CENTRE for ECONOMIC PERFORMANCE

CEP Discussion Paper No 723

May 2006

Basic Research and Sequential Innovation

Sharon Belenzon

Abstract
The commercial value of basic knowledge depends on the arrival of follow-up developments mostly from outside the boundaries of the inventing firm. Private returns would depend on the extent the inventing firm internalizes these follow-up developments. Such internalization is less likely to occur as knowledge becomes more general. This motivates the historical concern of insufficient private incentive for basic research. The present paper develops a novel empirical methodology of identifying unique patterns of knowledge flows (based on patent citations), which provide information about whether 'spilled' knowledge is reabsorbed by its inventor. Using comprehensive data on the largest 500 inventing firms in the US the classical problem of underinvestment in basic research is confirmed: spillovers of more general knowledge (and in this respect, more basic) are less likely to feed back to the inventing firm. This translates to lower private returns, as indicated by the effect of the R&D stock of the firm on its market value.

Keywords: basic knowledge, spillovers, patents and citations
JEL Classification: O31, O32 and O33

This paper was produced as part of the Centre's Productivity and Innovation Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements
I deeply appreciate the tremendous support of my PhD advisors Mark Schankerman and John Van Reenen. I thank Manuel Trajtenberg, Nick Bloom, Steve Bond, Bronwyn Hall, Iain Cockburn, John Haltiwanger, David Hendry, Sam Kortum, Paul Klemperer, Steve Redding, John Sutton and numerous seminar participants for helpful comments.

Sharon Belenzon is a Research Economist for the Productivity and Innovation Programme at the Centre for Economic Performance, London School of Economics.
Correspondence: sharon.belenzon@nuffield.ox.ac.uk

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

ISBN 0 7530 1949 3

1. Introduction

It is well accepted that significant advancements in scientific knowledge must come from basic research. Basic knowledge brings about follow-up developments that usually spread over a wide range of fields and are conducted outside the boundaries of the inventing firm. These follow-up developments substantiality enhance the commercial value of the basic knowledge[1]. In a context of sequential innovation, the literature refers to the outside follow-up developments of knowledge as knowledge spillovers (hereafter, spillovers).

For the inventing firm to capture substantial private rents it must internalize the spillovers of its basic knowledge, i.e., the inventing firm must benefit from the value enhancing features added to its basic knowledge by other agents. This internalization can take two forms: contractually and technologically. Under contractual internalization the inventing firm license its knowledge to using firms, where under technological internalization the spillovers created by the basic knowledge feed back into the future research of the inventing firm. The present paper focuses on technological internalization as a channel through which private rents are appropriated. An empirical methodology (based on patent citations) is developed to measure technological internalization and the extent it is correlated with the generality of knowledge and the market value of the inventing firm.

The main hypothesis of this paper is that as knowledge becomes more general, and in this respect more basic, the extent spillovers feed back to the inventing firm diminishes, since only firms with a wider technology base could achieve such internalization. The empirical prediction of this hypothesis is that there would be a negative correlation between the generality of knowledge (measured as the number of fields where follow-up research is inspired) and the extent this knowledge is reabsorbed by the inventing firm after external follow-up developments arrive. Yet, a competing hypothesis is that firms choose the "basicness" level of their knowledge: basic knowledge is invented only by firms with a wide technology base that allows internalizing private rents even when they are spread over

[1] An extreme form of basic knowledge is a General Purpose Technology (GPT). Helpman and Trajtenberg (1997) show that the economic value of GPT arrives only after follow-up developments take place (see also Bresnahan and Trajtenberg (1995) and Rosenberg and Trajtenberg (2004)).

many fields. The empirical prediction of this competing hypothesis is that the negative correlation between the generality of knowledge and internalization of spillovers will be mitigated if not completely muted (since firms that conduct basic research are those that are better able to technologically internalize it)[2].

Distinguishing between these two hypotheses is extremely important for analyzing the classical problem of underinvestment in basic research. Prior studies have adopted a production function approach to measure the returns to basic research and whether it is endogenously determined (Griliches (1986), Mansfield (1980)). The main finding coming from this literature is that there is a very large premium at the firm level on basic research. This is inconsistent with the hypothesis that firms choose between basic and applied research, since if this were the case we would expect private returns from both types of research to be equalized[3]. The present paper develops a complementary dynamic approach for studying the endogeniety of basic knowledge and the extent it is privately rewarded. The main advantage of this new approach is that it enables capturing the dynamic payoff associated with knowledge when innovation is sequential. The dynamic payoff of internalizing the follow-up developments of knowledge would be higher for basic knowledge. Yet, for basic knowledge such internalization is also less likely to occur.

Spillovers introduce two countervailing forces with respect to the incentive to innovate: on the one hand, spillovers encourage future research, but on the other hand, they discourage current research due to obsolescence of private rents (Schumpeter (1942), Aghion and Howitt (1992) and Segerstrom, Anant, and Dinopoulos (1990)). Most of our understanding of the incentive to innovate (of both early inventors and their followers) lies on how these two forces are reconciled. The conflict between these two forces is believed to be much stronger for basic knowledge (Nelson (1959), Arrow (1962)).

[2]There may still be a negative correlation between generality and technological internalization even if generality if endogenously chosen by firms due to the stochastic nature of research.

[3]In case there is a premium risk for basic research, private returns to basic research could be higher than to applied research, even when the type of research is endogenously determined. Yet, the estimated basic research premium is too high to represent such risk: Griliches (1986) reports that the private return to basic research is eight times the private returns to applied research.

The major contribution of this paper is in developing a novel empirical methodology, based on patents and citations, for testing whether appropriability is lower for basic knowledge in a dynamic framework of sequential innovation. Spillovers are measured as the sequential developments of knowledge coming from outside the inventing firm. Based on a complete characterization of the flow of knowledge underlying these spillovers, it can be determined whether they feed back into the inventing firm. This feeding back of spillovers is defined as *technological internalization*. Essentially, two types of spillovers are distinguished: *Internalized* and *Externalized*. Internalized spillovers are spillovers that feed back into the dynamic research of the inventing firm, whereas Externalized spillovers do not. Technological internalization is defined as the share of Internalized spillovers created by the invention. To the extent technological internalization is a channel through which private rents are appropriated by (early) inventors, the present paper adds a great deal to our understanding of the incentive to invent basic knowledge in a dynamic framework where private rents depend on external follow-up research.

In addition to technological internalization, the inventing firm can internalize private rents through a contractual channel. The literature has studied the theoretical aspects of contractual internalization in a framework of sequential innovation, mainly as a mechanism through which rents are shared between early innovators and their followers. Green and Scotchmer (1995), Scotchmer (1996) and Chang (1995) study the theoretical aspects of the effect of a second-generation invention on the rents captured on the first-generation invention. O'Donoghue (1998) study the inventive step requirement in patent protection and show how the inventive step can be chosen to minimize the trade-off between encouraging current research and discouraging future research.

Yet, a large body of research shows that contractual internalization can fail to provide sufficient private rents when transaction costs of contracting are high[4] . In this case, private rents could still be captured through the technological channel of internalization.

However, the theoretical and empirical literature has not yet investigated technological

[4]E.g., Eisenberg (1998), Grindley and Teece (1997), Hall and Ziedonis (2001), Lanjouw and Schankerman (2004) and Schankerman and Noel (2006).

internalization. Focusing on the technological channel of internalization is especially important in light of the role of basic knowledge in creating "pure" spillovers. According to the endogenous growth literature, "pure" spillovers, which occur when knowledge transfers freely across inventors and inspires follow-up research in numerous fields, allow the economy to depart from decreasing returns in the production of knowledge and achieve sustained economic growth (Romer (1986), Grossman and Helpman (1991)). Contractual internalization hinders the free access to knowledge (since the receivers of knowledge have to incur usage costs). Hence, contractual internalization should diminish economic growth, through restricting the increasing returns in knowledge production. Yet, under technological internalization, "pure" spillovers should not diminish in any obvious way, since private rents can be captured without limiting future research.

Finding a negative correlation between technological internalization and the "basicness" of knowledge would imply one of two things: either the incentive to invent basic knowledge is reduced (i.e., current research diminishes), or that the inventing firms must adopt the contractual channel to secure private rents (future research diminishes due to reduction in "pure" spillovers, whereas current research may diminish as well in case contractual internalization does not sufficiently reward the inventing firm). In both cases, lower technological internalization would reduce the pace of innovation and growth.

Henderson, Jaffe and Trajtenberg (1997)[5] show that patents and citations data can be used to measure the generality of knowledge: knowledge embodied in a patent is more *general* if the citations the patent receives spread over more technology fields. The present paper adopts generality as the main characteristic of basic knowledge and tests its correlation with technological internalization.

The essence of my empirical methodology for measuring technological internalization is as follows: knowledge is identified as a patent and knowledge flow is identified as a patent citation[6]. For each patent in the sample a "family-tree" is constructed, based on

[5] See also Hall and Trajtenberg (2005).

[6] Prior studies that empirically identified citations as knowledge flows are Jaffe, Henderson and Trajtenberg (1993), Caballero and Jaffe (1993) and Jaffe and Trajtenberg (1999).

the citations the patent receives. Figure 1 illustrates this methodology for a simple case of a sequence of three patents. Assume patent j cites patent i and patent k cites patent j. Hence, the "family-tree" of patent i includes both patent j and patent k, where, patent j is the 'child' of patent i and patent k is the 'grandchild' of patent i. Given this "family-tree", invention k is classified as an offspring of invention i, even though knowledge did not transfer directly from invention i to invention k. Applying this method to a high-order sequence of citations allows tracing the trajectory knowledge has followed, while spreading across inventions and firms. Based on these trajectories, it can be determined whether knowledge that leaves the inventing firm and is further advanced by other firms will have been reabsorbed by the inventing firm in a future period. (e.g., if patents i and k are held by the same firm whereas patent j is owned by another firm, the spillovers created by invention i are technologically internalized by the inventing firm).

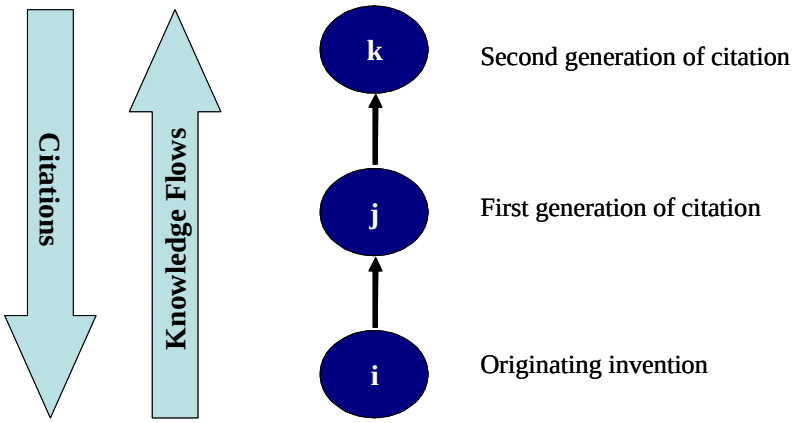

Figure 1: The "family-tree" of invention i

Based on the above methodology of identifying the diffusion pattern of knowledge, technological internalization is measured. An econometric specification of the effect of generality on technological internalization is estimated for all patents held by the largest 500 inventing firms in the US. There is strong evidence of a negative effect of generality on technological internalization. This finding supports the hypothesis that basic research is not endogenously chosen by firms and is less likely to be privately rewarded under the

dynamic consideration of technological internalization.

Finally, a market value equation is estimated to confirm that technological internalization is an important channel through which private rents are appropriated. The estimates from the value function are then used to quantify the impact of generality on private returns. A one standard deviation increase in technological internalization raises the market valuation of an additional one dollar spent on R&D by 50 percent, evaluated at the mean. Based on this estimate, a one standard deviation increase in the generality of knowledge lowers private returns by 4.8 percent. Moving from the most specialized to the most general knowledge (the two extreme points on the generality spectrum) lowers private returns by 15.3 percent, evaluated at the mean.

In summary, a novel empirical methodology is developed to measure internalization of private rents via a technological channel through which an inventor reabsorbs its knowledge that is "spilled" to other agents. This measure of appropriability is used to test the historical concern that basic knowledge is less privately rewarded. The econometric findings support this concern.

The rest of this paper proceeds as following: section 2 presents the methodology for measuring technological internalization, section 3 shows how generality is measured, section 4 describes the data, section 5 reports the findings and section 6 concludes.

2. Measuring technological internalization

This section describes the conceptual and empirical issues regarding measuring technological internalization. I start by showing how the technological contribution of an invention is identified. Then, spillovers are defined as the external exploitation of the technological contribution of the invention. Finally, it is shown how it is determined whether spillovers feed back into the inventing firm to generate technological internalization.

2.1. Technological contribution

Technological contribution is measured in two dimensions: the number of *lines of research* the invention originates and the 'quality' of these lines of research. A line of research is defined as *a sequence of inventions, where every invention is a follow-up development of its immediate ancestor.* This sequence of inventions is required to be unique over a given time period, i.e., not to be fully contained in a longer sequence of inventions. Define the first invention in the line of research as an *originating invention.* A line of research is assumed to be of a higher 'quality', if the number of subsequent developments of the originating invention along the line of research is higher.

More formally, the technological contribution of invention i, TC_i, is computed as the 'quality'-weighted count of the lines of research invention i originates, as following[7]:

$$TC_i \;=\; \sum_{k \in K_i} LR_k \times Q_k \tag{2.1}$$

Where, K_i is the set of lines of research originated in invention i, k indexes lines of research in this set, LR_k is a dummy that receives the value 1 for line of research k and zero otherwise, and Q_k is the 'quality' of line of research k, as measured by the number of inventions the line of research includes[8].

Applying this formulation to the diffusion patterns in figure 2 yields:

$$TC_A^1 \;= (1 \times 3) = 3 \tag{2.2}$$

[7] Belenzon (2005) shows that this method of measuring technological contribution is equivalent to an alternative approach of counting the number of offspring inventions and weighing each one by the number of direct citations received.

[8] Simply counting the number of inventions along a line of research may be an overestimate of the technological contribution of the originating invention. A subsequent invention which is a high generation of development of the originating invention is more likely to have benefited from other prior subsequent inventions along the line of research. Thus, I always discount every generation by a discount factor of δ per generation (which is assumed to be 15 percent), thus, $Q_k = \sum_{j=1}^{J} \delta^{j-1}$, where, J is the number of offspring inventions in line of research k. Since the choice of the discount factor is arbitrary, other values of δ are experimented with as robustness tests.

Where, TC_A^1 is the technological contribution of invention A under pattern 1. The term 1 in the brackets represents the singleton line of research $A \rightarrow B \rightarrow C \rightarrow D$ that is adjusted by its 'quality', which is 3 (since it includes three subsequent developments of invention A: B, C and D).

Similarly, the technological contribution of invention A under diffusion pattern 2, TC_A^2, is:

$$TC_A^2 = (1 \times 2) + (1 \times 2) = 4 \tag{2.3}$$

The term 1 in the first brackets represents the line of research $A \rightarrow B \rightarrow C$ that is adjusted by its 'quality', which is 2 (since it includes two subsequent developments of invention A: B and C). The term 1 in the second brackets represents the line of research $A \rightarrow B \rightarrow D$ that is adjusted by its 'quality', which is 2 as well (since it includes two subsequent developments of invention A: B and D).

From this is concluded that the technological contribution of invention A under diffusion pattern 2 is greater than its technological contribution under diffusion pattern 1 (intuitively, under both patterns of diffusion the number of subsequent developments is equal. However, there are more research opportunities under pattern 2, as indicated by the number of lines of research).

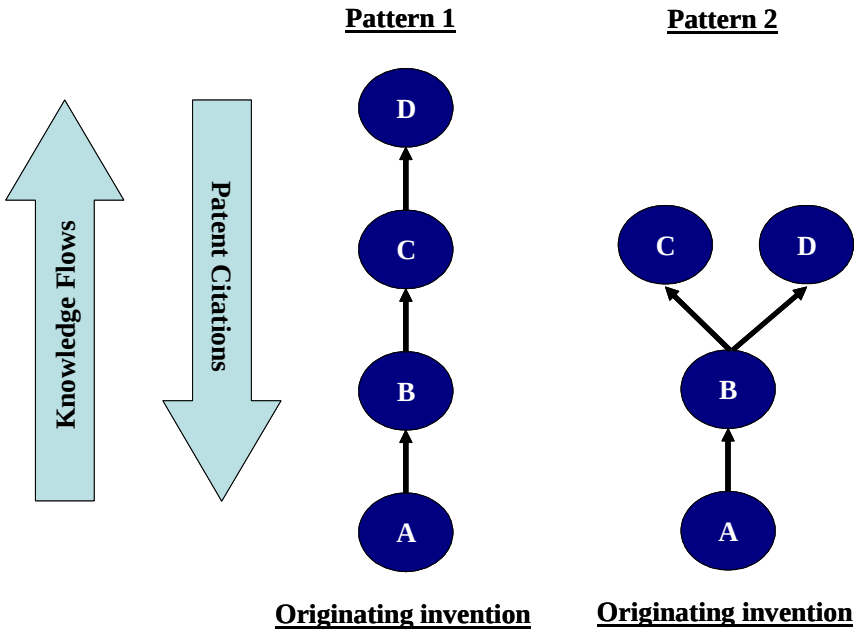

Figure 2: Technological contribution

Figure 2: *Circles in this figure represent inventions and arrows represent the direction of knowledge flow. Pattern 1 illustrates a singleton path of knowledge flow, which is* $A \rightarrow B \rightarrow C \rightarrow D$*, while diffusion pattern 2 illustrates two unique paths of knowledge flows, which are* $A \rightarrow B \rightarrow C$ *and* $A \rightarrow B \rightarrow D$*. Determining the technological contribution of invention* A *under the two diffusion patterns requires weighing these lines of research by their 'quality', by measuring their length in terms of the number of inventions they include.*

2.2. Spillovers

Spillovers are defined as the external exploitation of the technological contribution of an invention, where *external* refers to the set of firms that are different from the inventing firm. Following this definition, spillovers are measured as the number of external inventions along the lines of research the originating invention inspires.

For illustration, it is useful to examine a slightly more complicated diffusion pattern, as shown in figure 3. Capital letters represent inventions, where arrows represent the direction

of knowledge flow. This figure plots the diffusion pattern of the originating invention A, where the offspring inventions are B, C, D, E, F, G, H, I and J. To complete the presentation, the shape of each capital letter represents a different firm, i.e., a circle firm (the inventing firm), a triangle firm and a square firm.

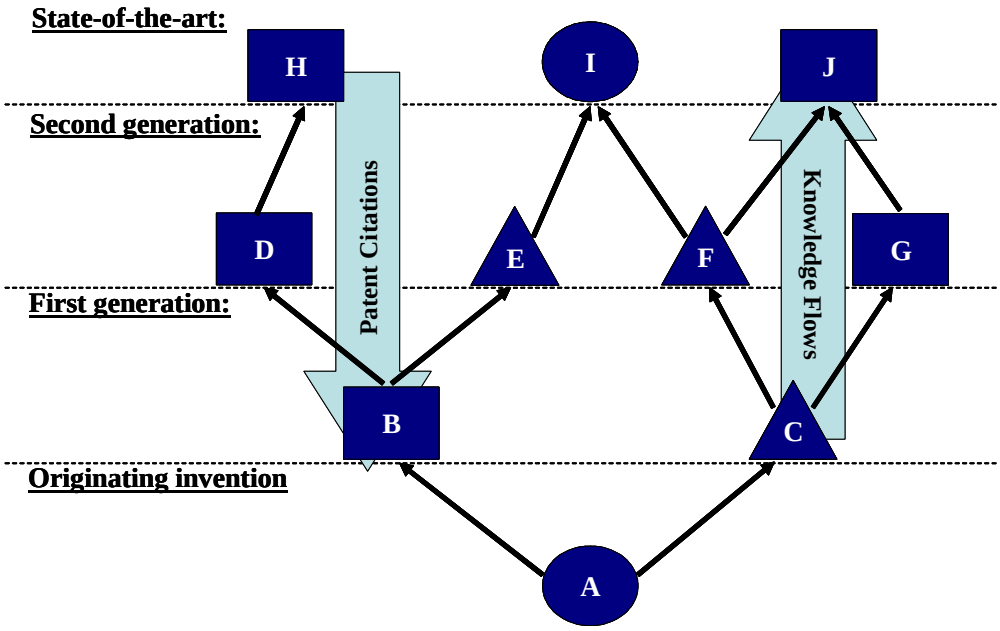

Figure 3: Measuring spillovers

Figure 3: *This figure illustrates the diffusion pattern of the originating invention A. Inventions are represented by a capital letter, while the firm that owns the inventions is represented by a shape (e.g., the inventing firm is the circle, since it owns the originating invention A). I define the spillovers created by invention A, given this diffusion pattern, as the number of inventions that are owned by the square and triangle firms (all the firms in the figure which are different from the inventing firm) along the lines of research invention A originates.*

Following the methodology presented above, in order to measure the technological

contribution of invention A, we need to identify the lines of research invention A originates and weigh them by their 'quality'. Since a line of research is defined as a singleton sequence of subsequent developments of the originating knowledge, there are five such lines of research: $A \rightarrow B \rightarrow D \rightarrow H$, $A \rightarrow B \rightarrow E \rightarrow I$, $A \rightarrow C \rightarrow F \rightarrow I$, $A \rightarrow C \rightarrow F \rightarrow J$ and $A \rightarrow C \rightarrow G \rightarrow J$. The technological contribution of invention A following equation (2.1) is given by:

$$TC_A = (1 \times 3) + (1 \times 3) + (1 \times 3) + (1 \times 3) + (1 \times 3) = 15 \tag{2.4}$$

Since spillovers are defined as the external inventions that compose the lines of research an invention originates, they are formulated as:

$$Spillovers_i = \sum_{k \in K_i} LR_k \times S_k \tag{2.5}$$

Where, i is an originating invention, K_i is the set of lines of research invention i originates, k indexes lines of research in this set, LR_k is a dummy that receives the value 1 for line of research k and zero otherwise and S_k is the number of external inventions included in line of research k. Following this formulation, the spillovers created by invention A are:

$$Spillovers_A = (1 \times 3) + (1 \times 2) + (1 \times 2) + (1 \times 3) + (1 \times 3) = 13 \tag{2.6}$$

Where, the second and third terms, (1×2) and (1×2), correspond to the fact that invention I is owned by the inventing firm. Thus, invention I is excluded from the spillovers measure for invention A (the spillovers along lines of research $A \rightarrow B \rightarrow E \rightarrow I$ and $A \rightarrow C \rightarrow F \rightarrow I$ are based only on inventions B, E, C and F)[9].

Finally, I aim at distinguishing between two types of spillovers: spillovers that contribute to the dynamic research of the inventing firm and spillovers that do not.

[9] In some patterns of diffusion, the first subsequent development of the originating knowledge is done by the inventing firm (which is identified as a *self-citation*). Hence, knowledge does not immediately spread to other inventors. In this case, the 'in-house' subsequent development is not measured as spillovers (where spillovers along such lines of research occur only if in a future generation knowledge leaves the boundaries of the inventing firm).

2.3. Internalized and Externalized lines of research

Two types of lines of research are identified: the first type is lines of research where the originating knowledge leaves the inventing firm and returns to this firm after having been further developed by other firms. The second type is lines of research where the originating knowledge leaves the inventing firm and does not return. Spillovers along the former type are *internalized* in the dynamic research of the inventing firm and, therefore, these lines of research are defined as *Internalized lines of research*. However, spillovers along the latter type do not contribute to the dynamic research of the inventing firm, therefore, these lines of research are defined as *Externalized lines of research*.

Hence, the spillovers of an invention can be written as:

$$Spillovers_i = \sum_{j \in Internalized_i} LR_j \times S_j + \sum_{t \in Externalized_i} LR_t \times S_t \tag{2.7}$$

Where i denotes an originating invention, $Internalized_i$ is the set of Internalized lines of research originated in invention i, $Externalized_i$ is the set of Externalized lines of research originated in invention i, j indexes lines of research in the $Internalized_i$ set and t indexes lines of research in the $Externalized_i$ set. I define the first term in the right-hand-side of equation (2.7) as $IntSpill_i$ and the second term in the right-hand-side of equation (2.7) as $ExtSpill_i$. Thus, equation (2.7) becomes:

$$Spillovers_i = IntSpill_i + ExtSpill_i \tag{2.8}$$

Technological internalization, $IntShare_i$, is defined as the ratio between $IntSpill_i$ and $Spillovers_i$.

To illustrate this decomposition, it is useful to refer back to figure 3. Out of the five lines of research that invention A originates, two are Internalized and three are Externalized. The set $Internalized_A$ is:

$$Internalized_A = \{A \rightarrow B \rightarrow E \rightarrow I, A \rightarrow C \rightarrow F \rightarrow I\}$$

Similarly, the set $Externalized_A$ is:

$$Externalized_A = \{A \rightarrow B \rightarrow D \rightarrow H, A \rightarrow C \rightarrow F \rightarrow J, A \rightarrow C \rightarrow G \rightarrow J\}$$

Given this decomposition, $IntSpill_A = (1 \times 2) + (1 \times 2) = 4$ (two external inventions in the first line of research and two external inventions in the second line of research in the $Internalized_A$ set). Similarly, $ExtSpill_A = (1 \times 3) + (1 \times 3) + (1 \times 3) = 9$ (three external inventions in each of the three lines of research in the $Externalized_A$ set). Thus, $IntShare_A$ is $\frac{4}{13}$.

2.4. Empirical methodology

Inventions are empirically identified as patents and knowledge flows as citations (where knowledge flows from the cited patent to the citing patent). Patents and citations data contain significant noise and bias[10]. Nonetheless, these data also offer unique information on the diffusion pattern of knowledge and sequential innovation, which I believe to be extremely useful for exploring the ideas developed in this paper.

Hence, the inventions in figures 2 and 3 are empirically identified as patents, whereas arrows are empirically identified as citations (e.g., an arrow from invention A to invention B in figures 2 and 3 reflects the fact that patent B cites patent A). The task I am facing is to effectively draw figure 3 for the sample of originating inventions[11].

A unique line of research is empirically identified as a *singleton sequence of citations* (where, each patent cites its direct ancestor). A sequence of citations is defined as singleton, if it is not fully contained in a longer sequence of citations for the given time period being explored. After extracting the lines of research for the sample of originating patents, each line of research is classified as either Internalized or Externalized[12].

[10] See, for example, Trajtenberg (1990) for the potential bias in patents as indicators for innovation output, and Trajtenberg, Jaffe and Fogarty (2001) for a study on the noise component in citations as indicators for knowledge flows.

[11] The design of this sample is explained below.

[12] The reader who is familiar with the economics of patents literature can find the definition of an

The period for which lines of research are constructed is restricted to 15 years after the grant year of the originating patent. For example, for a patent that was granted in 1975, the youngest patents in all the lines of research it originates cannot be granted after 1990. Further, citations along a line of research are added as long as the line of research has not already been classified as Internalized[13]. Thus, this methodology extracts all the unique trajectories where knowledge had left the boundaries of its inventor and returned to these boundaries in a time period of 15 years after the knowledge had been created[14], as well as all the unique trajectories where knowledge had left the boundaries of the inventing firm and did not return to these boundaries in the same time period[15].

3. Generality of patents

The main characteristic of basic knowledge is the extent it spurs follow-up research in many technology fields. Following Trajtenberg, Henderson and Jaffe (1997), patents are argued to be more general if the citations they receive spread over a larger number of fields.

The generality of patent i, denoted by G_i, is computed as one minus the HHI index

Internalized line of research similar to a self-citation. A self-citation is the case where a firm develops its prior knowledge directly (the first generation of citation the patent receives comes from the inventing firm itself). An Internalized line of research is the case where the firm *indirectly* develops its prior knowledge, after it has been developed by other firms. Thus, an Internalized line of research is a unique *indirect self-citation*, which I associate with a higher appropriability, as the existing literature does with self-citations (e.g., Hall, Jaffe and Trajtenberg (2005)).

[13] E.g., consider the Internalized line of research $A \rightarrow B \rightarrow E \rightarrow I$ that is presented in figure 3. Assume that patent I is cited by patent K, such that this line of research becomes $A \rightarrow B \rightarrow E \rightarrow I \rightarrow K$. The imposed restriction implies that only the line of research $A \rightarrow B \rightarrow E \rightarrow I$ will be extracted for the originating patent A.

[14] Since I refer to the grant year of the patent and not to its application year, the creation date of the patented knowledge is actually earlier. However, my algorithm builds on the fact that a citing patent cannot be cited before it cites. This crucial feature of the data can be exploited only by referring to the grant year of the patent (see Belenzon (2005) for detail on the algorithm).

[15] It is important to note that this methodology incorporates the case where knowledge is first developed sequentially 'in-house' by the inventing firm (i.e., self-citations). In numerous cases the inventing firm develops the first follow-up inventions of the originating knowledge. In such lines of research knowledge leaves the boundaries of the inventing firm via a higher order generation of citation. These lines of research are classified as Internalized or Externalized following the same criterion described above.

of concentration across the fields that cite patent i:

$$G_i = 1 - \sum_n \left(\frac{CR_{in}}{CR_i} \right)^2 \tag{3.1}$$

Where, n denotes citing fields, CR_{in} is the number of citations received by patent i from patents in field n and CR_i is the total number of citations received by patent i. Self-citations are excluded from G_i, due to the interest in characterizing follow-up research that is done outside the boundaries of the inventing firm[16]. The main technology breakdown used in the econometric analysis is based on the three-digit US Classification (Nclass), which includes 400 fields. G_i is based on citations received during the period 1975-1999.

Hall (2002) shows that G_i is downward biased in case patent i receives a small number of citations and suggests the following bias-corrected measure:

$$\widehat{G_i} = \left(\frac{CR_i}{CR_i - 1} \right) G_i \tag{3.2}$$

Since $\widehat{G_i}$ is based on technology field definitions, it is highly sensitive to measurement error in drawing the boundaries between fields. For example, in case in the Drugs sector, technology fields are defined more coarsely compared to the Computers sector, it is more likely for a patent to be more general in the Computers sector when the propensity of citations is stronger within sectors compared to between sectors. In order to mitigate this concern, $\widehat{G_i}$ is also constructed based on the following alternative technology classifications[17]: International-Patent-Class (742 cells), Sub- International-Patent-Class (3008 cells), Hall, Jaffe and Trajtenberg (HJT) subcategories (36 cells) and Manufacturing Industry *SIC-IPC* classification (37 cells).

Finally, knowledge should be more general if it transfers to fields that are more technologically remote from the field in which the knowledge was originally invented. Later in the paper, a more refine $\widehat{G_i}$ measure is developed to take into account the technological proximity between the citing fields and the field of the cited patent.

[16] The empirical results are robust to including self-citations in the construction of G_i.

[17] See Hall and Trajtenberg (2004).

4. Data

Patents and citations data are taken from the U.S. Patent and Trademark Office from the NBER archive. The sample of originating patents includes all cited patents held by the largest 500 patenting firms in the US between 1969 and 1980[18]. It is required that every firm remains active during the complete period for which the sequences of citations are constructed leaving the largest 492 inventing firms (all of which are active up to 1995, which is the last year an offspring patent can be added into a line of research). The set of originating patents includes 104,694 patents[19].

The sample of citing patents that participate in the sequences of citations includes about 600,000 patents that are held by all US Compustat firms in the USPTO[20]. These patents make around 1.7 million citations (either to the originating patents or to other citing patents[21]). Based on these citations, 13,107,634 lines of research (singleton sequences of citations) are extracted, which are originated in 97,921 inventions. 6,773 patents that appear in the initial set of originating patents do not originate Internalized or Externalized lines of research (these patents originate lines of research in which all the follow-up developments of the originating invention are done 'in-house'). 999,718 lines of research are classified as Internalized and are originated in 29,964 patents, while the remainder 12,107,916 lines of research are classified as Externalized and are originated in 97,212

[18] The year 1969 is the earliest year for which there is citations information for the patents held by the firms in the sample. Also, in practice I could extract the diffusion pattern of patents that were granted up to 1985, since the citations data goes up to 1999. However, there is a huge spike in the number of citations in 1995 (see figure A3), where the number of citations rises by around 800,000 in the period 1995-1999. In addition to the feasibly of extracting sequences of citations from these huge data, there is also a concern that the explosion in citations in this period is not associated with stronger learning and sequential innovation, but with changes in the patenting behavior of firms, which could contaminate the results.

[19] The set of originating patents includes 45 percent of all cited patents between 1969 and 1980 that are held by US Compustat firms that were matched to the USPTO by Hall, Jaffe and Trajtenberg (2001).

[20] Hall, Jaffe and Trajtenberg (2001) matched 2466 US Compustat firms to the USPTO. The citing patents of all these firms are allowed to take part in constructing the patterns of diffusion of the originating patents. The sample of citing patents includes about 30 percent of all citing patents in the USPTO (and 50 percent of the citing patents where the main inventor is a US resident).

[21] Where 599,884 patents make 1,760,143 citations to 573,373 patents in the sample.

patents[22] .

Detail on the algorithm developed to construct the diffusion data is provided in Belenzon (2005).

Table 1 describes the variation of lines of research across technology sectors and time. The largest number of lines of research per citation received by an originating patent is in the "Electrical and Electronics" sector. This may indicate a high technological complexity in this sector, where complexity refers to the various distinct ways along which knowledge can be sequentially developed. 7.6 percent of the lines of research are Internalized. This share appears to be rather stable over time, with an exception in "Drugs and Medicals". In the period 1978-1980 there is a large drop in the share of Internalized lines of research in this sector, which may be associated with the Biomed revolution that took place at the end of the 70's. I plan to investigate this separately in a future research.

[Table 1 about here]

Table A1 provides summary statistics for the main patent variables. The average technological internalization is 4.7[23] (i.e., on average, 4.7 percent of the spillovers created by a patent are defined as Internalized). The unconditional correlation between $IntShare$ and $\widehat{G_i}$ is -0.063 (with $p < 0.01$).

5. Estimation

The baseline specification links technological internalization to generality as following:

$$IntShare_i = \beta_0 + \beta_1 \widehat{G_i} + \beta_2 \, \text{Cites}_i + Z_i' \beta_3 + \tau_i + \phi_i + \eta_i + \epsilon_i \tag{5.1}$$

Where, i denotes the originating patent, $Cites_i$ is the number of citations patent i receives (over the period 1975-1999), Z_i is a vector of additional controls described below, η_i is a complete set of dummies for the inventing firms (the owner of patent i), τ_i is a

[22] The remaining 709 originating patents inspire only Internalized lines research (thus, all the subsequent generations of developments are done by the inventing firm).

[23] Belenzon (2005) shows that this percentage is rather stable over time and across fields.

complete set of dummies for the grant year of patent i, ϕ_i is a complete set of dummies for the field of patent i and ϵ_i is an *iid* error term.

Cites are added as a control for $\widehat{G_i}$, since both measures are based on counts of forward citations. As mentioned above, $\widehat{G_i}$ is likely to be higher when a patent receives more citations[24]. In case *Cites* has a negative effect on $IntShare$, β_1 will be downwards biased.

The set of grant year dummies, τ_i, is included since patents are pooled from different time periods (1969-1980). The main variable that is likely to substantially vary over time is *Cites* (see figure A3). This time trend may cause patents that are granted in later periods to appear on average more general, if $\widehat{G_i}$ is positively correlated with *Cites*.

The set of field dummies control for technology location: different fields may systematically vary in terms of patterns of diffusion, which could affect both $IntShare$ an $\widehat{G_i}$.

A complete set of firm dummies is also included. Although the regression is at the patent level, the underlying level of technological internalization is determined at the firm level and should be affected by firm-specific attributes. To the extent these attributes are correlated with $\widehat{G_i}$, β_1 would be biased. For example, firms that are more specialized in research could be better at internalizing the spillovers of their knowledge. If more specialized firms invent less general knowledge (as would be expected under the hypothesis that basic research is endogenously determined), β_1 will be downward biased.

Z_i includes three additional controls: *Complexity*, *PatShare* and *PatCon*.

$Complexity_i$ - $Complexity_i$ measures the degree to which the fields that cite patent i are diversified in terms of lines of research. Fields that include a higher average number of lines of research are argued to be more complex (as there are more unique ways to sequentially develop knowledge). $Complexity_i$ is calculated as following:

$$Complexity_i = \sum_n \omega_n \times Com_n$$

Where, n denotes technology fields that cite patent i, ω_n is the share of citations

[24]The bias-correction used in this paper aims to eliminate the downward bias in G_i when a patent receives only few citations. The correlation between $\widehat{G_i}$ and Cites is 0.07, compared to 0.23 for G_i (uncorrected).

patent i receives from technology field n and Com_n is the technological complexity in field n. Com_n is defined as the average number of lines of research per citation received by an originating patent (see table 1) and is based on the Nclass level.

A negative correlation between $IntShare_i$ and $Complexity_i$ would imply that it is harder to internalize own spillovers in case these spillovers are spread over more lines of research. This may indicate that specialization in research occurs not only between fields, but also within fields across lines of research. In the absence of within-field specialization, $Complexity_i$ should not negatively affect the degree of technological internalization.

$PatShare_i$ - $PatShare_i$ measures the overlap between the fields that cite patent i and the patent distribution of the inventing firm. A higher $PatShare_i$ implies a higher concentration of the research activity of the inventing firm across the citing fields. $PatShare_i$ is expected to be positively correlated with $IntShare$: the inventing firm would find it easier to internalize own spillovers where they are concentrated across fields to which the research of the inventing firm is more directed. $PatShare_i$ is calculated as the *HHI* index for the share of the inventing firm's patents in the technology fields that cite patent i (weighted by the share of citations patent i receives from every citing field):

$$PatShare_i = \sum_n \omega_n \times (Share_n)^2$$

Where, $Share_n$ is the share of patents the inventing firm has in technology field n and ω_n is as defined above.

$PatCon_j$ - $PatCon_j$ measures the research diversification of firm j as the *HHI* index of the concentration of the firm's patents across technology fields, as following:

$$PatCon_j = \sum_k (Share_k)^2$$

Where, j denotes the inventing firm, k denotes fields firm j operates in and $Share_k$ is the share of patents firm j has in field k out of the total patents firm j has (computed over the period 1969-1999). Since $PatCon$ is a firm-level measure (i.e., does not vary across patents within firms), its effect will not be identified in the presence of firm fixed-effects,

which are widely used in the econometric analysis. Yet, introducing $PatCon_j$ is interesting with regard to its correlation with $\widehat{G_i}$. To the extent firms decide the level of generality of their knowledge, $PatCon$ and $\widehat{G_i}$ would be negatively correlated: more specialized firms will choose more specialized knowledge.

5.1. Results

Table 2 summarizes the main estimation results. In column 1, equation (5.1) is estimated without firm fixed-effects (i.e., conditioning on cites received, fields dummies, year dummies and a dummy for $IntShare$ equals zero). The coefficient on $\widehat{G_i}$ (β_1) is negative and significant. This implies that patents that are cited by more fields exhibit less technological internalization, which supports the main hypothesis of this paper.

In column 2, $PatCon_j$ is added. The coefficient on $PatCon$ is positive and significant: a higher concentration in research increases technological internalization. The positive effect of $PatCon$ on $IntShare$ is a consistent explanation to the finding reported by Hall and Ziedonis (2001) of an increased specialization of entrant firms in the "Semiconductors" industry. In this industry sequential innovation plays a major role and the dynamic consideration of technological internalization is likely to be important. Furthermore, β_1 falls in absolute value when controlling for $PatCon$. This fall indicates a negative correlation between $PatCon_j$ and $\widehat{G_i}$ (the correlation is -0.147 with a $p\ value < 0.01$), i.e., firms that have a more diversified research capabilities invent more general knowledge. This is consistent with the hypothesis that firms choose the level of generality of their knowledge: in order to technologically appropriate significant private rents on general knowledge the inventing firm would need to conduct follow-up research in numerous fields. Knowing this, firms with more diversified research capabilities will choose to invent more general knowledge. Yet, β_1 remains negative and significant also after controlling for research diversification.

In column 3, a complete set of firm dummies is added to control for the attributes of firms that can affect both $IntShare$ and $\widehat{G_i}$. In this specification, only the variation across patents within inventing firms is exploited. With firm fixed-effects β_1 continues to

increase in absolute value, however, it remains negative and significant. Based on this specification, at the mean, moving from the 25th percentile to the 75th percentile in $\widehat{G_i}$ lowers $IntShare$ by 9 percent[25].

When exploiting only the variation across patents within firms, there may still be a patent-level variation in attributes that are correlated with both $\widehat{G_i}$ and $IntShare$. In case knowledge "spills"to technology fields that are more complex, where complexity is measured as the technology field average number of lines of research originated in a patent, it should become harder for the inventing firm to internalize a larger share of the spillovers it creates. $Complexity$ is added in column 4. $Complexity$ has a negative and significant effect on $IntShare$, as expected. Finding this negative effect implies that diversification in research is evident not only between technology fields but also within technology fields across lines of research (otherwise, technological internalization would not be harder to achieve when citing fields have a higher average number of lines of research).

In order to illustrate the range of the effect of $Complexity$, consider the following calculation: suppose knowledge "spills" only to one technology field ($\widehat{G_i}$ is zero). Consider two extreme fields in term of their complexity: Nclass 438 ("Semiconductor Device Manufacturing: Process"), which has a complexity measure of 112.3[26], and Nclass 139 ("Textiles: Weaving"), which has a complexity measure of 5.1. $IntShare$ would be higher in the latter pattern of diffusion by about 60 percent compared to the former[27].

Technological internalization should be easier to achieve in case the inventing firm is already active in research in the citing fields. To test this, column 5 adds $PatShare$, which measures the overlap between the research activity of the inventing firm and the fields that cite its knowledge. As expected, $PatShare$ has a positive and significant effect on $IntShare$. Thus, the extent the inventing firm is active in the fields its knowledge "spills" to, technological internalization would be higher. Evaluated at the mean, a one

[25] The predicted $IntShare$ (evaluated at the mean) is 5.077 when $\widehat{G_i}$ is at the 25th percentile. $IntShare$ drops to 4.594 when $\widehat{G_i}$ increases to the 75th percentile.

[26] I.e., 112.3 lines of research per citation received by an originating patent.

[27] When knowledge "spills" to Nclass 438, the predicted $IntShare$, evaluated at the mean, is 3.365, compared to 5.616, when knowledge "spills" to Nclass 139.

standard deviation increase in $PatShare$ raises $IntShare$ by 15 percent (from 4.7 percent to 5.4 percent).[28]

[Table 2 about here]

5.2. Robustness tests

5.2.1. Technological proximity between fields

$\widehat{G_i}$ does not take into account the 'distance' knowledge travels across fields: knowledge would be more general if it is cited by many fields that are also more technologically remote from the cited field. In this section $\widehat{G_i}$ is refined by weighting the citing fields according to their technological distance from the field of patent i, as indicated by the propensity of citations (fields that are closer to the field of patent i will receive a lower weight)[29]. Following Caballero and Jaffe (1993) and Jaffe and Ttajtenberg (1999), the propensity of citations is estimated by aggregating patents into "cells", based on characteristics of the citing and cited patents. The following equation is estimated by nonlinear least-squares:

$$\rho_{ss'tT} = \alpha_{ss'}\alpha_s\alpha_{s'}\alpha_T\alpha_t \exp(-\beta_1(T-t))\left(1-\exp\left(-\beta_2(T-t)\right)\right) \tag{5.2}$$

Where, s denotes the field of the citing patent, s' denotes the technology field of the cited patent, T is the grant year of the citing patent and t is the grant year of the cited patent. s includes 36 fields based on the HJT subcategory classification and s' includes the 6 main fields. $\alpha_{ss'}$ denotes a complete set of 215 dummies for all pair-wise combinations of the citing and cited fields ($36 \times 6 - 1$), α_s is a complete set of dummies for the citing fields (35 dummies), $\alpha_{s'}$ is a complete set of dummies for the cited technology fields (5

[28] Moreover, patents that create more spillovers could also be more general. In case spillovers are negatively correlated with $IntShare$, β_1 will be downward biased. In order to test this, I also add $Spillovers$ (the sum of $IntSpill$ and $ExtSpill$) into the right-hand-side of equation (5.1). β_1 increases to -1.033 with a standard error of 0.128, where there is no important change in the other coefficients. The effect of $Spillovers$ is negative and significant: at the mean, a one standard deviation increase in $Spillovers$ lowers $IntShare$ by 18 percent.

[29] It is also important to weight citing fields by the propensity to cite since larger fields are more likely to cite a given patent. In case a patent is surrounded by large technology fields, it can appear to be general simply because there is a higher probability it will be cited outside its own field.

dummies), T is a complete set of year dummies for the citing patent (24 dummies for the period 1975-1999) and t is a complete set of dummies for the grant year of the cited patents (7 dummies for the period cohorts of the cited patents[30]). $\rho_{ss'tT}$ is computed as:

$$\rho_{ss'tT} = \frac{C_{ss'tT}}{P_{sT}P_{s't}} \tag{5.3}$$

Where, $C_{ss'tT}$ is the number of citations from the citing field s at year T to the cited field s' at year t, P_{sT} is the number of citing patents in the cell and $P_{s't}$ is the number cited patents in the cell[31].

The main estimation results of equation (5.2) are summarized in table A5, which reports the estimated set of coefficients $\alpha_{ss'}$, $\widehat{\alpha_{ss'}}$. It is clearly evident that the propensity of citations is much stronger within fields in the same main technology sector, which implies that knowledge is less likely to travel across the boundaries of main technology fields. This highlights the sensitivity of G_i to measurement error in the definition of the boundaries of fields within the main technology fields. The next section tests this concern.

The propensity of citations between Nclass fields is estimated in two stages[32]: in the first stage, equation (5.2) is estimated to obtain the predicted propensity of citations between pairs of citing and cited fields as explained above ($\widehat{\alpha_{ss'}}$). In the second stage, the propensity of citations from Nclass n to Nclass n' is assumed to be proportional to $\widehat{\alpha_{ss'}}$. Thus, conditional on a citation coming from field s to field s', the probability this citation comes from a randomly drawn patent in Nclass $n \in s$ to Nclass $n' \in s'$ is:

$$p_{nn'} = \widehat{\alpha_{ss'}} \times p(n \in s \mid s) \times p(n' \in s' \mid s') \tag{5.4}$$

Where, $p(n \in s \mid s)$ is the probability that the citing patent belongs to field n, condi-

[30] The periods are: 1963-1969, 1970-1975, 1976-1980, 1981-1985, 1986-1990, 1991-1995 and 1996-1999.

[31] In order to deal with potential heteroskedasity and to improve efficiency, I always weight the observations by the reciprocal of the $\sqrt{(N_{ltg})\,(N_{LT})}$. This weighting does not importantly affect the results, however, it does improve the fit of the model (consistently with Jaffe and Trajtenberg (1999)).

[32] Potentially, one would allocate patents into cells in the most refine manner, i.e., at the Nclass level (since $\widehat{G_i}$ is based on the Nclass classification). However, this is not feasible computationally using this estimation approach, as there are 400 Nclass fields, which would require estimating $400 \times 400 - 1$ coefficients ($\alpha_{ss'}$).

tional on the citation coming from field s and $p(n' \in s' \mid s)$ is the probability that the cited patent belongs to field n', conditional on the citation being directed to field s'. $p(n \in s \mid s)$ is calculated as the share of patents in field $n \in s$ out of the total patents in field s (over the period, 1975-1999) and $p(n' \in s' \mid s')$ is calculated as the share of patents in field $n' \in s'$ out of all patents in field s' (over the period 1963-1998). Finally, the weights, $\omega_{nn'}$ that are assigned to the citing technology fields are computed as $\omega_{nn'} = \frac{1}{1+p_{nn'}}$[33].

The weighted measure of generality (WG_i) is:

$$WG_i = 1 - \sum_n \omega_{nn'} \left(\frac{CR_{in}}{CR_i} \right)^2 \tag{5.5}$$

WG_i follows the same bias correction as G_i (denoted by $\widehat{WG_i}$). Table A4 summarizes the main statistics for the variables used in estimating equation (5.2) and for WG_i.

Table 3 reports the estimation results for $\widehat{WG_i}$ for the equivalent specifications reported in table 2. The effect of $\widehat{WG_i}$ is negative and significant in all specifications. Compared to the estimation with $\widehat{G_i}$, the effect of *Compelxity* remains unchanged, whereas the effect of *PatShare* rises. Overall, the results are stable to the more refine measure of generality that also takes into account the distance knowledge has traveled across fields as inferred from the estimated propensity of citations[34].

[Table 3 about here]

5.2.2. Alternative breakdown of technology fields

The definition of generality builds on Nclass fields. In case there is a measurement error in this classification that is correlated with *IntShare*, β_1 will be biased. For example, the number of different Nclass fields in "Drugs and Medicals" is only 14, whereas the

[33] Other functional forms which are decreasing in the propensity to cite have been experimented with (e.g., $\omega_{nn'} = 1 - p_{nn'}$) to find a similar pattern of results.

[34] I also construct $\widehat{G_i}$ while considering only citations from Nclass fields that are not in the same main technology sector as patent i. The coefficient on $\widehat{G_i}$ in an equivalent specification to column 6 in table 2 is -1.274 with a standard error of 0.154. Similarly, The coefficient on $\widehat{G_i}$ in an equivalent specification when only considering citations from Nclass fields that are not in the same HJT subcategory fields as patent i is -1.673 with a standard error of -0.182.

number Nclass fields in "Electrical and Electronics" is 50. Thus, patents in "Electrical and Electronics" are likely to be more general than patents in "Drugs and Medicals", especially in light of the higher propensity of citations within these fields rather than between, as discussed above (the average of $\widehat{G_i}$ in "Electrical and Electronics" is 0.503, compared to 0.434 in "Drugs and Medicals"). In case $IntShare$ is higher for patents in "Drugs and Medicals" and the technology fields in "Drugs and Medicals" are defined too broadly, β_1 will be downward-biased.

In order to test this concern, the estimation results with additional four generality measures are reported, as described in section 3. Table 4 summarizes the estimation results. The negative and significant effect of $\widehat{G_i}$ on $IntShare$ is highly robust across the different field classifications. Regarding $Complexity$, it is always negative and significant, with the exception of the $SubIPC$ specification, where the effect of $Complexity$ disappears. Similarly, $PatShare$ is positive and significant, with the exception of the $SIC - IPC$ specification, where it is not significant.

[Table 4 about here]

5.2.3. Adding Originality

The third robustness test looks at an additional "basicness" characteristic: the originality of the patent (following Henderson, Jaffe and Trajtenberg (1993)). Originality measures the extent knowledge builds on many technology fields, under the conjecture that a more original patent integrates pieces of knowledge from many different areas of research. In this respect, more original patents are also more basic. Originality is constructed as following:

$$O_i = 1 - \sum_n \left(\frac{CM_{in}}{CM_i} \right)^2 \tag{5.6}$$

Where, n denotes fields that patent i cites, CM_{in} is the number of citations made by patent i to field n and CM_i is the total number of citations made by patent i.

Similarly to generality, originality is downward biased for patents that make a small number of citations. Thus, in all specifications where originality is included the number

of citations made by patent i (labeled as $BackCites$) is also included. The equivalent bias-correction is:

$$\widehat{O_i} = \left(\frac{CM_i}{CM_i - 1}\right) O_i \tag{5.7}$$

For originality, backward looking data is used. Since information on citations made is available only from 1975 onwards, the sample of originating patents now includes only the patents that were granted between 1975 and 1980.

The estimation results for the effect of originality are reported in table 5. In all specifications, there is a negative and significant effect of originality on $IntShare$. In columns 1 and 2 only $\widehat{O_i}$ and $BackCites$ are included with and without firm fixed-effects, respectively. There is no important change in the coefficient on $\widehat{O_i}$ when firm fixed-effects are added, which implies that $\widehat{O_i}$ is not strongly correlated with characteristics of the inventing firm. In column 3, $\widehat{G_i}$ and Cites are added. The coefficient on $\widehat{O_i}$ halves, however it remains significant. Hence, there is a positive correlation between originality and generality, which implies that inventions that synthesize knowledge from many technology fields (i.e., more original), also spread to more technology fields (i.e., more general). Yet, even when controlling for generality, the coefficient on originality remains negative and significant.

In columns 4 and 5 $Complexity$ and $PatShare$ are added with no major change in the results.

Overall, there is strong evidence that not only the generality of knowledge matters for technological internalization, but also its originality. There is no reason to suspect that more original patents will have lower technological internalization. One possibility would be that more original patents are also more general (the correlation between generality and originality is 0.3). I also add an interaction term between originality and generality to test whether the effect of originality comes only from the higher likelihood of being more general. The coefficient on the interaction term is positive but not significant (0.372 with a standard error of 0.472) and there is no important change in the coefficients on

either originality or generality (-0.559 with a standard error of 0.279 and -1.685 with a standard error of 0.296, respectively). From this is concluded that there is a separate channel through which more original patents exhibit less technological internalization, in addition to the higher likelihood of also being more general.

[Table 5 about here]

5.2.4. A Probit estimation

The final robustness test relates to the probability that a patent creates Internalized spillovers. Only about 30 percent of patents in the sample create Internalize spillovers (where 70 percent of the patents create only Externalized spillovers). Table 6 reports the estimation results of a Probit specification where the dependent variable is a dummy that receives the value of one if the patent creates Internalized spillovers and zero otherwise. The results are highly consistent with previous findings. As patents become more general and original, the probability of creating Internalized spillovers drops.

Thus, not only that generality and originality are negatively correlated with the share of Internalized spillovers, they are also negatively correlated with the probability of creating positive Internalized spillovers.

Complexity has a positive and significant effect in the Probit specification (where in previous estimations, its effect was significantly negative). This implies that the probability of creating positive Internalized spillovers is higher when knowledge "spills"to fields that are more diversified in terms of the possibilities they introduce for follow-up research. Thus, when the inventing firm has more possibilities for reabsorbing its "spilled" knowledge, the probability that some internalization of spillovers occurs rises.

[Table 6 about here]

5.3. Market value and technological internalization

The effect of the knowledge stock of the firm on its market value should incorporate the dynamic consideration of technological internalization. This section shows that $IntShare$ positively affect private returns to knowledge in a market value estimation framework. Es-

timating the effect of $IntShare$ on market value would also allow quantifying the negative effect of generality on private returns.

Since this section exploits the firm-level variation in technological internalization, $IntShare$ is aggregated to the firm-level by taking its mean over the set of originating patents held by the inventing firms. For ease of notations, $IntShare$ is not relabeled, however, in this section it refers only to the firm-level aggregate.

5.3.1. Accounting data

The accounting data (sales, R&D, capital, etc.) and market value data for the sample of inventing firms is taken from US Compustat for the period 1980-2001. The accounting data have been 'cleaned' to remove accounting years with extremely large jumps in sales, employment or capital signalling merger and acquisition activity, leaving a total of 476 firms and 9,454 observations.

Table A2 summarizes the descriptive statistics for $InShare$ as well as for the main accounting variables. About 40 percent of firms do not create Internalized spillovers at all, whereas all firms create Externalized spillovers (where only about 30 percent of patents create Internalized spillovers).

In order to estimate the effect of technological internalization on private returns, a simple version of the value function approach proposed by Griliches (1981)[35] is adopted. The market value of firm i at period t, V_{it}, takes the following form:

$$V_{it} = \kappa_{it} \left(A_{it} + \left(\gamma_0 + \gamma_1 IntShare_i \right) K_{it} \right) \tag{5.8}$$

Where, A_{it} denotes physical assets, K_{it} is the R&D stock (representing knowledge stock), γ is the shadow price of the R&D stock (higher values of γ indicate that the market valuation of the knowledge stock relative to physical stock rises)[36]. The term $\gamma_0 + \gamma_1 IntSpill_i$ captures the private returns to innovation, which are expected to rise with $IntShare$ (i.e., γ_1 is expected to be positive).

[35] See also Jaffe (1986), Hall et al (2005) or Lanjouw and Schankerman (2004).

[36] A constant returns in the market value function has been assumed, consistently with previous studies.

Taking logarithms and dividing by A_{it}, the left-hand-side of equation (5.1) becomes the traditional Tobin's average Q, where its deviation from unity depends on the ratio between the R&D stock to the tangible stock $\left(\frac{K}{A}\right)$, $IntShare$ and κ_{it}, as following:

$$\log\left(\frac{V_{it}}{A_{it}}\right) = \log\kappa_{it} + \log\left(1 + (\gamma_0 + \gamma_1 IntShare_i)\frac{K_{it}}{A_{it}}\right) \tag{5.9}$$

Finally, κ_{it} is specified as:

$$\log\kappa_{it} = X'_{it}\beta_0 + \beta_1 IntShare_i + \tau_t + \eta_i + \epsilon_{it} \tag{5.10}$$

Where, X_{it} is a vector of controls (such as industry and technology dummies, sales, patents stock, etc.), τ_t is a complete set of time dummies, η_i is the firm fixed-effect, which is discussed later in this section, and ϵ_{it} is an idiosyncratic error term. The linear terms of $IntShare$ is included mainly as a control for their interaction with the R&D stock. Since $IntShare$ has many zero values, a dummy for $IntShare$ equals zero is always included.

Thus, the following equation is estimated by non-linear least squares (where standard errors are clustered by firms):

$$\log\left(\frac{V_{it}}{A_{it}}\right) = X'_{it}\beta_0 + \beta_1 IntShare_i + \log\left(1 + (\gamma_0 + \gamma_1 IntShare_i)\frac{K_{it}}{A_{it}}\right) + \tau_t + \eta_i + \epsilon_{it} \tag{5.11}$$

5.3.2. Estimation results for Tobin's Q

All the Tobin's Q specifications include a complete set of two-digit industry dummies (78 dummy variables), a set of indicators for the share of patents the firm has in each of the six main technology sectors, a complete set of year dummies (20 dummy variables), a dummy variable that receives the value one if the R&D stock of the firm is zero and a dummy variable that receives the value one if $IntShare$ is zero[37].

[37] I control for firm fixed-effects by adopting the "mean scaling" approach developed by Blundell, Griffith and Van Reenen (1999). Their method assumes that computing the mean of Tobin's Q in a long enough pre-estimation period can be used as an initial condition to proxy for unobserved heterogeneity, if the first moment is stationary.

Table 7 reports the estimation results of equation (5.11). There is strong evidence of a positive effect of $IntShare$ on market values both interacted with the R&D stock and linearly. This supports technological internalization being an important channel through which private rents are appropriated in a dynamic framework of sequential innovation[38]. Belenzon (2006) reports numerous robustness tests that support and extend this finding (for brevity, they are no reported here)[39].

Based on the estimates reported in column 2, the elasticity of market value with respect to the R&D stock, evaluated at the mean, is 0.110[40]. This implies that an additional one dollar spent on R&D raises market value by 0.302 dollar (referred to as private returns). A one standard deviation increase in $IntShare$ raises private returns to 0.452 dollar (thus, a 50 percent increase).

Given this estimate, the effect of generality on private returns could be simply computed ($\frac{\partial V}{\partial \widehat{G}} = \frac{\partial V}{\partial IntShare} \times \frac{\partial ntShare}{\partial \widehat{G}}$). From column 5 in table 3, a one standard deviation decrease in $\widehat{G}$ raises $IntShare$ at the patent level by 0.344 (0.319×1.079). Suppose that the same increase occurs for $IntShare$ at the firm level (e.g., the generality of all patents held by the inventing firm drops by one standard deviation). At the mean, a 0.344 increase in $IntShare$ raises the valuation of an additional one dollar spend on R&D by 4.8 percent ($50 \times \frac{0.344}{3.524}$)[41]. Similar calculations show that when moving from the most general ($\widehat{G} = 1$)

[38] Belenzon (2006) estimates the effects of $IntSpill$ and $ExtSpill$: $IntSpill$ has a positive and significant effect and $ExtSpill$ has a negative and significant effect. Both are identified via their interaction with the R&D stock and linearly.

[39] The most important robustness test is to include the citations-weighted patents stock linearly and interacted with the R&D stock. Firms with a larger patents stock are more likely to randomly indirectly cite their previous patents (i.e., have a higher $IntShare$). In case a larger patent stock also positively affects private returns the coefficient on $IntShare$ will be upward biased. The coefficients on the linear and interacted terms of the citations-weighted patents stock are positive, yet only the linear term is significant (0.159 with a standard error of 0.037 and 0.020 with a standard error of 0.043, respectively). The coefficients on the linear and interacted terms of $IntShare$ remain positive and significant with no important change in their size (see also Belenzon (2006) table 7).

[40] The estimated elasticity is lower from that reported in previous studies. For example, Bloom, Schankerman and Van Reenen (2005) report an elasticity of 0.24, using a similar estimation sample without industry or technology effects.

[41] Moving from the 75th percentile to the 25th percentile in $\widehat{G}$ raises the market valuation of an additional dollar spent on R&D by 6.9 percent ($50 \times \frac{(0.765-0.282)}{3.524}$).

to the most specialized knowledge ($\widehat{G} = 0$) private returns rise by 15.3 percent ($50 \times \frac{1.079}{3.524}$), at the mean[42].

[Table 7 about here]

6. Summary and conclusions

This paper empirically tests the classical argument that inventors face insufficient private incentive for basic research (Nelson (1959), Arrow (1962)). In a dynamic framework of sequential innovation the commercial value of basic knowledge intensifies when follow-up developments arrive mostly from outside the boundaries of the inventing firm. Thus, for the inventing firm to capture substantial private rents it must internalize the spillovers its knowledge creates. The main hypothesis of this paper is that as knowledge becomes more basic technological internalization diminishes, since only firms with a wider technology base could achieve such internalization. The empirical prediction of this hypothesis is that there would be a negative correlation between the generality of knowledge and technological internalization. Yet, a competing hypothesis is that firms choose the "basicness" level of their knowledge: basic knowledge is invented only by firms with a wide technology base that allows internalizing private rents even when they are spread over many fields. The empirical prediction of this competing hypothesis is that the a negative correlation between the generality of knowledge and technological internalization will be substantially mitigated.

Using data on patents and citations a novel empirical methodology is developed that allows measuring the extent spillovers feed back into the inventing firm. Based on this methodology, the hypothesis that more general knowledge exhibits lower technological internalization is confirmed. This is inconsistent with basic research being endogenously chosen by firms (since firms that choose to conduct basic research are those that are better

[42] I also estimate the effect of the firm-level average of $\widehat{G}$ on market value (linearly and interacted with the R&D stock). Adding $\widehat{G}$ to column 5 in table 7 does not affect the coefficient on $IntShare$ in an important way. The coefficients on the linear and interacted terms of $\widehat{G}$ are negative but not significant (-0.071 with a standard error of 0.048 and -0.269 with a standard error of 0.221).

able to achieve high technological internalization).

A market value equation is estimated to confirm that technological internalization is an important channel through which private rents are appropriated. The estimates from the value function are then used to quantify the impact of generality on private returns. A one standard deviation increase in technological internalization raises the market valuation of an additional dollar spent on R&D by 50 percent, evaluated at the mean. Based on this estimate, a one standard deviation increase in the generality of knowledge lowers private returns by 4.8 percent. Moving from the most specialized to the most general knowledge (the two extreme points on the generality spectrum) lowers private returns by 15.3 percent, evaluated at the mean.

The findings of this paper are interpreted as supporting the classical problem of underinvestment in basic research. Private returns depend on the extent the inventing firm internalizes the spillovers its knowledge creates. Such internalization is less likely to occur as knowledge becomes more general, and in this respect more basic.

7. References:

Aghion, P. and Howitt, P. (1992), "A Model of Growth through Creative Destruction", *Econometrica* Vol. 60, No. 2, pp. 323 351

Belenzon, S. (2006), "Knowledge Flow and Sequential Innovation: Implications for Technology Diffusion, R&D and Market Value", *Oxford University, Department of Economics Discussion Paper*

Belenzon, S. (2005), "Knowledge Flows and Sequential Innovation: Implications for Technology Diffusion, Incentive for R&D and Firm Performance", *PhD Dissertation, University of London*

Bessen, K., Maskin, E. (2002), "Sequential Innovation, Patents and Imitation", *Department of Economics Working Paper 00-01, MIT*

Boldrin, M. and Levine, D.K. (2002), "The Case Against Intellectual Property", *The American Economic Review (Papers and Proceedings)* 92, pp. 209-212

Bresnahan, T. and Trajtenberg, M. (1995), "General Purpose Technologies - Engines of Growth?", *Journal of Econometrics*, Vol 65 (1), pp. 83-108

Caballero, R. J. and Jaffe, A. B. (1993), "How High are the Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth", *NBER Working Papers 4370, National Bureau of Economic Research*

Chang, H. F. (1995), "Patent Scope, Antitrust Policy, and Cumulative Innovation", *RAND Journal of Economics*, Vol 26, pp. 34-57

Cornelli, F and Schankerman, M. (1999), "Patent Renewals and R&D Incentives", *RAND Journal of Economics*, pp. 17-34

Gilbert, R., Shapiro, C. (1990), "Optimal patent length and breadth", *RAND Journal of Economics,* 21, pp. 106-112

Green, J., Scotchmer, S. (1995), "On the Division of Profit between Sequential Innovators", *The Rand Journal of Economics* 26, pp. 20-33

Griliches, Z. (1992), "The search for R&D spillovers", *Scandinavian Journal of Economics* 94, supplement, pp. S29-S47

Griliches, Z. (1986), "Productivity, R&D, and Basic Research at the Firm Level in the 1970s", *American Economic Review*, Vol. 76 (1), pp. 1 41-154

Grindley, P. C and Teece, D. J. (1997), "Managing Intellectual Capital: Licensing and Cross-Licensing in Semiconductors and Electronics", *California Management Review*, 39, pp. 1-34

Grossman, G. M. and Helpman, E. (1991), Innovation and Growth in the Global Economy, *MIT Press*, Cambridge, Mass

Hall, B. H. and Trajtenberg M. (2005), "Uncovering General Purpose Technologies using Patent Data", *In Antonelli, C., D. Foray, B. H. Hall, and W. E. Steinmueller (eds.), New Frontiers in the Economics of Innovation and New Technology*

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

Hall, B. H., Jaffe A.B. and Trajtenberg M. (2005), "Market Value and Patent Citations", *RAND Journal of Economics*, Vol. 36(1), pp. 16-38

Heller, M. A. and Eisenberg, R. S. (1998), "Can Patents Deter Innovation? The Anticommons in Biomedical Research", *Science*, 280, pp. 698-701

Helpman, E. and Trajtenberg, M. (1998), "Diffusion of General Purpose Technologies", *General Purpose Technologies and Economic Growth*, Cambridge: MIT Press

Helpman, E. and Trajtenberg, M. (1998), "A Time to Sow and a Time to Reap: Growth Based on General Purpose Technologies" *General Purpose Technologies and Economic Growth*, Cambridge: MIT Press

Jaffe, A. and Trajtenberg, M. (1999), "International knowledge flows: Evidence from patent citations", *Economics of Innovation and New Technology*, Vol. 8, pp. 105-136

Jaffe, A. and Trajtenberg, M (2002), "Patents, Citations and Innovations: A Window on the Knowledge Economy", *Cambridge, Mass, MIT Press*, August 2002

Jaffe, A., Trajtenberg, M. and Henderson, R. (1993), "Geographic localization of knowledge spillovers as evidenced by patent citations", *Quarterly Journal of Economics* 108 (3), pp. 577-598

Jaffe, A., Fogarty, S., Trajtenberg, M. (2000), "Knowledge spillovers and Patent Citations: Evidence from A Survey of Inventors", *American Economic Review*, pp. 215-218

Klemperer, P (1990), "How Broad Should the Scope of Patent Protection Be?", *RAND Journal of Economics*

Lanjouw, J. (1998), "Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value", *The Review of Economics Studies*, Vol. 65, pp. 671-710

Lanjouw, J. and Schankerman, M. (2004), "Protecting Intellectual Property Rights: Are Small Firms Handicapped?", *Journal of Law and Economics*, pp. 45-74

Mansfield, E. (1980), "Basic Research and Productivity Increase in Manufacturing", *American Economic Review*, 70 (5), pp. 863-873

Noel, M. and Schankerman, M. (2006), "Strategic Patenting and Software Innovation", *LSE mimeo*

O'Donoghue, T. (1998), "A Patentability requirement for Sequential Innovation", *Rand Journal of Economics*, 29 (4), pp. 654-679

Romer, P. M. (1986), "Increasing Returns and Long-run Growth", *Journal of Political Economy*, Vol. 94(5), pp. 1002-37

Rosenberg, N. and Trajtenberg, M. (2004), "A General Purpose Technology at Work: The Corliss Steam Engine in the late 19th Century US", *The Journal of Economic History*, Vol. 64 (1), pp. 61-99

Schankerman M. and Pakes A. (1986), "Estimates of the Value of Patent Rights in European Countries During the Post-1950 Period", *Economic Journal*, Vol.96, pp. 1052-1076

Scotchmer, S. (1991), "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law", Journal of Economic Perspectives 5(1), pp. 29-41

Scotchmer, S. (1996), "Protecting Early Innovators: Should Second-Generation Products be Patentable?", *The Rand Journal of Economics* 27, Summer, pp. 322-331

Scotchmer, S. (1999), "On the Optimality of the Patent Renewal System", *Rand Journal of Economics*, 30, pp. 181-196

Segerstrom, P. S., Anant, C. A. and Dinopoulos, E. (1990), "A Schumpeterian Model of

the Product Life Cycle", *American Economic Review*, 80, pp. 1077-1091

Trajtenberg, M. (1990), "A Penny for Your Quotes: Patent Citations and the Value of Innovations", *The Rand Journal of Economics*, 21(1), pp. 172-187

Trajtenberg, M. (1990), "Economic Analysis of Product Innovation: The Case of CT Scanners", *Harvard University Press*,Cambridge, MassS63-S84

A. Appendices

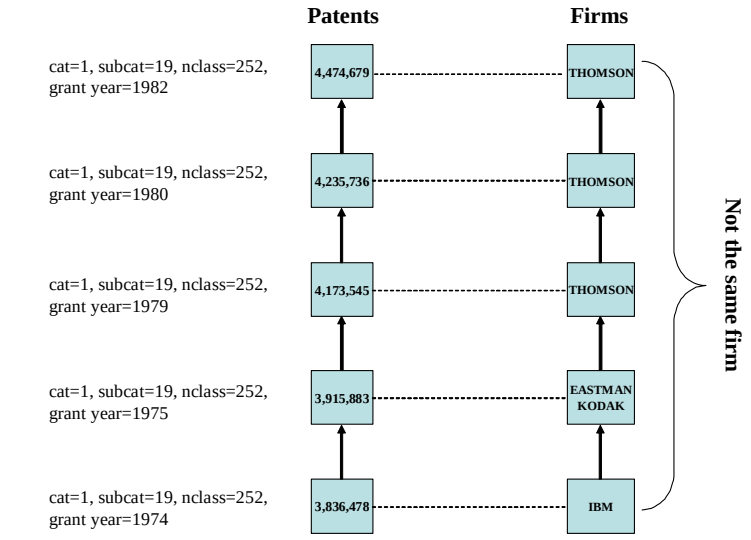

Figure A1: An example for an Externalized line of research

Figure A1: *This figure shows a unique line of research originated in invention 3,836,478, which is owned by IBM (the inventing firm). Since knowledge did not return to IBM in the period 1974-1989, this line of research is Externalized.*

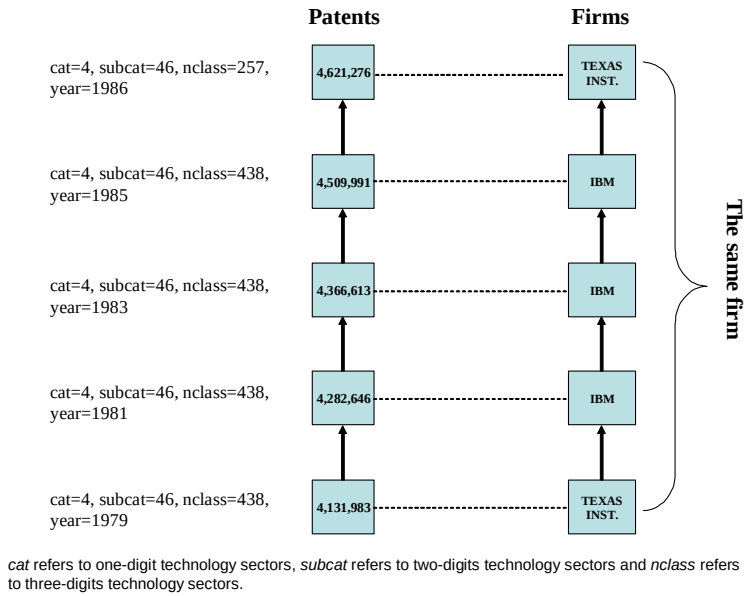

Figure A2: An example for an Internalized line of research

Figure A2: *This figure shows a unique line of research originated in invention 4,131,983, which is owned by Texas Instruments (the inventing firm). Since knowledge returned to Texas Instruments in the period 1979-1994, this line of research is Internalized.*

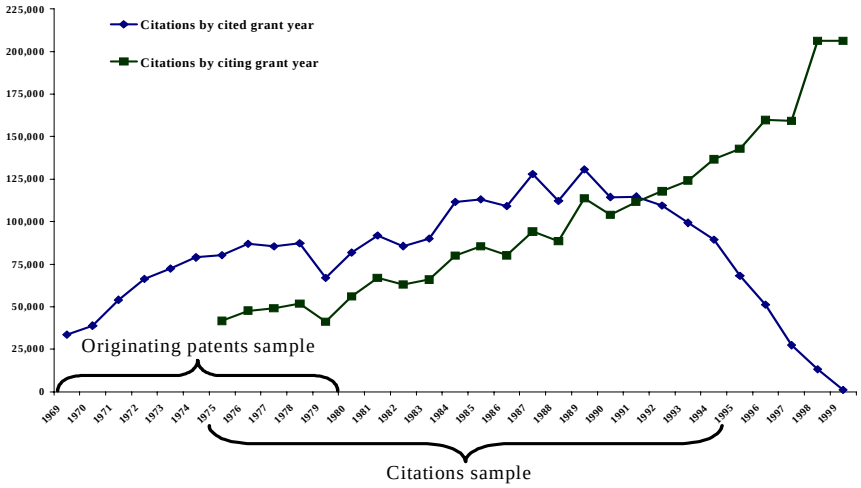

Figure A3: Citations sample

Figure A3: *This figure presents the number of citations made and received by patents in our sample. The upward sloping graph shows the number of citations made each year, where the U shaped curve shows the number of citations received each year.*

A.1. Data

The sample combines data from two datasets:

The NBER USPTO patents database includes detailed patenting and citations information for around 2,600 US firms (as described in Hall, Jaffe and Trajtenberg (2001)) and a list of all the citations made in the period 1975-1999.

The Compustat North-America dataset provides full accounts data for over 25,000 US firms from 1980 to 2001. This provides information on the key accounting information of R&D, fixed assets, employment, sales, etc.

I started by matching the Compustat accounting data to the USPTO data, and kept firms with 1 or more patents in the period 1969-1980 that received at least one citation from the 2,600 firms in the NBER USPTO data set between 1975 and 1995. This leaves a sample of 492 firms.

The accounting dataset has been 'cleaned' to remove accounting years with extremely large jumps (>+200% or <-66%) in sales, employment or capital signaling merger and acquisition activity, leaving a total of 476 firms and 9,454 observations.

The book value of capital is the net stock of property, plant and equipment (Compustat Mnemonic PPENT); Employment is the number of employees (EMP). R&D (XRD) is used to create R&D capital stocks calculated using a perpetual inventory method with a 15% depreciation rate (Hall et al, 2005). The citations-weighted patent stock was constructed by normalizing the number of patents the firm owns according to the number of citations it receives and the average number of citations to all patents in the same year. Given this normalized patents count the stock is constructed using the perpetual inventory method. The citations stock (used as a pre-estimation control) was constructed equivalently to the R&D stock. For Tobin's Q, firm value is the sum of the values of common stock, preferred stock, total debt net of current assets (Mnemonics MKVAF, PSTK, DT and ACT). Book value of capital includes net plant, property and equipment, inventories, investments in unconsolidated subsidiaries and intangibles other than R&D (Mnemonics PPENT, INVT, IVAEQ, IVAO and INTAN). Tobin's Q was set to 0.1 for values below 0.1 and at 20 for values above 20. See also Lanjouw and Schankerman (2004). Industry price deflators were taken from Bartelsman, Becker and Gray, 2000, until 1996 and from the BEA 4-digit NAICS Shipment Price Deflators afterwards. See Belenzon (2006) for more detail on the construction of the accounting data.

Table 1

Internalized and Externalized lines of research

	Number of lines of research[a]	Share of Internalized lines of research[b]			
	Total sample	Total sample	1969-1975	1976-1978	1979-1980
Pooled	46.8	7.6%	8.2%	7.6%	7.2%
Chemicals	28.8	6.2%	6.4%	6.3%	5.7%
Computers and Communications	30.2	7.6%	8.8%	7.1%	7.1%
Drugs and Medicals	16.8	15.0%	19.1%	16.8%	8.4%
Electrical and Electronics	78	7.4%	7.5%	7.1%	7.5%
Mechanicals	15.5	8.8%	9.1%	9.1%	7.9%

[a]Computed as the average number of lines of research per citations received by an originating patent for the entire period of the sample.

[b]Computed as the ratio between Internalized lines of research and the total number of lines of research.

Table 2

The effect of Generality on IntShare

Dependent variable: IntShare. OLS estimation

	(1)	(2)	(3)	(4)	(5)	(6)
Generality	-1.728*	-1.665*	-1.654*	-1.627*	-1.311*	-1.282*
	(0.215)	(0.210)	(0.204)	(0.198)	(0.194)	(0.187)
Cites	-0.099*	-0.102*	-0.098*	-0.098*	-0.097*	-0.098*
	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)
PatCon		7.752*				
		(1.940)				
Complexity				-0.021*		-0.021*
				(0.004)		(0.004)
PatShare					7.545*	7.559*
					(2.608)	(2.638)
Firm-fixed effects	No	No	Yes	Yes	Yes	Yes
Observations	92,032	92,032	92,032	92,032	92,032	92,032
R^2	0.321	0.322	0.336	0.337	0.337	0.338

Standard errors (in brackets) are robust to arbitrary heteroskedacity and serial correlation (clustered by firms).

All regressions include a complete set of two-digit technology field dummies (36), grant year dummies (10) and a dummy for IntShare equals zero.

* denotes a significance level of 5 percent.

Generality, Complexity and PatShare are based on the US Nclass classification.

Cites is the number of direct citations the originating patent receives (over the period 1975-1999). PatCon is a firm-level measure of research diversification (it is collinear with the firm fixed-effect). A higher PatCon implies a lower research diversification.

Table 3

The effect of the Weighted-Generality on IntShare

Dependent variable: IntShare. OLS estimation

	(1)	(2)	(3)	(4)	(5)
Generality	-1.249*	-1.197*	-1.179*	-1.096*	-1.079*
	(0.145)	(0.135)	(0.132)	(0.129)	(0.125)
Cites	-0.099*	-0.097*	-0.097*	-0.097*	-0.098*
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
Complexity			-0.021*		-0.021*
			(0.004)		(0.004)
PatShare				11.008*	10.933*
				(4.338)	(4.346)
Firm-fixed effects	No	Yes	Yes	Yes	Yes
Observations	92,032	92,032	92,032	92,032	92,032
R^2	0.321	0.336	0.337	0.336	0.337

Standard errors (in brackets) are robust to arbitrary heteroskedacity and serial correlation (clustered by firms).

All regressions include a complete set of two-digit technology field dummies (36), grant year dummies (10) and a dummy for IntShare equals zero.

* denotes a significance level of 5 percent.

Generality, Complexity and PatShare are based on the US Nclass classification.

Cites is the number of direct citations the originating patent receives (over the period 1975-1999).

Table 4

The effect of Generality on IntShare: alternative measures of Generality

Dependent variable: IntShare. OLS estimation

	(1)	(2)	(3)	(4)	(5)
	Nclass	IPC	SubHJT	SubIPC	SIC-IPC
Generality	-1.486*	-1.359*	-1.091*	-2.007*	-1.466*
	(0.188)	(0.209)	(0.152)	(0.263)	(0.196)
Cites	-0.098*	-0.098*	-0.099*	-0.095*	-0.099*
	(0.009)	(0.010)	(0.010)	(0.009)	(0.010)
Complexity	-0.021*	-0.005*	-0.052*	-0.001	-0.018*
	(0.004)	(0.001)	(0.007)	(0.005)	(0.005)
PatShare	10.580*	15.288*	8.200*	18.516*	-3.249
	(4.314)	(4.418)	(2.599)	(6.425)	(4.153)
Firm-fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	92,032	92,032	92,032	92,032	92,032
R^2	0.337	0.338	0.337	0.336	0.336

Standard errors (in brackets) are robust to arbitrary heteroskedacity and serial correlation (clustered by firms).

All regressions include a complete set of two-digit technology field dummies (36), grant year dummies (10) and a dummy for IntShare equals zero.

* denotes a significance level of 5 percent.

Cites is the number of direct citations the originating patent receives (over the period 1975-1999).

Table 5

The effect of Generality and Originality on IntShare

Dependent variable: IntShare. OLS estimation.

	(1)	(2)	(3)	(4)	(5)
Originality	-1.084	-1.049*	-0.523*	-0.442*	-0.445*
	(0.212)	(0.027)	(0.162)	(0.158)	(0.157)
Generality			-1.992*	-1.746*	-1.873*
			(0.338)	(0.311)	(0.312)
BackCites	-0.097*	-0.092*	-0.057*	-0.059*	-0.055*
	(0.023)	(0.023)	(0.021)	(0.020)	(0.020)
Cites			-0.084*	-0.084*	-0.093*
			(0.011)	(0.011)	(0.012)
Complexity				-0.029*	-0.028*
				(0.005)	(0.005)
PatShare				14.529*	14.575*
				(5.159)	(5.106)
Forward Lag					0.169*
					(0.028)
Firm-fixed effects	No	Yes	Yes	Yes	Yes
Observations	38,745	38,745	38,745	38,745	38,745
R^2	0.281	0.307	0.316	0.318	0.319

Standard errors (in brackets) are robust to arbitrary heteroskedacity and serial correlation (clustered by firms).

All regressions include a complete set of two-digit technology field dummies (36), grant year dummies (10) and a dummy for IntShare equals zero.

* denotes a significance level of 5 percent.

Generality, Originality, Complexity and PatShare are based on the US Nclass classification.

Cites is the number of direct citations the originating patent receives (over the period 1975-1999). BackCites is the number of citations made by the originating patent.

Since Originality and BackCites are backwards looking, the sample of originating patents covers only the originating patents granted between 1975 and 1980 (since data on citations made start at 1975).

Table 6

The effect of Generality and Originality the probability to internalize

Dependent variable: A dummy for a positive IntShare. Probit estimation.

	(1)	(2)	(3)	(4)	(5)	(6)
Generality	-0.123*	-0.129*	-0.109*	-0.111*	-0.119*	-0.103*
	(0.036)	(0.035)	(0.030)	(0.044)	(0.043)	(0.039)
Originality				-0.050*	-0.055*	-0.048*
				(0.026)	(0.025)	(0.024)
Cites	0.034*	0.034*	0.033*	0.036*	0.036*	0.036*
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
BackCites				0.020*	0.021*	0.020*
				(0.003)	(0.003)	(0.003)
Complexity		0.003*	0.003*		0.004*	0.004*
		(0.001)	(0.001)		(0.001)	(0.001)
PatShare			0.851*			0.656*
			(0.332)			(0.322)
Observations	92,032	92,032	92,032	38,745	38,745	38,745
R^2	0.123	0.124	0.125	0.117	0.118	0.119

Standard errors (in brackets) are robust to arbitrary heteroskedacity and serial correlation (clustered by firms).

All regressions include a complete set of two-digit technology field dummies (36) and grant year dummies (10).

* denotes a significance level of 5 percent.

Cites is the number of direct citations the originating patent receives (over the period 1975-1999). BackCites is the number of citations made by the originating patent.

Since Originality and BackCites are backwards looking, the sample of originating patents covers only the originating patents granted between 1975 and 1980 (since data on citations made start at 1975).

Table 7

The effect of IntShare on Tobin's Q

Dependent variable: log(Tobin's-Q). Nonlinear Least Squares.

	(1)	(2)	(3)	(4)	(5)
R&D stock/Assets	0.330*	0.120*	0.135*	0.141*	0.217*
	(0.101)	(0.064)	(0.026)	(0.026)	(0.040)
IntShare x (R&D stock/Assets)	5.624*	2.341*	1.702*	1.379*	1.311*
	(2.295)	(0.507)	(0.533)	(0.498)	(0.586)
IntShare			0.016*	0.020*	0.025*
			(0.004)	(0.006)	(0.007)
log(Sales)				0.035*	0.033*
				(0.004)	(0.004)
log(Industry Sales)				-0.005*	-0.011*
				(0.006)	(0.006)
Sales Growth					0.538*
					(0.018)
Pre-sample means[a]	No	Yes	Yes	Yes	Yes
Observations	9,454	9,454	9,454	9,454	9,015
R^2	0.294	0.496	0.496	0.499	0.504

Standard errors (in brackets) are robust to arbitrary heteroskedacity and serial correlation (clustered at the firm level). * denotes a significant level of 5 percent.

All regressions include 78 two-digits industry dummies, 4 technology indicators, a complete set of year dummies, a dummy variable for R&D stock equals zero and a dummy variable for IntShare equals zero.

[a]The set of pre-sample means includes: Market Share, Employees, Tobin's Q, Sales, Assets, R&D stock, Patents stock and Citations stock.

IntShare is the firm-level average of the patent-level IntShare.

Industry Sales is the aggregated sales of firms in the same for-digit SIC as the inventing firm (see Belenzon (2006) for detail). Sales Growth is the growth in the sales of the inventing firm.

Table A1

Patents' main characteristics

variable	Mean	Median	Std Dev	Minimum	Maximum
IntShare	4.689	0.00	12.69	0	100
Generality	0.432	0.50	0.26	0.0	0.93
Complexity	17.685	12.24	20.57	1.4	233
PatShare	0.055	0.02	0.09	0	1
Originality[a]	0.377	0.44	0.27	0	1
Cites	10.980	8.00	11.85	2	428
BackCites[a]	4.930	4.00	2.95	0.0	94
Backward Lag[a]	8.643	8.33	3.75	0.0	26
Forward Lag	7.67	7.00	4.07	0	15

[a]The backward looking variables: Originality, Backward Tech, Backward Citations and Backward Lag, are computed only for the patents that were granted between 1975 and 1980.

Generality, Originality, Complexity and PatShare are based on the US Nclass classification. Generality is not bias-corrected.

Cites is the number of direct citations the originating patent receives (over the period 1975-1999). BackCites is the number of citations made by the originating patent. Forward Lag is the average difference between the grant year of the patents that cite the originating patent and the grant year of the originating patent. Backward Lag is the average difference between the grant year of the originating patent and the grant year of the patents it cites.

Table A2

Descriptive statistics: accounting and patents variables

9,454 observations and 476 firms

Variable	Mnemonic	Mean	Median	Min	Max	Standard deviation
IntShare[1]		2.17	0.28	0	24.63	3.52
Tobin's Q	V/A	2	1.32	0.1	20	2.34
Market value, $m	V	4,689	592	0	485,566	16,782
R&D stock, $m	K	806	49	0	47343	3195
R&D stock / Assets	K/A	0.39	0.20	0	10	1
Capital stock, $m	A	3,090	392	2.13	199,303	9,736
Sales, $m		3,925	686	0	180,557	11,412
Patents stock		155	18	0.42	9,848	489
Patents stock weighted by citations		158	16	0.28	12,643	585

The statistics are computed over all the observations that were included in the estimation (1980-2001) and are given in thousands of 1996 USD.

IntShare is the firm-level average of the patent-level IntShare.

[1]For about 40 percent of firms IntShare is zero.

Table A3

Descriptive statistics for Generality (bias-corrected)

variable	Mean	Median	Std Dev	Min	Max
Generality (bias-corrected)					
Nclass	0.517	0.593	0.319	0	1
IPC	0.520	0.595	0.324	0	1
SubHJT	0.399	0.449	0.315	0	1
SubIPC	0.818	0.893	0.229	0	1
SIC	0.433	0.500	0.317	0	1
Cites[a]	10.967	8.000	11.850	2	428

The correlation between the alternative Generality measures

variable	Nclass	IPC	SubHJT	SubIPC	SIC
Nclass	1.000				
IPC	0.671	1.000			
SubHJT	0.802	0.612	1.000		
SubIPC	0.451	0.487	0.362	1.000	
SIC	0.721	0.610	0.672	0.610	1.000
Cites	0.072	0.059	0.060	0.113	0.054

[a]5,052 patents in the initial sample receive only one citation, thus, the bias-corrected measures are not defined. These patents were dropped from the sample.

[b]Refers to the Generality measure which is based on Nclass fields definition, where the citing fields are weighted by their proximity to the cited field (using citations proximity metric).

Cites is the number of direct citations the originating patent receives (over the period 1975-1999).

Table A4

Descriptive statistics for the propensity of citations estimation and weighted-generality

	Mean	Median	Std Dev	Min	Max
number of citations	89.7	14	255.9	0	5581
Potentially cited patents	11153.9	12066.0	6075.1	1037.0	21959.0
Potentially citing patents	2242.2	1801.0	1764.7	198.0	10300.0
cited grant year	1974	1974	7.6	1963	1994
citing grant year	1987	1987	5.8	1975	1995
citation frequency (10xe-5)	0.418	0.0791	1.170	0	19.130
lag in years	12.3	11.5	7.6	1	32
regression weight	4458.2	4108.4	2298.0	453.1	15039.2
Weighted-generality	0.629	0.750	0.365	0	0.996
Bias-corrected weighted-generality	0.793	0.946	0.424	0	1.5

The entries in the table refers to the citations "cell", as described in the text. Citations frequency is the dependent variable in equation (the estimation equation of the propensity of citations). The cited "cells" are include the dimensions of the grant year and main technology class of the cited patents. The citing "cells" include the dimensions of the grant year and HJT sub-category technology class of the citing patents. The lag in years is the difference between the grant year of the citing patent and the grant year of the cited patent for every "cell". Potentially cited and citing patents for a given "cell" are defined as following: the number of patents in a given main technology class in a given year for potentially cited patents and the number of patents in a given HJT sub-category technology class in a given year for the potentially citing patents.

Table A5

The propensity of citations between technology "cells"

Main technology class	HJT Subcategory class number	HJT Subcategory class name	Main technology class name					
			Chemical	Computers and Communications	Drugs and Medical	Electrical and Electronic	Mechanical	Others
Chemical	11	Agriculture,Food,Textiles	1.000	-0.972	-0.566	-0.961	-0.904	-0.718
Chemical	12	Coating Chemical	0.702	-0.872	-0.799	-0.690	-0.740	-0.656
Chemical	13	Gas Chemical	1.703	-0.949	-0.857	-0.719	-0.799	-0.745
Chemical	14	Organic Compounds	0.304	-0.994	-0.347	-0.992	-0.984	-0.977
Chemical	15	Resins Chemical	1.333	-0.977	-0.708	-0.975	-0.909	-0.839
Chemical	19	Miscellaneous-chemical	1.233	-0.879	-0.776	-0.746	-0.781	-0.803
		Average	***1.046***	***-0.940***	***-0.675***	***-0.847***	***-0.853***	***-0.790***
Computers and Communications	21	Communications	-0.971	11.993	-0.944	-0.126	-0.861	-0.936
Computers and Communications	22	Computer Hardware & Software	-0.965	16.012	-0.952	0.193	-0.634	-0.905
Computers and Communications	23	Computer Peripherials	-0.854	13.123	-0.971	-0.050	-0.713	-0.879
Computers and Communications	24	Information Storage	-0.971	12.416	-0.992	-0.329	-0.858	-0.952
		Average	***-0.940***	***13.386***	***-0.965***	***-0.078***	***-0.767***	***-0.918***
Drugs and Medical	31	Drugs	-0.368	-0.996	3.141	-0.993	-0.981	-0.950
Drugs and Medical	32	Surgery & Med Inst.	-0.858	-0.711	10.236	-0.595	-0.868	-0.793
Drugs and Medical	33	Biotechnology	-0.516	-0.978	4.103	-0.931	-0.977	-0.919
Drugs and Medical	39	Miscellaneous	-0.848	-0.933	6.904	-0.897	-0.860	-0.901
		Average	***-0.647***	***-0.905***	***6.096***	***-0.854***	***-0.921***	***-0.891***
Electrical and Electronic	41	Electrical Devices	-0.946	-0.139	-0.966	3.158	-0.892	-0.908
Electrical and Electronic	42	Electrical Lighting	-0.938	-0.523	-0.971	3.138	-0.877	-0.930
Electrical and Electronic	43	Measuring & Testing	-0.911	0.107	-0.833	2.907	-0.850	-0.902
Electrical and Electronic	44	Nuclear & X-rays	-0.905	-0.126	-0.914	3.299	-0.814	-0.916
Electrical and Electronic	45	Power Systems	-0.904	-0.331	-0.970	3.698	-0.715	-0.886
Electrical and Electronic	46	Semiconductor Devices	-0.797	-0.049	-0.993	5.256	-0.917	-0.954
Electrical and Electronic	49	Miscellaneous	-0.926	0.971	-0.924	3.443	0.031	-0.839
		Average	***-0.904***	***-0.013***	***-0.939***	***3.557***	***-0.719***	***-0.905***
Mechanical	51	Mat. Proc & Handling	-0.756	-0.785	-0.914	-0.862	1.025	-0.756
Mechanical	52	Metal Working	-0.840	-0.854	-0.945	-0.541	0.751	-0.793
Mechanical	53	Motors & Engines + Parts	-0.937	-0.737	-0.940	-0.691	1.366	-0.828
Mechanical	54	Optics Mech	-0.862	-0.094	-0.945	-0.413	1.262	-0.916
Mechanical	55	Transportation	-0.961	-0.782	-0.988	-0.829	0.001	0.882
Mechanical	59	Miscellaneous	-0.890	-0.611	-0.911	-0.831	0.968	-0.766
		Average	***-0.874***	***-0.644***	***-0.941***	***-0.695***	***0.895***	***-0.529***
Others	61	Agriculture,Husbandry,Food	-0.856	-0.947	-0.770	-0.902	-0.919	0.064
Others	62	Amusement Devices	-0.971	-0.643	-0.946	-0.895	-0.859	0.723
Others	63	Apparel & Textile	-0.910	-0.922	-0.803	-0.904	-0.846	0.778
Others	64	Earth Working & Wells	-0.811	-0.881	-0.967	-0.880	-0.708	1.342
Others	65	Furniture,House Fixtures	-0.959	-0.934	-0.794	-0.911	-0.792	0.766
Others	66	Heating Others	-0.791	-0.919	-0.969	-0.656	-0.806	0.900
Others	67	Pipes & Joints	-0.903	-0.966	-0.898	-0.823	-0.606	0.753
Others	68	Receptacles	-0.865	-0.938	-0.858	-0.906	-0.731	1.108
Others	69	Miscellaneous	-0.693	-0.718	-0.898	-0.760	-0.699	0.577
		Average	*-0.862*	*-0.874*	*-0.878*	*-0.848*	*-0.774*	*0.779*

The entries in the table are the estimated propensity of citations between "cells" of patent citations, as explained in the text. The rows represent the citing patents and the columns represent the cited patent. All entries are relative to the propensity of citations from "Agriculture,Food,Textiles" to "Chemicals" (which is normalized to unity). For example, the propensity of citations from a randomly drawn patent from "Semiconductor Devices" to a randomly drawn patent from "Electrical and Electronic" is 5.256 times the propensity of citations of the benchmark "cell" .

Table A6

Examples of patents with high Generality and low IntShare, and patents with low Generality and high IntShare

Patent	US Nclass	Grant year	IntShare	Generality	Cites	Inventing firm
High Generality and low IntShare						
3420032	52	1969	0	0.915	18	156
3551940	16	1971	0	0.909	18	229
3711081	269	1973	0	0.931	22	2449
3787351	523	1974	0	0.927	57	1466
3931090	524	1976	0	0.918	15	511
Low Generality and high IntShare						
3619207	426	1971	51.3	0.089	33	2221
3849721	324	1974	56.9	0.051	45	1753
3906090	424	1975	80.7	0.079	30	2452
3929987	424	1975	76.1	0.099	25	2452
4159011	123	1979	54.1	0.074	29	22

These patents exemplify the negative correlation between Generality and IntShare. I focus on patents that are highly cited since these patents are likely to generate a substantial diffusion "tree".

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

722	Daron Acemoglu Philippe Aghion Claire Lelarge John Van Reenen Fabrizio Zillibotti	Technology, Information and the Decentralization of the Firm
721	Sharon Belenzon	Knowledge Flow and Sequential Innovation: Implications for Technology Diffusion, R&D and Market Value
720	Rafael Gomez Konstantinos Tzioumis	What Do Unions Do to CEO Compensation?
719	Ralph Ossa	A Gold Rush Theory of Economic Development
718	Nick Bloom	The Impact of Uncertainty Shocks: Firm Level Estimation and a 9/11 Simulation
717	Holger Breinlich	Trade Liberalization and Industrial Restructuring through Mergers and Acquisitions
716	Nick Bloom John Van Reenen	Measuring and Explaining Management Practices Across Firms and Countries
715	Mirko Draca Stephen Machin John Van Reenen	Minimum Wages and Firm Profitability
714	Matteo Bugamelli Francisco Paternò	Do Workers' Remittances Reduce the Probability of Current Account Reversals?
713	Alex Bryson	Union Free-Riding in Britain and New Zealand
712	Marco Manacorda Carolina Sanchez-Paramo Norbert Schady	Changes in Returns to Education in Latin America: the Role of Demand and Supply of Skills
711	Claudia Olivetti Barbara Petrongolo	Unequal Pay or Unequal Employment? A Cross-Country Analysis of Gender Gaps

710	Hilary Steedman	Apprenticeship in Europe: 'Fading' or Flourishing?
709	Florence Kondylis	Agicultural Returns and Conflict: Quasi-Experimental Evidence from a Policy Intervention Programme in Rwanda
708	David Metcalf Jianwei Li	Chinese Unions: Nugatory or Transforming? An *Alice* Analysis
707	Richard Walker	Superstars and Renaissance Men: Specialization, Market Size and the Income Distribution
706	Miklós Koren Silvana Tenreyro	Volatility and Development
705	Andy Charlwood	The De-Collectivisation of Pay Setting in Britain 1990-1998: Incidence, Determinants and Impact
704	Michael W. L. Elsby	Evaluating the Economic Significance of Downward Nominal Wage Rigidity
703	David Marsden Richard Belfield	Performance Pay for Teachers Linking Individual and Organisational Level Targets
702	John Van Reenen	The Growth of Network Computing: Quality Adjusted Price Changes for Network Servers
701	Joas Santos Silva Silvana Tenreyro	The Log of Gravity
700	Alan Manning Joanna Swaffield	The Gender Gap in Early Career Wage Growth
699	Andrew B. Bernard Stephen Redding Peter K. Schott	Products and Productivity
698	Nicholas Oulton	Ex Post Versus Ex Ante Measures of the User Cost of Capital

The Centre for Economic Performance Publications Unit
Tel 020 7955 7673 Fax 020 7955 7595 Email info@cep.lse.ac.uk
Web site http://cep.lse.ac.uk