Companies and private R&D-labs account for a relatively low share of the papers in the SCI, which may be estimated around 5-10%.

2.2 Methodological steps

The objective of this study has been the construction and the analysis of a large scale dataset linking citing patents (inventors) and cited scientific publications (authors). Figure 1 illustrates in a highly schematic way the main methodological steps that have been followed in the course of the study. The work has gone through four basic phases. The first phase of the project involved the selection, extraction and parsing of non-patent literature (NPL) citations contained in patent documents classified in 10 broad technology fields.

In the second phase, the parsed records have been matched to ISI-SCI covered journal titles in order to identify those NPL citations that correspond to 'scientific' articles. This step allowed to determine the science intensity of each technology field and was therefore instrumental in the final selection of the five technology subfields chosen for the analysis of scientists-inventors network linkages.

The third phase involved the retrieval and further processing of the ISI-SCI covered articles cited in patents belonging to the five technology subfields chosen for the analysis. In this step, we defined and discussed the criteria to identify highly cited publications, i.e. those articles receiving a large number of citations in patents, and highly cited patents, i.e. those patents receiving a large number of citations from other patents. In addition to that, in this phase we also performed a benchmarking analysis to determine to what extent cited and

highly cited publications are only cited within the realm of patents, by comparing the number of citations they receive from other scientific publications with the average citation rate of publications within the same subject field and journal that are *not* cited by patents.

The fifth and final phase involved three basic tasks. First, we matched the list of authors' names reported in cited publications with the list of inventors' names reported in all patent documents in each technology field. This procedure allowed us to identify all those individuals that have produced both patented inventions and (cited) scientific publications. Second, we cleaned, processed and standardised the authors' affiliations reported in highly cited publications. Finally, we matched the list of standardised affiliations with the list of patent applicants, in order to identify those institutions that are responsible both for patented inventions and for highly cited publications.

The final dataset thus created is a quite complex, relational database that contains information on citing subjects (patents/inventors/applicants) and cited subjects (patents/publications/authors/affiliations). In what follows, we provide a detailed discussion of each phase of the work and of the resulting final dataset.

a) Phase 1: Selection of technology fields, extraction and parsing of NPL citations

Phase 1 of the study consisted of two basic steps. The first step was the definition of a starting dataset of patents, both for the EPO and the USPTO. The second phase involved the extraction and parsing of NPL citations contained in the selected patent documents. In what follows, we describe each step.

a1) Selection of technology fields and starting dataset on patents

The starting point of this study has been the preliminary selection of 10 broad technology fields that could represent good candidates for the analysis of the network linkages between scientists and inventors. In accordance with the Commission, we selected 10 technology fields on the basis of International Patent Classification (IPC) codes, aggregated according to the 30-fields nomenclature jointly developed by the Fraunhofer Gesellschaft-Institute fur Systemtechnik und Innovationsforschung (FhG-ISI, Karlsruhe, Germany) and the Observatoire des Sciences and des Techniques (OST, Paris).[2] The 10 selected technology fields are reported in table 1.

[2] For the 30 technology fields nomenclature, see http://www.obs-ost.fr/nomenclaturesfinal.pdf.

Figure 1. Methodological steps

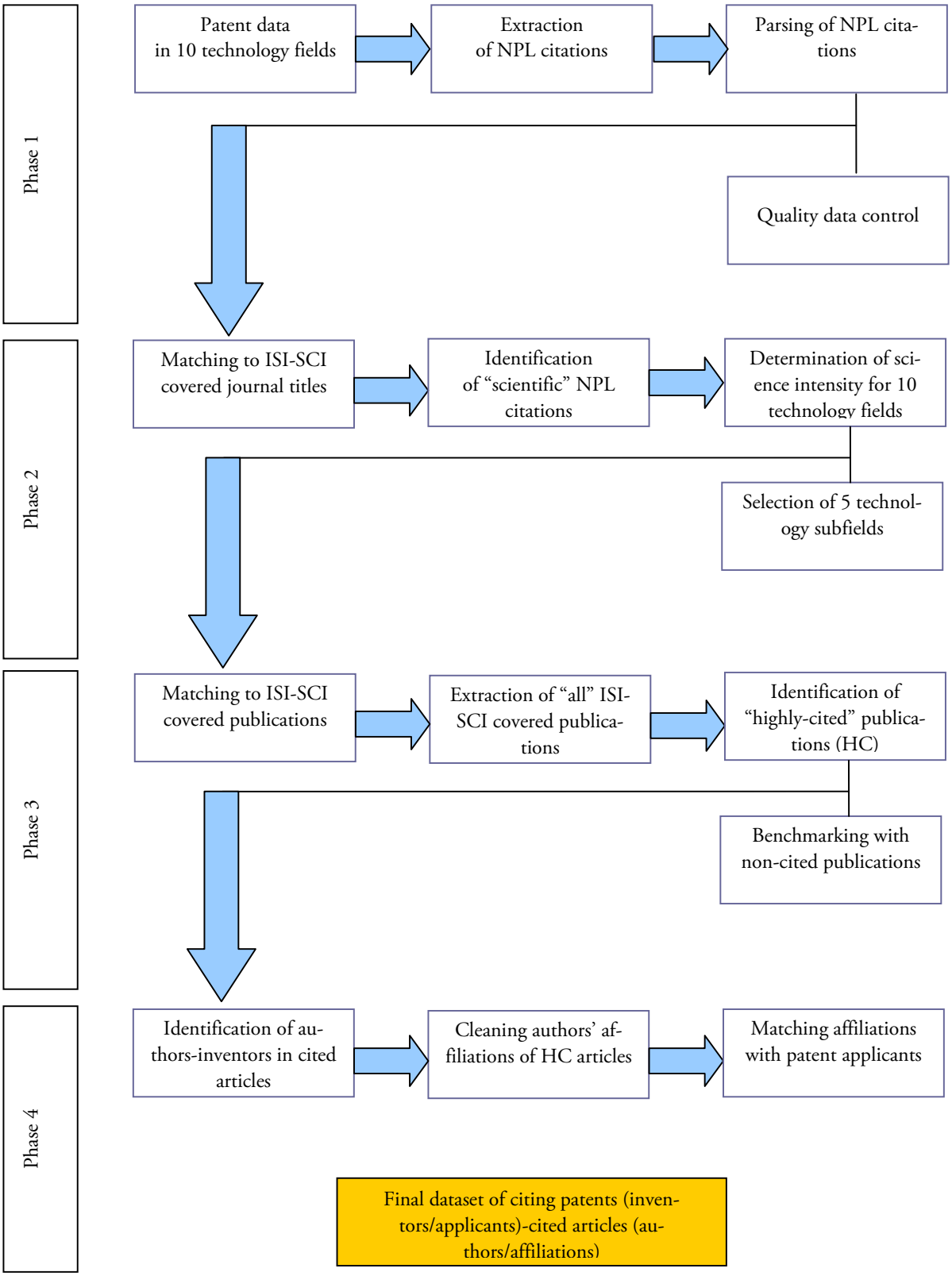


Table 1 – 10 selected technology fields on the basis of IPC codes

	Technology fields	IPC codes
1.	Telecommunications	H03B, H03C, H03D, H03H, H03K, H03L, H03M, G08C, H01P, H01Q, H04B, H04H, H04J, H04K, H04L, H04M, H04N1, H04N7, H04N11, H04Q
2.	Information Technology	G11C, G10L, G06
3.	Semiconductors	H01L
4.	Optics	G03B, G03C, G03D, G03F, G03G, G03H, H01S, G02
5.	Control Technology	G01B, G01C, G01D, G01F, G01G, G01H, G01J, G01K, G01L, G01M, G01N, G01P, G01R, G01S, G01V, G01W, G05B, G05D, G04, G07, G08B, G08G, G09B, G09C, G09D, G12
6.	Medical Technology	A61B, A61C, A61D, A61F, A61G, A61H, A61J, A61L, A61M, A61N
7.	Organic Chemistry	C07C, C07D, C07F, C07H, C07J, C07K
8.	Drugs	A61K
9.	Biotechnology	C07G, C12M, C12N, C12P, C12Q, C12R, C12S
10.	Environmental technology	A62D, B01D46, B01D47, B01D49, B01D50, B01D51, B01D53, B09, C02, F01N, F23G, F23J

For the USPTO patents, the selection criterion has been slightly different. In fact, whereas IPC codes are also available for USPTO patents, patent examiners at the USPTO classify patents according to the United States Patent Classification (USPC) system and have little familiarity with the IPC system. This implies that IPC codes assigned to USPTO patents are not fully reliable indicators of their technological domain. For this reason, the selection of the patents in the 10 technology fields reported in Table 1 has been implemented by adopting a classification of the USPC codes suggested by a few NBER researchers[3], appropriately tested and integrated using the concordance table between the USPC and the IPC codes available from http://www.uspto.gov/go/classification. The resulting list of USPC codes corresponding to the 10 technology fields is reported in the appendix.

Having defined the 10 technology fields, we have selected from the EP-CESPRI dataset all patent applications from 1990 to 2003 whose *primary technology class* was in any of these fields. Likewise, we have selected from the USPTO dataset all patents granted from 1990 to 2003 whose *primary technology class* was in any of these fields. The total number of patents and the average rate of growth in pat-

[3] Jaffe, A.B., M. Trajtenberg (2002). *Patents, Citations and Innovations: A Window on the Knowledge Economy.* Cambridge MA: MIT Press. See also, Hall B. H., A. B. Jaffe and M. Trajtenberg (2001). The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. NBER Working Paper 8498.

enting over the period 1990-2003 are reported, separately for the EPO and the USPTO, in the appendix for each technology field.

A few comments are needed to explain the problem of dating patents. Patent documents may be dated according to three basic principles: a) priority date; b) application date; c) grant date. The use of any of these dates depends on the objectives of the researcher. Broadly speaking, the priority date is the date closest to the time of the invention and it should be therefore used when the objective is to evaluate the time pattern of invention of specific countries and organisations. Yet, the use of the priority date has also a few drawbacks that limit its suitability in certain contexts. In the construction of the dataset for this study, we have followed a different criterion for EPO and USPTO patents. As for the EPO patents, the principle of application date has been followed, i.e. each patent has been assigned to the year of application to the EPO. The main reason for not choosing the priority date (i.e. the date closest to the time of the invention) is the substantial time lag that may separate the first application (i.e. priority) and the application to the EPO and thereby its publication. This lag is particularly long for patents that reach the EPO through the PCT route, i.e. for most US and Japanese patents. Using the priority date would therefore lead to a substantial underestimation of the patenting activity of some key countries, especially for the last years of the time series. In brief, our starting dataset comprises all patent applications to the EPO, whose application date was comprised between 1990 and 2003.

For USPTO patent documents, on the other hand, we have followed the practice of dating patents according to the date of grant. In this case, the reason for not choosing the priority date is that this information is available for a small fraction of all patents. Likewise, the reason for not choosing the application date is related to the long time lag separating the application and the grant date. Given the fact that until March 2001 the USPTO only published patents granted, the use of the application date would have led to a drastic reduction in the number of patents available for the analysis, particularly for the last years of the time series. In brief, our starting dataset comprises all patents granted by the USPTO, whose grant date was comprised between 1990 and 2003.

a2) Extraction and parsing of NPL citations

The second step of Phase 1 was the identification of citations to prior art patents and to non-patent literature contained in the selected patent documents. To this purpose, for each patent classified in

any of the 10 technology fields, we proceeded to extract all citations made to prior art patents (i.e. patent citations) and all citations made to non-patent literature references (i.e. NPL citations). In this respect, it must be pointed out that, while the selection of patent citations does not present major problems given that cited patents are quite easily identified by their publication number, the selection and use of NPL citations involves rather complex tasks given the fact that references to non-patent literature are provided in raw format as strings of characters, without any separation among the various items. This is illustrated in Figure 2, which reports a few typical examples of NPL citations coming from raw data sources. The first document refers to an EPO patent, publication number 379369. This document contains reference to two articles. The former is an article published in SCIENCE, which is journal covered in the SCI-ISI database. The latter is an article published in CHEMICAL ABSTRACTS, which is not covered in the SCI-ISI dataset. The second document refers to a USPTO patent document, publication number 5137812. This documents refers to an article published in ANALYTICAL BIOCHEMISTRY, which is a ISI-SCI covered journal title, and to an article published in APPLIED AND ENVIRONMENTAL MICROBIOLOGY, which is also a SCI-ISI covered journal.

Figure 2 – Selected examples of NPL citations

Citing patent	NPL citation
EP379369	SCIENCE, vol. 239, 29th January 1988, pages 487-491; R.K. SAIKI et al.: "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase"
EP379369	CHEMICAL ABSTRACTS, vol. 106, 25th May 1987, page 378, abstract no. 172466e, Columbus, Ohio, US; & US-A-888 058 (UNITED STATES DEPT. OF HEALTH AND HUMAN SERVICES) 05-12-1986
US5137812	Burnette, Anal. Biochem. 112, 195-203 (1981).
US5137812	Doyle et al., Applied and Environmental Microbiology, 53(10):2394-2396 (1987).

In order to separate NPL citations to scientific literature from NPL citations to other materials, we extracted and parsed the strings of text related to NPL citations for each of the ten technology fields. The overall procedure used to extract and parse NPL citations is illustrated in Figure 3. In the first place, we extracted all NPL citations from the set of patents classified in any of the 10 technology fields in the period 1990-2003. This returned 649,528 NPL records for the EPO and 2,097,532 NPL records for the USPTO. A parsing algorithm was then developed in-house and tested on this set of records. Actually, two algorithms were developed, one for the EPO and one for the USPTO, due

to the different layouts of NPL records.[4] The objective of the parsing process at this stage was to take the strings of text containing NPL citations and breaking them down into relevant items. More specifically, each record containing NPL citations was parsed into the following fields (some of which possibly blank):

- ID code (internally produced identification code of record)
- Authors
- Journal title
- Article title
- XP code (Internal EPO article reference code)[5]
- ISSN code (for articles-only available for EPO)
- ISBN code (for books-only available for USPTO)
- Volume
- Issue
- Publication month
- Publication year
- Starting page
- Ending page
- Editor

After parsing[6], NPL records were preliminary grouped (i.e. de-duplicated) by *all* parsed fields, to take into account the fact that a given NPL reference may be cited by more than one patent. This reduced the number of 'unique' NPL citations to 499,497 unique records for the EPO and 2,063,624 unique records for the USPTO[7]. Parsed references were then further grouped according to the following fields: non-blank article title, non-blank journal title, ISSN code and publication year. This returned 350,735 unique NPL records for the EPO and 718,256 unique NPL records for the USPTO. The reduction in the number of NPL records is mostly due to the plenitude of citations that refer to publications such as Chemical Abstracts of Japan or IBM Technical Bulletin. The resulting dataset of parsed NPL citations represented the basic input for the next phase of the study.

---

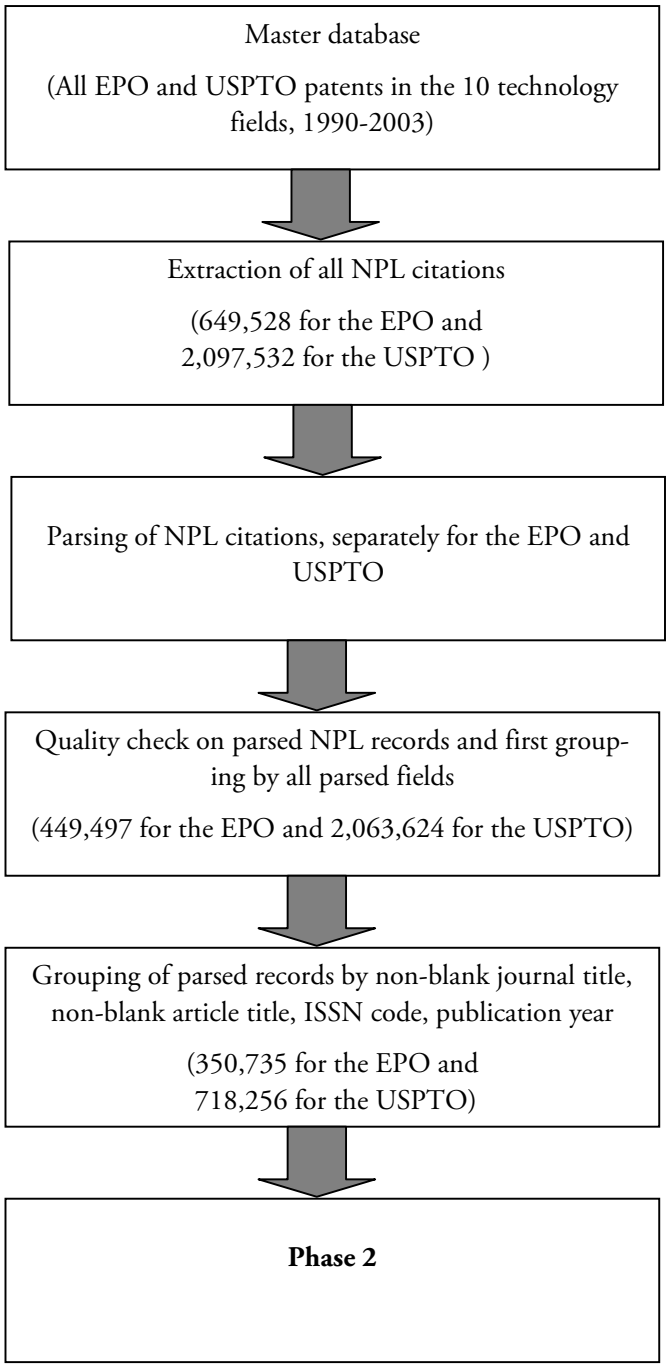
[4] The parsing algorithms were developed by Gianluca Tarasconi and Hrannar Johnson.

[5] This code is available only for the EPO NPL citations. However, a major problem with this identification code is that apparently the same article may get different codes, making it quite unreliable.

[6] A thorough data quality check was implemented on the output of the parsing procedure.

[7] Here 'unique' means that NPL records cited in more than one patent have been de-duplicated according to one or more common fields.

Figure 3. Selection, extraction and parsing of NPL citations



b) Phase 2: Identification of scientific citations and selection of technology subfields

Phase 2 of the study involved three fundamental steps. First, for each technology field, we identified among all NPL citations, those citations referring to scientific publications. Second, for each technology field a preliminary statistical analysis was carried out to examine the degree of science intensity. Finally, on the basis of such analysis we identified 5 technology subfields for the subsequent phases of the study. In what follows, we examine each step in turn.

b1) Identification of citations to scientific publications in NPL citations

The first step of Phase 2 was devoted to identify within the set of parsed NPL citations, those records that correspond to scientific paper references. To this end, for each parsed NPL citation, we took the *journal title* field and matched it with the list of journal titles covered by ISI-SCI. The matching procedure involved both electronic and manual processing of data.

In the first place, we generated the Cartesian product of the journal titles contained in NPL citations and the journal titles covered by ISI-SCI, i.e. we have joined each row of the table containing NPL journal titles to every row containing ISI-SCI journal titles. For each pair of titles thus resulting, we then computed various measures of *distance* between the two strings of text. In particular, we implemented three measures of distance using SAS v.9: the generalised edit distance, the Levensthein distance and the asymmetric spelling distance.

Pairs of journal titles for which the distance was lower than a certain threshold (i.e. with a high degree of similarity between the journal title in the NPL citation and the journal title in the SCI-ISI dataset) have been manually checked. Whenever we found that a NPL citation referred to a journal covered by ISI-SCI, the journal title in the NPL citation was standardised and the correct ISSN code has been assigned. It is important to point out that this procedure was very time-consuming, but allowed us to identify *all NPL citations*, which have been successfully parsed and are covered by ISI-SCI. In other words, we can confidently say that the probability that a parsed NPL citation covered by ISI has not been captured is extremely low.

The procedure described above allowed us to split the NPL citations dataset into two distinct subsets of data: on one hand, the NPL citations corresponding to scientific articles, and on the other hand, the NPL citations corresponding to other kinds of materials (books, manuals, technical bulletins etc.).

Out of 350,735 unique (i.e. de-duplicated) NPL citations contained in EPO patents, we identified 212,228 unique NPL citations to ISI covered journals. In a similar way, out of 718,256 unique NPL citations contained in USPTO patents, we identified 500,547 unique NPL citations covered by ISI journals. Considering jointly the EPO and the USPTO, the number of distinct SCI-ISI journals for which we found citations in patents amounts to 755 titles.

Box 1- Data quality control

The identification of ISI-SCI covered publications in NPL citations is a complex and time consuming task. Although there are various possible ways in which one can test the success of the procedures implemented, the most obvious one is to calculate the share of all scientific articles cited in patents that one is able to correctly identify and retrieve. Yet, this is quite a problematic kind of test, given that one should first identify *all* SCI-covered references (i.e. precisely what is unknown) in order to say anything about the success ratio of the matching procedure. A possible alternative approach is to take a small sample of all NPL citations, manually identify and retrieve all scientific references contained in them, and compare these with the results obtained by implementing the matching procedure. To assess the quality of the procedures used in this study, however, we have taken a different route by comparing our results with those obtained in the study *Linking Science to Technology* conducted by INCENTIM for the DG Research in 2003.

Before illustrating the comparison, it must be pointed out that the focus of that study was quite different from the focus of the present study. In addition, the parsing and matching procedures were also quite different. The study conducted by INCENTIM started from the identification of *potential* scientific references in NPL citations using a set of keywords, such as 'journal' and 'volume', and then applied a parsing procedure to the set of identified records. We started instead from parsing all NPL citations and then proceeded to match journal titles with SCI-covered journal titles in order to identify scientific references. A further difference that should be noted is that, whereas the procedure implemented by INCENTIM researchers is highly automatised, our methodological approach involves several manual checkings. As a consequence, the two procedures are probably suited to different types of problems. A highly automatised procedure is likely to be suited to the analysis of large numbers of NPL citations, whereas our procedure is more suited when the focus is upon specific technologies and, as in our case, on analysis at the level of individual papers and authors.

The following table illustrates the number of citations to scientific papers contained in EPO patents for two technology fields, by comparing our parsing and matching procedure and the procedure followed by INCENTIM. As the time coverage of the two studies is different, we conducted the comparison only for the years they have in common[8]. Moreover, we compared two technology fields for which the definition in terms of IPC codes is the same. We note that, in general, our approach tends

[8] The figures reported in the table are taken, respectively, from volume 8 (Table 55, p. 89) and volume 5 (Table 62, p. 87) of the Report *Linking Science to Technology - Bibliographic References in Patents*, downloadable from http://cordis.europa.eu/indicators/kul_report.htm.

to produce a larger number of identified scientific citations, although the difference is particularly significant in the case of Telecommunications. It is likely that the difference between the two studies is due to the different parsing and matching strategies followed. In particular, the strategy of identifying potential scientific citations on the basis of keywords such as 'journal' and 'volume' has the drawback that many journal titles cited in patents do not contain the word 'journal'. Likewise, many NPL citations do not contain the word 'volume' (or 'vol.'). As a consequence, relying upon such pre-identification strategy involves the risk of underestimating the actual number of scientific citations.

Years	1992	1993	1994	1995	1996
Fields					
Telecommunications					
CESPRI	1181	1219	1356	1596	1634
INCENTIM	76	48	56	161	212
Biotechnology					
CESPRI	3929	4013	4759	5512	6105
INCENTIM	2669	1938	1981	2149	2086

Given the results reported above, we believe that our work of parsing and matching NPL citations respects high standards of quality and that coverage of our dataset is more than satisfactory.

b2) Analysis of science intensity of 10 technology fields

The dataset thus compiled was used to analyse the science intensity and various other aspects of the science-technology linkages for each of the 10 technology fields originally selected[9].

Table 2 and 3 report the total number of scientific paper citations, the fraction of patents citing scientific articles and the degree of science intensity, respectively for the EPO and the USPTO patents, for each of the 10 technology fields. The average science intensity is defined as the number of cita-

[9] It is worth remarking that at this stage of our procedure, for each NPL citation we were able to distinguish between citations to articles published in ISI covered journals and citations to other material. Yet, for the set of citations to scientific articles, the only information available at this stage, was the standardised title of the journal. This means that we were still unable to identify with a unique ID code the same article cited in more than one patent. In addition, we did not possess any information on authors' names, affiliations and publication year of the cited articles. However, this dataset allowed us to calculate some basic statistics about the science intensity of the different technology fields and was therefore instrumental in the identification of the 5 technology subfields for the final phases of the study. In fact, we could safely assume that a given patent will not cite the same article more than once, while it is certainly possible that the same patent cites two different articles from the same ISI journal. In other words, for each patent, we could calculate how many citations it made to scientific articles by simply summing up the number of NPL references containing a 'journal title' covered in the ISI dataset.

tions to scientific articles per 100 patents, both including and excluding in the calculation patents that do not cite scientific articles.

Table 2 – EPO average science intensity by technology field (1990-2003)

Technology fields	# of citations to ISI scientific publications	% of all patents citing scientific ISI articles	Average science intensity (only patents citing scientific articles)	Average science intensity (all patents)
1. Telecommunications	21213	19.6	142.0	27.8
2. Information Technology	14112	19.9	145.9	29.0
3. Semiconductors	8837	21.8	181.3	39.5
4. Optics	12106	16.2	194.1	31.4
5. Control Technology	23705	17.8	209.0	37.2
6. Medical Technology	3755	5.0	162.0	8.1
7. Organic Chemistry	54429	39.0	260.5	101.7
8. Drugs	49373	43.3	319.7	138.4
9. Biotechnology	81296	78.2	382.1	298.9
10. Environmental technology	669	4.2	135.6	5.7

Table 3 – USPTO average science intensity by technology field (1990-2003)

Technology fields	# of citations to ISI scientific publications	% of all patents citing scientific ISI articles	Average science intensity (only patents citing scientific articles)	Average science intensity (all patents)
1. Telecommunications	37821	9.2	219.8	20.2
2. Information Technology	41193	12.0	211.2	25.4
3. Semiconductors	39290	17.9	274.0	49.2
4. Optics	24135	8.5	272.0	23.1
5. Control Technology	36948	10.4	324.8	33.8
6. Medical Technology	46960	11.4	376.5	42.8
7. Organic Chemistry	107573	26.8	642.5	172.4
8. Drugs	235199	34.5	773.9	267.5
9. Biotechnology	252214	57.0	976.4	557.2
10. Environmental technology	6738	6.9	329.4	22.7

The first point to note is that the fraction of patents that cite scientific literature greatly differs across technologies. The highest propensity to cite science is found in biotechnology, drugs and organic chemistry. On the other hand, a very small fraction of all patents in medical technology and environmental technology tends to rely on scientific papers. Similar differences emerge from examining the average number of cited articles per patent.

Looking at EPO data, we observe that patents in biotechnology cites on average almost 3 articles, which becomes almost 4 if one excludes from the calculation patents that do not cite scientific articles. Medical technologies and environmental technologies are the fields with the lowest science intensity. However, it is also interesting to note that if one takes into account only patents that actually cite scientific articles, the average number of cited articles increases dramatically and does not differ significantly from that of other technological fields. Put differently, very few patents in these domains cite scientific literature, but the science intensity of those few patents does not differ from that found in other technological domains[10].

A comparison between EPO and USPTO data reveals that the share of USPTO patents that cite ISI-covered articles is significantly lower than the corresponding share at the EPO. This result may suggest either differences in the quality of patents between the two offices or the existence of a different citation style between patent examiners at the EPO and at the USPTO. In addition, if one restricts the attention to patents that actually cite scientific articles (second column), the average science intensity at the USPTO is significantly higher than the corresponding value at the EPO for all technology fields examined. Combining the two observations, one can say that a relatively lower fraction of USPTO patents tends to cite scientific articles, but the average number of cited articles by patents that actually do so is much higher than the corresponding value for the EPO patents.

Beside the science intensity, we also calculated the share of all scientific citations accounted for by the most important journals for each technology field. Table 4 and 5 report the number of distinct ISI-

[10] We also conducted a thorough analysis of the number of cited articles, average science intensity and fraction of patents citing scientific articles at the 4-digit IPC level for each technology field. Detailed results are available in the website of this study.

journals cited by patents in the 10 technology fields as well as the share of citations accounted for by the 4 most cited journals, respectively for the EPO and the USPTO.[11]

Box 2 – Citation rules

The results reported in the text are consistent with the differences in the citation rules and patent examination procedures between the EPO and the USPTO. Citations are references either to previous patents (issued by the same patent office or by other offices) or other literature (mainly, scientific literature) to be found on the so called 'search report' attached by patent examiners to patent applications. Search reports by EPO examiners are separate documents one can find attached to patent applications, once published. As for USPTO, no separate 'search report' is published; however, the examiner's citations, as opposed to the applicant's, are listed separately on the front page of the patent document (Karki, 1997). More citations can be found in other sections of both the EPO and the USPTO patent documents, such as those dedicated to describing the invention or the novelty claims. However, these are much less easily available in electronic format, and much more erratic in their frequency.

Citations help both the examiner and the applicant to judge the degree of novelty and the inventive step of each application. After receiving the search report the applicant should have enough information to decide whether to go on pursuing the patent (which requires paying additional fees) or to give up, because the risk of rejection has been proved too high. Citations on the search report also form the basis for future search activities, especially by opponents wishing to challenge the patent's validity in court.

The USPTO requires applicants to disclose all the prior art they are aware of and deem relevant to this end ('duty of candour' rule), so we presume that many citations, although filtered by the examiner, were first proposed by the designated inventors. Formally, USPTO applications may come only from individual inventors who assign their rights to legal persons such companies and other organizations only after the patent has been granted. So, ideally, all the prior art cited in observance of the 'duty of candour' rule come from the inventors themselves. Of course this is not the case: it is the inventors' employers who actually manage the application procedure, with their legal and patent intelligence aids actually choosing the prior art to be cited (even truly independent inventors rely upon such aids).

The EPO does not impose any requirement of that kind, so that all the citations come straight away from the patent examiners. The EPO places great emphasis on the thoroughness and parsimony of its 'patentability search' procedure: the examiners report only the prior art that really threatens the patentability of the invention. In contrast, the USPTO provides a broader 'documentary search', aimed at collecting any reference which the applicant or the examiner suggest to be somehow useful in understanding the application contents (Akers, 2000). The following statements confirm this difference:

[11] The list of the four most important journals and the four most important subject fields in terms of citations received in patents is reported in the appendix, separately for the EPO and the USPTO, for each of the 10 technology fields.

According to the EPO philosophy a good search report contains all the technically relevant information within a minimum number of citations. [Michel and Bettels, 2001; p.189]

The USPO examiner's] purpose is to identify any prior disclosures of technology … which might be similar to the claimed invention and limit the scope of patent protection … or which, generally, reveal the state of the technology to which the invention is directed [OTAF (1976), as cited in Hall, Jaffe, and Trajtenberg, 2001; pp.14-15]

When it comes to counting the number of citations per patent, the USPTO stands out as an exception: the average number of citations reported on its patents is much higher than similar figures for the EPO. According to Michel and Bettels (2001), UPSTO patents cite on average about 13 other patents, and about 3 non-patent documents, whilst the same figures for EPO patents are 4 and one. For USPTO patents applied for in 1990, Agrawal, Cockburn and McHale (2003) calculate 10.2 average citations; our own calculations for EPO patents reveal about 2.8 citations received over 10 years of life (Breschi et al., 2003). However, when one compares the search reports issued by the USPTO and the EPO for international patent applications subject to Patent Cooperation Treaty (PCT), all of these differences disappear, with the USPTO figures converging towards EPO values. It is the 'duty of candour' rule and the 'documentary search strategy' which make the difference: when examining PCT patent applications, in fact, both the USTPO and the EPO have to stick to the same set of rules issued by the World Intellectual Property Organization (WIPO), and differences in the citation figures disappear. In addition, Hall, Jaffe, and Trajtenberg (2000) make clear that some kind of 'citation inflation' phenomenon may have affected USPTO patents in recent times, due to the booming patenting activity of US companies, which has placed an increasing burden on patent examiners. Clashing against time-constraints and the USPTO rules for the 'documentary search' strategy, this burden may have forced the examiners to be less and less selective in picking up the right references to place on their reports.

In conclusion, the messages one can obtain from EPO citations are much less 'noisy' than those from the USPTO ones. With EPO patents we can safely presume that all the citations have been chosen by the examiner, no matter whether the inventors knew about them in advance. With USPTO patents confusion reigns about who is entirely responsible for the front page citations: it is only since January 2001 that indications have become available on whether individual citations come from the examiner or the inventor. In addition, cited-citing patent couples retrieved from EPO databases may be legitimately supposed to be 'closer', both in time and as for technological content, than those coming from USPTO data.

Both the total number of cited ISI journals and the share of the four most cited journals widely differ across technologies. The number of distinct journals cited by patents is relatively low for environmental technology, semiconductors and optics. In addition, the share of all citations accounted for by the four most cited journals is remarkably high especially for semiconductors and optics. At the other end of the spectrum, biotechnology, drugs and control technology tend to cite a much larger set of journals and the share of the four most cited journals is significantly lower, therefore indicating a

wider dispersion of all citations across different journals. Comparing tables 4 and 5 also highlights a few differences. In the case of medical technology and environmental technology, the total number of distinct ISI journal cited by USPTO patents is much larger than the set of journals cited by EPO patents. Moreover, for information technology the share of citations accounted for by the top 4 journals is significantly higher at the USPTO than at the EPO, while the opposite occurs for optics.

Table 4 – Share of scientific citations of top 4 cited journals (EPO)

Technology fields	# of ISI journals	Share of top 4 cited journals
1. Telecommunications	453	23.1
2. Information Technology	823	11.1
3. Semiconductors	328	41.0
4. Optics	435	43.5
5. Control Technology	2060	10.5
6. Medical Technology	857	17.3
7. Organic Chemistry	1705	23.8
8. Drugs	2245	10.7
9. Biotechnology	1876	21.1
10. Environmental technology	193	24.0

Table 5 – Share of scientific citations of top 4 cited journals (USPTO)

Technology fields	# of ISI journals	Share of top 4 cited journals
1. Telecommunications	591	22.1
2. Information Technology	774	22.5
3. Semiconductors	406	42.9
4. Optics	516	31.7
5. Control Technology	1577	13.6
6. Medical Technology	1690	11.1
7. Organic Chemistry	1700	20.3
8. Drugs	2314	16.4
9. Biotechnology	2025	25.9
10. Environmental technology	649	25.8

b3) Selection of 5 technology subfields

On the basis of the analysis presented above, the final step of Phase 2 of the study involved the selection of 5 technology fields to use in the next phases of the research. Given the overall objective of the

study, i.e. analysing the network linkages among scientists and inventors, the first and most obvious criterion that has driven our choice has been the selection of fields in which the propensity of patents to cite scientific articles was high. This selection criterion ensures in fact that the fields selected present (at least potentially) a high degree of interactions among individuals involved in science and individuals involved in industrial research.

Beside this general principle, however, our choice has been also influenced by a few other considerations. In the first place, given the purpose of evaluating the nature and strength of network linkages among authors and inventors, we selected technology fields to ensure that a certain degree of technological and scientific *coherence* exists between the activities of academic scientists, on one hand, and industrial technologists, on the other hand. This selection strategy minimises the risk that the set of (highly cited) authors and the set of (highly cited) inventors are engaged in different areas of research and are therefore less likely to be connected by any network linkage. In this respect, the 10 technology fields identified for the first phase of the study were just *too broad*, relating to different and rather heterogeneous domains of research. Selecting the 5 technology fields from this set would have therefore generated the risk that scientists and inventors were not connected by any linkage, not because the lack of collaboration or interaction among them, but because they are engaged in different fields of research and therefore belong to different research communities.

Second, our choice was dictated by considerations of feasibility. The total number of articles cited in the most science-intensive technology fields is in fact rather large. For example, EPO patents in the drugs field, which is one of the sectors presenting the highest degree of science intensity, cited 51,706 distinct scientific articles over the period 1990-2003. Assuming an average of three authors per paper, the number of distinct authors was likely to be so large that it would have been unfeasible to clean and process all relevant data. Moreover, the very large number of nodes in the network of scientists-inventors would have presented computational problems in the calculation of basic network measures.

Given the three broad criteria discussed above, i.e. *science-intensity*, *coherence* and *feasibility*, we identified 5 technology *subfields*, by considering 4-digit IPC classes comprised within the 10 technology

fields originally selected.[12] The 5 technology subfields and the corresponding IPC codes that identify them are reported in Table 6, while Table 7 report the degree of science intensity of the 5 subfields chosen in comparison with the science intensity of the broader fields from which they have been selected. [13]

Table 6 – 5 technology subfields on the basis of IPC codes

Technology fields	IPC codes
1. Transmission of digital information	H04L
2. Speech analysis and image data processing	G10L, G06T
3. Semiconductors	H01L
4. Lasers	H01S
5. Biotechnology (measuring, testing, diagnostics)	C12Q, G01N33 (/53,54,55,57,68,74,76,78,88,92)

Table 7 – EPO average science intensity by technology field and selected subfields (1990-2003)

Technology fields	Average science intensity (only patents citing scientific articles)
Telecommunications	142.0
1. Transmission of digital information	**206.8**
Information Technology	145.9
2. Speech analysis and image data processing	**218.8**
3. Semiconductors	**181.3**
Optics	194.1
4. Lasers	**333.9**
Biotechnology	382.1
5. Biotechnology (measuring, testing, diagnostics)	**450.4**

[12] The choice of the 5 technology subfields involved a mixture of quantitative analysis of available data, subjective assessment based on experience with patent statistics and technological fields, and review of existing studies. In particular, for the identification of the Biotech subfield we adopted suggestions contained in the OECD Report "A Framework for Biotechnology Statistics" (Paris, 2005). Likewise, for the identification of the Telecom and ICT subfields we followed suggestions contained in the Report "Europe's strengths and weaknesses in Information Society Technologies, A patent analysis", (Fistera Thematic Network, IST-2001-37627, 2005).

[13] As we did for the selection of the 10 broad technology fields, the identification of the 5 technology subfields has taken IPC codes as reference point. This procedure required therefore the identification of the US patents and the USPC codes corresponding to the selected technological subfields. This has been accomplished by using the USPC-IPC concordance table, which is available from http://www.uspto.gov/go/classification/. The USPC codes corresponding to the selected IPC classes are available upon request.

In general, we note that all the fields selected present a science intensity, which is significantly higher than the corresponding science intensity for the broader technology field to which they belong. For example, in the case of lasers each patent citing scientific articles cites on average 3.3 articles, as compared to 1.94 articles for patents in the larger field of optics. It should be also noted that the science intensity of lasers is comparable to that of biotechnology as a whole, and it is higher than that of organic chemistry and drugs (cf. Table 2 above). This amounts to say that within each broad technology field (e.g. optics) there is a large heterogeneity in terms of science intensity across subfields. Our choice has been to pick up subfields whose science intensity is higher than the average science intensity in their respective broader fields.

Besides satisfying the general principles discussed above, the selected subfields also present a few additional characteristics. First, they show a highly dynamic trend in patenting. As we expect the interaction between science and technology to be stronger in sectors with high rates of growth in patenting, this constitutes a further element supporting the choice made. Second, new innovative firms account for a large share of total patenting activity in all the five fields selected. As we expect the scientific background of such companies to play an important role in their performance, we believe that this represents another aspect which militates in favour of our choice.

Finally, it should be also noted that the selected subfields account for a fairly large number of all citations made to scientific articles in their respective broader technology field. This is shown in Table 8, which reports, for each technology subfield, and separately for the EPO and the USPTO, the number of citations to scientific articles as well as the share of all citations to scientific articles within the broader technology field to which they belong. Thus, for example, we observe that EPO patents in the subfield of *lasers* account for about 38% of all scientific citations made by EPO patents in the broader field of *optics*. More generally, we observe that, possibly with the exception of 'transmission of digital information' and 'speech analysis and image data processing', the chosen subfields appear to be well representative of the wider technological domains from which they have been selected.

Table 8 – Coverage of selected subfields in terms of scientific citations (1990-2003)

Technology subfields	# of citations to scientific articles	Share of citations of the broader technology field (%)
	EPO	
Transmission of digital information	6191	29.2
Speech analysis and image data processing	2841	20.1
Semiconductors	8837	100.0
Lasers	4570	37.7
Biotechnology (measuring, testing, diagnostics)	21631	26.6
	USPTO	
Transmission of digital information	4939	13.1
Speech analysis and image data processing	6034	14.6
Semiconductors	36627	93.2
Lasers	9524	39.5
Biotechnology (measuring, testing, diagnostics)	100364	39.8

c) Phase 3: Extraction of ISI-publications and identification of highly cited articles

Phase 3 of the study focused on the five technology subfields described above. This phase involved three basic steps. First, for each technology subfield, we matched the identified paper citations with the source data coming from ISI, in order to extract the full set of information regarding each paper citation (i.e. authors' names, affiliations etc.). Second, we defined the criteria to identify highly cited publications and highly cited patents. Finally, we carried out a benchmarking analysis to examine to what extent cited and highly-cited publications are only cited within the realm of patents or conversely they are also cited by other scientific publications. In what follows, we discuss these various steps in detail.

c1) Extraction and matching with ISI-covered publications

The first step of phase 3 consisted in matching for each technology subfield the identified paper citations with the source data coming from ISI. To this purpose, we proceeded as follows. For each technology subfield, we identified the *most important* journals in terms of number of citations received from patents and extracted from the ISI dataset all articles published therein over the period 1969-

2003. The choice of focusing only on the most important journals in each subfield was dictated by considerations of feasibility, given the amount of publications that had to be extracted from the ISI dataset, and by the fact that the most cited articles are likely to appear in journals that receive overall a large number of citations. The number of distinct journals cited in patents, both at the EPO and at the USPTO, is in fact very large (see above Tables 4 and 5), but most journals account for just one or few citations from patents. Given the criteria followed to define highly cited publications (see below), these journals are likely not to be particularly relevant and have been therefore excluded. Specifically, for each technology subfield we only considered journals, which received at least 5 citations from patents over the period 1990-2003.

For each journal considered, we compared the list of *article titles* cited in patents with the list of *article titles* as reported in the ISI dataset. More specifically, for each specific journal (e.g. Journal of Biological Chemistry) we created all possible pairs of article titles cited in patents and article titles published, by generating the Cartesian product of the two vectors, and compared the two strings of text by calculating various measures of distance among them. In particular, we implemented three measures of distance using SAS v.9: the generalised edit distance, the Levensthein distance and the asymmetric spelling distance. Pairs of article titles for which the distance was lower than a certain threshold were manually checked and the same ID code was assigned to pairs of articles that turn out to be the same publication. This was a rather long and time-consuming task, which involved a fairly large amount of manual checking. However, it allowed to identify and extract information on *all* cited publications published in the most important journals in each field.

Table 9 reports summary statistics on the output of this task. More specifically, it reports information for each technology subfield and separately for the EPO and the USPTO on the coverage of the dataset in terms of total number of citations to ISI covered articles and in terms of share of all scientific citations of each subfield for which we found a match with ISI publications. As one can see, the coverage of the dataset is more than satisfactory.

Table 9 – Coverage of selected subfields after matching with ISI (1990-2003)

Technology subfields	# of citations to scientific articles	Share of citations of the subfield (%)
	EPO	
Transmission of digital information	4245	68.6
Speech analysis and image data processing	1689	59.4
Semiconductors	6700	75.8
Lasers	3756	82.1
Biotechnology (measuring, testing, diagnostics)	13940	64.4
	USPTO	
Transmission of digital information	3155	63.8
Speech analysis and image data processing	2185	36.2
Semiconductors	24909	68.0
Lasers	7878	82.7
Biotechnology (measuring, testing, diagnostics)	74231	73.9

For example, with reference to semiconductors, the overall number of citations to scientific publications from patent applications in the period 1990-2003 is equal to 8837 for the EPO and 36627 for the USPTO. Of these, we were able to match publications data from ISI for 6700 citations (75.8%) in the case of EPO and 24909 (68.0%) citations for the USPTO. The data we have collected are also likely to contain the most cited publications, which was one of the purposes of the study.

c2) Definition of highly cited publications

The second step of Phase 3 involved a thorough statistical analysis of the citation patterns from patents to scientific publications in order to define the criteria for the identification of highly cited publications. The identification of the highly cited articles required to make a choice concerning: i) the time window within which counting the number of citations received by scientific articles; ii) the threshold above which articles qualify as highly cited. To this purpose, we adopted the following procedure.

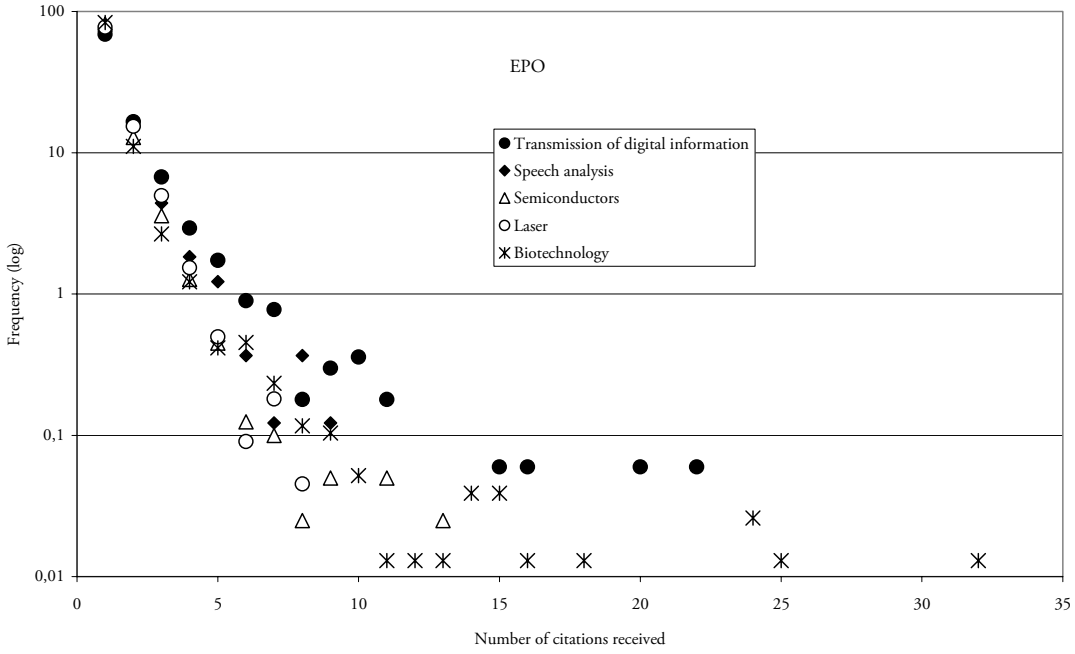
First, we defined a *citation window* of *t* years after the publication date of the cited articles. The need to define a citation window arises from the fact that articles published far in the past have simply had more time to receive citations from patents compared to most recently published articles. Defining a

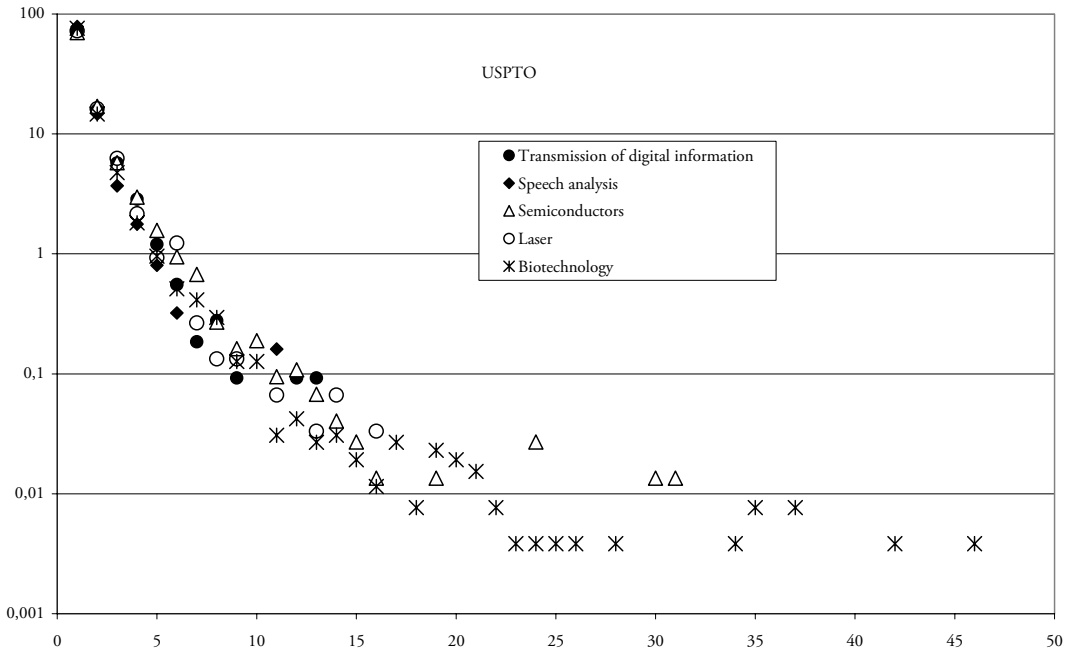
time window solves this problem by allowing all articles the same period of time to receive citations independently on the year of publication. More specifically, we selected three windows of time, respectively of 3, 5 and 7 years, within which we counted the number of citations received in patents by individual publications. The citation window has been defined as the difference between the application date of the citing patent and the publication date of the cited article. Thus, for example, the 3-years citation window includes only those publications for which the difference between the year of the citing patent and the year of the cited publication is less than or equal to three years.

For each time window, we then calculated the distribution of the number citations of *per* publication, for each technology subfield and separately for the EPO and the USPTO. It is very important to remark that in calculating such distributions, we took into account the fact that for the most recently published articles the available time-window may be shorter than the selected time window. For example, when considering the 3-years time window, all articles published between 2001 and 2003 had to be excluded from the calculations, as the available time window between their publication year and the application date of most recent citing patents, i.e. 2003, was shorter than 3-years.

For each technology subfield, and separately for the EPO and the USPTO, we have tabulated the distribution of citations, by calculating the following variables: the number of scientific articles receiving a given number (i.e. 1, 2, 3, etc.) of citations, the percentage of all cited articles receiving a given number of citations, the cumulative distribution, i.e. the percentage of all cited articles receiving *less* than (or equal to) a given number of citations, the *inverted* cumulative distribution, i.e. the percentage of all cited articles receiving *more* than (or equal to) a given number of citations, and the number of articles receiving *more* than (or equal to) a given number of citations. For reasons of space, these tabulations are not presented in this report, but they are available in the website constructed for this study. For the sake of summarising, below we have only reported the graphical distribution of the number of citations received by publications in a time window of 5 years since their publication date, separately for the from EPO and the USPTO (Figure 4). The x-axis reports the number of citations, whereas the y-axis reports the log of the frequency of articles receiving a given number of citations. For all technology fields examined, the distribution of citations to scientific articles appears highly skewed with the vast majority of articles receiving only one or two citations.

Figure 4 – Frequency distribution of the number of citations received by scientific articles in a 5-years time window since their publication date





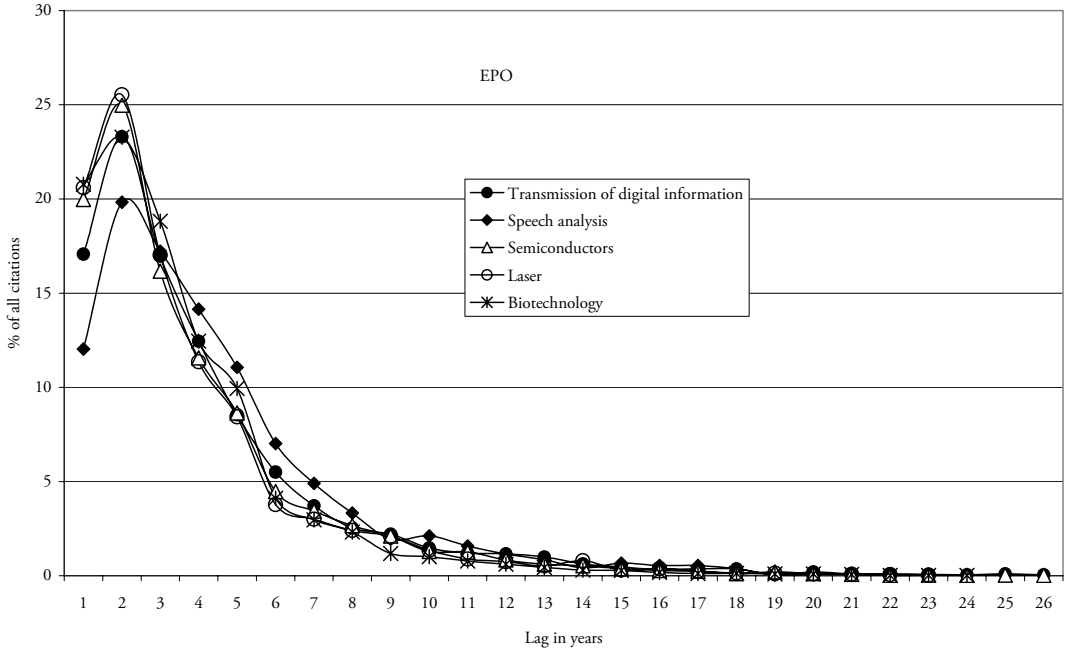
With reference to EPO data, the percentage of all articles receiving only one citation in a time window of 5 years from the date of publication goes from 83% for biotechnology to 69% for transmission of digital information. Always with reference to EPO data, the percentage of all articles receiving two citations in a time window of 5 years since their publication date ranges from 16,6% for transmission of digital information to 11.0% for biotechnology. Looking instead at the right tail of the distribution, very few articles receive a large number of citations. The distributions for the USPTO data look rather similar to those for the EPO. Taking again a 5-years time window, the percentage of all articles receiving only one citation goes from 78% for speech analysis to 70% for semiconductors.

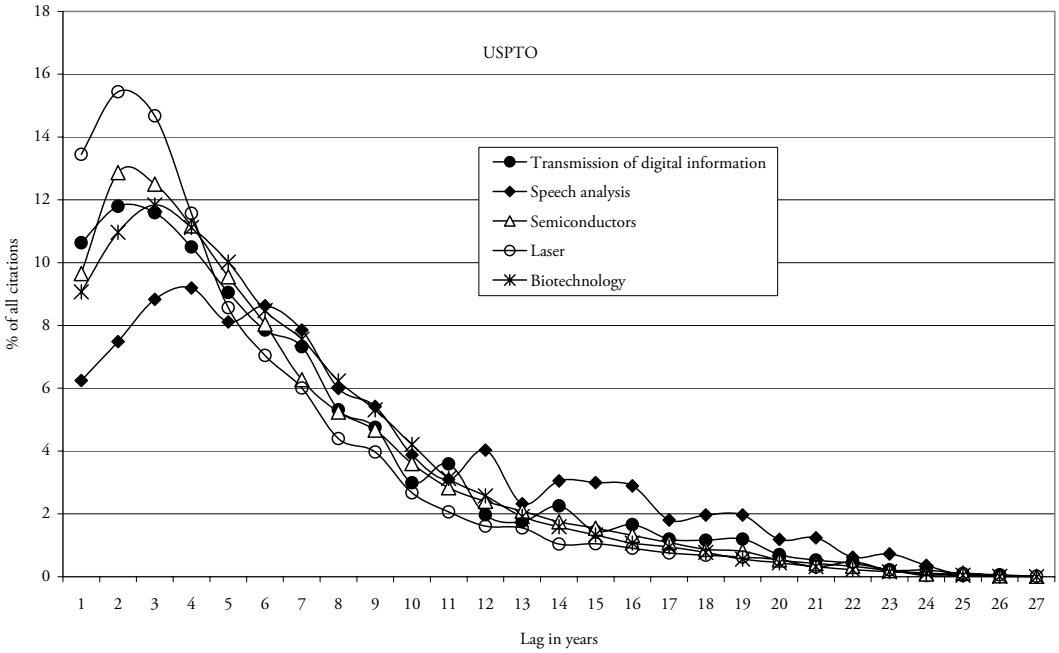
Overall, publications receiving less than or equal to two citations account for about 90% of all cited articles in our sample. This is, of course, a crucial aspect to take into account for the definition of *highly cited publications*. It implies that receiving at least three citations puts an article in the upper 10% of the citation distribution.

In order to choose among the three options for the time window, we calculated for each technology subfield the time lag between the publication year of cited articles and the application year of citing patents. The distribution of time lags between patents and publications is reported in Figure 5 respectively for the EPO and the USPTO.

In both cases, we observe that the distributions tend to peak around two years. In other words, the modal time lag between scientific publications and citing patents is around two years. However, the shape of two distributions looks rather different. In the case of the EPO, we note a sharp decline in the frequency of citations with a long time lag: for all the technology fields considered here around 80% of all citations have a time lag between citing patents and cited publications lower than 5 years. In the case of the USPTO, the right tail of the distributions declines more slowly: the percentage of citations with a time lag lower than 5 years is comprised between 49% for biotechnology and 61% for semiconductors. The difference in the shape of the two distributions is probably due to the fact that USPTO patents cite on average a larger number of scientific papers per patent, some of which have been published far in the past. As the focus of this study was on the network linkages between scientists and inventors, and given that network ties are likely to decay over time, we believed that a reasonable compromise in this context was to assume a 5-years time window.

Figure 5 – Lag in years between citing patents and cited publications (EPO and USPTO)
Percentage of all citations from patents to publications, 5 technology subfields





As for the definition of a citation threshold, we considered two possible alternatives: i) defining a *percentage* threshold (e.g. 2%) to delimit the percentile of publications that qualify as 'highly cited'; ii) defining an *absolute* number of citations (e.g. 5) above which publications qualify as 'highly cited'. The major drawback of the first approach is that for some of the fields examined the number of articles comprised within the top percentile chosen may be rather small. For example, with reference to the field of lasers, if highly-cited papers are defined as those articles in the top 1% of the distribution one ends up with less than 18 articles out of more than 2200 articles cited by EPO patents. Also the alternative approach has some limitations. The major one is that once defined the absolute number of citations that qualify highly cited articles, the percentage of all cited articles receiving more than the chosen number of citations may differ across fields. For example, if highly-cited papers are defined as those receiving at least 5 citations within a 5-years time window, one ends up with 78 papers for transmission of digital information, corresponding to 4,65% of all cited papers in this field; on the other hand, only 1,51% (121) of all articles cited by biotechnology patents receive at least 5 citations.

Putting together the results presented above and looking at the distributions of citations for the five subfields, we believed that a reasonable compromise in the context of the present study was to define highly cited articles as those *publications receiving four or more citations within a time window of five years*. It should be however added that the exact definition and meaning of what constitutes a highly cited publication is not crucial, at least with respect to EPO data. As we will explain below, for EPO data, we have in fact chosen to analyse network linkages by considering *all* papers and authors cited in patents, independently on the number of citation received from patents.

Table 10 reports the number and percentage of articles that qualify as highly cited according to this criterion (middle panel), respectively for the EPO and the USPTO, comparing the results with possible alternative definitions of highly cited publications. According to the criterion proposed here, the percentage of all cited articles that qualify as highly cited is comprised between 2.09% in the case of semiconductors and 7.58% in the case of transmission of digital information for the EPO, and between 3.05 in the case of speech analysis and 7.19% in the case of semiconductors for the USPTO.

Table 10 – Highly cited publications (Publications with 3, 4, 5 or more citations within a 5 years time window)

Technology field	% of all cited articles	Number of articles	% of all cited articles	Number of articles
	EPO		USPTO	
	3 or more citations			
Transmission of digital information	14.33	240	11.09	120
Speech analysis and image data processing	8.34	69	6.75	42
Semiconductors	5.68	228	12.96	960
Laser	7.32	162	11.30	340
Biotechnology	5.44	420	9.31	2424
	4 or more citations			
Transmission of digital information	7.58	127	5.36	58
Speech analysis and image data processing	4.03	33	3.05	19
Semiconductors	2.09	84	7.19	533
Laser	2.35	52	5.05	152
Biotechnology	2.78	215	4.55	1186
	5 or more citations			
Transmission of digital information	4.65	78	2.49	27
Speech analysis and image data processing	2.20	18	1.28	8
Semiconductors	0.82	33	4.22	313
Laser	0.81	18	2.89	87
Biotechnology	1.56	121	2.73	713

With the exception of transmission of digital information for the EPO and semiconductors for the USPTO, the percentage of all cited articles that qualify as highly cited is below 5% for all fields examined. In other terms, the proposed selection criterion identifies highly cited articles within the 95th percentile of the distribution. In the absence of any a priori definition of what constitutes a highly cited article, we believed this was a reasonable criterion for the purposes of the present study.

c3) Definition of highly cited patents

The identification of highly cited patents has followed a procedure similar to the one described above for publications. First of all, we defined a *citation window* of *t* years after the application date of the cited patent. The citation window has been defined as the difference between the application date of the citing patent and the application date of the cited patents. Using this time window, we then counted the number of citations received by each patent and calculated the overall distribution of

citations between patents. More specifically, we have chosen three windows of time- respectively of 3, 5 and 7 years.[14]

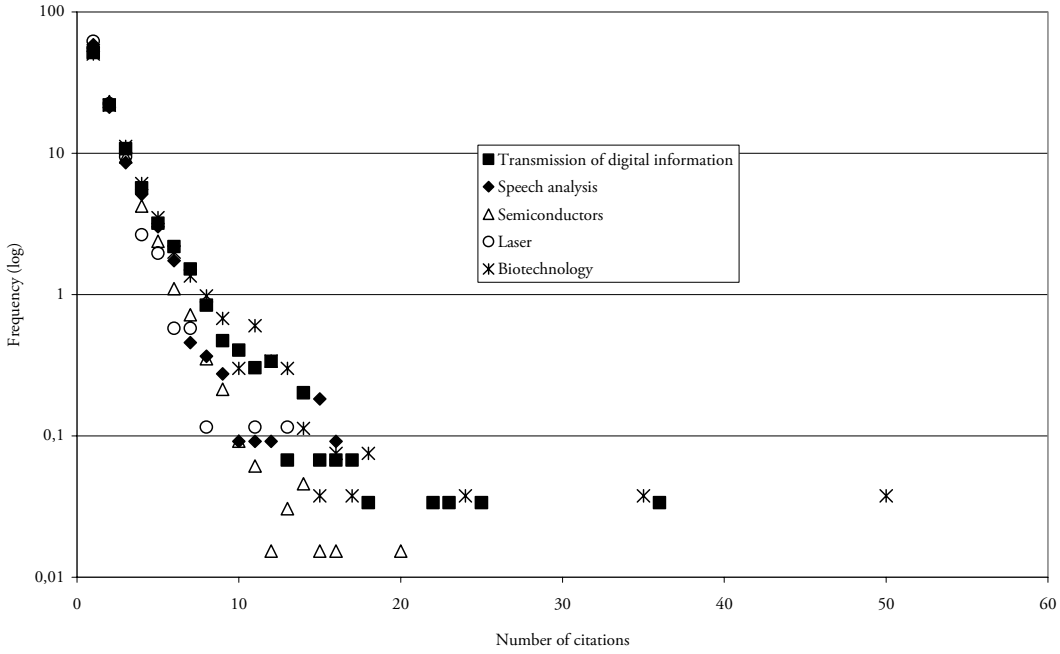
Figure 6 below reports the distribution of the number of citations received by patents within a time window of 5 years since their application date for EPO patents in the 5 technology subfields[15]. From a visual inspection of the graph below, it is immediate to observe that also in the case of patent citations the distribution of the number of citations look rather skewed, although the degree of skewness appears to be lower than in the case of scientific articles. Thus, for example, while the articles receiving more than three citations from EPO biotech patents within a 5-years window represent around 5% of all cited articles, patents receiving more than three citations from EPO biotech patents within the same time window represent 28% of all cited patents. In other terms, the right tail of the distribution appears relatively fatter for patent citations than for citations from patents to scientific articles. This suggests that the citation threshold for EPO patents should be set at a higher value than the threshold chosen for cited publications.

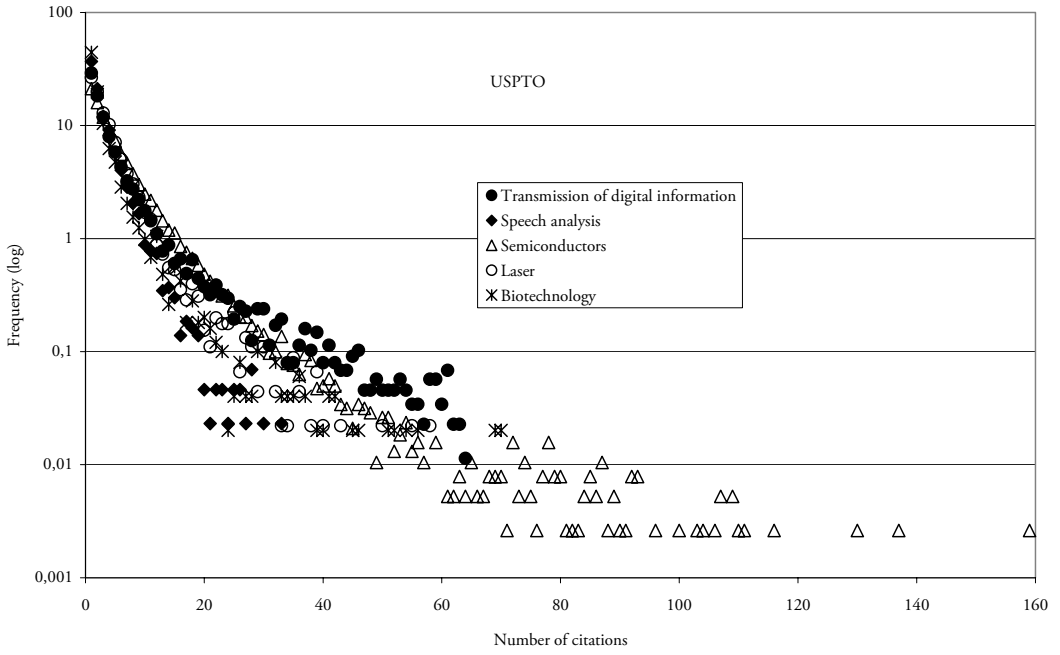
Looking at the distribution of patent citations for the USPTO (Figure 6) reveals that the shape of the distribution differs quite remarkably to the distribution for the EPO. In particular, the absolute number and percentage of patents receiving a very large number of citations is much greater at the USPTO compared to the EPO. This results in a distribution of citations that appears declining more slowly, i.e. the right tail of the distribution contains a larger number of documents. For example, with reference to semiconductors, while the percentage of all cited patents receiving more than 4 citations is equal to 9.3% at the EPO, this value rises to 51.0% at the USPTO. Likewise, the percentage of all semiconductors cited patents receiving more than 5 citations is equal to 5.0% at the EPO as opposed to 41.8% at the USPTO. The difference is even more striking if one takes the corresponding absolute values. Thus, the number of patents receiving more than 4 citations is equal to 608 for the EPO as opposed to 19458 for the USPTO. A similar pattern may be observed for all the subfields considered here.

[14] Detailed tabulations are not reported here, but are available from the website of the study.

[15] We do not report here the full set of distributions for various time windows. They are available in the website of this study.

Figure 6 – Distribution of the number of citations received by patents within a 5 years time window, 5 technology subfields (EPO and USPTO)





The reasons for this difference are twofold. On the one hand, the higher absolute number of patent applications at the USPTO as compared to the EPO. On the other hand, the different examination practices at the two patent offices, which result in a much higher number of citations reported on average on USPTO patent documents compared to EPO patent documents.

For the purposes of the present project, the major problem arising from these different patterns of citations is that the definition of what constitutes a highly cited patent must necessarily differ for the two patent offices. Looking again at the data reported in Figure 6, the percentage of semiconductors patents receiving 5 citations or more is around 5% for the EPO. Assuming that percentage as the threshold to identify highly-cited patents, this would imply selecting USPTO patents receiving 20 citations or more. There is a further problem that needs to be discussed. Even assuming a different (absolute) citation threshold for the two patent offices, the absolute number of resulting patents for the USPTO is remarkably larger. With reference to the previous example, the number of semiconductors patents receiving 5 citations or more is equal to 331 for the EPO; on the other hand, the number of semiconductors patents receiving 20 citations or more is equal to 1747 for the USPTO. Given that the next steps of the project involved cleaning the names, affiliations and addresses of highly-cited authors and inventors and linking them, we limited the examination of highly-cited inventors only to the case of EPO patents.

As for publications, we have also calculated the time lag between citing and cited patents. This is reported in Figure 7 for the EPO[16]. The graph shows that between 70% (in biotech) and 79% (in telecom) of all patent citations have a time lag between citing and cited patents lower than or equal to five years[17]. Moreover, the peak of the distribution is situated around two years for all the fields examined here[18].

---

[16] A comparison between the EPO and the USPTO (not reported here) reveals that the distribution of citation lags are also quite different between the two patent offices. In particular, the distribution of citation lags at the USPTO appears to decline relatively more slowly than the corresponding distribution for the EPO, thereby suggesting that patent citations at the former patent office refer to relatively older documents than patent citations at the latter patent office.

[17] The corresponding range for the USPTO is comprised between 56% for biotech and 78% for telecom

[18] The corresponding peak of the distribution for the USPTO is located around three years for all sectors, except lasers and transmission of digital information.

Figure 7 – Lag in years between citing and cited patents, Percentage of all citations from patents to patents, 5 technology subfields (EPO)

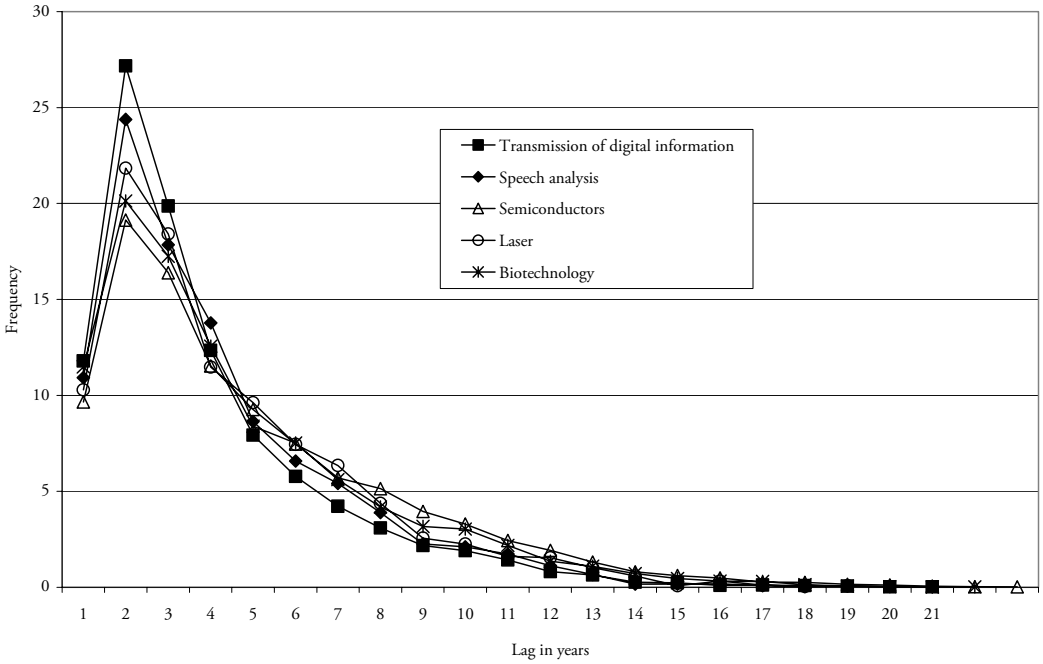


Table 11 – Highly cited patents, EPO (Patents with 4, 5, 6 or more citations within a 5 years time window since application date)

Technology field	% of all cited patents	Number of patents
	4 or more citations	
Transmission of digital information	15.6	463
Speech analysis and image data processing	11.5	126
Semiconductors	9.3	608
Laser	6.1	53
Biotechnology	16.6	441
	5 or more citations	
Transmission of digital information	9.9	294
Speech analysis and image data processing	6.4	70
Semiconductors	5.1	331
Laser	3.5	30
Biotechnology	10.5	279
	6 or more citations	
Transmission of digital information	6.7	199
Speech analysis and image data processing	3.3	37
Semiconductors	2.7	175
Laser	1.5	13
Biotechnology	7.0	186

On the basis of these data, we assumed a citation window of five years as a reasonable criterion for the purpose of counting the number of citations received by each individual patent.

Putting together the results presented above, we decided to limit the analysis of highly cited inventors to EPO patents and to define as *highly cited patents those (EPO) patents that have received six citations or more within a time window of five years since their date of application*. The percentage of all cited patents and the number of patent documents in the sample of highly cited patents according to this selection criterion are reported in Table 11, which also provides a comparison with respect to alternative criteria. The share of patents qualifying as highly cited is comprised between 7.0% for transmission of digital information and 1.5% in the case of lasers.

c.3) Benchmarking analysis of cited publications

The final step of Phase 3 concerned a benchmarking analysis that was conducted in order to examine whether cited and highly-cited publications are only cited within the realm of patents or whether they tend to be also highly cited by other scientific articles. The existing evidence on this issue is rather scant. One of the few studies on this topic is the one conducted by Gittelman and Kogut (2003). They analyse publications and patents of 116 biotechnology firms during the period 1988-1995 and show that important scientific papers (i.e. papers that receive many citations from other papers) are negatively associated with high-impact innovations (i.e. patents that receive many citations from other patents). According to them, this result points to conflicting logics between scientific and industrial communities, which are characterized by "different rules that govern the logic by which a good paper or a valuable patent is selected and replicated".

Our objective in this report was slightly different. We did not aim to test whether high-quality scientific papers are associated with high-impact patents. Rather, our aim was to assess whether and to what extent there is a positive correlation between the citations that a paper receives from patents and the citations that it receives from other publications.

Our benchmark analysis was conducted separately for each technological subfield. More specifically, our basic methodology consisted of comparing the average number of citations that publications cited in patents receive from other publications with the corresponding average for publications that are not cited in patents. The latter represent what we may call the *control group*. A major problem to be

addressed in this context regards exactly the selection of the control group. Roughly speaking, this should include publications that are as more *coherent* as possible with the publications cited in patents. By coherence here we mean the fact that the control group should be selected to ensure that differences in citation rates do not arise from underlying differences in factors, such as the publication date, the specific topic of the publication and the knowledge base. To this purpose, we adopted the following methodology. First of all, for each technology subfield, we selected all publications cited in patents and grouped them into four cohorts according to their publication year. Each cohort includes publications whose year of publication is within a three-year time window. More specifically, the four cohorts of publications are 1987-89, 1990-92, 1993-95 and 1996-98. For each cohort of publications, we computed the average number of citations received from other scientific papers and compared it with the average number of citations received by publications not cited in patents (control group). The control group was selected by including *all* publications in the same cohort, but which were *not cited* in patents and that were published

1) either in the same journal of the cited publications
2) or in the same subject field of the cited publications.[19]

In other words, the citation rates of publications cited in patents were compared with the citation rates of publications not cited in patents, but published in the same journal or in the same subject field. Of course, this selection procedure does not eliminate altogether all possible differences between cited and non-cited publications. Yet, it is likely that the problem becomes more important in the case of relatively broad scientific areas and generalist journals. As we will show below, the most important journals in each field are mostly specialist journals, rather than generalist journals. For this reason, we believe that most of the differences in citation rates between cited and non-cited publications are likely to reflect their underlying relevance for technological developments rather than other factors and that the analysis conducted here provides at least some new evidence on this issue.

Before presenting the results, two further remarks are needed. In the first place, the benchmarking analysis was carried out only with reference to scientific publications cited in EPO patents (and re-

[19] The ISI subject field is a journal-level classification developed by ISI Thompson, which groups journals according to the topics covered.

lated control groups). Secondly, given that for each technology field, the list of journals (and subject fields) in which publications cited in patents have been published includes a great number of different titles (and subject fields), we decided to delimit our analysis to the four most important journals and subject fields in terms of citations received from patents, for each technology subfield. Table 12 reports the top four journals by total number of citations received from patents for each of the five technology fields, whereas Table 13 reports the top four subject fields.

Table 12 – Top 4 journals by number of citation received from patents, Five technological subfields – EPO 1990-2003

Subfields/Journals	Citations received
Transmission of digital information	
COMPUTER NETWORKS AND ISDN SYSTEMS	306
IEEE COMMUNICATIONS MAGAZINE	304
IEEE JOURNAL ON SELECTED AREAS IN COMMUNIC	290
IEEE TRANSACTIONS ON COMMUNICATIONS	239
Speech analysis	
IEEE COMPUTER GRAPHICS AND APPLICATIONS	451
SPEECH COMMUNICATION	124
COMPUTERS & GRAPHICS	73
IEEE TRANSACTIONS ON MEDICAL IMAGING	49
Semiconductors	
APPLIED PHYSICS LETTERS	1241
IEEE TRANSACTIONS ON ELECTRON DEVICES	326
JOURNAL OF APPLIED PHYSICS	302
OPTICS LETTERS	275
Laser	
ELECTRONICS LETTERS	691
APPLIED PHYSICS LETTERS	579
IEEE PHOTONICS TECHNOLOGY LETTERS	431
OPTICS LETTERS	275
Biotechnology	
NUCLEIC ACIDS RESEARCH	796
SCIENCE	659
JOURNAL OF BIOLOGICAL CHEMISTRY	585
NATURE	451

Table 13 – Top 4 subject fields by number of citation received from patents, Five technological subfields – EPO 1990-2003

Subfields/Subject fields	Citations received
Transmission of digital information	
ENGINEERING, ELECTRICAL & ELECTRONIC	1645
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	471
COMPUTER SCIENCE, CYBERNETICS	389
COMPUTER SCIENCE, INFORMATION SYSTEMS	154
Speech analysis	
ENGINEERING, ELECTRICAL & ELECTRONIC	353
COMPUTER SCIENCE, SOFTWARE ENGINEERING	241
ACOUSTICS	174
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	102
Semiconductors	
ENGINEERING, ELECTRICAL & ELECTRONIC	1682
PHYSICS, APPLIED	1657
MATERIALS SCIENCE, MULTIDISCIPLINARY	328
PHYSICS, FLUIDS & PLASMAS	289
Laser	
ENGINEERING, ELECTRICAL & ELECTRONIC	1628
PHYSICS, APPLIED	670
OPTICS	437
PHYSICS, FLUIDS & PLASMAS	57
Biotechnology	
BIOCHEMISTRY & MOLECULAR BIOLOGY	3318
CHEMISTRY, ANALYTICAL	797
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	740
GENETICS & HEREDITY	736

Looking in particular at journals, it is worth noting that the distribution of patent citations is uneven across journals, even in the same technological class: for example, in Semiconductors, articles published in *Applied Physics Letters* receive almost 4 times the number of citations gathered by articles published in the second most important journal, i.e. *IEEE Transaction on Electronic Devices*. The laser field exhibits a similar pattern: the first journal receives on average more than twice citations as the fourth. With respect to Biotechnology, we note the presence of generalist journals like *Science* and *Nature* among the top four journals in the field. In this case, we suggest taking our results with cau-

tion, given that the comparison between cited and non-cited publications is likely to include articles that belong to rather different scientific areas.

In what follows, we shortly discuss the basic findings that emerge from such analysis, by reporting the average number of citations for the cohort of articles published in 1996-98 (Tables 14 and 15).[20]

Table 14 – Benchmarking analysis, Average number of scientific citations received by publications cited and not cited in patents (EPO, cohort 1996-98)

Subfields/Journals	Not cited	Cited
Transmission of digital information		
COMPUTER NETWORKS AND ISDN SYSTEMS	2,5 (475)	3,2 (91)
IEEE COMMUNICATIONS MAGAZINE	4,7 (564)	22,4 (66)
IEEE JOURNAL ON SELECTED AREAS IN COMMUNIC	15,2 (434)	37,2 (44)
IEEE TRANSACTIONS ON COMMUNICATIONS	9,2 (662)	35,8 (33)
Speech analysis		
SPEECH COMMUNICATION	4,9 (166)	6,0 (31)
COMPUTERS & GRAPHICS	2,1 (217)	5,4 (16)
IEEE TRANSACTIONS ON MEDICAL IMAGING	21,6 (281)	57,3 (20)
IEEE COMPUTER GRAPHICS AND APPLICATIONS	3,3 (318)	9,7 (7)
Semiconductors		
APPLIED PHYSICS LETTERS	22,4 (7213)	69,7 (176)
IEEE TRANSACTIONS ON ELECTRON DEVICES	10,1 (1032)	28,6 (42)
JOURNAL OF APPLIED PHYSICS	12,7 (8213)	35,7 (50)
IEEE ELECTRON DEVICE LETTERS	12,6 (477)	25,8 (39)
Laser		
ELECTRONICS LETTERS	5,3 (4611)	14,7 (77)
APPLIED PHYSICS LETTERS	23,3 (7315)	43,7 (74)
IEEE PHOTONICS TECHNOLOGY LETTERS	9,0 (1558)	16,3 (95)
OPTICS LETTERS	18,3 (1853)	35,7 (51)
Biotechnology		
NUCLEIC ACIDS RESEARCH	32,0 (2379)	155,3 (147)
SCIENCE	65,1 (8165)	502,9 (107)
JOURNAL OF BIOLOGICAL CHEMISTRY	49,9 (14550)	89,3 (263)
NATURE	59,0 (9161)	475,9 (104)

Note: number of publications in parenthesis.

[20] Fuller details about other cohort of publications are reported in the website of the study. Yet, they broadly confirm the results discussed in the text.

Table 15 – Benchmarking analysis, Average number of scientific citations received by publications cited and not cited in patents (EPO, cohort 1996-98)

Subfields/Journals	Not cited	Cited
Transmission of digital information		
ENGINEERING, ELECTRICAL & ELECTRONIC	5.3	16.0
COMPUTER SCIENCE, HARDWARE ARCHITECTURE	4.6	17.0
COMPUTER SCIENCE, CYBERNETICS	0.4	2.5
COMPUTER SCIENCE, INFORMATION SYSTEMS	4.9	47.1
Speech analysis		
ENGINEERING, ELECTRICAL & ELECTRONIC	4.9	18.1
COMPUTER SCIENCE, SOFTWARE ENGINEERING	8.9	21.4
ACOUSTICS	8.4	7.7
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	14.1	34.3
Semiconductors		
ENGINEERING, ELECTRICAL & ELECTRONIC	6.1	17.3
PHYSICS, APPLIED	13.9	58.7
MATERIALS SCIENCE, MULTIDISCIPLINARY	9.9	26.9
PHYSICS, FLUIDS & PLASMAS	8.8	82.9
Laser		
ENGINEERING, ELECTRICAL & ELECTRONIC	6.7	17.3
PHYSICS, APPLIED	14.0	58.7
OPTICS	7.9	25.2
PHYSICS, FLUIDS & PLASMAS	9.1	114.7
Biotechnology		
BIOCHEMISTRY & MOLECULAR BIOLOGY	28.8	121.8
CHEMISTRY, ANALYTICAL	17.9	38.1
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	21.8	76.9
GENETICS & HEREDITY	26.6	122.7

Broadly speaking, our results provide evidence of a strong positive correlation between citations coming from patents and citations coming from scientific publications. Both analysis at the level of specific journals and at the level of subject fields show that scientific publications cited in patents belonging to our five technology subfields receive, on average, a far larger number of citations from other publications than articles, which are not cited in patents.

On the one hand, this result validates in some way the methodology followed in this study. Publications cited by patents are not only cited in the realm of technology, but they are also heavily cited by other scientific publications. This means that we are not analysing a random and unchecked sample of publications, but we are probably focusing upon the most important publications in each field. On the other hand, we also think that one should not derive too broad conclusions on the correlation between scientific and industrial "value" of new ideas. There are in fact some caveats that must be taken into account when interpreting these results. First, the number of citations received by articles is an imperfect proxy of the quality of a paper. Scientists often cite other colleagues' work for a variety of reasons, which are not necessarily related to the quality of their publications. Second, one must also consider that the patent examination process itself has probably an impact on the observed correlation: while searching for existing prior art, patent examiners often rely upon a delimited set of scientific publications to build their lists of non-patent references. Since these publications are likely to be the most visible and easy-to-find contributions, the observed correlation could be probably partly spurious.

In particular, results for biotechnology should be treated with caution: among the top 4 journals in terms of citations received by patents, two are multidisciplinary ones (*Nature* and *Science*). Since these journals span across a variety of scientific domains, it is likely that the articles they host do not meet our coherence requirements. Different scientific disciplines have different citation patterns, so that we are probably comparing biotech-related publications cited in patents with an excessively vast and heterogeneous typology of articles. Nonetheless, our results are quite clear-cut: among articles published in *Nature* and *Science*, being cited by biotech patents increases the number of citations from scientific literature by a factor of 10, which is far the largest figure among the whole set of journals examined. The remaining two journals, being more focused on biotechnologies, provide a more reliable test of the correlation between patent citations and scientific citations: once again, articles published in *Nucleic Acid Research* and *Journal of Biological Chemistry* are significantly more cited by other articles if they are also cited by at least one patent. Similar results are obtained if one looks at subject fields, rather than journals. Actually, the spread between cited and non-cited publications is larger in this field than in any other field examined here, which provides further evidence of the blurring boundaries between science and technology in this area.

Despite the limitations mentioned above, the analysis still provides an interesting insight: there are signs of convergence in the evolutionary logics adopted by scientists and inventors in the selection of new knowledge. This effect is stronger in technological areas of intense S&T interaction, where new frontier research often provides "ready-made" inputs for technological deployment: biotechnology, lasers and semiconductors.

d) Phase 4: Identification of authors-inventors and cleaning affiliations

The last phase of the study involved the identification of authors-inventors and the task of cleaning and processing authors' affiliations and addresses.

As far as the former task is concerned, for each technology subfield we compared the names of inventors with the names of authors of cited scientific publications and matched them in order to identify those individuals that have produced both patented inventions and (cited) scientific publications. In this respect, a major problem we had to deal with was related to the fact that the patent datasets report name and surname of each inventor, whereas the SCI dataset reports only the first letter of name and the surname of each author. As a consequence, the risk in performing a simple matching by surname and first letter of name is that different individuals are identified as the same person, thereby leading to an overestimation of the number of authors- inventors. To solve this problem, we carried out a desktop research, which involved a large amount of manual checking and the use of several sources of information. The primary source was the affiliation of the author as reported in the cited publication and the affiliation of the inventor as reported in the patent document. In addition to this, we also used further sources of information, including Internet, university and company websites etc. In performing this task, we adopted a conservative approach, by matching two individuals (i.e. authors and inventors) only in those cases in which we were reasonably confident that they corresponded to the same person.

It is important to point out that for the EPO, we matched the names of *all* inventors in each technology subfield with the names of *all* authors of cited scientific publications, independently on the number of citations received by these papers. Although, this work exceeded what was contained in our original project proposal, we thought that it was nonetheless necessary to have a complete and reliable picture of the network linkages among patent inventors and paper authors.

For the USPTO, on the other hand, we matched the names of *all* inventors in each technology subfield with the names of authors responsible for highly cited publications. The reason for limiting the matching to authors of highly cited papers was mainly due to the fact that the number of papers (and authors) cited in USPTO is too large to allow a complete matching. For this reason, in what follows we will provide a few basic statistics for the network linkages involving USPTO patents, whereas a full and sophisticated network analysis will be restricted to the case of EPO patents.

In addition to matching the names of authors-inventors, we also cleaned and processed information concerning the affiliation and the address of both subjects, by focusing upon the set of highly cited patents and highly cited publications for the USPTO and the EPO. Affiliations have been cleaned, standardised and classified into five different types: universities (U), companies (C), public research organizations (PRO), government agencies (G) and other research organizations (SP). Once again, this task involved quite a large amount of desktop research and the use of different sources of information.

Finally, we parsed, cleaned and processed information on the geographical address of patent inventors and paper authors. Using the address reported either in the patent document or in the (cited) scientific publication, we parsed the corresponding record and extracted information on the city. We then implemented a PHP/Javascript program, which exploited the Google Maps online service to extract GPS coordinates for every possible affiliation of authors and inventors. This information was used to test the impact of spatial proximity on the probability of a citation tie between patents and publications (see below, section 3.2).

e) The final dataset of citing patents/inventors and cited publications/authors

The output of the four phases described above was a complex relational database that contains information on citing patents/inventors and cited publications/authors. More specifically, for each of the 5 technology subfields considered here, the dataset contains:

1) all patent applications in the period 1990-2003
2) all patents cited by the set of patents under 1)
3) all publications cited by the set of patents under 1)

For each patent application included in 1) (and for each patent cited by those patents), the dataset contains information on:

1) the patent applicant (name, type and ID code)
2) the applicant address (country and city)
3) the patent inventors (names and ID codes)
4) the inventors' address (country, city and GPS coordinates)
5) main and supplementary IPC codes
6) priority, publication and grant dates

For each publication cited by patents under 1), the dataset contains information on:

1) the title of article (and its ID code)
2) the article authors (names and ID codes)
3) the authors' affiliation (name, type and ID code)
4) the affiliation address (country, city and GPS coordinates)
5) the journal title
6) the publication year
7) the number of citations made and received

In addition to this, a further set of tables connect through ID codes the set of inventors with the set of authors, and thus provides information on those individuals that have taken part both in the community of technologists and in the community of inventors. This information will play a crucial role in the analysis of network linkages among scientists and inventors (see below, section 3.2).

Table 16 provides a few summary statistics on the coverage of the final dataset.

Table 16 – Final dataset – Summary statistics (1990-2003)

	EPO	USPTO
1) *# of patent applications 1990-203*		
Transmission of digital information	19580	15174
Speech analysis and image data processing	6931	8156
Semiconductors	23965	63016
Laser	3791	6951
Biotechnology	10565	13574
2) *# of patents cited by 1)*		
Transmission of digital information	5461	44698
Speech analysis and image data processing	1732	25322
Semiconductors	10539	115916
Laser	1373	16923
Biotechnology	4266	30638
3) *# of patents cited by 1) within a 5 years time window*		
Transmission of digital information	2969	21195
Speech analysis and image data processing	1094	11331
Semiconductors	6541	110762
Laser	866	11866
Biotechnology	2658	9702
4) *# of publications cited by 1)*		
Transmission of digital information	2409	1828
Speech analysis and image data processing	1172	1175
Semiconductors	5059	12642
Laser	2698	4548
Biotechnology	10448	44461
5) *# of inventors of 1)*		
Transmission of digital information	10248	35473
Speech analysis and image data processing	3758	18374
Semiconductors	18798	155372
Laser	10119	17051
Biotechnology	41273	38083
6) *# of authors of 4)*		
Transmission of digital information	4687	3031
Speech analysis and image data processing	2524	1858
Semiconductors	13224	22827
Laser	5929	7909
Biotechnology	40105	99866

3. RESULTS

This section is devoted to discussing the main results derived from the analysis of the citing-cited dataset described above[21]. The section is divided into two main parts. The first part is devoted to a statistical analysis of the dataset, with a specific focus upon knowledge flows from science to technology, as measured by citations from patents to scientific publications. The second part is devoted to a sophisticated network analysis of the ties linking inventors of patented inventions and authors of cited scientific publications

3.1 Knowledge flows from science to technology

The analysis starts from the examination of the share of cited and highly cited publications held by different areas. This information is reported in Tables 17 and 18, respectively for the EPO and the USPTO. Each table is divided into three panels. The top panel reports the share of *all* citations to publications cited in patents[22] for each technology subfield held by organisations located in four broad areas: European Union[23], United States, Japan and Rest of the World. In calculating citations, for each publication we considered only citations received from patents in a time window of 5 years since their publication date.[24] The reason for calculating the share of all citations in this way is to allow a comparison with the share of highly cited publications, given the fact that this type of articles has been defined according to this criterion (see above). The bottom panel reports the share of citations to publications cited in patents, but restricting the calculation only to highly cited publications. Finally, the bottom panel simply reports the ratio between the latter and the former shares. A ratio greater than 1 just means that a certain area holds a share of *highly cited* publications which is higher than its share of citations to *all* publications.

---

[21] Christian Catalini and Lorenzo Novella have provided invaluable research assistance in the elaboration of results reported in this section.

[22] It is quite important to remark that in order to locate publications to the four geographical areas, we have used the affiliation country of authors as reported in each publication. Moreover, we have adopted a *whole counting* method in order to count citations. For example, if publication X, co-authored by a European author and by a US author, has been cited by patent Y, we counted one citation from Y to X in Europe and one citation from Y to X in the US.

[23] The European Union includes the 25 Member States.

[24] Since citing patents go from 1990 to 2003, this also means that we excluded from the computation articles whose publication date was before to 1985 and articles whose publication date was after 1998. It is also important to remark that the tables report the share of *citations* and not of *cited* articles. The same article may in fact be cited more than once, leading to more than one citation.

Table 17 – Share of cited and highly cited publications by area (EPO, 1990-2003)

	Transmission	Speech analysis	Semiconductors	Laser	Biotechnology
			1) All cited publications		
EU25	26.9	32.1	19.6	23.9	29.8
Japan	12.4	11.1	24.7	21.3	6.3
Rest of world	14.7	17.1	9.6	9.4	10.6
United States	45.9	39.7	46.1	45.5	53.4
			2) Highly cited publications		
EU25	28.3	55.7	10.1	11.4	24.9
Japan	10.2	7.7	36.6	22.1	2.7
Rest of world	9.4	10.3	3.8	5.2	8.8
United States	52.1	26.4	49.6	61.3	63.6
			Ratio (2/1)		
EU25	1.1	1.7	0.5	0.5	0.8
Japan	0.8	0.7	1.5	1.0	0.4
Rest of world	0.6	0.6	0.4	0.5	0.8
United States	1.1	0.7	1.1	1.3	1.2

Table 18 – Share of cited and highly cited publications by area (USPTO, 1990-2003)

	Transmission	Speech analysis	Semiconductors	Laser	Biotechnology
			1) All cited publications		
EU25	15.8	19.9	12.7	20.7	22.3
Japan	8.1	7.2	18.0	18.0	4.8
Rest of world	16.0	11.7	8.6	7.6	8.8
United States	60.1	61.2	60.7	53.6	64.2
			2) Highly cited publications		
EU25	11.0	18.7	9.7	14.7	19.7
Japan	2.9	3.3	19.2	22.9	3.6
Rest of world	9.1	9.8	6.6	6.7	7.8
United States	76.9	68.3	64.5	55.7	68.9
			Ratio (2/1)		
EU25	0.7	0.9	0.8	0.7	0.9
Japan	0.4	0.5	1.1	1.3	0.8
Rest of world	0.6	0.8	0.8	0.9	0.9
United States	1.3	1.1	1.1	1.0	1.1

An inspection of the tables reveals a number of interesting results.

If we look at EPO data (Table 17), out of 5 technology fields, Europe shows a relative strength only in two sectors (transmission of digital information and speech analysis), whereas in semiconductors, lasers and biotechnology its share of highly cited publications is systematically lower than its overall share of cited publications. The European shares of cited and highly cited publications at the USPTO are lower than the corresponding shares at the EPO (Table 18). In addition to this, we also observe that its share of highly cited publications at the USPTO is lower than its share of all cited publications for all technology fields considered here.

As far as the other areas are concerned, the US leadership is quite evident, especially in the fields of biotechnology, lasers and TDI. With reference to EPO data, the share of citations to *highly cited* publications is, respectively, 64%, 50% and 52%, compared to a share of citations to *all* cited publications of, respectively, 53%, 45% and 46%. Not surprisingly, the US share of citations is higher, both for all cited and for highly cited publications, if one looks at USPTO data. However, the sectoral patterns of relative strength seem to be quite consistent across the two patent offices.

With few exceptions, the share of citations accounted for by Japanese authors is lower than the share of US and European authors. Of course, language barriers and the under-representation of Japanese authors in the ISI-SCI dataset may partly account for this result. Yet, Japan shows a consistent pattern of relative strength across the two patent offices in the fields of lasers and semiconductors.

Broadly speaking, the empirical evidence seems to show that European science is relatively under-represented in publications that provide key contributions to technological developments. A key issue in this respect is to what extent the fact that Europe does not feature prominently among highly cited publications is due to the underlying quality of its scientific production or, conversely, it has to be ascribed to weak transfer mechanisms from science to technology. Even though sharp conclusions cannot be derived on the basis of our data, we believe they help to shed some light on this crucial question. To this purpose, we have divided the sample of all cited publications in four subgroups according to the total number of citations received in patents. The first three subgroups contain publications that have received, respectively, one, two and three citations.

Table 19 – Average number of citations in scientific literature by number of citations in patents (EPO, 1990-2003)

Subfield	# of citations in patents	Average number of citations in scientific literature	
		EU25	United States
Transmission of digital information	1	10.8	20.0
	2	25.1	26.4
	3	18.7	27.8
	>3	23.2	53.3
Speech analysis	1	18.1	32.3
	2	15.9	47.3
	3	22.6	13.1
	>3	48.2	17.0
Semiconductors	1	38.6	45.9
	2	82.7	88.9
	3	121.0	58.9
	>3	488.3	108.5
Laser	1	24.0	30.6
	2	25.6	45.1
	3	47.4	46.9
	>3	28.8	86.1
Biotechnology	1	108.0	149.1
	2	191.4	219.9
	3	234.4	385.3
	>3	389.9	574.7

The fourth subgroup contains publications that have received four or more citations. These are publications that we have defined as highly cited. Each group of publications has been further divided according to the nationality of authors' affiliations. In particular, we have focused the attention on publications authored by European and US scientists. Finally, for each subgroup of publications we have calculated the average number of citations received in the scientific literature. The results of this tabulation are reported in Tables 19 and 20, respectively, for the EPO and the USPTO. Two main things are worth noting. In the first place, we observe the existence of a rather strong correlation between the number of citations received in patents and the number of citations received in the scientific literature.

Table 20 – Average number of citations in scientific literature by number of citations in patents (USPTO, 1990-2003)

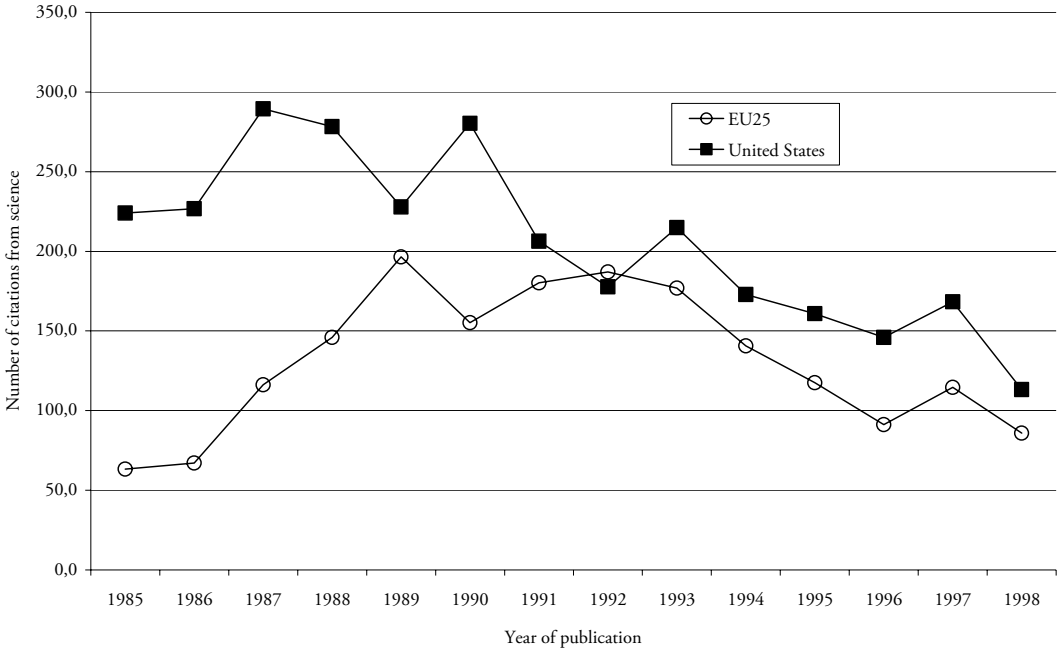
Subfield	# of citations in patents	Average number of citations in scientific literature	
		EU25	United States
Transmission of digital information	1	17.2	25.5
	2	31.7	59.7
	3	29.1	57.7
	>3	21.4	57.8
Speech analysis	1	20.7	33.6
	2	12.4	38.2
	3	22.0	176.3
	>3	51.0	47.0
Semiconductors	1	33.4	45.1
	2	40.0	65.0
	3	38.4	79.0
	>3	154.1	109.1
Laser	1	25.6	39.5
	2	27.7	40.8
	3	59.5	64.6
	>3	65.1	78.3
Biotechnology	1	116.4	141.0
	2	171.4	187.2
	3	235.8	322.2
	>3	272.6	496.2

This is especially true for highly cited publications, and in the fields of semiconductors and biotechnology. This result is quite consistent with the findings emerged in our benchmarking analysis and seems to indicate that 'high quality' scientific output finds its way in a large number of technological developments.

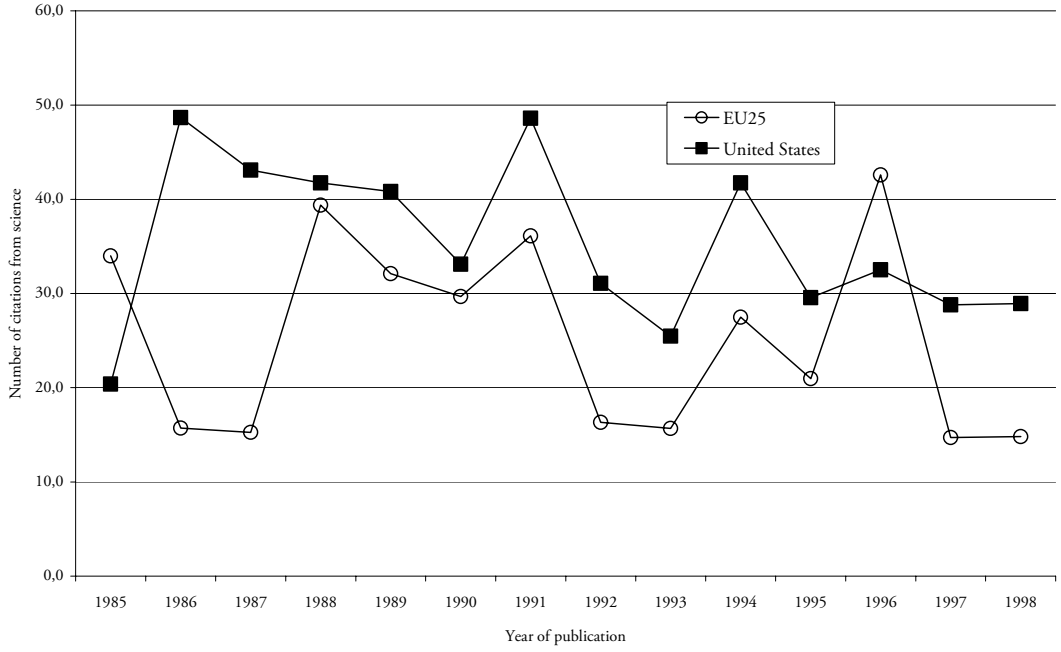
The second main point worth noting is that, for each subgroup of publications, articles produced by European authors receive a lower average number of scientific citations than articles produced by US authors, with the only exception of semiconductors and speech analysis. The differences in the average citation rates are particularly striking for highly cited publications and in the fields of lasers and biotechnology.

Figure 8 – Average number of citations from scientific literature of publications cited in patents by year of publication, Biotechnology

EPO



USPTO



Given the key strategic importance of this latter field, we have repeated the same tabulation for this specific subfield and we have computed the average number of scientific citations for each cohort of publications cited in patents, according to their publication year. The rationale for doing this is that the quality of cited publications, as measured by the number of scientific citations received, might change over time and across the two geographical areas[25]. The tabulation has been carried out considering all cited publications and the results are reported in Figure 8, separately for the EPO and the USPTO.

Looking first at EPO data, we observe a remarkable negative gap in the scientific citation rate of European publications as compared to US ones until the beginning of the 1990s. The gap starts narrowing after 1990 and over the last decade. Yet, we observe that European articles cited in patents and published in 1996 still receive on average 100 citations in the scientific literature as compared to about 150 for US articles.

The evidence emerging from USPTO data is less clear cut, although the line relative to European publications remains below the line relative to the US for most years, with the exception of 1985 and 1996.

As argued above, we believe that it would be quite difficult to draw strong conclusions on the basis of this empirical evidence. On the one hand, although the fields considered here are quite representative of broader technological domains, a larger and more systematic effort of data collection should be undertaken before generalising our findings. On the other hand, a key issue that should be examined more in depth is for what reasons European publications that receive a large number of scientific citations do not reach a level of application and diffusion into technological developments comparable to those of US publications.

A further and related set of issues that we have considered concerns the contribution of different types of organisations to the production of highly cited publications. As discussed in the methodological section, we have cleaned and processed information on authors' affiliation as reported in publications and we have classified them into five categories: universities (U), companies (C), public research organizations (PRO), government agencies (G) and other research organizations (SP).

[25] Of course, the problem of comparing the average citation rate across different cohorts of publications is that older cohorts will exhibit higher citation rates than younger cohorts simply because they had more time to be cited. Our focus here is not on comparing citation rates across cohorts, but in comparing citation rates across geographical areas by cohort.

Table 21 – Percentage contribution of different types of organisations to the production of highly cited publications by geographical area (EPO, 1990-2003)

	C	G	PRO	SP	U
			Transmission of digital information		
EU25	38.6	2.2	8.8	11.8	38.6
Japan	86.1			5.0	8.9
United States	65.9	2.4		3.5	28.2
			Speech analysis		
EU25	39.4		6.3	4.9	49.3
Japan	76.2				23.8
United States	48.8		4.7		46.5
			Semiconductors		
EU25	44.9	16.3	10.2	8.2	20.4
Japan	71.8		8.6	2.3	17.2
United States	58.3			1.7	40.0
			Laser		
EU25	22.9		11.4		65.7
Japan	93.3	6.7			
United States	64.7	6.0	2.4	2.4	24.6
			Biotechnology		
EU25	16.2	5.7	22.4	7.8	47.9
Japan	20.8				79.2
United States	28.0	3.3	11.1	4.3	53.2

Legend: companies (C), government agencies (G), public research organizations (PRO), other research organizations (SP), and universities (U).

Tables 21 and 22 report, respectively for the USPTO and the EPO, the share of citations to highly-cited publications within each geographical area (excluding the residual area represented by all other countries than EU, US and Japan) accounted for by different types of organisations. Looking at data across *subfields* reveals the existence of some differences in the relative importance of different types of institutions. In particular, the role played by universities is generally greater in biotechnology than in any other of the five subfields considered here.

Table 22 – Percentage contribution of different types of organisations to the production of highly cited publications by geographical area (USPTO, 1990-2003)

	C	G	PRO	SP	U
			Transmission of digital information		
EU25	64.2			9.4	26.4
Japan	100.0				
United States	75.3		1.3	2.6	20.8
			Speech analysis		
EU25	26.1		17.4		56.5
Japan	100.0				
United States	52.4	4.8	4.8		38.1
			Semiconductors		
EU25	28.5	16.7	12.8	8.9	33.1
Japan	85.3		2.2	3.1	9.5
United States	51.6	1.5	3.8	0.3	42.9
			Laser		
EU25	27.2	2.5	9.9		60.5
Japan	85.3			8.3	6.3
United States	59.1	2.1	3.4	3.5	31.9
			Biotechnology		
EU25	8.0	6.7	34.3	5.8	45.2
Japan	8.2	4.7	27.6	10.9	48.6
United States	20.5	2.0	13.8	7.7	55.9

Legend: companies (C), government agencies (G), public research organizations (PRO), other research organizations (SP), and universities (U).

However, the most interesting aspect emerging from such tables relates to the differences across *areas* in the relative role of different organisations. In this respect, the most striking result is perhaps that European companies contribute to a significantly lower extent than their US counterparts in the production of publications highly cited in patents. Thus, for example, US companies account for 28% and 21% of all highly cited publications produced by US organisations in the field of biotechnology, respectively at the EPO and at the USPTO. The corresponding figures for European companies are, respectively, 16% and 8%. Quite interestingly, the share of European universities is also lower than the corresponding share of US universities; yet, this is in some way balanced by a larger share of European public research organisations as compared to US

PROs. These results therefore indicate the existence of major structural differences in the organisation of the system of scientific research across the two areas, which would probably deserve further investigation.

The lower contribution of European companies to the production of highly cited publications is not limited to the field of biotechnology. The evidence reported in Tables 21 and 22 shows quite clearly that the share of highly cited publications accounted for by private companies is systematically lower in Europe than in the US and Japan for all the technology fields examined here. On the other hand, the contribution of the public system of scientific research, i.e. universities and public research organisations, is generally comparable to, and often larger than the contribution of the corresponding system in the US. We believe that this result is rather consistent with the findings that emerge from the analysis of the network linkages among scientists and inventors (see below). There, we show that the key mechanism that channel the transmission of scientific research into industrial innovation is represented by the network of collaborative (i.e. co-authorship) relations among scientific researchers and industrial technologists. In that respect, a crucial role in bridging the two communities is played by a specific category of individuals, i.e. authors-inventors. To the extent that this specific type of individuals is relatively absent in Europe, this represents, in our view, a major obstacle to the successful diffusion of knowledge from the realm of science to that of technology.

To corroborate our findings, we have also calculated for each technology subfield the share of all highly cited publications accounted for by different types of organisations and geographical areas. Results are reported in Tables 23 and 24, respectively for the EPO and the USPTO. The first point to note relates to the US predominance in all the technology fields. If we look at EPO data, the combined share of all highly cited publications accounted for by US companies and universities is about 55% in lasers, 52% in biotechnology, 50% in TDI, and 49% in semiconductors. With reference to the USPTO, the same shares are, respectively, 51%, 53%, 77% and 61%.

As far as Europe is concerned, we note that the public system of scientific research (i.e. universities and public research organisations) accounts for about 16% of all highly cited publications in biotechnology, both at the EPO and at the USPTO. This is the third largest share, after US companies and the US public system of research. On the other hand, European companies account for only 4% of all highly cited publications at the EPO as compared to 18% of US companies.

Table 23 – Percentage contribution of different types of organisations by geographical location to the production of highly cited publications (EPO, 1990-2003)

	Transmission	Speech analysis	Semiconductors	Laser	Biotechnology
U (United States)	15.1	14.7	19.7	15.1	34.5
C (United States)	35.3	15.4	28.8	39.9	17.8
U (EU25)	10.2	25.6	2.1	8.5	11.3
PRO (United States)		1.5		1.5	7.3
PRO (EU25)	2.3	3.3	1.1	1.5	5.3
U (Rest of world)	6.4	7.3	2.9		3.9
C (EU25)	10.2	20.5	4.6	3.0	3.8
SP (United States)	1.9		0.8	1.5	2.8
PRO (Rest of world)	0.6	1.5	0.8	3.3	2.5
G (United States)	1.3			3.7	2.2
U (Japan)	1.0	1.8	6.3		2.1
SP (EU25)	3.1	2.6	0.8		1.9
G (EU25)	0.6		1.7		1.4
C (Rest of world)	0.6				1.3
SP (Rest of world)	0.6				1.2
C (Japan)	10.1	5.9	26.3	20.7	0.6
G (Rest of world)					0.1
G (Japan)				1.5	
PRO (Japan)			3.2		
SP (Japan)	0.6		0.8		
Total	100.0	100.0	100.0	100.0	100.0

Legend: companies (C), government agencies (G), public research organizations (PRO), other research organizations (SP), and universities (U). The geographical origin of each organizational type is reported among brackets.

The performance of European companies looks more satisfactory only in two fields: TDI and speech analysis. Particularly, in the former subfield, if we combine the share of cited publications of universities, companies, government agencies and other research organisations, Europe accounts for about one fourth of all cited publications.

Finally, it is worth pointing out the role of Japanese companies in two specific subfields: semiconductors and lasers. The share of all highly cited publications held by this type of organisations is 26% and 17% in semiconductors, respectively at the EPO and at the USPTO, and 21% and 20% in lasers.

Table 24 – Percentage contribution of different types of organisations by geographical location to the production of highly cited publications (USPTO, 1990-2003)

	Transmission	Speech analysis	Semiconductors	Laser	Biotechnology
U (United States)	16.6	26.0	27.5	18.0	38.5
C (United States)	60.1	35.8	33.0	33.3	14.1
PRO (United States)	1.0	3.3	2.4	1.9	9.5
U (EU25)	2.9	10.6	3.4	8.9	9.1
PRO (EU25)		3.3	1.3	1.5	6.9
SP (United States)	2.1		0.2	2.0	5.3
U (Rest of world)	2.1	6.5	3.2	0.9	3.6
PRO (Rest of world)	3.3	3.3	2.5	2.5	2.3
U (Japan)			1.8	1.5	1.8
C (EU25)	7.1	4.9	2.9	4.0	1.6
G (United States)		3.3	1.0	1.2	1.4
G (EU25)			1.7	0.4	1.4
SP (EU25)	1.0		0.9		1.2
PRO (Japan)			0.4		1.0
SP (Rest of world)				1.0	0.7
C (Rest of world)	0.8		0.4	1.5	0.4
SP (Japan)			0.6	1.9	0.4
C (Japan)	2.9	3.3	16.5	19.5	0.3
G (Rest of world)			0.3		0.2
G (Japan)					0.2
Total	100.0	100.0	100.0	100.0	100.0

Legend: companies (C), government agencies (G), public research organizations (PRO), other research organizations (SP), and universities (U). The geographical origin of each organizational type is reported among brackets.

A further important issue that our data allow to examine regards to what extent patents originating in a certain area (e.g. Europe) cite scientific publications generated in other areas (e.g. US). In order to investigate this issue, we have tabulated, for each of the five technology fields and for the each of the four geographical areas, what fraction of all citations made by patents of organisations located in a certain area are directed to publications produced by organisations located either in the same area or in other areas. This is illustrated in Table 25, which provides a breakdown of knowledge flows, as captured by citations from patents to scientific publications, by geographical origin of citing patents and geographical origin of cited publications. In the calculation, we have included patent citations to *all* scientific publications by field, i.e. both cited and highly cited publications.

Table 25 – Flows of knowledge by origin of citing patents and origin of cited publications, EPO (1990-2003, percentage values)

		Transmission of digital information				
		Cited publications				
		EU	JP	Other	US	Total
Citing patents	EU	31.9	11.5	14.8	41.8	100
	JP	21.3	17.3	16.1	45.3	100
	Other	23.6	10.9	15.9	49.6	100
	US	20.2	8.6	15.3	55.9	100
		Speech analysis				
		Cited publications				
		EU	JP	Other	US	Total
Citing patents	EU	36.0	8.7	15.2	40.1	100
	JP	28.0	14.5	16.5	41.0	100
	Other	30.4	7.6	20.7	41.3	100
	US	27.4	10.9	17.9	43.9	100
		Semiconductors				
		Cited publications				
		EU	JP	Other	US	Total
Citing patents	EU	34.5	16.7	9.3	39.5	100
	JP	15.3	34.9	8.9	40.9	100
	Other	22.6	15.7	24.7	37.0	100
	US	14.2	18.6	10.3	56.9	100
		Laser				
		Cited publications				
		EU	JP	Other	US	Total
Citing patents	EU	35.1	15.0	8.5	41.4	100
	JP	17.4	39.7	8.3	34.6	100
	Other	28.1	11.3	26.0	34.6	100
	US	16.9	15.4	9.4	58.4	100
		Biotechnology				
		Cited publications				
		EU	JP	Other	US	Total
Citing patents	EU	43.7	5.5	9.8	41.0	100
	JP	21.9	29.3	9.9	38.9	100
	Other	24.6	4.0	27.2	44.2	100
	US	24.0	5.1	8.8	62.0	100

The most important point to note in this table is that the propensity of European patents to cite US scientific publications is relatively larger than the propensity of US patents to cite European publications. In other words, we observe an asymmetry in knowledge flows between EU and the US, with a larger amount of knowledge flowing from the US to Europe than vice versa. Thus, for example, of all citations made by European patents in biotechnology 41% are directed to US publications, and 44%

to European publications. Likewise, of all citations made by US patents in biotechnology 24% are directed to European, and 60% to US publications. The propensity of US inventors to rely upon the domestic science base seems therefore to be significantly greater than the propensity of European inventors to exploit their domestic science base[26]. Similar patterns may be found also for the other four technology fields examined here. In particular, it is worth noting that, beside the US, Japan represents also an important source of scientific knowledge for European patents in the fields of lasers and semiconductors.

Broadly speaking, to the extent that patent citations to scientific literature may be used as a measure of the industrial importance of scientific knowledge, one may be tempted to conclude that the European research system has a major weakness in the ability to translate its knowledge inputs into technologically relevant outputs.

The previous analysis has been based upon citations from patents to scientific literature. It is interesting to compare the findings reported above with the analysis of patent citations to prior art patents.

Table 26 – Share of cited and highly cited patents by area (EPO, 1990-2003)

	Transmission	Speech analysis	Semiconductors	Laser	Biotechnology
			All cited patents		
EU25	40.3	30.8	23.8	31.9	40.6
Japan	21.2	35.6	40.5	35.3	7.4
Rest of world	7.6	3.9	3.1	2.8	8.1
United States	30.9	29.7	32.6	30.0	43.9
			Highly cited patents		
EU25	40.5	32.0	22.5	32.5	35.6
Japan	18.6	35.1	42.0	31.8	7.6
Rest of world	8.3	3.3	3.3	2.9	8.6
United States	32.5	29.7	32.2	32.8	48.1

[26] It is worth noting that our results differ quite significantly from those reported in Verbeek et al. (2003). Using EPO data, they find that in the broader field of biotechnology, the share of European publications in the citations made by European patents is about 59%, whereas the share of European publications in the citations made by US patents is around 39%. Although the difference with our results may be due to a different definition of biotechnology in terms of IPC codes included, the gap seems to be too large to be imputed only to that fact.

Table 26 reports the share of cited and highly cited patents held by different areas[27]. The top panel reports the share of *all* citations to patents[28] for each technology subfield held by organisations located in four broad areas: European Union, United States, Japan and Rest of the World. In calculating citations, for each patent we considered only citations received in a time window of 5 years since their application date.[29] The reason for calculating the share of all citations in this way is to allow a comparison with the share of highly cited patents, given the fact that this type of patents has been defined according to this criterion (see above). The bottom panel reports the share of citations to highly cited patents (i.e. patents that have received six citations or more in a time window of 5 years since the application date).

Compared to our findings on cited and highly cited publications, the data do not show any significant difference between the shares of cited and highly cited patents for any of the four areas geographical areas. The only exception is represented by biotechnology, in which Europe holds about 41% of citations to all cited patents and 36% of citations to highly cited patents, whereas the US account for 44% of citations to all cited patents and 48% of citations to highly cited ones.

More interesting results emerge if we compare the share of cited (and highly cited) publications and the share of cited (and highly cited) patents (see above Table 17). In this respect, it is worth noting that the position of Europe looks more favourable if one looks at cited patents than at cited publications. Thus, for example, in the field of biotechnology, Europe accounts for 30% and 25% of, respectively, cited and highly cited publications, as compared to 41% and 36% of cited and highly cited patents. Likewise, in the field of TDI, the share of all cited and highly cited publications for Europe is of, respectively, 27% and 28%, as compared to 40%, both for cited and highly cited patents. A similar pattern may be found in the case of semiconductors, albeit the difference is less striking, and in the field of lasers. An exception is represented instead by the subfield of speech analysis. This is the only domain in which the share of cited and highly cited publications for Europe is higher than its share of cited and highly cited patents.

---

[27] For the analysis of patent citations, we have reported only data from the EPO.

[28] In order to locate patents to the four geographical areas, we have used the address of the applicant as reported in the patent document.

[29] Since citing patents go from 1990 to 2003, this also means that we excluded from the computation patents whose application date was before to 1985 and patents whose publication date was after 1998. It is also important to remark that the table reports the share of *citations* and not of *cited* patents. The same patent may in fact be cited more than once, leading to more than one citation.

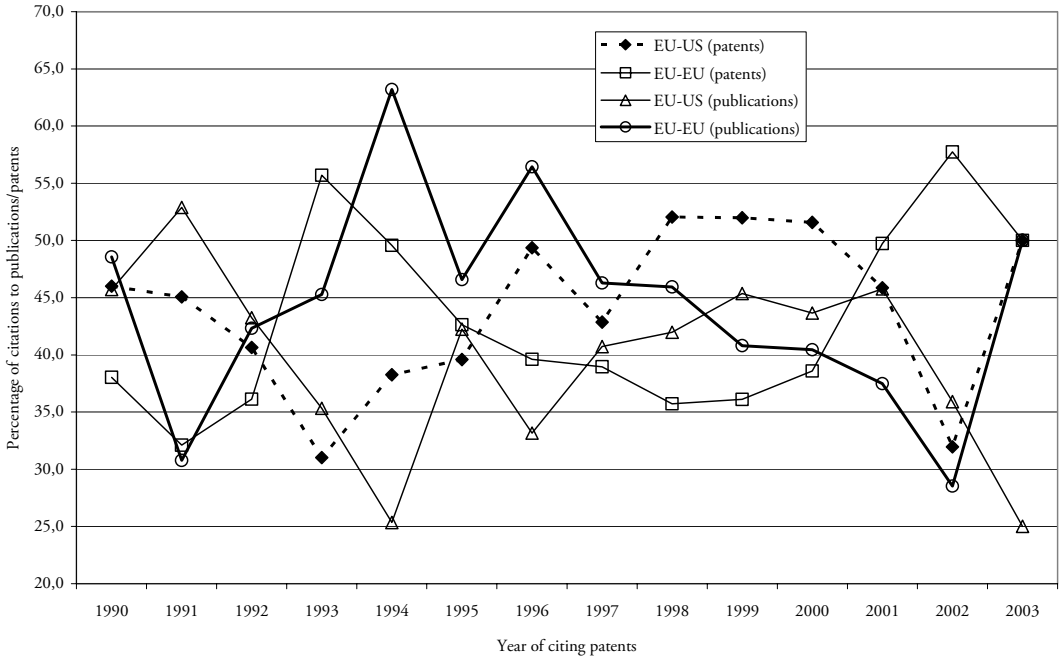
Table 27 – Flows of knowledge by origin of citing patents and origin of cited patents, EPO (1990-2003, percentage values)

		Transmission of digital information				
		Cited patents				
		EU	JP	Other	US	Total
Citing patents	EU	41.3	18.4	4.7	35.6	100
	JP	29.5	39.5	3.0	28.0	100
	Other	26.7	14.4	8.7	50.2	100
	US	19.8	15.1	4.4	60.6	100
		Speech analysis				
		Cited patents				
		EU	JP	Other	US	Total
Citing patents	EU	36.6	23.1	3.6	36.7	100
	JP	16.2	48.9	3.8	31.1	100
	Other	15.3	40.0	16.5	28.2	100
	US	18.8	23.2	4.3	53.7	100
		Semiconductors				
		Cited patents				
		EU	JP	Other	US	Total
Citing patents	EU	40.0	27.8	3.4	28.8	100
	JP	11.6	63.3	1.7	23.4	100
	Other	20.1	32.4	16.1	31.4	100
	US	16.6	30.2	1.8	51.4	100
		Laser				
		Cited patents				
		EU	JP	Other	US	Total
Citing patents	EU	44.5	23.8	4.0	27.8	100
	JP	13.6	66.0	1.7	18.7	100
	Other	22.2	31.1	11.1	35.6	100
	US	18.4	28.2	5.9	47.5	100
		Biotechnology				
		Cited patents				
		EU	JP	Other	US	Total
Citing patents	EU	41.3	6.0	7.9	44.9	100
	JP	17.8	38.8	7.7	35.8	100
	Other	20.2	4.3	29.8	45.7	100
	US	17.6	4.9	8.4	69.1	100

As done for publications, we have also tabulated the flows knowledge by geographical origin of citing patents and geographical origin of cited patents. Results are reported in Table 27. Also in this case, we note that in general the propensity of European patents to rely upon technological developments in the US is higher than the corresponding propensity of US patents to rely upon European patents, for all the five technology fields considered.

Some interesting results emerge if we compare the propensity to cite domestically produced patents to the propensity to cite domestically produced scientific publications (see above Table 25). As far as Europe is concerned, we note that inventors tend to rely more upon domestically produced technology than domestically produced science. The only exception is represented by biotechnology, where European inventors cite relatively less European patents than they do cite European scientific publications; in particular, they tend to cite US scientific publications more than they cite US patents. Concerning the US, it is interesting to note that in two important fields, such as lasers and semiconductors, the share of citations to domestically produced patents is lower than the share of citations to domestically produced scientific publications, meaning that US inventors in these fields tend to rely more upon internationally produced scientific knowledge than they do with respect to technology.

Figure 9 – Flows of knowledge between EU and the US by year of citing patents, Biotechnology



Finally, we have compared the time trend in the patterns of knowledge flows between EU and the US for the field of biotechnology, given the strategic importance of this field and its peculiar characteristics. In particular, we have calculated the fraction of all citations of European patents to other European (US) patents, and compared it to the fraction of all citations of European patents to European (US) scientific publications, by

application year of the citing patent. This is reported in Figure 9. The most evident trend emerging from the figure is the declining tendency of European patents to rely upon domestic science (solid, thick line). The share of all citations to European publications declines quite sharply from a peak of around 64% for patents applied in 1994 to less than 30% for patents applied in 2002.

3.2 Analysis of the network linkages among scientists and inventors

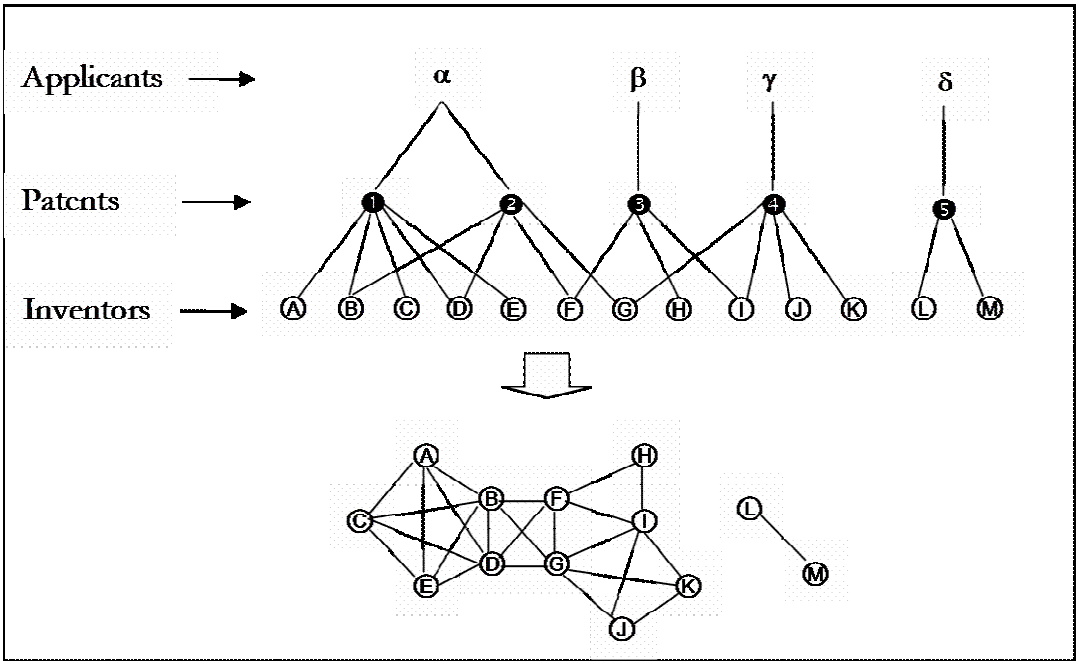
The rest of this report is devoted to a sophisticated network analysis of the linkages among scientists and inventors in the five technology fields considered. We start by providing a detailed discussion of the methodology that has been adopted to examine network linkages. In particular, we show how co-invention and co-authorship data can be exploited to map the complex web of social ties among inventors and authors, and measure a number of 'structural properties' of such a web, typical of social network analysis. Then, we proceed to characterise the main structural properties of the network that links inventors and authors. Finally, we propose an econometric model that estimates the role of social linkages on the probability of a citation tie between patents and publications.

a) Methodology

In order to analyse the network linkages among scientists and inventors, we have exploited the information on co-authorship and co-invention contained in our dataset. In particular, we assume that two inventors (authors) who have been collaborating in the production of a patented invention (scientific publication) are connected by a network tie, which means that they are linked by some kind of knowledge exchange and share a common knowledge base.

Figure 10 reports a hypothetical example, which illustrates the main idea. Let us suppose we have five patent documents (1 to 5), coming from four different applicants (α,β,γ,δ). Applicant α is responsible of two applications (1,2), while applicants β, γ and δ of one each. Patents have been produced by 13 distinct inventors (A to M). In the language of graph theory, the top part of the figure reports the *affiliation* network of patents, applicants and inventors. An affiliation network is a network in which actors (inventors) are joined together by common membership to groups of some kind (patents). Affiliation networks can be represented as a graph consisting of two (or more) distinct kinds of vertices, one representing the actors (e.g. inventors) and the other the groups (e.g. patents).

Figure 10 – Tripartite graph of applicants, patents and inventors



Let us focus on patent 1. This patent document has been produced by five inventors, i.e. A, B, C, D, and E. This fact is represented in the top part of the figure by the lines that connect each node corresponding to the five inventors to the node that correspond to patent 1. We can then reasonably assume that, due to the collaboration in a common research project, the five inventors are 'linked' to each other by some kind of knowledge relation. The existence of such a linkage can be graphically represented by drawing an undirected arrow between each pair of inventors, as in the bottom part of figure 10. Repeating the same exercise for each patent and each team of inventors, we end up with a map representing the network of linkages among all inventors. In the language of graph theory the bottom part of figure 10 represents the unipartite (or one-mode) graph of actors joined by undirected edges, i.e. two inventors who participated in the same patent, in our case, being connected by an edge.[30]

[30] Of course, the same pair of inventors might be linked by more than one edge, to the extent that they collaborated in the production of more than one patent. In this case, one could represent the graph as a valued network, i.e. attaching to the edges a value corresponding to the number of times a pair of inventors has been collaborating. In what follows, however, we will adopt the assumption that a line between two inventors is ei-

Using the graph just described, we can derive various measures of "connectedness" and "social distance" among inventors. Let us introduce a few basic concepts.

- CONNECTEDNESS. Inventors may belong to the same component or they may be located in disconnected components. A component of a graph is defined as a subset of the entire graph, such that all nodes included in the subset are connected through some path. More precisely, a component of a graph is a subset of nodes, for which one can find a path between all pairs of nodes within the subset, but no paths towards the nodes outside. In our specific context, a node must be interpreted as an individual inventor. In Figure 10, for example, inventors A to K belong to the same component, whereas inventors L and M belong to a different component.

- GEODESIC DISTANCE The geodesic distance is defined as the minimum number of steps (or, more formally, 'edges') that separate two distinct inventors in the network. In Figure 10, for example, inventors A and C have geodesic distance equal to 1, whereas inventors A and H have distance 3. This means that the linkage between them is mediated by two other actors (i.e. B and F). In other terms, even though inventor A does not know directly inventor H, she *knows who* (inventor B) knows who (inventor F) knows directly inventor H. The geodesic distance between a pair of inventors belonging to two distinct components is equal to infinity (there is no path connecting the two inventors).

- DEGREE CENTRALITY. Some inventors stand out for the number of links they exhibit: they have not just signed a high number of patents, but have also worked along with a large number of co-inventors (that is in large teams, on with many different teams). We expect these inventors to be chief researchers in large R&D departments, or senior academic researchers with a long tradition of consultancy to or joint research with industrial firms. For example, in Figure 10, inventor B has worked with no less than six co-inventors, signing two patents (1 and 2) both of them produced by relatively large teams (six and four people, respectively). In her absence, the overall connectedness of the component she belongs to would

ther present or absent. In other words, we will work with binary (i.e. not valued) networks. Please also note that the position of nodes, and the length of lines in the graph do not have any specific meaning.

be much lower, that is distances between inventors would be higher. Social network analysis refers to this property as high "degree centrality".

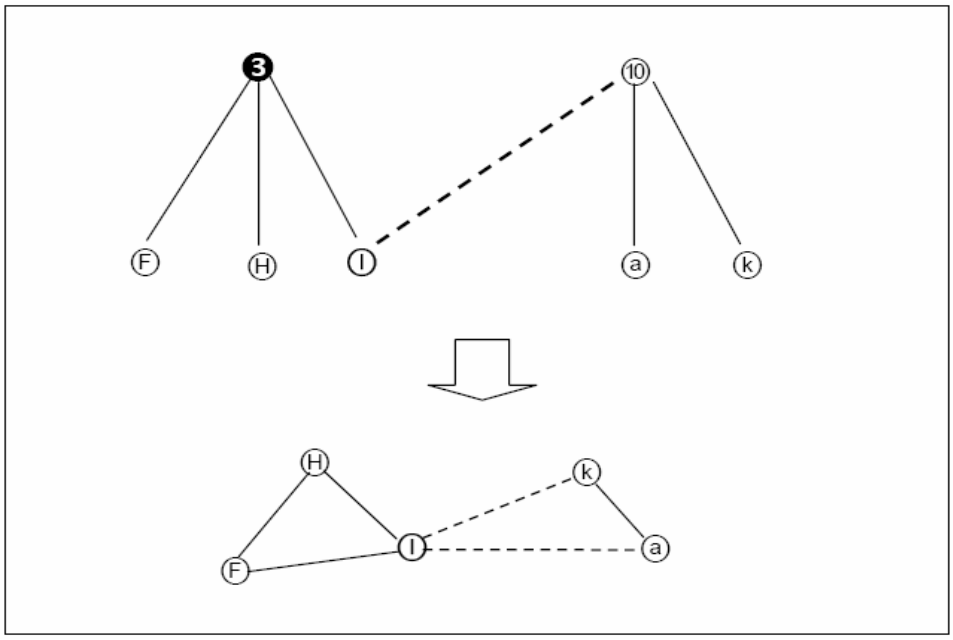
- BETWEENNESS CENTRALITY. Some inventors may have a particularly important role in connecting different components. They can be either by "mobile" inventors, that is industrial researchers moving across firms, or, once again, academic researchers whose ties with industry are not limited to just one company. For example, in Figure 10, inventor F worked for both company α and β, thus connecting the sub-component listing inventors from A to G with the sub-component listing inventors H–K. In her absence the component "A–K" would be split in two. Social network analysis refers to this property as high "betweenness centrality".

What we have said so far refers to inventors of patent documents. However, the same methodology can be applied to the network of scientists. With reference to Figure 10, one should replace "patents" with "articles" and "inventors" with "scientists". The resulting one-mode network will represent the co-authorship network of authors of scientific papers.

The methodology described above has been adopted by several authors in recent years to analyse the properties of the networks of inventors (Balconi et al, 2004; Breschi and Lissoni, 2006; Singh, 2005) as well as the properties of the networks of scientific authors (Newman, 2000, 2001). To date, however, there has not been any attempt to combine the two networks and analyse them jointly. To the best of our knowledge, this is the first large-scale attempt to carry out this type of analysis.

How the two networks (i.e. scientists and inventors) can be connected? A crucial role in this respect is played by a specific type of individuals that we label here as *authors-inventors*. These are individuals that participate in both communities, by producing patents and by publishing scientific papers. Figure 11 illustrates the idea with a hypothetical example. Let us take patent 3 in Figure 10. This patent has been produced by three inventors, F, H and I. Let us suppose that inventor I has also published a scientific paper, coded as paper *10*, with authors *a* and *k*. The bottom part of Figure 11 shows that inventors F and H are *indirectly* connected to authors *a* and *k*, through I. In other words, inventor I plays the crucial role of bridging the community of inventors and the community of scientists.

Figure 11 – The network of scientists and inventors



Having defined the basic methodology, we now turn to describing how our dataset has been used to build and analyse the network of scientists and inventors. Before doing that, however, it is important to point out that the analysis has been limited to EPO data. The reason is that, as discussed above, we have cleaned and processed data for *all* inventors and *all* authors of cited publications only for EPO data[31].

For each technology subfield, we have proceeded as follows.

In the first place, we have selected all patent applications and the related inventors, and we have built the corresponding network of inventors. The network has been built for each year t in a cumulative way starting from 1978 (i.e. first year for which we have EPO data) by adding each year new nodes (i.e. inventors) and new linkages (i.e. patents). To the extent that the importance of linkages among inventors (and scientists) decays over time, one could alternatively think to remove old patent applications (and publications) in order to construct the network of social linkages among inventors (and

[31] The analysis of the network of scientists and inventors has to include all actors involved in the network. Although the original project proposal was limiting the scope of the analysis to the authors of highly cited papers, we decided to pursue a larger scale work of cleaning data on all authors of cited papers. However, given the amount of work involved, we could do this only for EPO data.

scientists). Yet, in the absence of rules to establish the decay of social links, we simply assumed that once formed, social linkages last forever (at least for the time period considered here). Thus, for example, the network of inventors in year t=1995 includes all inventors and their co-invention linkages from 1978 to 1994.

A similar approach has been applied to the network of scientists. In building such a network, however, we have included only authors of publications *cited* in patents of the five technology subfields. Thus, for each year *t* (i.e. 1995), the network of scientists includes only authors of publications that have been cited by patents of a certain technology field and that have been published before time *t*. Finally, we have merged the two networks into a single network, which includes both co-invention and co-authorship relationships. In other words, for each technology subfield, the network of scientists and inventors include both authors of cited papers and inventors of patents.

The time period we considered for analysing network is from 1991 to 2000. Table 28 reports the total number of inventors and authors in the network for each technology subfield. We note that in all technology subfields, with the exception of biotechnology the number of authors of publications cited in patents is always lower than the number of inventors.[32]

Table 28 – Total number of authors and inventors in the network (1991-2000)

Technology fields	Inventors	Authors
1. Transmission of digital information	15435	3489
2. Speech analysis and image data processing	5967	2045
3. Semiconductors	34996	11450
4. Lasers	5259	4913
5. Biotechnology (measuring, testing, diagnostics)	16283	34510

b) Connectedness

We started our analysis by examining the overall degree of connectedness among inventors, and among authors and inventors. To this purpose, we have calculated the number of distinct components in each network and we have identified the largest one in terms of number connected nodes.

[32] For social network analysis, we used two major software programs. PROC IML SAS 9.1, and in particular the IML modules and libraries built by James Moody, and Pajek, a freeware software tool for network analysis and visualisation, built by Vladimir Batagelj and Andrej Blado.

Table 29 reports the total number of inventors as well as the fraction of all inventors included in the largest component, taking the network at year 2000 (i.e. the last year of our time series). Table 30 reports the same information but for the network of inventors *and* authors.

Table 29 – Network of inventors, Largest component (2000)

Technology fields	Size of largest component	All inventors	% of all inventors in the largest component
1. Transmission	422	15435	2.73
2. Speech analysis	50	5967	0.84
3. Semiconductors	3226	34996	9.22
4. Lasers	153	5259	2.91
5. Biotechnology	456	16283	2.80

Table 30 – Network of inventors and authors, Largest component (2000)

Technology fields	Size of largest component	All inventors and authors	% of all inventors & authors in the largest component
1. Transmission	1999	18924	10.56
2. Speech analysis	104	8012	1.30
3. Semiconductors	16152	46446	34.78
4. Lasers	5228	10172	51.40
5. Biotechnology	27148	50793	53.45

Looking first at Table 29, we observe that in all technology fields, the degree of connectedness among inventors is extremely limited. The only (partial) exception is represented by semiconductors, where 9.22% of all inventors are either directly or indirectly connected to each other. In all other fields, the network of inventors appears highly disconnected with many components of small size. This is not surprising, given the (relatively) low mobility rates of inventors across (patenting) organisations.

More surprising and interesting appears the result reported in Table 30. In three, important technology fields, such as semiconductors, lasers and biotechnology, the degree of connectedness is extremely high: 35%, 51% and 53%, respectively, of all authors and inventors are either directly or indirectly connected, via co-invention or co-authorship, to each other in a large connected component. We believe this is an extremely important result as it suggests that the two communities of researchers are much more socially connected that one would normally presume. Moreover, it also suggests that the

community of inventors itself is much more connected than the data from Table 29 would indicate. Although not directly connected to each other, inventors are indirectly connected through scientific authors and through authors-inventors, i.e. individuals that participate in teams of inventors *and* in teams of scientists.

This hypothesis is supported by data reported in Table 31. The table reports the share of all inventors that are included in the largest component of the network of authors and inventors.

Table 31 – Network of inventors and authors, Fraction of all inventors in the largest component

Technology fields	Inventors in the largest component	All inventors	% of all inventors & authors in the largest component
1. Transmission	1564	15435	10.1
2. Speech analysis	59	5967	1.0
3. Semiconductors	8392	34996	24.0
4. Lasers	2124	5259	40.4
5. Biotechnology	5255	16283	32.3

For example, with reference to biotechnology we note that 5255 inventors, which represent 32% of all inventors in this field, are connected to each other in the largest connected component of the network of inventors and authors. If we take only co-invention relationships (i.e. the network of inventors only, as reported in Table 29), the largest connected component includes just 456 inventors, which represent 2.8% of all inventors in this field. This means that looking only at co-invention relationships grossly underestimates the extent of connectivity among inventors and may lead to wrong conclusions.

The analysis above refers to the last year available in our time series. The following figures illustrate the evolution over time of the largest connected component for each technology subfield. In particular, they show the evolution of the largest component as a fraction of all inventors (i.e. network of inventors) and of all inventors and authors (i.e. network of inventors *and* authors). From an inspection of the figures, we note that in all the five subfields the share of inventors in the largest connected component of the network of inventors is always extremely limited.

Figure 12 – Evolution of the largest connected component for network of inventors, and for the network of inventors and authors

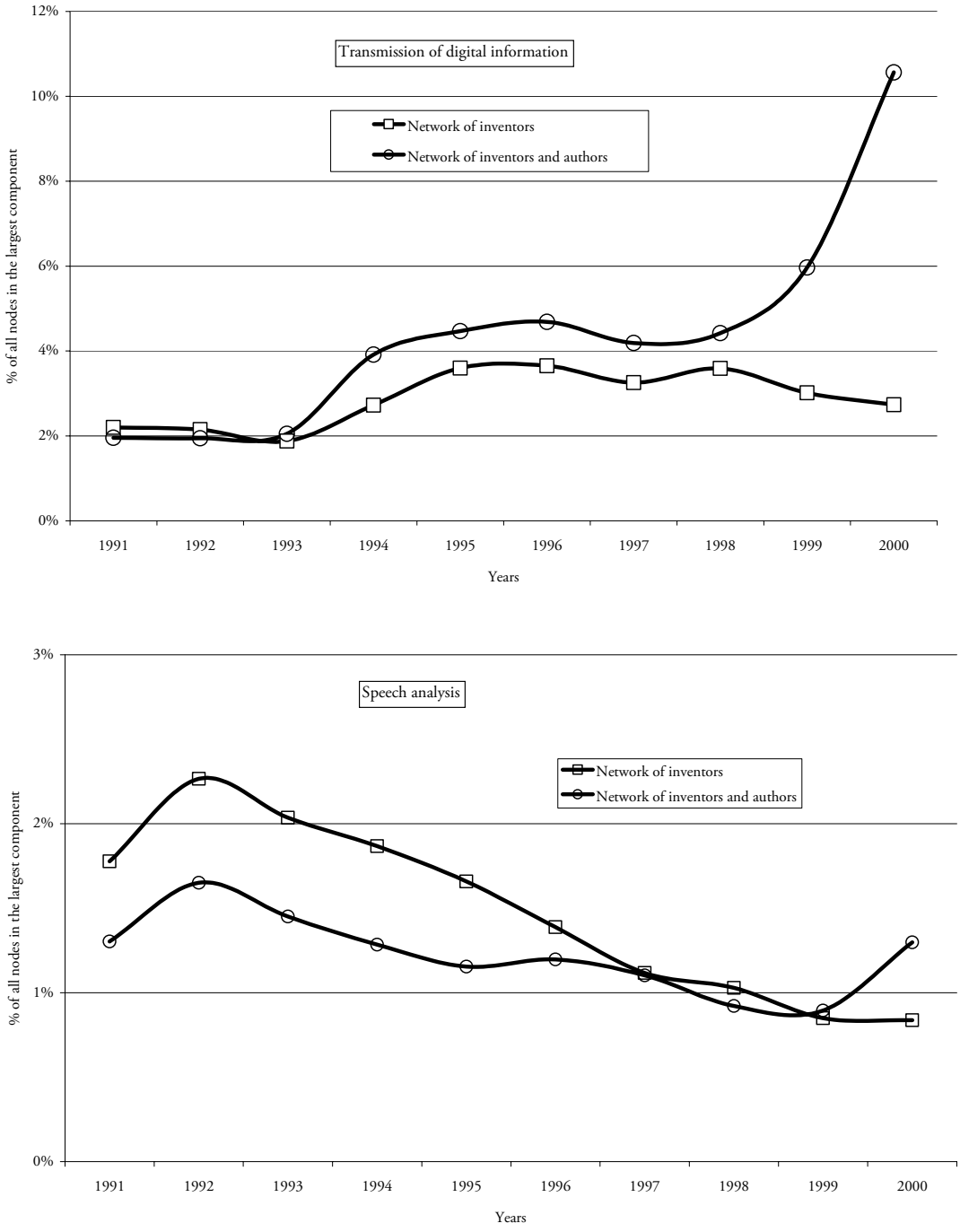
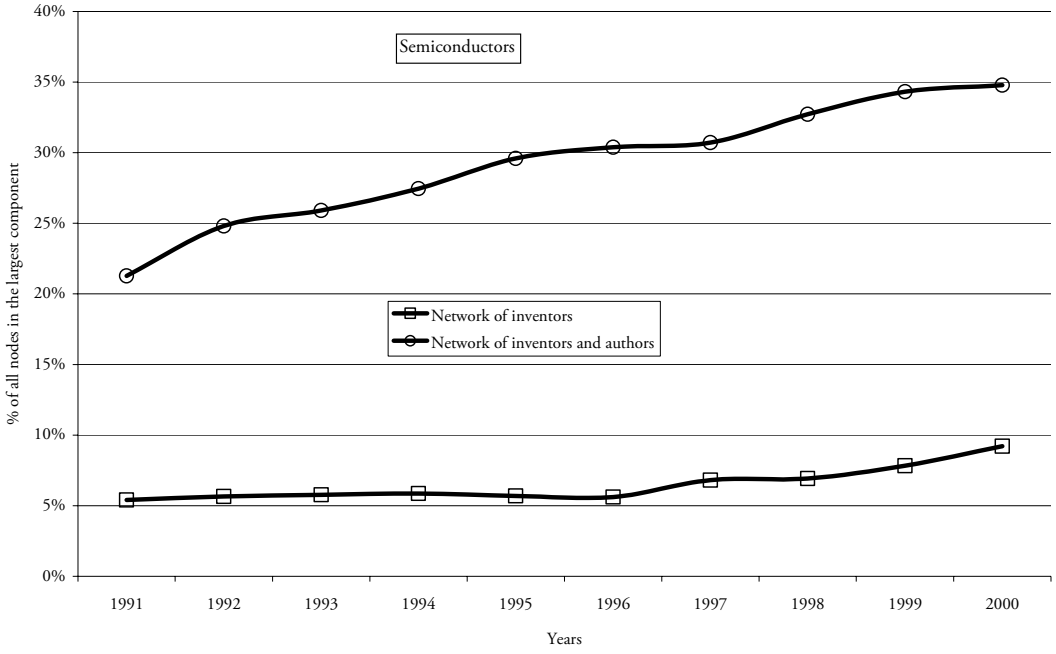


Figure 12 – cont.



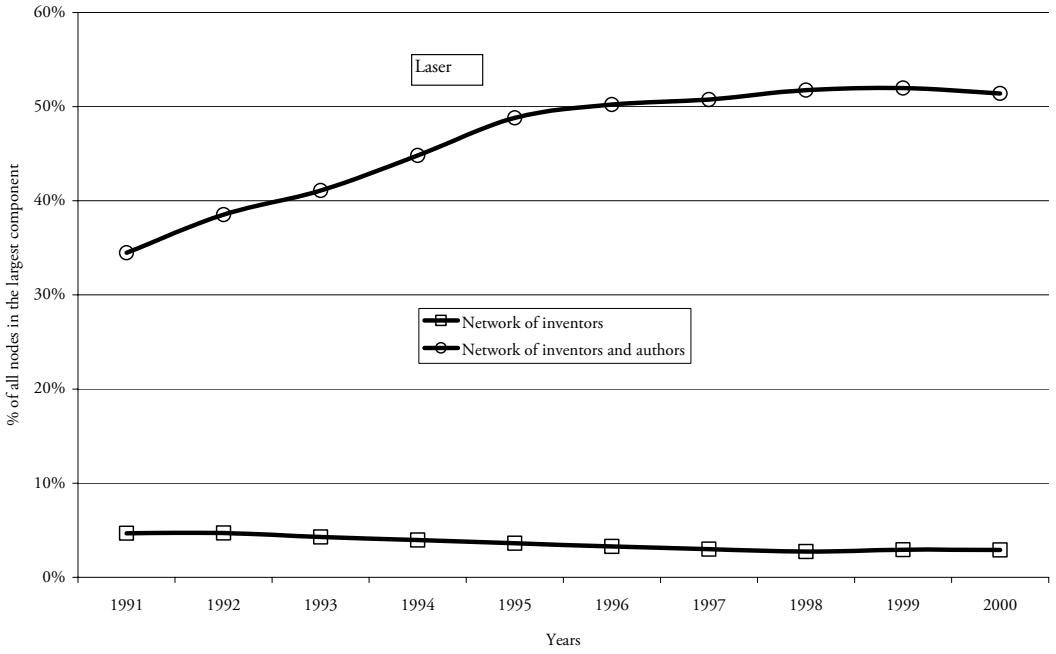
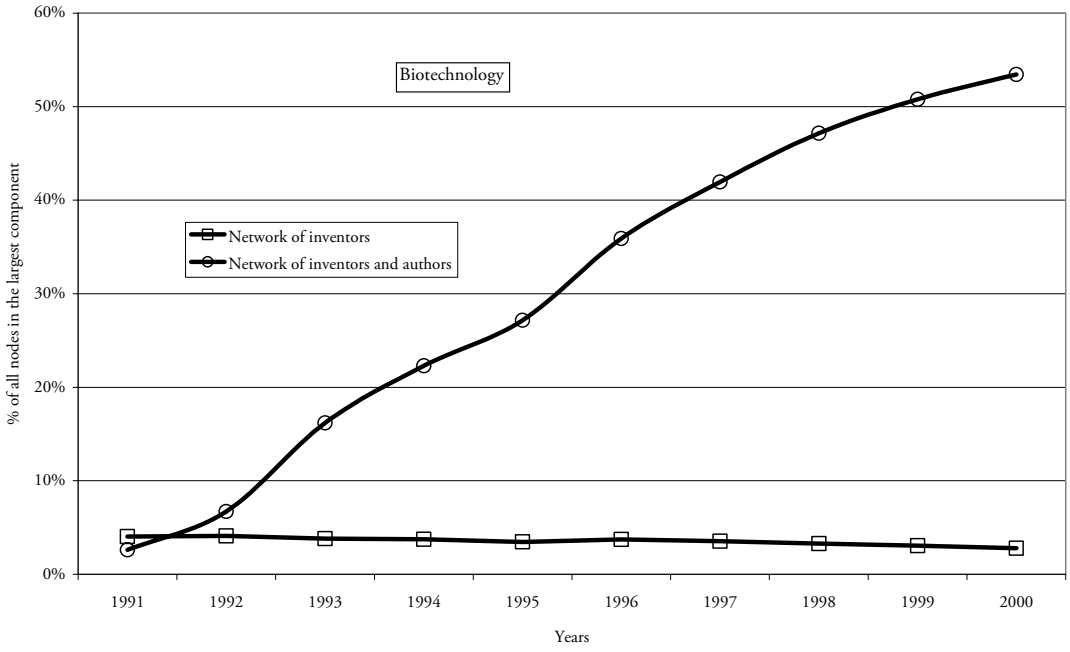


Figure 12 – cont.



On the other hand, if we look at the network of inventors *and* authors, the share of all nodes accounted for by the largest connected component is quite high in three out five technology subfields. Actually, in semiconductors and lasers, the size of the largest component tends to grow over time, but starting from relatively high levels. In 1991, the largest components account for about 22% and 35% of all authors and inventors, respectively in semiconductors and lasers. On the other hand, the size of the largest component rises steeply in biotechnology, but starting from relatively low levels. In 1991, the largest component accounts for only 3% of all authors and inventors, while this fraction goes up to almost 54% in 2000.

c) Is the network of scientists and inventors a "small world"?

The high degree of connectedness of the network of scientists and inventors in semiconductors, lasers and biotechnology is typical of graphs that have the property of being "small worlds". Broadly speaking, a small world network is a graph in which nodes are grouped around tightly linked local cliques (i.e. most nodes in each clique are neighbours of one another), but every node can be reached from every other in the network by a small number of steps. This type of structure is thought particularly

important both for the generation and the diffusion of knowledge. On one hand, the high degree of density and redundancy of linkages within local cliques ensures the formation of a common language and communication codes that enhances reciprocal trust and supports the sharing of complex and tacit knowledge among actors, thereby increasing the rate of diffusion of knowledge. On the other hand, the shortcuts linking local cliques to different and weakly connected parts of the network ensures a rapid diffusion and recombination of new ideas throughout the network and allow to keep a window open on new sources of knowledge, thereby mitigating the risk of lock-in that could arise in the context of densely connected cliques (Cowan and Jonard, 2003).

Formally, a small world graph is characterised by two main properties. First, it presents a high degree of clustering. Second, it shows a short average distance among pairs of nodes.

As far as the first property is concerned, this implies that a small-world network will have many subgraphs that are characterized by the presence of connections between almost any two nodes within them. An index that captures this idea is the so-called clustering coefficient, C, which can be formally defined as follows: for any node i one picks the k_i other nodes with which the node in question is linked. If these nodes are all connected to one another (i.e. they form a fully connected clique), there will be $k_i(k_i -1)/2$ links between them, but in reality there will be much fewer. If one denotes with K_i the actual number of links that connect the selected k_i nodes to each other, the clustering coefficient for node i is then $C_i = 2K_i/k_i(k_i -1)$. The clustering coefficient for the whole network is obtained by averaging C_i over all nodes in the system. The clustering coefficient C thus tells how much of a node's collaborators are, on average, willing to collaborate with each other.

As far as the second property is concerned, the average distance among nodes in the network is defined as follows. For any pair of nodes, i and j, in the network, the ability to communicate with each other depends on the length of the shortest path l_{ij} (i.e. the minimum number of edges), which links them. The average of over all pairs of nodes, denoted as $d=\langle l_{ij} \rangle$, is called the average separation (distance) of the network, characterising the network interconnectedness. In other words, the average distance measures the number of steps that have to be taken in order to connect two randomly se-

lected nodes. A low average distance therefore implies a (potentially) high speed of diffusion of information and knowledge throughout the network.[33]

For each technology subfield, we have calculated the clustering coefficient only for the largest connected component and for the year 2000. Results are reported in Table 32. They show that the coefficient C takes extremely high values, both for the network of inventors and for the network of inventors and authors. Thus, for example, with reference to biotechnology, results indicate that on average 70% of an inventor's co-inventors also collaborate each other (second column). Likewise, 77% of the co-inventors or co-authors of a given individual also collaborate with each other. Overall, the degree of clustering (or cliquishness) is very high in all fields, and much larger than the value one would observe in classical random graph.[34]

Table 32 – Clustering coefficient, Largest component (2000)

Technology fields	Network of inventors	Network of inventors and authors
1. Transmission	0.746 (0.348)	0.759 (0.315)
2. Speech analysis	0.752 (0.332)	0.858 (0.257)
3. Semiconductors	0.799 (0.296)	0.721 (0.365)
4. Lasers	0.718 (0.334)	0.598 (0.411)
5. Biotechnology	0.700 (0.368)	0.770 (0.327)

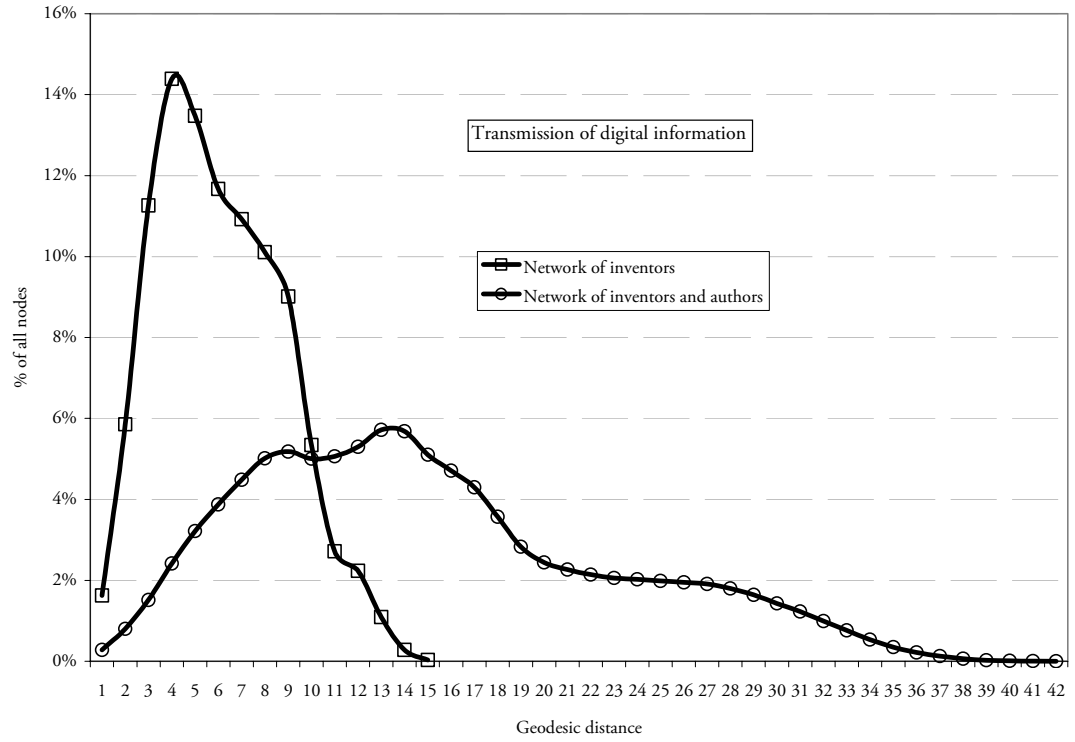
Note: standard errors among brackets

As far as the average distance among inventors is concerned, we calculated the distribution of geodesic distances among any pair of inventors, and any pair of inventors and authors, only for the largest component and for the year 2000. The distribution of geodesic distances is reported in Figure 13, separately for the network of inventors only, and for the network of inventors and authors. The average distance for the two networks is instead reported in Table 33.

[33] An alternative measure that is sometimes used to evaluate the degree of connectedness of a network is its *diameter*, which is defined as the maximum separation of pairs of nodes in the network, namely the greatest distance one will ever to have to go to connect two nodes together.

[34] In a classical random graph, the clustering coefficient is $C = z/N$, where z is the average degree of nodes and N is the total number of nodes. For brevity, we do not report here the values of the clustering coefficient in random graphs with the same parameters of our networks. Yet, they are an order of magnitude lower than the corresponding values in our networks for all technology fields.

Figure 13 – Distribution of geodesic distances among pairs of nodes, Largest component (2000)



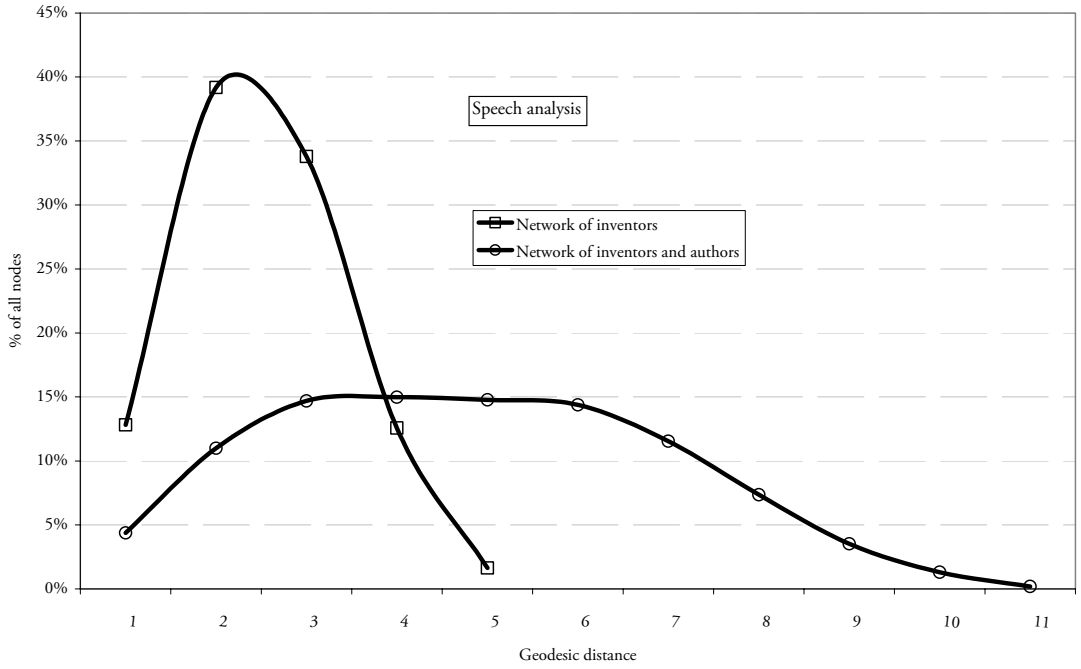
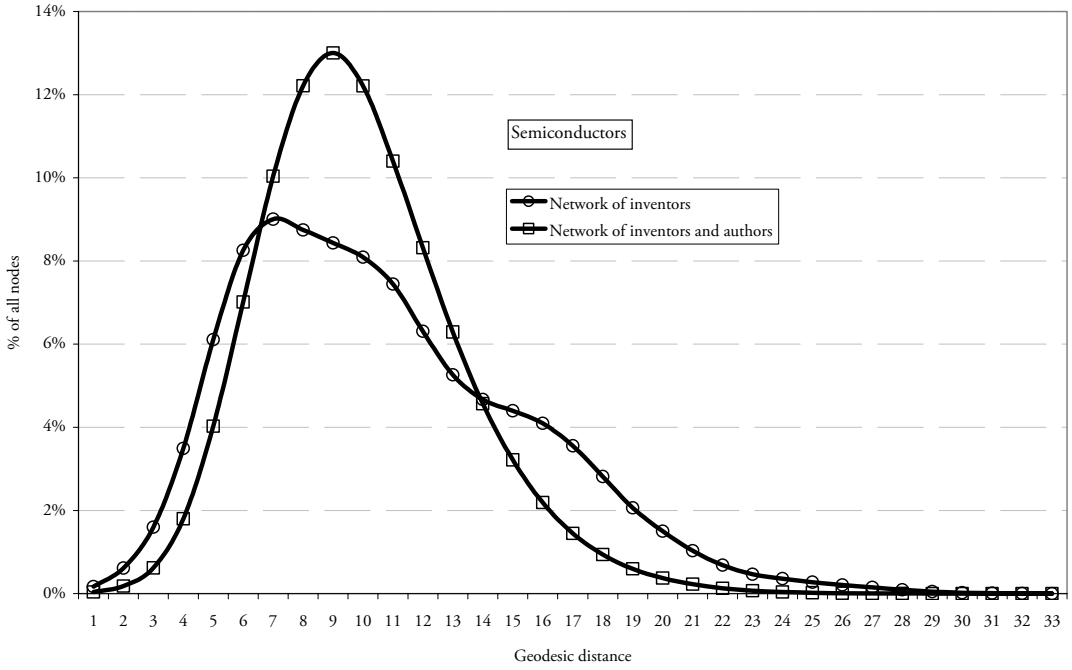


Figure 13 – cont.



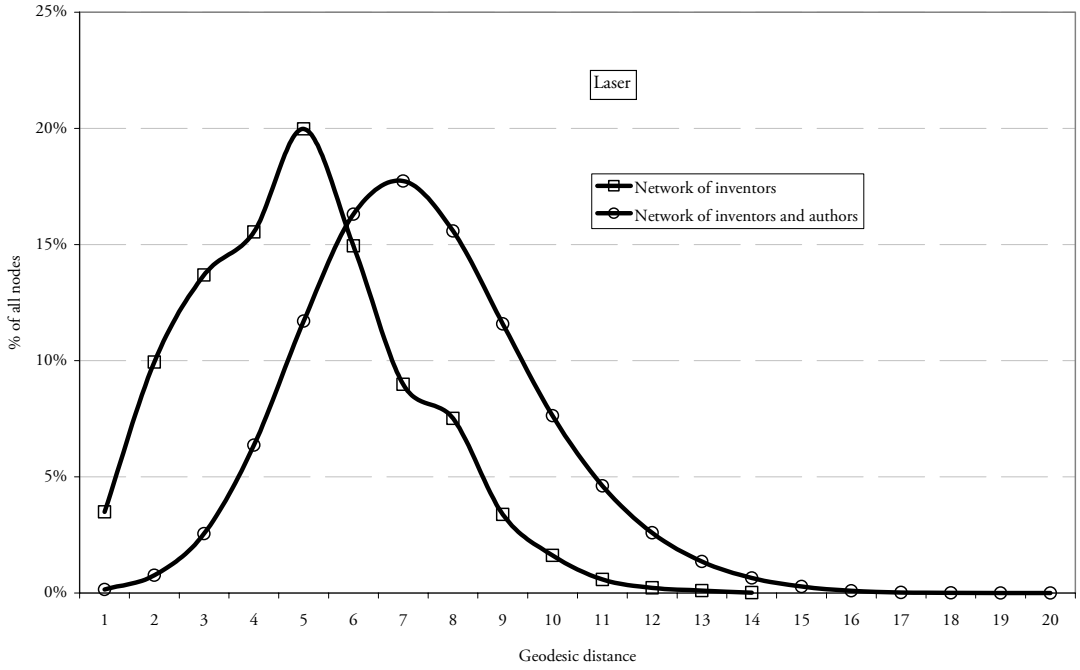


Figure 13 – cont.

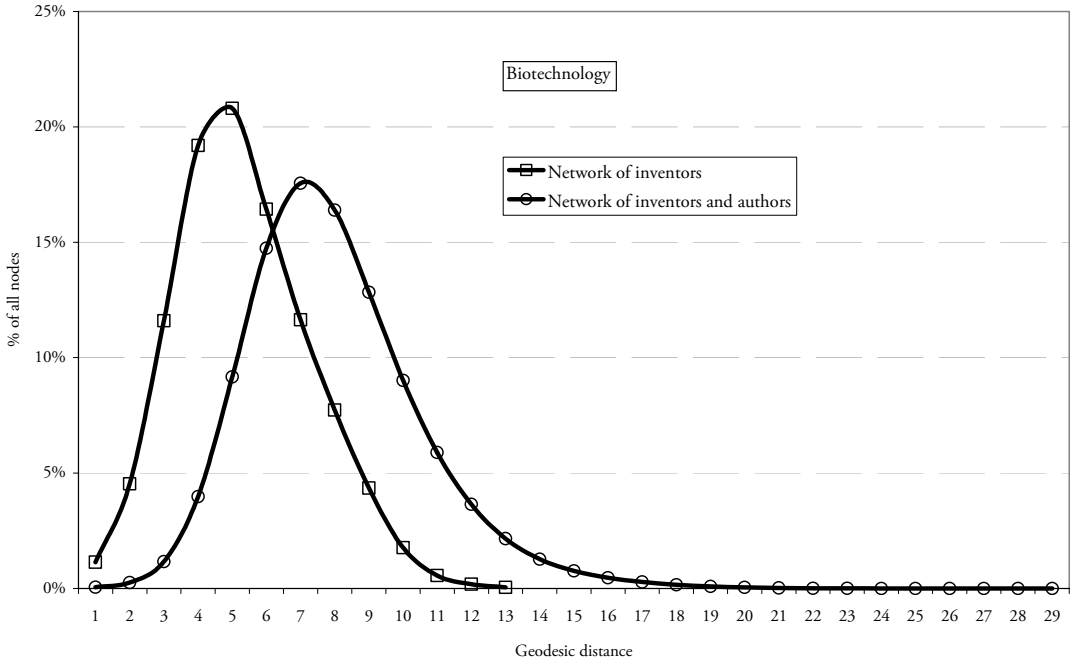


Table 33 – Average geodesic distance, Largest component (2000)

Technology fields	Network of inventors	Network of inventors and authors
1. Transmission	6.09	15.13
2. Speech analysis	2.51	4.86
3. Semiconductors	10.71	9.94
4. Lasers	4.96	7.35
5. Biotechnology	5.34	7.99

With the exception of semiconductors, the distribution of the geodesic distances of networks that include authors is moved to the right compared to the distribution of the network that include only inventors. Correspondingly, the average distance is slightly larger. On average, however, we observe that the average distance among inventors, and among inventors and authors, is relatively short and compatible with a small world structure, perhaps with the exception of transmission of digital information where the average distance in the network of authors-inventors is larger. Thus, for example, if we examine biotechnology, it takes 8 steps on average to reach a randomly chosen inventor (author) from any other inventor (author). This means that the network of authors and inventors work at least potentially as an effective means of knowledge transmission and diffusion.

d) The position of nodes in the network of authors and inventors

The results reported above suggest that inventors of patents and authors of (cited) scientific publications are highly inter-connected to each other, at least in three important technology fields, such as semiconductors, lasers and biotechnology. We argued above that a crucial role in ensuring a high degree of connectivity between the two communities is played by a specific type of individuals that we have labelled as authors-inventors. Researchers that do publish scientific articles and patent new inventions bridge academic and industrial worlds, by pouring new scientific knowledge into the inventors' domain. By connecting with authors-inventors, industrial researchers (i.e. inventors) can keep track of scientific advances relevant for their activities. As a consequence, we also expect that, by embodying stocks of tacit and valuable knowledge, authors-inventors are more likely to attract other inventors' collaborations. To state it in terms of social network analysis, we expect that authors-inventors are more *central* and *in-between* than 'simple' inventors.

To test this idea, we have computed two measures of node centrality, which are widely used to assess the extent to which nodes occupies a central position in the information flows that take place within a network. The first measure we computed is the *degree centrality*. This is simply defined as the number of edges that connect a given node to other nodes in the network. A node with a high degree centrality maintains contacts with many other nodes in the network. A central actor, according to this measure, occupies a structural position that acts as a source or conduit for larger volumes of informa-

tion exchange. In contrast, peripheral actors maintain few or no relations and are thus located at the margins of the network.

The second statistic used to investigate the position of authors-inventors is the so-called betweenness centrality. It measures how many times a node lies "between" two others, so that it must be activated to enable a knowledge exchange among them. Formally, for a graph G:=(V,E), with V vertices and E edges, the betweenness centrality for node n is defined as:

$$C_B(n) = \sum_{s \neq t \neq n \in V} \frac{\sigma_{st}(n)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths from s to t, $\sigma_{st}(n)$ the number of shortest paths from s to t that pass through node n. Nodes with high betweenness centrality are thus nodes that occur on many shortest paths between other vertices. According to our previous discussion, we expect that authors-inventors will lie more between individuals than simple inventors, because they attract a high number of collaborators, acting as "hubs" for large portions of the network.

Table 34 reports the average value of the degree centrality index, respectively, for inventors, authors and authors-inventors. The values are calculated for largest component of the network of inventors and authors in the year 2000.

Table 34 – Degree centrality, Network of inventors and authors, Largest component (2000)

Technology fields	Inventors	Authors	Authors-Inventors
1. Transmission	3.82 (2.92)	3.87 (3.03)	7.95 (5.61)
2. Speech analysis	3.60 (2.82)	3.52 (2.06)	6.58 (3.99)
3. Semiconductors	4.32 (3.26)	6.87 (4.91)	13.24 (10.23)
4. Lasers	3.95 (2.86)	6.96 (5.13)	14.86 (10.97)
5. Biotechnology	4.21 (2.78)	12.40 (24.77)	16.07 (21.88)

Note: standard deviation among brackets

Looking at the three fields where the size of the largest component as a fraction of all nodes is significant, i.e. semiconductors, lasers and biotechnology, we note that the average degree centrality of simple inventors is always lower than the average degree centrality of authors of cited scientific publica-

tions, which in turn is lower than the average degree centrality of authors-inventors. A simple t-statistics (not reported here for brevity) reveals that difference in the means across the three groups is statistically significant at the 99% level. We may therefore conclude that authors-inventors tend to collaborate on average with a significantly larger number of other inventors *and* authors, than do simple inventors and authors[35].

Comparing the values across fields reveals, quite interestingly, that the degree centrality of simple inventors and of authors-inventors is quite comparable, at least as far as semiconductors, lasers and biotechnology are concerned. On the contrary, the degree centrality of authors in biotechnology is significantly higher than degree centrality of authors in semiconductors and lasers. This result is likely to depend on the larger average number of authors per paper in the field of biotechnology, as compared to the fields of semiconductors and lasers.

Table 35 reports the average value of the betweenness centrality index for the network of inventors and authors calculated on the largest component in the year 2000. Table 36 reports instead a t-test statistic for the difference in the average betweenness centrality among inventors, authors and inventors-authors. Results show that in all fields examined, with the exception of speech analysis, the average betweenness centrality of authors-inventors is significantly higher than the betweenness centrality of simple inventors and authors. In turn, the betweenness centrality of authors is higher than the betweenness centrality of simple inventors. As reported in Table 35, the differences in the mean values of betweenness centrality across groups are statistically significant at the 99% level.

Overall, the results reported above suggest that authors-inventors play a crucial role in bridging the two communities of scientists and inventors. Their peculiar function of knowledge brokers in the network makes them more in-between than simple inventors and ensures a rapid diffusion of knowledge and ideas from one domain to the other.

[35] As far as authors are concerned, it must be pointed out again that the networks examined here only includes publications cited in patents, and not other publications that cited authors might have published with other co-authors. Therefore, the degree centrality of authors must be understood and interpreted with reference to the way in which the network has been built.

Table 35 – Average betweenness centrality (x100), Network of inventors and authors, Largest component (2000)

Technology fields	Inventors	Authors	Authors-Inventors
1. Transmission	0.22 (0.98)	0.85 (4.66)	2.8 (6.15)
2. Speech analysis	1.12 (4.03)	2.68 (7.59)	12.03 (16.57)
3. Semiconductors	0.03 (0.14)	0.05 (0.24)	0.26 (0.58)
4. Lasers	0.02 (0.09)	0.08 (0.33)	0.41 (0.84)
5. Biotechnology	0.01 (0.05)	0.02 (0.12)	0.11 (0.13)

Note: standard deviation among brackets

Table 36 – T-test on the difference in the average value of betweenness centrality

Technology fields	Inventors vs. Authors	Inventors vs. Authors-Inventors	Authors vs. Authors-Inventors
1. Transmission	-3.07*	-6.55*	-4.43*
2. Speech analysis	-1.20	-2.83	-2.36
3. Semiconductors	-6.45*	-13.34*	-12.05*
4. Lasers	-9.79*	-13.66*	-11.25*
5. Biotechnology	-10.85*	-14.54*	-12.82*

Note: * difference statistically significant at the 99% level.

Given the strategic importance played by authors-inventors, we computed the distribution of these individuals in the largest component by geographical area and compared it with the corresponding share of simple inventors in the largest component. Results are reported, respectively, in Tables 37 and 38, for three technology subfields, i.e. semiconductors, lasers and biotechnology. The first point to note is that the European and US individuals account for the largest share of authors-inventors and simple inventors, in semiconductors and biotechnology. Thus, for example, in biotechnology Europe and the US together account for about 94% of simple inventors and 93% of authors-inventors included in the largest connected component. On the contrary, other areas, especially Japan, accounts for almost half of all inventors and authors-inventors in lasers.

Second, we also note that the share of US authors-inventors is significantly larger than the US share of simple inventors in all three technology fields. To the extent that authors-inventors play as brokers of knowledge from the domain of science to that of technology, we believe this finding has very im-

portant implications for our understanding of the gap between Europe and the US in the effectiveness to translate the results of scientific research into commercially useful applications.

Table 37 – Share of authors-inventors in the largest component by geographical area (2000)

Technology fields	EU25	US	Others
Semiconductors	26.9 (311)	60.5 (698)	12.6 (145)
Lasers	20.5 (148)	35.5 (314)	44.0 (389)
Biotechnology	27.6 (515)	65.0 (1214)	7.4 (138)

Note: absolute values among brackets.

Table 38 – Share of simple inventors in the largest component by geographical area (2000)

Technology fields	EU25	US	Others
Semiconductors	22.5 (1625)	53.4 (3868)	24.1 (1745)
Lasers	25.0 (307)	29.5 (366)	45.7 (567)
Biotechnology	33.1 (1123)	60.5 (2049)	6.4 (216)

Note: absolute values among brackets.

Likewise, we observe that both in biotechnology and lasers, the share of Europe of all authors-inventors in the largest component is remarkably lower than its share of simple inventors. Once again, we believe this is an important result for our understanding of the mechanisms through which scientific output translates into technological developments. To the extent that authors-inventors play a crucial role in brokering knowledge across the two domains, proximity to such individuals, and more generally proximity among authors of scientific publications and inventors of patented inventions may be a fundamental factor affecting the effective diffusion of scientific knowledge.

Next section addresses this issue explicitly, by analysing the role of social proximity among inventors and authors in determining the probability of a citation tie between patents and scientific articles.

e) The role of social and geographical proximity on citation ties

The aim of this section is to analyse what factors may affect the probability of observing a citation tie between a patent and a scientific publication. More specifically, we aim to test two contrasting hypotheses.

On the one hand, a large body of recent empirical literature has argued that *spatial proximity* between senders and receivers of knowledge flows greatly enhances the effectiveness through which knowledge is transmitted. According to this spatial proximity argument, the tacit and contextual nature of (scientific) knowledge makes publication a rather inadequate vehicle to transfer it; scientific publications diffuse only the codified part of the knowledge they embody, while the tacit component can be transferred only through personal interactions with the authors that possess it. As the probability of meeting the owners of such tacit knowledge as well as the effectiveness of any inter-personal contact with them decrease with spatial distance, geographical proximity becomes crucial in the transfer of scientific knowledge from the domain of science to that of technology. Following this argument, we would expect that the probability that a patent cites a specific publication decreases with the geographical distance among inventors of the citing patent and authors of the scientific publication.

On the other hand, a few other authors have recently advanced a different view, according to which the crucial kind of proximity that matters in facilitating the transmission of knowledge is not the spatial one, but the *social* one. According to the *social proximity* argument, tacit knowledge can be, and very often is codified, by developing appropriate vocabularies and codebooks. What is tacit then is not the knowledge itself, but the messages that transport that knowledge, in the sense that only a small portion of the relevant codebook is usually referred to explicitly when transmitting knowledge. This implies two things. First, that the language (i.e. the vocabulary) used for exchanging knowledge is the language of a rather restricted and close community of experts (i.e. an epistemic community) and that the common codebook may be used as a powerful exclusionary device, even for local actors who live and work side by side with the community members. Second, it also implies that tacit messages may be sent over long distances by means of a variety of communications media (both written and orally), to the extent that the receiver possesses the necessary codebook to interpret the tacit knowledge contained in them. Even when dispersed in space, epistemic communities will share more jargon and trust among each other than with any other outsider within their present local communi-

ties. According to this view, therefore, (scientific) knowledge flows through social networks that link researchers and we would expect that the probability that a patent cites a specific publication decreases with the *social distance* among inventors of citing patents and authors of the scientific publication, irrespective of their geographical location. In other words, what matters in the effective transmission of knowledge is the epistemic (i.e. social) proximity among scientists and inventors and not their spatial proximity.

In order to test the hypotheses briefly discussed above, we have adopted a regression approach to estimate the probability of observing a citation tie between patent-paper pairs. In what follows, we discuss the methodological approach and the results of our estimates.

Sampling design

Indicating with P(K,m) the probability that patent *K* cites publication *m*, we aim to test whether such a probability depends on the degree of social and spatial proximity among authors and inventors behind the citing patent and the cited scientific publication, after controlling for other possible factors that may affect the probability of a citation.

In principle, one could approach the problem by estimating the factors that affect the probability of a citation tie, by looking at all possible pairs of potentially citeable publications and potentially citing patents, using a logistic regression to estimate the effects of covariates. Yet, this approach is practically not feasible, given that this would require to deal with very large data matrices[36]. For this reason, we adopted an endogenous stratification sampling strategy (or choice-based sampling procedure) (Breschi and Lissoni, 2004; Stolpe, 2002).

In particular, for each technology field *j*, we followed a 4-steps procedure:

1) We selected a cohort of cited publications, e.g. 1990, by publication year. Let m_{tj} be the m^{th} cited publication in cohort *t* in technology class *j*;

2) For each subsequent cohort of patents, i.e. patents with application year equal to $T=t+s$, e.g. 1995, we generated all potential pairs between them and the cited publications at year *t*. Let

[36] A simple random sampling of all potential pairs of patents and publications would not work either, given that actual citations are an extremely low fraction of all potential pairs.

K_{Tj} be K^{th} patent in cohort $T>t$ in technology j; the pair (m_{tj},K_{Tj}) identifies a potential citation from patent K to publication m;[37]

3) From the set of potential citations generated in this way, we selected all actual citations or "cases" (i.e. pairs of patents-publications that correspond to actual citations);

4) For each "case", i.e. actual citation, we selected two "control' pairs", i.e. two patent-paper pairs that do not correspond to any actual citation. The control pairs are therefore similar in all respects to the cases (i.e. patents belong the same technology field and have the same application year, and publications have been cited by patents in same technology field and have the same publication year), except for the fact that a citation tie exists for the cases, whereas it does not for the controls.

The four steps have been, of course, repeated several times, one for each cohort of cited publications and for each cohort of patents. The dependent variable in our model is therefore a binary variable, which takes value 1 for all cases of actual citations, and value 0 for the controls.

Given the choice-based sampling procedure followed, we did not use a simple logistic regression to estimate our model, but adopted a "weighted exogenous sampling maximum likelihood" (WESML) procedure. The idea behind this method is to weight each observation in the sample by the number of population elements it represents (i.e. the inverse of its sampling probability). In our case, we assigned a weight of 1 to all observations corresponding to actual citations, as all of them have been sampled. Observations corresponding to controls have been weighted by the inverse of the fraction of all patent-paper pairs, with that specific combination of application-publication year.

Explanatory variables

Given that our interest is in estimating the effect of social and spatial distance on the probability of a citation tie, we first discuss in which way we measure such variables.

[37] Let n_{tj} the number of publications in cohort t cited by patents in technology field j and N_{Tj} the number of patents in cohort T in technology field j. The number of possible pairs of patents-publications (i.e. potential citations) is therefore given by $n_{tj}*N_{Tj}$.

As far as the social proximity is concerned, this has been constructed as follows. Given a patent-paper pair (m_{ij},K_{Tj}) the "social distance" between them at time (T-1), i.e. the period just before the potential citation, is defined as the shortest among the geodesic paths connecting the inventors of the patent and the authors of the publication in the network of inventors and authors[38]. The social *distance variable* defined in this way has therefore the characteristics of a categorical variable and for estimation purposes it is convenient to transform it into a set of dummies. In particular, we defined seven of them, mutually exclusive and exhaustive of the social distance possibilities:

a) d_0*:* this variable takes value 1 whenever at least one individual in the team of patent inventors also appears in the team of paper authors (and 0 else). In the case of actual citations, this corresponds to what may labelled as a *personal self-citation,* i.e. an inventor citing her own scientific work. In social network terminology, we can say that in this case the geodesic distance between the patent and the publication is 0;

b) d_1*:* this variable takes value 1 whenever one or more inventors of patent have *previously* collaborated with at least one paper author (and 0 else). In other words, this dummy variable captures *prior direct collaborations* among inventors and authors in the production of either patents or publications. In the terminology of social network analysis, the geodesic distance between the patent and the publication is 1;

c) d_2*:* this variable takes value 1 whenever one (or more) inventors of a patent share a common collaborator with at least one (or more) paper authors (and 0 else). In other words, this dummy variable captures *prior indirect collaborations* among inventors and authors through *common acquaintances*, i.e. common collaborators. In terms of social network analysis, the geodesic distance between the patent and the publication is 2;

d) d_3: this variable takes value 1 whenever the shortest path connecting the team of inventors and the team of authors is equal to 3 (and 0 else);

[38] It should be pointed out that in the calculation of the geodesic distances among all possible pairs of inventors and authors we had to solve complex computational problems arising from the size of the involved matrices. For example, the largest component of the network of inventors and authors in the year 2000 involves 27148 nodes. Even restricting the attention to the largest component, computing the distance among all possible pairs of nodes implies building a matrix with more than 700 millions of cells.

e) $d_{4\text{-}6}$: this variable takes value 1 whenever the shortest path connecting the team of inventors and the team of authors is comprised between 4 and 6 (and 0 else);

f) $d_{>6}$: this variable take value 1 if the shortest path connecting the team of inventors and the team of authors is greater than 6, but finite, i.e. patent inventors and paper authors belong to the same connected component in the network of inventors and authors (and 0 else);

g) *disconnected*: this variable takes value 1 if patent inventors and paper authors are not reachable as they belong to disconnected components, i.e. the social distance between them is infinity (and 0 else).

In the estimations reported below, the reference group is 'disconnected', i.e. patent-paper pairs whose respective inventors and authors are not reachable.

As far as the spatial proximity is concerned, we implemented a small PHP/Javascript program, which exploits the Google Maps service to extract the geographical coordinates (i.e. latitude and longitude) from the inventors' addresses reported in patent documents, and from affiliations' addresses as reported in publications for paper authors. Since a patent-publication pair typically features multiple authors and inventors, we decided to compute the spatial distance between all possible pairs of authors and inventors reported in the patent-publication pair; among them, we took the lowest value as the spatial distance between the patent and the publication.[39]

In addition to proximity variables, we included in the regression fixed effects for the year of citing patents, and a variable (*time lag*) that measures the time lag in years between the citing patent and the cited publication. Given that the time lag between citing patents and cited publications follow a non linear pattern (see above), we expect that this variable will have a non linear effect on the probability of a citation.

[39] The haversine formula has been used to calculate the geographical distance. It can be computed as follows: $HavDist = EarthRadius \cdot c$, where $c = 2 \cdot a \cdot \tan 2\left(\sqrt{a}, \sqrt{1-a}\right)$ and $a = \left[\sin\left(\frac{lat2 - lat1}{2}\right)^2\right] + \cos(lat2) \cdot \left[\sin\left(\frac{lon2 - lat1}{2}\right)^2\right]$

Regression results

As a first step, we have estimated a model in which the only explanatory variables included are the (log of) geographical distance and the time lag between the patent and the publication. Results are reported in Table 39. As the coefficients of logit estimates do not have a direct economic interpretation, in the following tables we have reported odds-ratios. They are easily obtained by exponentiating logit coefficients. As expected, the geographical distance has a statistically significant negative effect on the probability of a citation tie between patents and publications, as shown by the fact that the odds ratio takes a value significantly lower than 1. The probability that a patent builds upon a scientific publication decreases with the spatial distance among inventors and authors. On the contrary, the coefficient of the time lag variable is not statistically different from zero (i.e. the odds-ratio is 1).

Table 39 – The impact of geographical distance on citation ties

	Semiconductors	Lasers	Biotechnology
Spatial distance	0.848[a]	0.799 [a]	0.735 [a]
	(0.008)	(0.012)	(0.006)
Time lag	0.993[ns]	0.985[ns]	1.002[ns]
	(0.009)	(0.013)	(0.007)
Number of observations	15881	9345	30506
Log-likelihood	-88.7	-250.2	-304.1
Pseudo-R2	0.013	0.031	0.043

Notes: Robust standard errors in parenthesis. [a] significant at the 1% level; [b] significant at the 5% level; [ns] not significant. Coefficients for year fixed effects nor reported.

To test the robustness of this result, we have included in our model the set of dummies capturing the extent of social distance among authors and inventors. Results are reported in Table 40. We note that the negative impact of spatial distance, although still statistically significant, drops quite remarkably, as shown by the fact that the odds ratio is now closer to 1. This means that, once we control for the social distance, the spatial location of authors and inventors matters relatively less in explaining knowledge flows from science to technology.

Table 40 – The impact of social distance on citation ties

	Semiconductors	Lasers	Biotechnology
Spatial distance	0.920[a]	0.888 [a]	0.815 [a]
	(0.008)	(0.017)	(0.007)
Time lag	1.003[ns]	1.003[b]	1.026 [a]
	(0.009)	(0.012)	(0.007)
Social distance d=0	786.123 [a]	315.24 [a]	542.76 [a]
	(616.93)	(210.30)	(365.27)
Social distance d=1	19.419 [a]	6.719 [a]	33.738 [a]
	(9.706)	(2.016)	(33.299)
Social distance d=2	10.713 [a]	1.870 [a]	17.068 [a]
	(4.570)	(0.447)	(17.581)
Social distance d=3	5.566 [a]	1.292 [ns]	0.800 [ns]
	(1.793)	(0.234)	(0.707)
Social distance 4≤d≤6	2.039 [a]	1.538 [a]	0.555 [ns]
	(0.245)	(0.150)	(0.233)
Social distance >6 (finite)	1.531 [a]	1.171 [b]	0.364 [a]
	(0.118)	(0.107)	(0.133)
Number of observations	15881	9345	30506
Log-likelihood	-84.409	-232.50	-280.51
Pseudo-R2	0.060	0.100	0.117

Notes: Robust standard errors in parenthesis. [a] significant at the 1% level; [b] significant at the 5% level; [ns] not significant. Coefficients for year fixed effects nor reported. The baseline category left out is represented by patent-.publications pairs not socially connected (disconnected).

On the other hand, we do observe that the dummy variables capturing the social distance among authors and inventors have a positive and statistically significant impact on the probability of a citation tie between patents and publications. Thus for example, with reference to semiconductors, patents and publications, which have been produced by individuals that have previously collaborated, i.e. that are at a social distance 1, are 19 times more likely to be also connected by a citation tie.

We also note that the citation probability falls quite sharply with social distance, starting from very high levels at low social distances. Thus, always with reference to semiconductors, we observe that the citation probability of a patent-publication pair at social distance 1 is 19 times more likely to result in a citation, as compared to non-connected pairs, whereas the citation premium decreases to 10 times for pairs at social distance 2. In this respect, it is also interesting to note some differences across the

three technology fields examined. In particular, we note that the citation premium decreases relatively faster in lasers and biotechnology, as compared to semiconductors.

Table 41 – The impact of geographical distance on citation ties

	Semiconductors	Lasers	Biotechnology
Time lag	0.993[ns]	0.990[ns]	1.000[ns]
	(0.008)	(0.012)	(0.006)
Citing patent – cited publication			
Europe-Europe	1.683[a]	1.695[a]	1.575[a]
	(0.145)	(0.180)	(0.078)
Europe-United States	1.011[ns]	1.006[ns]	0.819[a]
	(0.074)	(0.100)	(0.038)
Europe-Japan	0.941[ns]	0.877[ns]	0.802[b]
	(0.087)	(0.108)	(0.078)
United States-Europe	0.829[b]	0.810[ns]	0.953[ns]
	(0.078)	(0.096)	(0.043)
United-States-United States	1.506[a]	1.701[a]	1.650[a]
	(0.100)	(0.162)	(0.067)
United States-Japan	0.853[ns]	0.779[ns]	0.944[ns]
	(0.070)	(0.095)	(0.076)
Japan-Europe	0.900[ns]	0.742[b]	0.662[a]
	(0.079)	(0.095)	(0.068)
Japan-United States	0.890[ns]	0.718[a]	0.583[a]
	(0.059)	(0.075)	(0.050)
Japan-Japan	1.551[a]	1.989[a]	3.784[a]
	(0.112)	(0.211)	(0.508)
Number of observations	15881	9345	30506
Log-likelihood	-89.254	-254.50	-313.65
Pseudo-R2	0.006	0.014	0.016

Notes: Robust standard errors in parenthesis. [a] significant at the 1% level; [b] significant at the 5% level; [ns] not significant. Coefficients for year fixed effects nor reported.

In these two fields, the citation premium even disappears for social distances greater than 2. While extremely important for transmitting knowledge, the effectiveness of social connections seems to decay very rapidly with social distance. Alternatively, we can presume that long-distance links decay rapidly over time, and do not convey anymore any knowledge flow.

In order to test further the effect of geographical vs. social distance on the probability of citations, we have re-estimated our model by replacing the variable measuring the spatial distance among authors and inventors with a set of dummy variables, capturing whether the inventors of citing patents and

the authors of scientific publications are located in the same geographical area. Thus, for example, the variable Europe-Europe takes value 1 whenever at least one inventor of the patent and one author of the scientific publications are located in Europe (and 0 else).

Table 41 reports the estimates of a model, which includes only this set of dummy variables and the time lag between the patent and the publication. As expected, the probability of a citation tie decreases for patent-publication pairs whose authors and inventors are located in different geographical areas. This is shown by odds ratios significantly larger than 1 for those variables indicating co-location. Thus, for example, with reference to biotechnology, the probability that a European patent builds upon a European publication is 57% higher than expected, thereby suggesting that knowledge flows tend to be spatially bounded at the continental level. On the contrary, cross-continental citations are less likely to occur, although the odds ratios for most of such variables are not statistically different from one.

Finally, Table 42 reports estimates of a model which includes social distance effects, in addition to dummies for co-location. Again, we note that, once we control for social proximity, the explanatory power and the statistical significance of the geographical variables tend to become less strong or even to vanish, as shown by the fact that the odds ratios of variables capturing co-location effects take values closer to 1. On the other hand, the odds ratios of dummy variables for social distance are statistically significant (at least up to distance 2) and remarkably larger than 1.

Our interpretation of these results is that being spatially close to the source of scientific knowledge is not a sufficient condition to benefit from any kind of knowledge flow. Social networks of collaboration among scientists and inventors explain a great deal of the knowledge transfer that takes place from the realm of science to that of technology. In this respect, scientific knowledge may easily flow over long distances as long as the sender and the receiver of such knowledge are connected through a short chain of collaborators.

Table 42 – The impact of social distance on citation ties

	Semiconductors	Lasers	Biotechnology
Time lag	1.004 [ns]	1.030[b]	1.021[a]
	(0.009)	(0.012)	(0.007)
Social distance d=0	1336.739 [a]	506.554[a]	1329.934 [a]
	(1007.6)	(331.45)	(880.10)
Social distance d=1	25.293 [a]	9.360[a]	30.585 [a]
	(13.051)	(2.608)	(32.230)
Social distance d=2	17.255 [a]	2.474[a]	16.594 [a]
	(7.344)	(0.556)	(17.502)
Social distance d=3	6.857 [a]	1.398[ns]	0.675 [ns]
	(2.225)	(0.253)	(0.632)
Social distance 4≤d≤6	2.315 [a]	1.602[a]	0.565 [ns]
	(0.285)	(0.165)	(0.233)
Social distance >6 (finite)	1.611 [a]	1.189[b]	0.411[a]
	(0.129)	(0.107)	(0.131)
Citing patent – cited publication			
Europe-Europe	1.431[ns]	1.179[ns]	1.372[a]
	(0.142)	(0.208)	(0.087)
Europe-United States	0.992 [ns]	1.035 [ns]	0.854 [a]
	(0.074)	(0.100)	(0.042)
Europe-Japan	0.894 [ns]	0.922 [ns]	0.816 [b]
	(0.082)	(0.112)	(0.079)
United States-Europe	0.835 [ns]	0.848 [ns]	1.006 [ns]
	(0.078)	(0.099)	(0.050)
United-States-United States	0.919 [ns]	1.122 [ns]	1.276 [a]
	(0.072)	(0.136)	(0.056)
United States-Japan	0.799 [b]	0.770[b]	0.981 [ns]
	(0.067)	(0.093)	(0.080)
Japan-Europe	0.936 [ns]	0.789 [ns]	0.679 [a]
	(0.082)	(0.098)	(0.068)
Japan-United States	0.903 [ns]	0.752 [a]	0.592 [a]
	(0.061)	(0.078)	(0.051)
Japan-Japan	1.491[a]	1.655[a]	3.727 [a]
	(0.111)	(0.187)	(0.498)
Number of observations	15881	9345	30506
Log-likelihood	-84.382	-232.87	-283.77
Pseudo-R2	0.060	0.098	0.109

Notes: Robust standard errors in parenthesis. [a] significant at the 1% level; [b] significant at the 5% level; [ns] not significant. Coefficients for year fixed effects nor reported. The baseline category left out is represented by patent-.publications pairs not socially connected (disconnected).

4. CONCLUSIONS

This report has offered a large-scale empirical appraisal of the social connections linking academic scientists and industrial researchers in five science intensive technology fields, namely transmission of digital information (telecommunications), speech analysis (ICT), semiconductors, lasers, and biotechnology (measuring, testing, diagnostics). In spite of different objectives and systems of incentives, our results show that the two communities of researchers are socially connected to a much larger extent than one would normally presume. A key role in connecting the two communities is played by specific individuals, i.e. authors-inventors, that act as gatekeepers bridging the boundaries across the two domains. A further important result emerging from our study is that social networks of collaboration among scientists and inventors work as effective conduits of knowledge flows from the realm of science to that of technology. In this respect, our analysis shows that social proximity to authors of scientific publications is a much more fundamental factor affecting the knowledge transfer from scientific research to technological applications than just geographical proximity.

The increasing inter-dependency between science and technology has made the theme of university–industry knowledge transfer a key research issue both in economics and management studies, as well as a top entry in the science and technology policy agenda of many countries. In the context of Europe, a general and widespread belief is that the mechanisms leading to the transfer of scientific knowledge into technological applications are somehow impaired and less effective than in other areas of the world, notably the United States. This conjecture has led to interpreting the European lag in some key high tech sectors, such as electronics and biotechnology, as a consequence of its inability to convert its scientific strength into economic profitable innovations. This phenomenon has also deserved the name of "European Paradox" to stress the fact that European strength in the production of high quality scientific output is not matched by the ability of European private companies to benefit from such output.

The existence and the extent of a European weakness in the transfer of knowledge from the domain of scientific research to technological applications is normally predicated on the basis of bibliometric indicators on the quantity and quality of scientific output. To date, however, very few studies have attempted to investigate in depth the actual mechanisms through which knowledge produced within

the boundaries of academic organisations get transferred and translated into technological developments. This study has contributed to filling this gap by proposing a large scale quantitative analysis of the social connections linking academic scientists and industrial researchers in five science intensive technology fields.

To this purpose, the study has exploited a complex, relational dataset reporting full bibliographical information on patent applications and on scientific publications cited in those patent documents. The key analytical tool used to investigate the linkages connecting academic scientists and industrial researchers has been represented by social network analysis. Specifically, information on co-authorship and on co-invention has been exploited to assess the extent of connectedness among the two social communities of researchers. Likewise, citations from patent documents to scientific publications have been used as proxy for the knowledge flows from the realm of science to that of technology.

The main findings of the study may be summarised as follows.

High quality scientific publications find their way into a large number of technological developments. Publications that are (highly) cited in patents are not only cited in the realm of technology, but they are also heavily cited by other scientific publications. Besides validating the methodological choice of using patent citations to scientific publications as proxy of knowledge flows from science to technology, this finding suggests that there is not necessarily a conflicting logic between scientific and industrial communities. In this respect, however, it should be also noted that European scientific publications cited in patents receive a lower average number of citations in scientific literature than the corresponding articles published by US authors. This evidence seems to suggest that high quality European publications face more obstacles in translating into technological applications than comparable scientific output in the US.

European science is relatively under-represented among publications that provide key contributions to technological developments. The share of European organisations among scientific publications that are *highly cited* in patents is systematically lower than the its share of all cited publications. This gap is particularly evident in fields such as lasers, semiconductors and biotechnology. This result suggests that European scientific output translates into a lower number of technological developments, thereby providing further support to the conjecture about the existence of weaknesses in the process of knowledge transfer from science to technology.

Private companies account for a large share of scientific publications highly cited in patents. The role played by different types of institutions in the production of scientific publications highly cited in patents varies across technology fields, with universities accounting for a large share particularly in biotechnology. However, a key result emerging from our analysis is that private companies account for a quite large fraction of highly cited publications in all technology fields. In particular, the share of highly cited publications held by private companies is remarkably larger than their share of all scientific publications, which according to other studies may be estimated around 5-10%. This result suggests that corporate labs contribute to a large extent to the scientific research that is incorporated into technological applications.

The European private companies' contribution to the production of scientific publications highly cited in patents is significantly lower than the contribution of private companies located in other areas, notably the United States. A major weakness of the European systems of research, as compared to other geographical areas, especially the United States, is related to the low degree of involvement of private companies in the conduct of research leading to scientific publications cited in patents. Whereas the contribution of the public system of scientific research, i.e. universities and public research organisations, is generally comparable to, and often larger than the contribution of the corresponding system in the US, the fraction of scientific publications accounted for by the private system of research is considerably lower. To the extent that the ability of private companies to profit from scientific output generated in the sphere of science depends on the possession of absorptive capabilities and especially on the existence of boundary-spanning individuals, we believe this characteristic represents one of the major obstacles to the effective diffusion of knowledge from the realm of science to that of technology.

The propensity of European technology to build upon US scientific publications is generally higher than the propensity of US technology to rely upon European science. An analysis of the knowledge flows across geographical areas by origin of citing patents and origin of cited publications reveals that European patents tend to cite US scientific publications to a larger extent than US patents tend to cite European scientific papers. In other terms, the empirical evidence shows the existence of an asymmetry in knowledge flows between Europe and the US, with a larger amount of knowledge flowing from the US to Europe than vice versa. Likewise, we observed that the propensity of US inventors to rely upon the domestic science base is significantly greater than the propensity of European inventors to exploit their domestic science base.

The two communities of academic scientists and industrial researchers are highly connected to each other. The social network analysis shows that the network of co-inventors is highly disconnected with many components of small size. This means that most collaborators of each inventor come from the same organisation and that few connections exist among teams of industrial researchers. However, when one looks at co-invention and co-authorship relations simultaneously, the key result is that the two communities of researchers are significantly more socially connected that one would probably expect. In three crucial technology fields, such as semiconductors, lasers and biotechnology, 35%, 51% and 53%, respectively, of *all* authors and inventors are either directly or indirectly connected, via co-invention or co-authorship, to each other in a large connected component. In addition to that, 24%, 40% and 32% of all inventors are either directly or indirectly connected (i.e. reachable) to each other. Besides indicating that academic scientists and industrial researchers are highly connected, these results suggest that the community of inventors itself is much more connected than data on co-invention only would lead us to presume. Although not directly connected to each other, inventors are indirectly connected through scientific authors and through authors-inventors, i.e. individuals that participate in teams of inventors *and* in teams of scientists.

Authors-inventors play a key role in connecting the communities of scientists and inventors and act as gatekeepers across the two realms. A crucial role in ensuring high degrees of connectivity between the two communities of researchers is played by a specific type of individuals that we have labelled as authors-inventors. They are researchers that do publish scientific articles and patent new inventions, thereby participating into both communities. Social network analysis reveals that such individuals are characterised by a higher degree centrality, i.e. they tend to collaborate on average with a significantly larger number of other inventors *and* authors, than do simple inventors and authors, and by a higher betweenness centrality, i.e. they play a crucial function of knowledge brokers in the network that makes them more in-between than simple inventors and authors, and ensures a rapid diffusion of knowledge and ideas from one domain to the other.

Europe is characterised by a relatively low number and share of science-technology gatekeepers, i.e. authors-inventors. Given the key role played by authors-inventors in bridging the realms of science and technology, we believe that a key finding of our study is that the share of European inventors playing this specific function is lower than its share of simple inventors. To the extent that authors-inventors play

as brokers of knowledge from the domain of science to that of technology, we believe this finding has very important implications for understanding of the gap between Europe and the US in the effectiveness to translate the results of scientific research into commercially useful applications. Proximity to such individuals, and more generally proximity among authors of scientific publications and inventors of patented inventions is in fact a fundamental factor affecting the effective diffusion of scientific knowledge (see below). In addition to this, we do also believe that this result is consistent with our finding that a major European weakness is related to the feeble commitment of private companies in the production of scientific publications relevant for technological developments, given that authors-inventors are most likely to come from such organisations.

The network of academic scientists and industrial researchers has the properties of a "small world". The social network of academic scientists and industrial researchers is characterised by topological properties typical of "small world" graphs. On the one hand, it presents high degrees of *local cliquishness*, i.e. an individual's collaborators tend to collaborate each other; on other hand, it also presents a low average distance among individuals, i.e. any random pair of individuals is separated by a low number of steps. This means that the network of authors and inventors work at least potentially as an effective means of knowledge transmission and diffusion.

Social proximity among (academic) scientists and industrial researchers is the most important factor affecting the probability that a patented invention will build upon a scientific publication. In this study, we estimated an econometric model for the probability that a patent-paper pair is linked by a citation tie. Our findings reveal that such a probability is apparently affected in a negative way by the geographical distance that separate patent inventors and paper authors. Yet, the effect of spatial distance vanishes once we control for the social distance among them. Inventors that are socially *closer* to authors of scientific publications are much more likely to build upon such publications than are inventors located at a larger social distance. In other terms, knowledge transfer from science to technology takes place mostly through social networks of collaboration among scientists and inventors.

Results of this study provide further empirical support to the conjecture that the mechanisms driving the transfer of scientific outputs into technological applications in Europe are somehow impaired and less effective than in other areas of the world, notably the United States. At the same time, they also point out that social networks of (academic) scientists and industrial researchers account for much of

the observed patterns of knowledge diffusion from science to technology. In particular, the study has shown that a crucial role in connecting the two communities of researchers is played by a specific type of individuals, i.e. authors-inventors, that act as gatekeepers and channel information and knowledge between groups with different objectives and incentives. In this respect, a major European weakness is related to the comparatively lower involvement of private companies in the conduct of basic and applied research leading to scientific publications and to the consequently lack of authors-inventors that are able to bridge and connect the realms of science and technology. In other words, it is possible that part of the European backwardness in this field is due to a less connected research area. We do believe that increasing such a connectivity should feature prominently in a policy agenda aiming to spur the rate of knowledge transfer from science to technology. In this respect, the mobility of inventors (i.e. industrial researchers) and academic scientists across regions, countries, and organisations represents, in our view, a major policy objective in order to achieve higher degrees of social connectivity among the two communities of research.

References

Agrawal A.K., Cockburn I.M., McHale J. (2003), "Gone But Not Forgotten: Labor Flows, Knowledge Spillovers, and Enduring Social Capital", *NBER Working Paper* 9950

Balconi M., Breschi S., Lissoni F. (2004), "Networks of inventors and the location of academic research: An exploration of Italian data", *Research Policy* 33(1): 127-45.

Breschi S., Lissoni F. (2001a) 'Knowledge Spillovers and Local Innovation Systems: A Critical Survey', *Industrial and Corporate Change* 10/4, 2001a, pp. 975-1005.

Breschi S., Lissoni F. (2001b) 'Localised Knowledge Spillovers vs. Innovative Milieux: Knowledge 'Tacitness' Reconsidered', *Papers in regional science* 80/3, 2001.

Breschi S., Lissoni F., Malerba F. (2003), 'STI-NET patent and patent citations database. Methodology and preliminary analyses', mimeo.

Breschi S., Lissoni F. (2004), "Knowledge networks from patent data: methodological issues and research targets", in Moed H., Glänzel W., Schmoch U. (Eds.) *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*, Springer, Berlin.

Breschi S., Lissoni F. (2006), "Mobility and Social Networks: Localised Knowledge Spillovers Revisited", *Annales d'Economie et Statistique*.

Cockburn I.M., Henderson R.M. (1998), "Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery", *Journal of Industrial Economics* XLVI, pp.157-182.

Cockburn I.M. (2004), *Tracking Knowledge Flows: Opportunities and Challenges in Using Patents and other Bibliometric Data*, lecture given at the European Summer School on Industrial Dynamics, Institut d'Etudes Scientifiques de Cargese, August 28-September 4

Cockburn I.M., Kortum S., Stern S. (2002), "Are all patent examiners equal? The impact of characteristics on patent statistics and litigation outcomes", *NBER Working Paper* 8980

Collins P., Wyatt S. (1988), "Citations in patents to the basic research literature", *Research Policy* 17: 65-74.

Cowan, R., David, P.A., Foray, D. (2000). "The explicit economics of knowledge codification and tacitness". *Industrial and Corporate Change* 9, 211–254.

Cowan, R., Jonard, N. (2000). "The dynamics of collective invention", MERIT WP # 00-018. Maastricht Economic Research Institute on Innovation and Technology, Maastricht University.

Dasgupta P., David P.A. (1994), "Toward a New Economics of Science", *Research Policy* 23, pp. 487-521.

Dosi G., Llerena P., Sylos-Labini M. (2005) "Science-Technology-Industry Links and the "European Paradox": Some Notes on the Dynamics of Scientific and Technological Research in Europe", *LEM Papers Series* 2005/02, Scuola Superiore Sant'Anna, Pisa.

Feldman M.P. (1999), "The New Economics of Innovation, Spillovers and Agglomeration: A Review of Empirical Studies", *Economics of Innovation and New Technology* 8, pp.5-25.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2000). Market Value and Patent Citations: A First Look," NBER Working Paper W7741.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools. NBER Working Paper 8498.

INCENTIM (2003), *Linking Science to Technology – Bibliographic References in Patents*, Project Report (http://www.cordis.lu/indicators).

Jaffe A.B. (1989), "Real effects of academic research", *American Economic Review* 79, 957-70.

Jaffe A.B., Trajtenberg M., Henderson R. (1993), "Geographic localisation of knowledge spillovers as evidenced by patent citations", *Quarterly Journal of Economics* 108: 577-598

Jaffe, A.B., M. Trajtenberg (2002). Patents, Citations and Innovations: A Window on the Knowledge Economy M.I.T. Press.

Karki M.M.S. (1997), 'Patent Citation Analysis: A Policy Analysis Tool', *World Patent Information* 19, pp. 269-272.

Luwel M. (1999), "Is the Science Citation Index US biased", *Scientometrics* 46: 549-62.

Melin, G., Persson, O., 1996. Studying research collaborations using co-authorship. *Scientometrics* 36, 363–377.

Merton, R.K. (1957). "Priorities in scientific discovery: a chapter in the sociology of science". *American Sociological Review* 22 (6), 635–659.

Mowery D., Ziedonis A. (2001), "The Geographic Reach of Market and Non-Market Channels of Technology Transfer: Comparing Citations and Licences of University Patents", NBER Working Paper 8568.

Narin F., Hamilton K.S., Olivastro D. (1997). "The increasing linakge between US technology and science", *Research* Policy 26: 317-330.

Newman M.E.J. (2000), "Who is the best connected scientists? A study of scientific co-authorship networks", *SFI Working Paper* 00-12-64, Santa Fe

Newman M.E.J. (2001), "The structure of scientific collaboration networks", *Proceedings of the National Academy of Science USA* 98, 404-409.

Singh J. (2003), "Inventor Mobility and Social Networks as Drivers of Knowledge Diffusion", mimeo, Harvard Business School.

Steinmueller, E. (2000). "Does information and communication technology facilitate 'codification' of knowledge?". *Industrial and Corporate Change*.

Stern S. (1999). "Do scientists pay to be scientists?". NBER Wp 7410.

Stolpe M. (2002), "Determinants of knowledge diffusion as evidenced in patent debates: the case of Liquid Crystal Display technology", *Research Policy* 31, pp. 1181-1198

Thompson P., Fox-Kean M.(2005), "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment", American Economic Review (forthcoming)

Van Looy B., Zimmermann E., Veugelers R., Verbeek A., Mello J., Debackere K. (2003), "Do science-technology interactions pay off when developing technology? An exploratory investigation of 10 science-intensive technology domains", *Scientometrics* 57(3): 355-67.

Verbeek A., Debackere K., Luwel M., Andries P., Zimmermann E., Deleus F. (2002), "Linking science to technology: Using bibliographic references in patents to build linkage schemes", *Scientometrics* 54(3): 399-420.

Wasserman S., Faust C. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Watts D.J. (2003), *Six Degrees: The Science of a Connected Age,* W.W. Norton & Company.

Watts, D.J., Strogatz, S.H. (1998). "Collective dynamics of 'small world' networks". *Nature* 393, 440–442.

Zucker L.G., Darby M.R., Armstrong J. (1998), 'Geographically localized knowledge: Spillovers or markets?', *Economic Inquiry* 36, pp. 65-86

Table A1 - 10 selected technology fields on the basis of USPC codes

	Technology fields	USPC codes
1.	Telecommunications	178, 329, 331, 332, 333, 340, 341, 342, 343, 347, 348, 358, 360, 367, 370, 375, 377, 379, 381, 385, 386, 455
2.	Information Technology	235, 327, 345, 365, 382, 400, 463, 473, 708, 709, 710, 711, 712, 713, 714, 704
3.	Semiconductors	257, 326, 438, 505
4.	Optics	349, 351, 352, 353, 355, 359, 396, 430
5.	Control Technology	033, 073, 109, 116, 177, 194, 236, 307, 323, 324, 356, 368, 374, 380, 434, 436, 701, 702, 453
6.	Medical Technology	128, 422, 433, 600, 601, 602, 604, 606, 607, 623, 119
7.	Organic Chemistry	127, 530, 534, 536, 540, 544, 546, 548, 549, 552, 554, 556, 558, 560, 562, 564, 568, 570, 585
8.	Drugs	424, 514
9.	Biotechnology	435, 800
10.	Environmental technology	055, 095, 110, 126, 210, 261, 588

Table A2 – 10 technology fields, average annual growth rate of patenting activity

	Technology fields	EPO (1990-2001)	USPTO (1990-2003)
1.	Telecommunications	12.2	8.7
2.	Information Technology	10.6	12.7
3.	Semiconductors	4.3	14.0
4.	Optics	3.1	5.1
5.	Control Technology	4.8	6.4
6.	Medical Technology	7.7	7.3
7.	Organic Chemistry	1.6	0.8
8.	Drugs	8.7	6.4
9.	Biotechnology	9.1	12.6
10.	Environmental technology	3.2	0.5
All fields		6.8	7.7

Note: for the EPO, average growth rate was calculated over the period 1990-2001, given the drop in the number of patent applications *published* after 2001.

Figure A1 – 10 technology fields, total number of patent applications 1990-2003 (EPO)

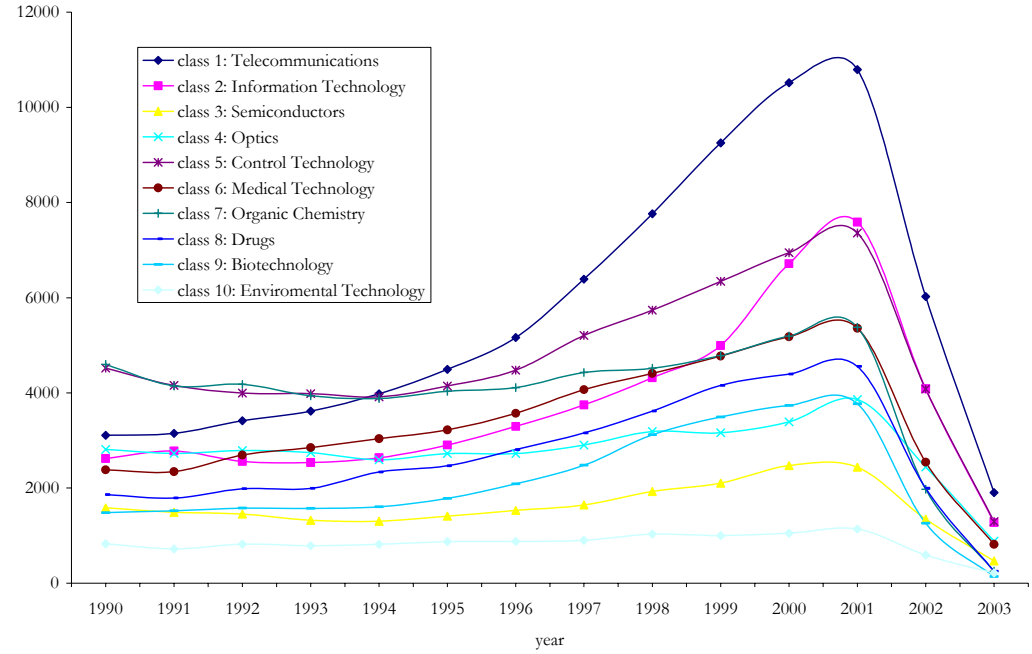


Figure A2 – 10 technology fields, total number of patents granted 1990-2003 (USPTO)

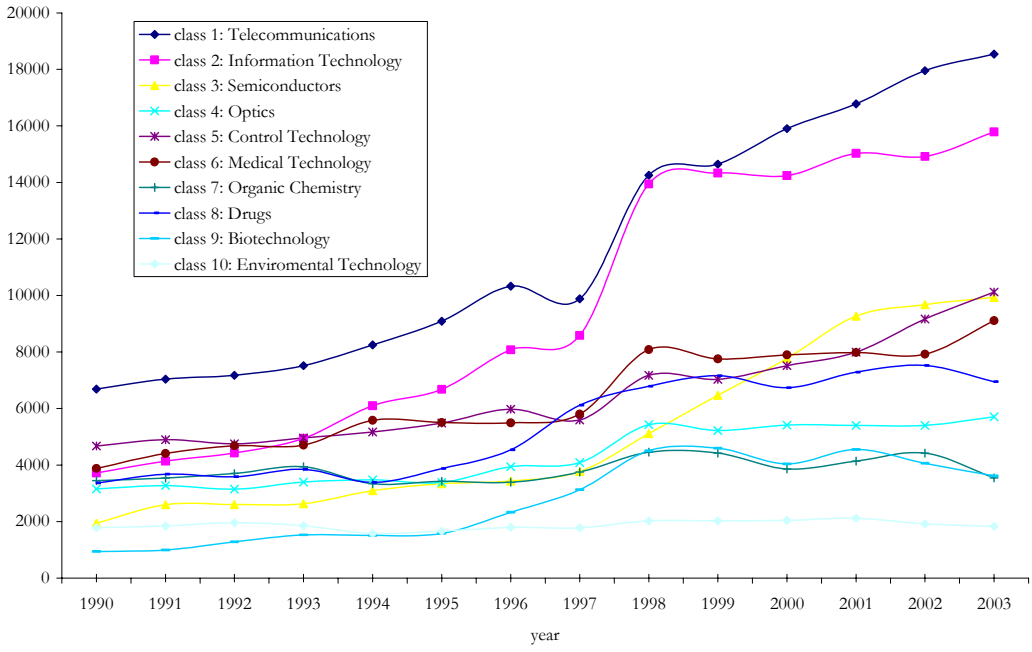


Table A3 – EPO four most cited ISI-journals by technology field (% share of citations)

Telecommunications	
IEEE TRANSACTIONS ON COMMUNICATIONS	6.3
ELECTRONICS LETTERS	6.3
IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS	5.4
IEEE COMMUNICATIONS MAGAZINE	5.2
Information technology	
COMPUTER NETWORKS AND ISDN SYSTEMS	3.3
COMPUTER	2.7
COMMUNICATIONS OF THE ACM	2.6
IEEE TRANSACTIONS ON COMPUTERS	2.5
Semiconductors	
APPLIED PHYSICS LETTERS	19.4
JAPANESE JOURNAL OF APPLIED PHYSICS	9.0
IEEE TRANSACTIONS ON ELECTRON DEVICES	6.9
JOURNAL OF THE ELECTROCHEMICAL SOCIETY	5.7
Optics	
ELECTRONICS LETTERS	14.5
APPLIED PHYSICS LETTERS	12.6
IEEE PHOTONICS TECHNOLOGY LETTERS	9.1
OPTICS LETTERS	7.4
Control technology	
ANALYTICAL CHEMISTRY	3.6
MAGNETIC RESONANCE IN MEDICINE	2.3
PROCEEDINGS OF THE US NATIONAL ACADEMY OF SCIENCES	2.3
REVIEW OF SCIENTIFIC INSTRUMENTS	2.3
Medical technology	
IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING	5.8
MEDICAL & BIOLOGICAL ENGINEERING & COMPUTING	5.0
BIOMATERIALS	3.9
PROCEEDINGS OF THE IEEE	2.6
Organic chemistry	
JOURNAL OF MEDICINAL CHEMISTRY	8.8
JOURNAL OF ORGANIC CHEMISTRY	5.5
TETRAHEDRON LETTERS	4.9
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	4.5
Drugs	
PROCEEDINGS OF THE US NATIONAL ACADEMY OF SCIENCES	3.6
JOURNAL OF BIOLOGICAL CHEMISTRY	2.7
JOURNAL OF MEDICINAL CHEMISTRY	2.5
JOURNAL OF IMMUNOLOGY	1.9
Biotechnology	
PROCEEDINGS OF THE US NATIONAL ACADEMY OF SCIENCES	7.5
JOURNAL OF BIOLOGICAL CHEMISTRY	6.8
SCIENCE	3.5
NUCLEIC ACIDS RESEARCH	3.4
Environmental technology	
WATER RESEARCH	9.0
DESALINATION	5.3
JOURNAL WATER POLLUTION CONTROL FEDERATION	5.0
CHEMIE INGENIEUR TECHNIK	4.8

Table A4 – EPO four most cited subject fields by technology field (% share of citations)

Telecommunications	
ENGINEERING, ELECTRICAL & ELECTRONIC	42.6
TELECOMMUNICATIONS	24.4
COMPUTER SCIENCE, INFORMATION SYSTEMS	7.8
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	6.5
Information technology	
ENGINEERING, ELECTRICAL & ELECTRONIC	22.3
COMPUTER SCIENCE, SOFTWARE ENGINEERING	14.8
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	13.8
COMPUTER SCIENCE, INFORMATION SYSTEMS	8.1
Semiconductors	
PHYSICS, APPLIED	42.6
ENGINEERING, ELECTRICAL & ELECTRONIC	21.1
PHYSICS, CONDENSED MATTER	7.6
MATERIALS SCIENCE, MULTIDISCIPLINARY	6.4
Optics	
ENGINEERING, ELECTRICAL & ELECTRONIC	30.5
OPTICS	27.2
PHYSICS, APPLIED	27.2
INSTRUMENTS & INSTRUMENTATION	1.6
Control technology	
ENGINEERING, ELECTRICAL & ELECTRONIC	8.0
BIOCHEMISTRY & MOLECULAR BIOLOGY	7.5
CHEMISTRY, ANALYTICAL	7.4
INSTRUMENTS & INSTRUMENTATION	6.7
Medical technology	
ENGINEERING, BIOMEDICAL	16.3
MEDICAL INFORMATICS	6.0
RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING	5.9
SURGERY	4.3
Organic chemistry	
CHEMISTRY, ORGANIC	15.1
BIOCHEMISTRY & MOLECULAR BIOLOGY	14.3
CHEMISTRY, MULTIDISCIPLINARY	12.9
CHEMISTRY, MEDICINAL	11.1
Drugs	
PHARMACOLOGY & PHARMACY	11.8
BIOCHEMISTRY & MOLECULAR BIOLOGY	10.6
IMMUNOLOGY	6.9
MULTIDISCIPLINARY SCIENCES	5.2
Biotechnology	
BIOCHEMISTRY & MOLECULAR BIOLOGY	24.5
MULTIDISCIPLINARY SCIENCES	10.1
CELL BIOLOGY	8.0
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	7.1
Environmental technology	
ENVIRONMENTAL SCIENCES	13.9
ENGINEERING, CHEMICAL	13.5
ENGINEERING, ENVIRONMENTAL	12.9
WATER RESOURCES	11.4

Table A5 – USPTO four most cited ISI-journals by technology field (% share of citations)

Telecommunications	
ELECTRONICS LETTERS	7.7
IEEE TRANSACTIONS ON COMMUNICATIONS	7.2
SIGNAL PROCESSING	3.7
JOURNAL OF LIGHTWAVE TECHNOLOGY	3.6
Information technology	
IEEE JOURNAL OF SOLID-STATE CIRCUITS	6.9
SIGNAL PROCESSING	6.1
COMPUTERS & GRAPHICS	5.1
IEEE TRANSACTIONS ON COMPUTERS	4.4
Semiconductors	
APPLIED PHYSICS LETTERS	19.6
IEEE TRANSACTIONS ON ELECTRON DEVICES	8.3
JAPANESE JOURNAL OF APPLIED PHYSICS	7.8
JOURNAL OF APPLIED PHYSICS	7.1
Optics	
APPLIED OPTICS	11.0
APPLIED PHYSICS LETTERS	8.4
ELECTRONICS LETTERS	6.7
OPTICS LETTERS	5.6
Control technology	
APPLIED OPTICS	4.5
ANALYTICAL CHEMISTRY	4.4
APPLIED PHYSICS LETTERS	2.4
REVIEW OF SCIENTIFIC INSTRUMENTS	2.4
Medical technology	
CIRCULATION	3.8
RADIOLOGY	3.7
IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING	2.0
SCIENCE	1.6
Organic chemistry	
PROCEEDINGS OF THE US NATIONAL ACADEMY OF SCIENCES	6.1
JOURNAL OF BIOLOGICAL CHEMISTRY	5.7
TETRAHEDRON LETTERS	4.4
SCIENCE	4.0
Drugs	
PROCEEDINGS OF THE US NATIONAL ACADEMY OF SCIENCES	5.0
JOURNAL OF BIOLOGICAL CHEMISTRY	4.4
JOURNAL OF MEDICINAL CHEMISTRY	3.7
SCIENCE	3.3
Biotechnology	
PROCEEDINGS OF THE US NATIONAL ACADEMY OF SCIENCES	8.9
JOURNAL OF BIOLOGICAL CHEMISTRY	6.8
SCIENCE	5.2
NATURE BIOTECHNOLOGY	5.1
Environmental technology	
JOURNAL OF CHROMATOGRAPHY	10.5
ANALYTICAL CHEMISTRY	7.8
JOURNAL OF MEMBRANE SCIENCE	4.0
APPLIED AND ENVIRONMENTAL MICROBIOLOGY	3.5

Table A6 – USPTO four most cited subject fields by technology field (% share of citations)

Telecommunications	
ENGINEERING, ELECTRICAL & ELECTRONIC	42.0
TELECOMMUNICATIONS	16.7
OPTICS	9.0
PHYSICS, APPLIED	8.8
Information technology	
ENGINEERING, ELECTRICAL & ELECTRONIC	33.5
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	14.1
COMPUTER SCIENCE, SOFTWARE ENGINEERING	14.1
TELECOMMUNICATIONS	5.6
Semiconductors	
PHYSICS, APPLIED	42.6
ENGINEERING, ELECTRICAL & ELECTRONIC	23.9
ELECTROCHEMISTRY	5.1
MATERIALS SCIENCE, COATINGS & FILMS	4.6
Optics	
OPTICS	25.4
PHYSICS, APPLIED	23.8
ENGINEERING, ELECTRICAL & ELECTRONIC	20.6
PHYSICS, MULTIDISCIPLINARY	3.7
Control technology	
ENGINEERING, ELECTRICAL & ELECTRONIC	9.8
PHYSICS, APPLIED	8.5
CHEMISTRY, ANALYTICAL	7.9
OPTICS	7.1
Medical technology	
CARDIAC & CARDIOVASCULAR SYSTEM	8.6
SURGERY	8.3
RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING	6.8
ENGINEERING, BIOMEDICAL	5.6
Organic chemistry	
BIOCHEMISTRY & MOLECULAR BIOLOGY	21.4
CHEMISTRY, ORGANIC	10.1
CHEMISTRY, MULTIDISCIPLINARY	8.4
CELL BIOLOGY	5.9
Drugs	
BIOCHEMISTRY & MOLECULAR BIOLOGY	13.9
PHARMACOLOGY & PHARMACY	7.3
IMMUNOLOGY	7.0
CHEMISTRY, MEDICINAL	4.8
Biotechnology	
BIOCHEMISTRY & MOLECULAR BIOLOGY	26.1
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	8.1
CELL BIOLOGY	8.0
MULTIDISCIPLINARY SCIENCES	7.8
Environmental technology	
CHEMISTRY, ANALYTICAL	21.0
ENGINEERING, CHEMICAL	10.7
BIOTECHNOLOGY & APPLIED MICROBIOLOGY	6.7
POLYMER SCIENCE	6.2