Available online at www.sciencedirect.com

SCIENCE DIRECT®

Research Policy 33 (2004) 1615–1634

research policy

www.elsevier.com/locate/econbase

Science and the diffusion of knowledge

Olav Sorenson[a,∗,1], Lee Fleming[b,1]

[a] Anderson Graduate School of Management, UCLA, 110 Westwood Plaza, Box 951481, Los Angeles, CA 90095-1481, USA
[b] Harvard Business School, Boston, MA 02163, USA

Received 13 February 2004; received in revised form 8 September 2004; accepted 20 September 2004

Abstract

Scientists, social scientists and politicians frequently credit basic science with stimulating technological innovation, and with it economic growth. Despite a substantial body of research investigating this general relationship, relatively little empirical attention has been given to understanding the mechanisms that might generate this linkage. This paper considers whether more rapid diffusion of knowledge, brought about by the norm of publication, might account for part of this effect. We identify the importance of publication by comparing the patterns of citations from future patents to three groups of focal patents: (i) those that reference scientific (peer-reviewed) publications, (ii) those that reference commercial (non-scientific) publications; and (iii) those that reference neither. Our analyses strongly implicate publication as an important mechanism for accelerating the rate of technological innovation: Patents that reference published material, whether peer-reviewed or not, receive more citations, primarily because their influence diffuses faster in time and space.

Keywords: Research; Publication; Reference

1. Introduction

Social scientists have long believed that the development of science has driven the rapid economic growth of the Western world. Both de Toqueville (1848) and Marx (1844), for example, pointed to science as a key progenitor of technological progress, and concomitantly of national wealth. At least three distinct lines of empirical research have offered evidence in support of this claim: Macroeconomic studies have linked GDP growth in the United States to higher scientific employment (Sveikauskas, 1981) as well as to increased private and public expenditures on research and development (e.g., Mansfield, 1972; Adams, 1990).[2] At

∗ Corresponding author. Tel.: +1 310 825 7348; fax: +1 310 825 1581.
E-mail address: osorenso@anderson.ucla.edu (O. Sorenson).
[1] Both authors contributed equally to this work.

[2] Not all findings, however, have supported such a link. Cross-national comparisons of the effects of public research on GDP growth, for example, have found a negative relationship between the two (Shenhav and Kamens, 1991; Schofer et al., 2000).

0048-7333/$ – see front matter
doi:10.1016/j.respol.2004.09.008

the firm level, management scholars have provided evidence that high technology companies that adopt a scientific research orientation (i.e. encouraging the self-direction of research, conference attendance, and publication) outperform those that do not (Henderson and Cockburn, 1994; Gambardella, 1995). And at the level of the invention, studies examining the number of future citations received by a patent – considered a measure of its value – find higher citation rates among patents originating from university labs (Jaffe and Trajtenberg, 1996; Mowery and Ziedonis, 2002). By focusing on the level of the invention, this study builds most closely on this third line of research.

Within this literature, three broad classes of mechanisms have been forwarded as explanations for the value of science to society.[3] One type of argument claims that science differs from other forms of research and development in terms of the issues on which it focuses, with scientists preferring to tackle more fundamental problems with a wider range of potential applications (Henderson et al., 1998; Mowery and Ziedonis, 2002). A second perspective maintains that science enhances the process of investigation itself: winner-take-all rewards to first discovery promote maximal effort (Merton, 1957; Dasgupta and David, 1994); scientific norms allow for the effective management of research activities (Owen-Smith, 2001); and the development of theories transforms search from trial-and-error learning to a more directed form of problem solving (Nelson, 1982; cf. Fleming and Sorenson, 2004). On the other hand, regardless of its influence on the problem solving process, science might also prove valuable by accelerating the diffusion of knowledge following its discovery. Rather than hoarding knowledge for private gain, the norm of communism (also referred to as openness) in science compels discoverers to disseminate it to others (Merton, 1942). To the extent that this dissemination reduces the degree of redundant effort, it too should increase the efficiency of research activities (Nelson, 1959; Arrow, 1962; Dasgupta and David, 1994).

Despite the large number of studies examining the general relationship between science, technological progress and economic growth, relatively little empirical attention has been given to the mechanisms that educe these effects.[4] This paper considers the plausibility of one factor, the norm of publication, as a link between science and rates of technological innovation. Though the norm of communism encourages scientists to disseminate knowledge through multiple channels – including conferences, personal correspondences, and the training of students – one channel in particular strikes us as unusually important: publication (Bernal, 1939). Absent publication, much of the knowledge gained from research remains private and must diffuse through interpersonal networks. Since these social networks typically remain localized in both physical and social space, transmission occurs most frequently among those in close proximity, and only slowly reaches more distant parties. Publication accelerates the flow of information by releasing the diffusion process from the limited set of contacts available through social networks.

To examine whether science accelerates the diffusion of knowledge, this study analyses the forward citations that patents receive from future patents. We study citations from these future patents to a set of focal patents that reference – or do not reference – various types of published materials. In particular, we compare three groups of focal patents: those that draw on published scientific (peer-reviewed) research, those that refer to other types of (non-scientific) published knowledge, and those that do not reference any non-patent publications. If the published information on which some of these patents draws diffuses more rapidly, we expect: (i) patents referencing either scientific or non-scientific published sources to receive more citations than those that do not reference these materials; (ii) that this citation premium exists even prior to the granting of the patent; (iii) that this citation premium should decline over time (as network diffusion substitutes for communication through publication); and that patents referencing either type of published ma-

[3] We do not explicitly define the practice of science because, as Merton (1942) noted, it encompasses a variety of norms and practices. By focusing on one type of behavior associated with science, that of publication, our analysis attempts to unpack the importance of this mechanism.

[4] A few notable exceptions exist: Owen-Smith (2001), for example, in a qualitative study details the ways in which the norms of science may offer effective organizing principles for research and development labs. Stern (2004) provides evidence that autonomy in research direction and the ability to publish may act as an effective incentive for high caliber researchers. And Fleming and Sorenson (2004) demonstrate with patent data that science appears particularly useful when applied to complex and interdependent technologies.

terial should receive future patent citations from (iv) more geographically dispersed inventors, and (v) more socially dispersed inventors, than patents that do not build on published data.

Our results confirm all of these effects; patents that reference published materials – whether scholarly or not – receive more citations, earlier, and from more geographically and socially distant inventors, consistent with the idea that publication accelerates the diffusion of information. Though these results usefully provide empirical evidence for the importance of information diffusion as a mechanism underlying the link between scientific research and innovation rates, they strike us as unsurprising. The more intriguing finding is that peer-reviewed journals do not appear to differ *in any way* in their effects on the rate and dispersion of future citations from commercial publications (e.g., marketing materials) and non-scientific journals. This fact suggests that the more rapid dissemination of knowledge may account for not just some, but perhaps even the largest share, of the apparent value of science in stimulating rates of invention. We discuss this possibility and its implications in greater depth in the final section.

2. Science, openness, and technological advance

Robert K. Merton, a sociologist, encouraged researchers to consider science as an institution. Through their training and both positive and negative career incentives, scientists internalize a set of values that guides their behavior. Though Merton (1942) identified several central norms in the scientific community – including universalism, communism, disinterestedness, and organized skepticism – this study focuses on the potential relevance of the norm of 'communism' to the rate of technological advancement, as a result of its influence on the process of knowledge diffusion.

'Communism' refers to the idea that individual scientists believe that their property rights over discoveries extend only to the credit associated with finding them (Merton, 1942). Other scientists may freely use knowledge generated by their predecessors as long as they give tribute to the original discoverer. At first glance, this norm might appear to weaken the incentives for innovation. But in fact, scientists receive powerful, though somewhat indirect, incentives to innovate because rewards – both in terms of recognition (e.g., citations, prizes) and resources (e.g., research grants, endowed chairs) – in the field invariably accrue to those that discover things first (Merton, 1957; Dasgupta and David, 1994). Moreover, the 'winner-take-all' nature of these races for primacy likely engenders particularly intense competition (Stephan, 1996).

Together, the norm of communism and the incentives surrounding primacy create an intense desire among scientists to publish research (Merton, 1942). Publication provides one means of conforming to the norm of communism, making one's ideas available to the scientific community; hence, scientists likely seek to publish their ideas as a course of habit. In addition to relaying new ideas to peers, however, publicly disseminating research establishes and allows one to defend a claim to primacy. Those wishing to garner accolades and riches in this community must therefore publish to establish their achievements. Moreover, universities, research institutions, and even firms engaged in scientific research typically monitor publications as an important criterion for promotions, providing further incentives to publish.

The act of publishing importantly alters the nature of the diffusion process.[5] Absent publication, knowledge must typically pass from researcher to researcher through face-to-face interaction (see Rogers, 1995, for a review). Publication, on the other hand, involves broadcast diffusion, conveying its information to any that can, and choose to, absorb it. To enable this shift from network to broadcast diffusion though, researchers must frequently codify tacit knowledge (Dasgupta and David, 1994). In enabling this codification, the institutional characteristics of science again play an important role: communities of scientists share highly specialized vocabularies and grammars that more readily encode complex information.[6] In addition, common bodies of tacit knowledge, disseminated through training and interactions with other

[5] In addition to journals, the scientific community has also generated a wide array of organizations that facilitate the flow of communication: conferences, departments, academies, etc. These associations influence the interaction patterns among researchers because they form conduits that shape the daily activities of researchers. By forming networks that bridge geographic space, these organizations also help to expand the range of network diffusion (Sorenson and Stuart, 2001; Owen-Smith and Powell, 2004).

[6] Of course, by their very nature, these specialized languages also exclude those outside the community (Eamon, 1985; Abbott, 1988).

scientists, mean that researchers need not fully specify the information they wish to transmit; rather, they can rely on readers to fill the gaps (Senker, 1995).

This accelerated diffusion of knowledge through publication may stimulate the rate of technological innovation. Inventors frequently draw on existing knowledge as a baseline for informing their directions of investigation (Rosenberg, 1990). By making knowledge freely available, publication opens newly acquired information to a much wider audience of inventors who could potentially build on it, allowing the community to develop technologies more rapidly. It also reduces the likelihood of duplicated effort (Bernal, 1939; Dasgupta and David, 1994); researchers can build on the work of others instead of "reinventing the wheel." Scientists themselves implicitly recognize the importance of this cumulative research effort, as one observes in statements such as Newton's "If I have seen farther, it is standing on the shoulders of giants." In addition, the codification of knowledge required for publication may further enhance the usefulness of information by making it more amenable to the application of search technologies that reduce the cost of identifying relevant information, and by ensuring that information survives its discoverer (Cowan and Foray, 1997).

3. Empirical strategy

To explore whether communication plays a role in the relationship between science and the rate of technological innovation, this study examines patents, their references to non-patent prior art (publications), and their forward citations (the prior art citations that they receive from future patents).[7] Patents provide a window on the generation and diffusion of knowledge through their references to other materials. They also, however, should provide a particularly difficult domain in which to find benefits associated with the public disclosure of knowledge. The very justification for offering patents, and the limited term monopolies that they grant, stems partially from the idea that society benefits from providing incentives for inventors to disclose the basis of their discoveries (Kitch, 1977). And indeed, patent offices make application materials publicly available when they award a patent; hence, the patent itself likely accelerates the diffusion of the knowledge embodied in it. Our ability to identify an additional effect to publication therefore depends either on a propensity for inventors to keep some portion of the underlying knowledge secret, or from the failure of potential inventors to monitor constantly all new patent awards. In the first case, publications might reveal some of the information not included in the paper, and in the second, publications may alert inventors to relevant knowledge available in patents. Though both seem likely, one should nonetheless read our analyses only as relatively conservative estimates of the importance of publication.

Patents report two types of "prior art": (i) those earlier patents on which the new invention builds, and (ii) other materials (e.g., scientific, technical, marketing, and other publications) that influenced the inventor. Inventors have an incentive to minimize these references. As patents represent a property right, both citations to other patents and references to prior publications can potentially reduce the scope of their claims and limit the effectiveness of any future legal actions defending them. The patent review process, however, places a check on the failure to cite prior art; based upon personal expertise and automated search techniques, patent examiners add relevant citations to applications in the review process (Carr, 1995).[8]

Our analysis assumes that references to both other patents and non-patent prior art identify cases where an invention builds, to some extent, on the knowledge embodied in these sources. We do not claim that these non-patent references inform other inventors of the focal inventions themselves; rather, we suspect that both patents draw on some latent (published or unpublished)

[7] Although other researchers have investigated the relationship between patents and non-patent references, they have typically explored small samples of references to the scientific literature using case study methods (for an exception, see Narin et al., 1997).

[8] Patent examiners appear to have much less influence on the assignment of non-patent prior art. In a small sample, Tijssen (2001), for example, found that most of these references came from the inventors themselves. In a series of interviews with patent holders, we asked them about why they had included such references. Inventors' accounts typically focused on two issues: many simply thought of citing relevant material – whether a patent or a paper – as the "right thing to do."A large proportion also felt that these non-patent references would add legitimacy to their applications, and hence, increase the likelihood of their patent being granted. Few, on the other hand, simply reference their own published papers; in our sample, only 3% of the inventors also appeared as authors on the non-patent references that their patents cited.

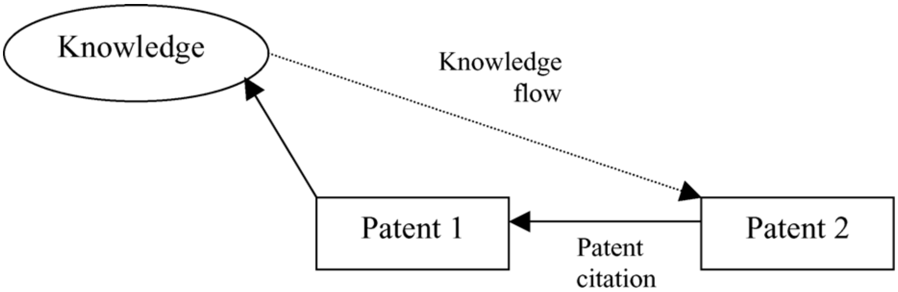

Fig. 1. Model of knowledge flows.

body of knowledge. Publication – whether in a peer-reviewed journal or not – broadens the set of potential inventors aware of this underlying information, thereby increasing the pool of individuals able to build on a particular knowledge base. Fig. 1 illustrates the conceptual model underlying our analysis. The boxes represent patents, while the oval denotes a piece of non-patent (potentially published) knowledge. In cases where patent 1 builds on published work, other inventors have access to this knowledge. We expect that this access increases the likelihood that they build on that knowledge, and hence cite patent 1 (as in the case of patent 2).

One might reasonably question, however, whether an inventor, such as in patent 2 in the figure, citing a prior patent (patent 1) actually has awareness of the non-patent prior art referenced by that prior patent. To assess the validity of this claim, we surveyed a sample of patent holders (for sampling details, see Fleming and Sorenson, 2004). When a patent holder cited prior patent art that had itself referenced non-patent material, we inquired as to their awareness of this non-patent prior art (i.e., in the figure, we would have asked the holder of patent 2 of their awareness of the non-patent references made by patent 1). In response to the question – "Were you aware of this/these publication(s) prior to or during the time of the invention?" – more than half (62%) indicated an awareness of at least some of the *specific* references listed on the patents that they cited. Moreover, when asked about their more general knowledge, the proportion of inventors noting an awareness of several similar articles increased to 71%.[9]

[9] Specifically, the question asked: "Were you aware of similar literature prior to or during the time of the invention (that is, while you did not read these particular articles, the titles, subjects, authors and journals look familiar to you)?".

Hence, for at least a large share of the inventors, our assumption appears sound.

3.1. Citation rates

How then does drawing on published sources (as opposed to private knowledge) influence the citations received by a particular patent? More rapid diffusion, as a mechanism, suggests that we should look for several patterns. The most basic evaluation involves looking simply at the number of citations received by patents that referenced published research relative to those that did not. To the extent that diffusion enlarges the pool of inventors that might potentially build on a piece of knowledge, publication likely increases the number of inventors working on closely related technologies. Hence, one would expect patents referencing published research to receive more citations.

On the other hand, these citations might also reflect the quality of the underlying knowledge (e.g., Trajtenberg, 1990). If scientific methods enhance the quality of invention, that too could generate a citation premium for patents that build on scientific research (Fleming and Sorenson, 2004). We gain purchase on excluding this second possibility by examining the effects of two distinct classes of non-patent references. Patents frequently draw upon knowledge published in refereed scientific journals, but they also sometimes reference other types of published materials (e.g., technical reports, marketing materials, and non-refereed sources). Though the higher number of citations received by patents drawing on peer-reviewed scientific articles may stem either from diffusion or from the higher quality of this knowledge, this potential confound does not afflict this second class of non-patent references. Although these other types of non-refereed materials should also enhance the diffusion process, references to them say little about the quality of the underlying knowledge. Hence, to the extent that increases in the number of citations arise from the faster diffusion of knowledge rather than variation in patent quality, references to both types of published materials should have similar effects:

H1. Patents that reference either scientific or non-scientific published material receive more citations from future patents than those that do not.

We can also make somewhat more nuanced predictions by considering the temporal distribution of these future citations. As noted above, patents themselves also disclose information. To some extent, the publication of the patent therefore acts as a substitute for other forms of public disclosure. The information embodied in the patent application, however, does not become available until the government grants and publishes the patent.[10] This timing effect potentially offers additional evidence. If references to published materials increase the citations received by an invention as a result of more rapid knowledge diffusion, one would expect references to published materials to have an even larger effect on the citations to an invention *prior* to its being granted a patent:

H2. References to published materials have an even larger positive effect on the number of citations received by patents prior to their grant dates.

Over time, additional substitutes arise for the broadcast diffusion provided by articles. Researchers present their results at conferences and in seminars. They train assistants and students who may build on this knowledge base, or who may act as carriers transferring it to others. Even the movement of information through interpersonal contact networks, such as through one-to-one correspondences, meetings, or collaborations, gradually diffuses knowledge to potential inventors. Knowledge moreover can travel at an increasing rate through social networks as those directly aware of the original invention pass it on to their contacts, who transmit it to their associates, and so on. Since all of these forms of transmission substitute over time for the diffusion of knowledge through publication, we further expect:

H3. The positive effect on the number of citations received by patents referencing published materials, relative to those that do not, declines over time.

3.2. Citation distribution

In addition to influencing the number of forward citations, diffusion processes should also affect the spatial distribution of these citations. When private information travels through interpersonal ties, it cannot escape the spatial limitations of the social network; however, when information becomes publicly available, its transmission should transcend these boundaries, flowing more freely to loosely- or un-connected individuals. By removing the flow of information from these constraints, publication should therefore accelerate the diffusion of ideas through space.

Several studies, beginning with Jaffe et al. (1993) have demonstrated that patent citations tend to localize in (geographic) space (cf. Singh, 2004). A compelling explanation for this effect comes from the literature on social networks and diffusion. In the absence of broadcast channels of diffusion, private and tacit information flows through interpersonal relations. Two factors contribute to propinquity in the structure of these networks: First, the likelihood that any two individuals meet in the course of their day-to-day activities – thereby having an opportunity to form a relation – declines rapidly with distance (Stouffer, 1950; Hawley, 1971; cf. Sorenson and Stuart, 2001). Second, even when a tie does occur, the costs of maintaining it increase, as actors must bridge increasingly wide expanses to interact (Zipf, 1949; Allen, 1977). As a result, the social networks linking inventors together connect most densely when these actors reside in close geographic and social proximity to one another (Bossard, 1932; Lazarsfeld and Merton, 1954), and consequently flows of private and tacit knowledge remain relatively localized.

By removing the diffusion process from these spatial constraints, publication should accelerate the flow of information along those dimensions on which social networks tend to cluster. Our data allow us to examine

[10] This practice has changed for recent applications. In 2001, the U.S. Patent and Trademark Office began publishing applications before deciding whether or not to award a patent. This new policy, however, does not affect the patents analysed here.

changes in the rate of information dispersal along three dimensions: geography, organizational boundaries, and technological communities. Let us first consider the geographic dimension. Actors in close physical proximity more frequently share a connection, thanks to their likelihood of interacting at neighborhood meetings, in coffee shops, at soccer leagues, school meetings, etc. (Festinger et al., 1950). The local character of labor markets also engenders a greater density of social relations among actors within a region, as they meet and interact in the workplace. These geographically localized social relations facilitate the movement of knowledge within regions, but stymie its flow to more distant locations. Publication, however, allows information to escape these networks to reach a broader range of inventors, hence:

H4. Patents that reference either scientific or non-scientific published material receive citations from more geographically distant inventors.

Both within and across regions, social relations also tend to cluster within organizational boundaries. The coordination of productive activities requires employees to interact with one another on a regular basis forming task-based relations – that may nonetheless carry other types of information. And even purely social relations arise more commonly within organizations. Employees of a firm typically occupy common buildings that engender serendipitous interactions with others sharing common interests and values that could lead to lasting social relations (cf. Allen, 1977). Moreover, incentives and social norms reinforce the salience of these boundaries in knowledge transmission. Even in cases where social relations do cross firm boundaries, in the absence of a common employer, knowledge holders may refuse to transfer their wisdom to others, thereby trapping private knowledge within the confines of the organization in which it originated. By making such information freely available, publication ought to increase dramatically the likelihood that individuals outside the firm can access and build upon it:

H5. Patents that reference either scientific or non-scientific published material more likely receive citations from inventors outside the firm.

Finally, absent publication, information also tends to flow primarily within technological domains. On the one hand, technological domains categorize problems. One would therefore expect most knowledge to have relevance to a particular class of technical problems, and consequently for most citations to cluster within these boundaries. But these domains also bound communities; university departments, corporate labs, professional societies and their meetings frequently organize around these technological areas. Like firms and regions, these technological communities also structure the interactions of the individuals that belong to them, as researchers more frequently interact with those tackling similar problems. As a result, even when one community discovers knowledge that has a direct and obvious application to a problem in another domain, one would expect it to reach that other domain slowly. Publication may, however, allow knowledge to cross these disciplinary boundaries more rapidly; both because it divorces the information from the constraints of the social networks, and because the ability to use automated search technology afforded by codification may prove particularly valuable across technological communities:

H6. Patents that reference either scientific or non-scientific published material receive citations from more technologically distant inventors.

4. Data and analysis

To explore our hypotheses, we made use of a sample of utility patents constructed in the course of prior research (for details, see: Fleming, 2001; Fleming and Sorenson, 2004). This sample offered several advantages with respect to the research question examined here: It covers a narrow time range (May–June 1990) of patents to limit the degree to which time-varying heterogeneity in the level of activity across fields influences our results.[11] It includes sufficient information on forward citations to examine the diffusion process. And, most important for our present purpose, a trained researcher had already categorized by type

[11] A comparison of our descriptive statistics to those found in Narin et al. (1997) does not suggest any significant differences between our sample and their 1987–1988 and 1993–1994 panels. Hence, we have no reason to believe that this selection limits the generalizability of our results.

each of the 16,728 non-patent references appearing in our sample.[12]

The researcher sorted these references into seven mutually exclusive and exhaustive categories (see Table 1 for the distribution):

Scientific Index journal: These publications appear in the *Scientific Index*. The journals in this category include both the familiar high prestige journals, such as *Science* and *Physica*, and a multitude of more obscure or non-English references – examples in the sample include *Chermetinformatsia* and *Cryogenic Engineering*. Although the quality of the journals indexed here undoubtedly varies, inclusion in the index denotes some level of acceptance in the scientific community. Therefore, we focused on the patents referencing these journals as representing the fruits of scientific research.

Conference proceedings: Though most of these conference proceedings refer to meetings for the presentation of scientific research, the standards at these meetings may not meet those necessary to secure publication in a peer-reviewed journal. One example of this type of reference appears in patent 4,922,432, which cites a paper entitled "The CMU design automation system – an example of automated data path design," in the *Proceedings of the 16th Design Automation Conference*. Automated data path design appeared many years earlier in an engineering textbook (Mead and Conway, 1980); hence, the paper might fail the peer review process of a journal for lack of novelty.

Technical report: Most of the items in this category refer to research institute publications, such as the Battelle Institute's "Final Report on High-Performance Fibers II, An International Evaluation to Group Member Companies."

Corporate publication – technical: Corporate technical publications typically refer to product specifications or schematics. Technical Bulletin BH183 (series) by Howell Instruments Inc. of Fort Worth, TX, U.S.A., "3"-Dia. Digital Indicators", provides an example of one in our data.

Book: Although the book once represented the primary mechanism for disseminating scientific knowledge, academics – especially those in the physical and biological sciences – have increasingly turned to journal articles as the preferred outlet for publishing research (Bazerman, 1988). Moreover, books face varying standards for publication depending on the imprint and market. Several books also appear simply to provide references for establishing facts. Titles in this sample include *Epoxy Resins*, *Kenkyuska's New Japanese-English Dictionary*, and *Oils, Fats and Fatty Foods*.

Corporate publication – non-technical: This category typically refers to marketing literature. Obviously, these publications do not aspire to the pretense of presenting technical, or even accurate, information. Some examples in our data include: "The Complete Guide to Roof Windows and Skylights", the Starrett Product Catalog and an advertisement entitled "Antec-Screen Printing Equipment Engineered for the Hands".

Non-index journal: This category subsumes all periodicals that do not appear in the *Scientific Index*. In practice, it contains popular magazines and industry newsletters, such as *Concrete and Cement Age*, *Guns and Ammo* and the *Harvard Business Review*, which do not ascribe to the standards of scientific research journals.

Some patents reference more than one type of literature. Table 2 reports the frequency with which two of these types of non-patent references appeared on the same patent. *Scientific Index* journals, conference proceedings and technical reports cluster together, suggesting a similarity between these sources in the information they carry.[13] Consistent with what one would expect if these sources reflect the influence of science, these materials also tend to appear on patents assigned to universities. By comparison, non-technical corporate materials and non-index journals, while clustering together, rarely appear on university patents.

To discriminate between publication and the effects of other mechanisms that might associate science with the rate of technological innovation, we compared citation patterns across patents that reference these various sources. Patents that reference journals in the

[12] One individual coded the entire sample. She began by comparing every reference to journals listed in the *Scientific Index*. She then proceeded based on the descriptions below. To assess the clarity and reliability of this classification schema, a second coder independently assigned a random sample of 100 references to these categories. The second coder agreed with the first on 96% of these cases. Given the consistency across independent coders, the classification scheme appears to identify distinct and meaningful categories.

[13] In an unreported analysis, we used all three types of references as indicators of science, as opposed to only those from journals in the *Scientific Index*. Doing so did not substantively change any of the results.

Table 1
Publication types, frequency and future citations[a]

Publication type	Total patents	Percent of sample	Average citations	
			5 year	10 year
University assignees	245	1.4	5.04*	9.67*
Scientific Index journal	3118	18.1	4.86*	8.77*
Conference proceedings	483	2.8	5.85*	11.46*
Technical report	334	1.9	4.95*	10.17*
Corporate publication–technical orientation	337	2.0	5.57*	11.35*
Book	922	5.3	4.16*	7.86*
Corporate publication–non-technical	711	4.1	4.52*	8.41*
Non-index journal	298	1.7	5.00*	10.75*
No references to publications	12,769	74.0	3.57	6.17

[a] 17,264 patents; the total across all categories exceeds this number because many patents reference more than one type of non-patent prior art.

* Indicates a mean significantly different from the baseline, patents without references to publications, at the $p<0.01$ level.

Table 2
Co-occurrence matrix[a]

	Patent category						
	2	3	4	5	6	7	8
1. University	151* (3.4)	29* (4.2)	30* (6.4)	7(1.4)	47* (3.6)	9 (0.9)	7 (1.7)
2. Scientific Index		345* (4.0)	194* (3.2)	177* (2.9)	538* (3.2)	131 (1.0)	142 (2.6)
3. Conference proceedings			75* (8.0)	61* (6.5)	116* (4.5)	35 (1.8)	38* (4.6)
4. Technical report				42* (6.4)	99* (5.6)	30 (2.2)	28* (4.9)
5. Corporate technical					79* (4.4)	71* (5.1)	37*(6.4)
6. Book						59 (1.6)	60* (3.8)
7. Corporate non-technical							54* (4.4)
8. Non-index							

[a] Each cell reports the number of patents with non-patent references to more than one type of material. For example, the number in the upper left cell indicates that 151 of the patents assigned to universities referenced a paper in a journal listed in the Scientific Index. The number in parentheses shows the ratio of actual co-occurrences to those expected by a random distribution (e.g., patents with university assignees reference conference proceedings 4.2 times as often as the average patent).

* Indicates a cell count significantly different from the marginal distribution at the $p<0.01$ level.

Scientific Index draw from published materials, but this peer-reviewed research may differ in multiple ways from other sources of knowledge. We therefore compared the distribution of their citations to those of patents that reference publications that clearly do *not* arise from scientific research: non-index journals and corporate non-technical publications. We examined the effects of these references on two dependent variables: (i) the number of citations received, and (ii) their distribution in space.

4.1. Citation rates

We began by considering simple mean citation rates. For each patent in the sample, we tracked all future patents that cite it through two time periods: the end of June 1995 (5 years following the sample grant dates), and the end of June 2000 (10 years following the sample grant date). Prior research suggests that most of the information available in forward citations appears in the first 5 years (Trajtenberg, 1990; Lanjouw and Schankerman, 2004). In both cases, we assigned citing patents to periods according to their application dates.

Table 1 reports the average number of citations received depending on the type of materials referenced by the patent. Patents referencing scientific (i.e. peer-reviewed) literature received more citations than patents that do not cite any non-patent references.[14] On average, these patents received 4.86 citations in the 5 years (or 8.77 in 10 years) following their grant-

Table 3
Variables in models of citation counts

	Full sample		Science	Corp non-tech	Non-index journal
	Mean	S.D.	Mean	Mean	Mean
Citations (5 year window)	3.84	4.94	4.86	4.52	5.00
(without self cites)	3.25	4.19	4.37	4.06	4.35
University assignee	0.01	0.12	0.05	0.01	0.02
Scientific index	0.65	2.52	3.61	0.64	2.78
(dummy coding)	0.18	0.38	1	0.18	0.47
Corporate non-technical	0.09	0.64	0.09	2.10	0.47
(dummy coding)	0.04	0.20	0.04	1	0.18
Non-index journal	0.03	0.25	0.08	0.13	1.55
(dummy coding)	0.02	0.13	0.05	0.08	1
Recent technological area	3.96	0.65	4.17	3.79	3.87
% NPR in subclass	0.21	0.22	0.45	0.25	0.35
Activity control	1.19	0.41	1.27	1.20	1.22
Number of classes	1.78	0.95	1.90	1.79	1.84
Number of subclasses	4.21	3.31	4.97	4.41	4.77
Number of backward cites	7.63	6.99	7.90	12.88	12.14

ing versus the 3.57 (6.17) citations received by patents that do not reference published sources, a premium of 36% (42%). Though patents referencing scientific articles received more cites, consistent with H1, *every* other form of publication also appears to accord the patents that reference it an increase in forward citations. Some of these categories, such as conference proceedings, also confound publication with other mechanisms that differentiate science from non-science, but at least two types of publications – the corporate non-technical and non-index journals – provide clean measures of the effects of publication. Patents that reference non-technical corporate publications received 4.52 citations on average, a 27% premium, while those that refer to non-index journals received five citations – 40% more than patents that do not reference any publications. Moreover, *t*-tests could not reject the possibility that *Scientific Index*, corporate non-technical and non-index journals all have equivalent effects on the future citations received by the patents that reference them.

The results suggest that building upon publicly available knowledge raises future citation rates, but these averages do not account for other factors that might distinguish patents that arise from science-based research from those that do not. To control for other sources of heterogeneity, we modeled the number of citations each of our 17,264 patents received using negative binomial regression.[15] The models included several important control variables. *University assignee* denotes those patents that belong to universities. It presumably controls to some degree for other ways in which academic research may differ from science-based research taking place outside of universities. Since inventions vary in their proximity to the technological frontier, we included *recent technical area* – the average of the patent numbers of the focal patent's prior art (higher numbers indicate more recent technology) – as a control.[16] An *activity control*, the number of citations a patent will likely receive conditional on its class memberships, accounted for differences in citation patterns across technological domains.[17]

[14] Although we focused on citations to non-patent references to test our hypotheses, to ease the comparison of our results to prior research (e.g., Henderson et al., 1998; Mowery and Ziedonis, 2002), we have also noted whether the patent comes from a university. As in earlier studies, universities received more citations to their patents than other assignees. Universities, however, accounted for only a small number of the patents citing the scientific literature (university patents also did not uniformly cite scientific articles; see Table 2).

[15] Because counts, such as the number of future citations, cannot fall below zero, linear regression can yield inefficient and biased coefficient estimates. Although researchers often use Poisson regression to model count data, our sample violates the assumption of a one-to-one mean/variance ratio, making the negative binomial a more appropriate specification (Cameron and Trivedi, 1986).

[16] This variable makes use of the fact that the USPTO assigns patent numbers sequentially. This assignment pattern means that the patent numbers correlate at .98 to the time since grant date in days.

[17] The USPTO classifies patents into one or more of roughly 400 technological classes. If all patents fell into a single class, we could

Table 4
Correlation matrix for variables in models of citation counts

	2	3	4	5	6	7	8	9	10	11	12
1. Cites (5 year window)	0.94	0.02	0.10	0.03	0.04	0.21	0.10	0.30	0.07	0.11	0.12
2. Cites (5 year window, no self cite)		0.03	0.08	0.04	0.04	0.18	0.09	0.29	0.07	0.10	0.10
3. University assignee			0.13	0.01	−0.00	0.04	0.15	0.01	0.03	0.02	−0.00
4. Scientific Index (dummy coding)				0.11	0.00	0.15	0.50	0.10	0.05	0.10	0.04
5. Non-index journal (dummy coding)					0.09	−0.02	0.08	0.01	0.01	0.02	0.09
6. Corporate non-technical (dummy)						−0.06	0.04	0.00	0.00	0.01	0.16
7. Recent technological area							0.28	0.38	0.04	0.07	−0.16
8. NPR in subclass (%)								0.20	0.08	0.13	−0.07
9. Activity control									−0.02	0.02	0.01
10. Number of classes										0.52	0.07
11. Number of subclasses											0.09
12. Number of backward cites											

The models also included controls for the *number of classes* and *the number of subclasses* into which the patent falls. Patents assigned to multiple classes may apply to a broader range of technologies, thereby expanding the number of future patents that might cite them. The *number of backward citations* also entered the models; patents that cite more prior art may describe more incremental inventions (Podolny and Stuart, 1995). Tables 3 and 4 describe the variables used in the count models.

Table 5 presents the results of these regressions. Model 1, which included only the controls, interestingly reveals that university assignees received 51% more (non-self) citations than similar patents from other types of institutions.[18] The second model added a dummy for having at least one reference to an article in the *Scientific Index*.[19] Patents referencing one of these papers received 19% more citations on average from other assignees. And after accounting for the effects of these publications, the citation premium enjoyed by university patents declined by 24%. Model 3 incorporated dummy variables to account for references to non-index journals and corporate non-technical publications. Consistent with our expectations, references to these publications predicted similar increases in forward citations: a reference to either a corporate non-technical publication or a non-science journal increased the expected citations from other assignees by 16%. *t*-Tests, in fact, indicate that we could not reject the possibility that all three types of publications generated the same size effect. The results therefore strongly support H1; if mechanisms other than publication accounted for a portion of the citation premium associated with the application of science, one would have expected patents referencing science to have had a larger coefficient than those referencing other types of publications.

We added two variables in model 4: An interaction between references to science and university assignee tested whether references to peer-reviewed publications had a different effect when they came from academic institutions. The insignificant coefficient estimate suggests that they did not; references to publications had equivalent effects for both academic and non-academic patent holders. The second variable allowed us to consider the degree to which technological domains might vary in systematic ways in the degree to which inventors draw on published materials (e.g., Fleming and Sorenson, 2004). To control for these effects, we created a variable, *proportion referencing NPR (non-patent references) in subclass*, using the 5

simply use fixed effects for each class; however, most patents (92%) fall into more than one class. Therefore, this amounts to a weighted mixture of the expected number of citations for a particular class based on citation activity from 1985 to 1990 (see Fleming, 2001, for details).

[18] The multiplicative specification of the negative binomial model makes multiplier rates (i.e. percentage changes) – obtained by taking the exponential of the coefficient estimate – the most intuitive means of assessing effect magnitude.

[19] Unreported models used the count of the number of publications referenced as a covariate. These models produced substantially identical results. We reported the models using the dummy coding for two reasons: (i) additional publications have little incremental effect; and (ii) the counts differ substantially across types of publications making it difficult to compare the coefficient estimates of the models with article counts.

Table 5
Negative binomial regression of forward patent citation counts[a]

	Model 1 (no self)	Model 2 (no self)	Model 3 (no self)	Model 4 (no self)	Model 5 (all cites)
Scientific Index (dummy)		0.174** (0.021)	0.167** (0.021)	0.119** (0.024)	0.121** (0.024)
Scientific Index X university assignee				−0.113 (0.134)	
Corporate non-technical (dummy)			0.148** (0.040)	0.139** (0.040)	0.105** (0.040)
Non-index journal (dummy)			0.151** (0.060)	0.141* (0.060)	0.144** (0.060)
University assignee	0.409** (0.065)	0.309** (0.065)	0.307** (0.066)	0.304** (0.103)	0.203** (0.067)
Recent technological area	0.137** (0.010)	0.138** (0.009)	0.142** (0.010)	0.139** (0.010)	0.122** (0.010)
Activity control	0.855** (0.021)	0.835** (0.021)	0.832** (0.021)	0.815** (0.021)	0.782** (0.021)
Number of classes (logged)	0.064** (0.020)	0.062** (0.020)	0.063** (0.020)	0.060** (0.020)	0.052** (0.020)
Number of subclasses (logged)	0.215** (0.016)	0.208** (0.016)	0.208** (0.016)	0.204** (0.016)	0.200** (0.016)
Number of backward citations (logged)	0.122** (0.011)	0.126** (0.011)	0.117** (0.011)	0.128** (0.012)	0.153** (0.012)
Proportion referencing NPR in subclass				0.203** (0.046)	0.139 (0.046)
Constant	−0.979** (0.045)	−0.989** (0.046)	−0.990** (0.046)	−1.01** (0.045)	−0.778** (0.045)
Alpha	0.779 (0.013)	0.774 (0.013)	0.772 (0.013)	0.771 (0.013)	0.802 (0.012)
Log-likelihood	−38965.5	−38930.3	−38919.1	−38908.8	−40980.2
N	17,264	17,264	17,264	17,264	17,264

[a] All models used a 5-year window for counting forward citations. With the exception of model 5, the dependent variable in all models included only those future citations coming from patents belonging to a different assignee from the focal patent (i.e. not self-citations).

* $p \leq 0.05$.

** $p \leq 0.01$; S.E. shown in parentheses.

years of patents preceding the sample (1985–1990). This measure records the number of patents in the same subclass as the focal patent that included any non-patent references, divided by the total number of patents in that subclass. The results remained robust to the inclusion of this measure, though the estimates from model 5 – which replicated the results for all citations (i.e. including self-citations) – indicate that research areas in which science plays an active role have an unusually low-level of self-citation (consistent with H5). This result suggests that using measures excluding self-citations to assess spillovers biases analysis in favor of universities, where much of the work developing a technology for market may occur at the behest of a licensed party (outside the university). Though not reported here, this general pattern remained robust to a wide variety of specifications: using 10-year counts of citations for the dependent variable; including self-citations in the dependent variable; and estimating with patent class or firm-level fixed effects.

Table 6 delves into the timing of these effects. Model 6 offers what one might consider the acid test for diffusion. The dependent variable included only those patents whose application dates *preceded* the grant date of the patent they cited. In other words, the future patent cited the focal patent before the focal patent application had been granted and published. The issuance of the patent itself therefore could not account for the transmission of information across inventors. It also excluded self-citations. Even under these carefully controlled conditions, references to any type of published literature significantly increased the number of citations received (though for non-refereed articles only at the $p < 0.10$ level). Moreover, consistent with H2, references to *Scientific Index* articles had an even larger effect on future citations before the granting and publication of the patent (t-test = 2.68, relative to model 3).[20] In model 7, we investigated this effect further by including a (logged) count of the total number of non-patent references in addition to the dummy variables for the three types of principal interest.[21] The results again suggest that the three publication types do not vary in how they influence forward citations.

[20] To compare coefficients across models, we assume that the coefficients come from independent distributions with means at the point estimate and a standard error equivalent to that of the estimates.

[21] We added a constant of one to avoid taking the log of zero.

Table 6
Negative binomial regression of forward patent citation counts[a]

	Model 6 Pre-grant (no self)	Model 7 Pre-grant (no self)	Model 8 0–1 year (all cites)	Model 9 1–3 years (all cites)	Model 10 3–5 years (all cites)
Scientific Index (dummy)	0.272** (0.033)	0.080 (0.057)			
Corporate non-technical (dummy)	0.141* (0.069)	−0.092 (0.075)			
Non-index journal (dummy)	0.165 (0.097)	0.053 (0.100)			
Number of non-patent references (logged)		0.149** (0.035)	0.125** (0.015)	0.101** (0.015)	0.103** (0.017)
University assignee	0.350** (0.102)	0.285** (0.102)	0.234** (0.081)	0.191* (0.081)	0.233* (0.093)
Recent technological area	0.156** (0.020)	0.157** (0.019)	0.134** (0.013)	0.137** (0.013)	0.126** (0.015)
Activity control	1.12** (0.033)	1.12** (0.033)	0.952** (0.025)	0.751** (0.025)	0.669** (0.029)
Number of classes (logged)	0.118** (0.034)	0.116** (0.034)	0.085** (0.025)	0.076** (0.025)	0.014 (0.028)
Number of subclasses (logged)	0.264** (0.026)	0.260** (0.026)	0.196** (0.020)	0.181** (0.019)	0.212** (0.022)
Number of backward citations (logged)	0.080** (0.020)	0.076** (0.020)	0.126** (0.014)	0.150** (0.014)	0.170** (0.017)
Constant	−3.49** (0.088)	−3.49** (0.088)	−2.19** (0.061)	−1.69** (0.059)	−1.84** (0.067)
Alpha	0.847 (0.040)	0.828 (0.022)	0.830 (0.022)	0.952 (0.021)	1.256 (0.028)
Log-likelihood	−14577.3	−14566.9	−25649.5	−28190.7	−24938.6
Mean cites/patent	0.45	0.45	1.17	1.51	1.16

[a] Categories based on application dates (i.e. the dependent variable for model 8 counts all forward citations from patents filed after, but less than 1 year from, the grant date of the focal patent).
* $p \leq 0.05$ (17,264 cases).
** $p \leq 0.01$; S.E. shown in parentheses.

Models 8 through 10 split forward citations into three groups according to their application date: ≤1 year, 1–3 years, and 3–5 years after the grant date of the patent they cite. Because the publication types do not appear to matter (model 7), and because the cell counts become small as we split the time periods, we used the logged count of non-patent references to investigate the timing effects. As expected in H3, the magnitude of the citation premiums accorded to patents referencing published literature declined from the first year following the grant date to later years. Though one might worry that the temporal distribution of patent cites drives this result, this seems quite unlikely as the decline in the magnitude of the publication effect occurs within the range of the data with an increasing unconditional citation rate. *t*-Tests, however, indicated that we could not reject the possibility that the premiums do not decline over time.

4.2. Citation distribution

To analyse the diffusion of citations in geographic and social space, we modeled the probability that a future patent cites a given focal patent as a function of distance, publication, and a variety of control variables. Dyads of future patents and focal patents thus became the unit of analysis. To avoid the potential problems associated with estimating non-independent cases, we assembled the data for this analysis using a matched sample design; in other words, we paired a set of future patents that cited our focal patents with a second set that did not.[22] Our sample included all 60,999 citations that actually occurred in our sample of 17,264 patents. In addition, we matched each of the 17,264 focal patents with four future patents that did not cite it (but that could have).[23] Though this generated a data set of 130,055 dyads, our analyses restricted the sample used for estimation to the 72,801 cases, where both inventors resided in the U.S. (the findings nonetheless remained robust to the inclusion of foreign inventors).

[22] Estimating results on a complete matrix of possible dyads can contribute to network autocorrelation and inefficient estimates (Sorenson and Stuart, 2001).

[23] We chose four patents for the 'control' group, so that the sample would have a roughly equal proportion of realized and unrealized dyads. To address the fact that focal patents enter the data more than once, we report robust standard errors.

To account properly for the effects of the sampling procedure, our estimations employed the rare events logistic regression method suggested by King and Zeng (2001) (2001; cf. Sorenson and Stuart, 2001). Logistic regression can yield biased estimates when the proportion of positive outcomes (citations in this case) in the sample does not match the proportion in the population. Our matching procedure, for example, generated a sample with a higher proportion of citations than found in the population as a whole. As a result, uncorrected logistic regression would underestimate the factors that predict a positive outcome (King and Zeng, 2001).[24]

The analysis of dispersion requires measures of distance. All patents list the home address of the inventor on the front page of the patent application. To locate inventors, we matched the inventors' three-digit zip codes to the latitude and longitude of the centers of the areas in which they lived based on information from the U.S. Postal Service.[25] We calculated the distance between two points using spherical geometry; taking the natural log of this value accounted for the fact that the relevance of a mile declines with distance.

One could also think of distance in the technological sense. To create a measure of the distance between two patents in technological space, we calculated the overlap in subclass assignments between each focal patent, i, and each potentially citing patent, j:

$$o_{ij} = 1 - \frac{s_i \cdot s_j}{|s_i|},$$

where s represents a vector of subclass assignments, with each cell being a binary indicator variable of membership in a subclass (i.e. 1 denoting assignment to the subclass). The measure ranges from 0 to 1, with larger values representing more distant technologies.

In addition to the distance measures and the control variables used in the negative binomial models, the citation probability models included several additional controls. *Same assignee* essentially indicates self-citations. *Same zip* takes a value of one when the inventors of both patents reside within the same three-digit zip code. *Same class* denotes dyads where both patents belong to the same primary class. And *time (application date)* measures the difference in years between the grant date of the focal patent and the application date of the potentially citing patent.

The results of these regressions appear in Tables 7 and 8. The first column (model 11) simply shows that both patents from universities and those that reference articles from journals in the *Scientific Index* have a higher probability of receiving a citation from any future patent. Models 12 and 13 investigated the dispersion of these cites across both geographic and technological space by interacting *Scientific Index* with both of the distance measures. Including terms to account for the distance dramatically improved the model; all of the coefficients imply that proximity increases the likelihood that a future patent cites the focal patent. As we expected, references to the scientific literature reduce this localization: patents referencing a *Scientific Index* article (i) more likely received citations from geographically distant patents (in support of H4), (ii) more commonly escaped organizational boundaries (consistent with H5), and (iii) more frequently received citations from technologically remote patents (as predicted in H6).

The magnitude of these effects appears quite substantial. One cannot easily interpret coefficients in isolation in logistic regression because their effects change depending on the levels of the other variables. Our calculations estimated the effect of changing publication and distance, setting all other variables to their mean levels across the population analysed. The relative risk that a forward patent cited the focal patent declined by 88% from a location just outside the focal patent's zip code to one 2000 miles away, if a patent references no publications. For patents referencing *Scientific Index* publications, however, the change in relative risk over the same distance dropped to 44%. With regard to technological distance, the relative risk of citation dropped by 98% from a future patent with complete overlap

---

[24] King and Zeng (2001) demonstrate that the following expression describes the bias in β generated by estimating logistic regression on a sample with an oversampling of rare events: $\text{bias}(\hat{\beta}) = (X'WX)^{-1}X'W\xi$, where $\xi = 0.5\,Q_{ii}[(1+w_1)\hat{\pi}_i - w_1]$, the Q are the diagonal elements of $Q = X(X'WX)^{-1}X'$, $W = \text{diag}\,\{\hat{\pi}_i(1-\hat{\pi}_i)w_i\}$, and w_1 represents the fraction of ones (citations) in the sample relative to the fraction in the population.

[25] The USPTO includes five-digit zip information, but we chose to reduce the measurement error by using cleaned data. CHI, an information provider, has called every patent holder to verify the inventor's location; however, it only records this information at the three-digit level. Where the patent lists more than one inventor, we assigned a location using the residence of the first inventor listed. Models where we randomly selected a location from the listed inventors, however, yielded equivalent results.

Table 7
Rare events logit models of the likelihood of a focal patent receiving a citation from a future patent[a]

	Model 11	Model 12	Model 13	Model 14
Scientific Index (dummy)	0.307** (0.019)	−0.354 (0.355)	−0.899* (0.428)	−4.38** (0.528)
Scientific Index X log (distance)		0.206** (0.072)	0.451** (0.079)	4.03** (0.152)
Scientific Index X log (distance) X time				−0.813** (0.031)
Scientific Index X subclass overlap		2.00** (0.831)	4.38** (0.708)	8.02** (0.716)
Scientific Index X subclass overlap X time				−0.312** (0.096)
Scientific Index X same assignee			−1.80** (0.372)	−1.30** (0.463)
University assignee	0.252** (0.057)	−0.514* (0.261)	−1.42** (0.296)	−5.01** (0.366)
Log (distance)		−0.203** (0.053)	−0.173** (0.062)	−0.029 (0.064)
Subclass overlap		−4.09** (0.396)	−3.85** (0.298)	−3.70** (0.300)
Same assignee		0.503** (0.124)	0.516** (0.143)	0.283 (0.165)
Same zip		2.49** (0.487)	3.11** (0.426)	1.60* (0.824)
Same class		4.33** (0.416)	4.57** (0.260)	4.64** (0.455)
Activity control			0.914** (0.343)	0.841* (0.364)
Number of classes			0.261** (0.101)	0.177 (0.096)
Number of subclasses			−0.109** (0.037)	−0.208** (0.037)
Number of backward citations			−0.011 (0.021)	−0.036 (0.020)
Recent technological area			0.275 (0.153)	0.570** (0.135)
Time (application date)				0.312** (0.096)
Constant	−10.5** (0.008)	−9.03** (0.342)	−12.6* (0.530)	−14.4** (0.632)
Log-likelihood	−50097.5	−14068.3	−13279.2	−9497.7

[a] 76,807 cases, 51.7% represent ties (vs. 0.00059% in population).
* $p \leq 0.05$.
** $p \leq 0.01$; S.E. shown in parentheses.

Table 8
Rare events logit models of the likelihood of a focal patent receiving a citation from a future patent[a]

	Model 15	Model 16
Non-index journal (dummy)	−6.98** (0.791)	
Non-index X log (distance)	1.16** (0.119)	
Non-index X different class	3.21** (1.29)	
Corporate non-technical (dummy)		0.836** (0.306)
Corporate non-technical X log (distance)		0.134** (0.035)
Corporate non-technical X subclass overlap		3.59** (0.835)
University assignee	−0.260 (0.347)	−0.207 (0.359)
Log (distance)	−0.252** (0.073)	−0.246** (0.074)
Subclass overlap	−3.61** (0.310)	−3.76** (0.316)
Same assignee	0.401* (0.170)	1.11** (0.165)
Same zip	2.55** (0.485)	2.58** (0.495)
Same class	4.32** (0.308)	4.22** (0.306)
Activity control	0.766** (0.351)	0.712** (0.354)
Number of classes	0.166 (0.189)	0.212 (0.187)
Number of subclasses	−0.045 (0.079)	−0.085 (0.072)
Number of backward citations	−0.011 (0.019)	−0.022 (0.019)
Recent technological area	0.057 (0.192)	0.229 (0.193)
Constant	−10.9** (0.938)	−11.4** (0.928)
Log-likelihood	−9944.6	−9944.7

[a] 76,807 cases, 51.7% represent ties (vs. 0.00059% in population).
* $p \leq 0.05$.
** $p \leq 0.01$; S.E. shown in parentheses.

in subclass assignments to one with no overlap, a ratio that declined to 85% for focal patents referencing a *Scientific Index* journal. And with respect to firm boundaries, the relative risk of citation actually shifted from predicting a higher rate of self-citation, for patents that do not reference non-patent materials, to favoring citations from outside the firm for focal patents referencing a *Scientific Index* publication.

Time also plays an important role in diffusion processes. To test H3 further, we interacted the interactions between a *Scientific Index* reference and distance with time (model 14). The estimates revealed that the degree to which references to published material increased the likelihood of citation from distant sources declined over time, as one would expect if other mechanisms gradually substitute for published communication. In fact, the estimates suggest that after 5 years references to scientific material no longer increase the likelihood of citation from geographically distant patents. Similarly, the degree to which these references enhance citations across technological distance declines to zero after 25 years (though one should view this calculation with caution as it lies well outside the range of the data used to estimate it). The dyadic analysis thus supported H3.

Table 8 shows that similar patterns of results held for non-technical corporate publications and non-refereed journals. Consistent with the idea that publication, as opposed to other mechanisms, generates the link between science and the rate of technological innovation, these other types of publications also sped the diffusion of citations across space – supporting H4, H5 and H6.

5. Discussion

Our results support the idea that at least a portion of the citation premium accorded to science- and university-based patents stems from the fact that they draw on publicly available knowledge. Although patents from universities and those that make reference to scientific publications receive more citations than those that do not reference non-patent material, patents that reference *any* type of publication – including items such as popular magazines and marketing materials – exhibit a similarly large number of citations from future patents. Hence, the number of citations received by patents referencing the scientific literature appears to reflect the fact that these references indicate a broad dissemination of knowledge through publication. The spatial distribution of these future citations further supports this conclusion. Patents making reference to publications more frequently receive citations from geographically distant inventors, from inventors outside the firm, and from those operating in different technological areas. Together these results provide strong evidence for the notion that science accelerates innovation because its norms of openness and publication speed the diffusion of knowledge.

One might reasonably ask, however, just how much of the value of science we can attribute to the more rapid dissemination of knowledge. Two results from our analysis point to a rather large effect. First, when comparing the differences in forward citations associated with different types of non-patent references, we could not reject the hypothesis that all types of publications generated the *same* level of effects. To the extent that peer-reviewed publications capture other mechanisms, such as differences in the problem solving process, our results suggest that the importance of these other mechanisms pales in comparison to the benefits garnered from faster information diffusion. Second, in our analyses of citation dispersion, the main effect of references to *Scientific Index* articles turns negative after accounting for social and geographic distance. A direct interpretation suggests that in the absence of diffusion, these inventions would have produced *fewer* forward citations – and presumably less societal value – than patents building on scientific research.[26] Though the magnitude of these results may seem surprising, they correspond to those derived from other identification strategies: Reitzig (2004), for example, using a structural estimation approach, which allows him to distinguish between technical value (novelty) and non-technical market value, finds no evidence that non-patent references indicate either higher technical or higher non-technical value.

Although we believe that the results strongly implicate communication as a factor in linking science to the rate of technical innovation, some alternative explanations warrant consideration. The most obvious concerns the possibility that publications might reflect

[26] Of course, the winner-take-all nature of patent races makes it difficult to interpret the main effects from the diffusion analysis, since they reflect the *distribution* of future activity rather than the rate.

higher quality inventions (despite Reitzig's, 2004, evidence to the contrary). Though one might expect higher quality inventions to also receive a larger number of citations, this explanation fails to explain the broader pattern of results. First, one would need to believe that both refereed articles as well as all other types of publications reflect higher quality. Though possible, the story behind such a relationship would probably claim that patent holders who have a better sense of the value of their inventions more likely publicize them. That account nevertheless appears unlikely here for two reasons: (i) few of the inventors reference their own published materials (about 3%); and (ii) publications *decrease* the likelihood of self-citation (but certainly if patent holders recognized the value of these inventions they would build on them themselves). Even if one accepts that all types of publications might reflect differences in value, one must still construct a story for why that value should localize in space; variations in the rate at which citations move across geographic and social boundaries fits with a diffusion-based account, but not with one that argues for quality differences across these patents.

Others contend that science differs in the *type* of knowledge it produces, rather than on some hierarchical dimension of quality. The diffusion effects may account for this claim as well. Henderson et al. (1998), for example, argue that universities produce patents that differ from industry-based research in the generality of the knowledge generated; in other words, university research applies to a broader range of inventions. Their evidence for this difference comes from the number of future citations patents receive outside of their own technological class. Differences in the diffusion of knowledge would generate the same pattern of effects: Networks localize in social space as well as geographic space; researchers more likely know other researchers within their own field. Information traveling through social networks therefore diffuses slowly across technological fields. Publication removes the dissemination of knowledge from the constraints of these networks, thereby increasing the flow of information across technological fields much as it does across geographic space (H6), an expectation supported by the citation distribution analysis.

Our findings more generally suggest that researchers should use citation counts as a dependent variable with particular caution. Many studies have found a strong relationship between forward citations and patent value (e.g., Trajtenberg, 1990; Lanjouw and Schankerman, 2004; Reitzig, 2004). Uniformly, however, value explains only a fraction of the variance in cumulative citations in these studies. Though other factors also likely play a role, our results demonstrate that communication differentials account for a portion of this unexplained variance. Therefore, though we would agree that, on average, future citations correlate with the value of an invention, our findings indicate that people should exercise substantial care when drawing conclusions in the opposite direction: that factors that increase future citations imply higher value.

A third potential alternative might account for our geographic diffusion findings. When citations localize, two factors could explain it: On the one hand, information might diffuse slowly from its place of origin as it moves through (highly localized) social networks. On the other hand, citations might cluster simply because everyone researching the problem resides in a concentrated geographic area (Jaffe et al., 1993). We addressed this issue in three ways. First, we created a distance control by averaging the mean distance between all dyadic pairs within a class. The inclusion of this term did not affect the interaction terms between publication and spatial distance. Second, we estimated a set of (unreported) models predicting the distance between citing and cited patent including fixed-effects by class to capture differences in the distribution of activity. These models also found that patents referencing published sources received cites from more distant inventors. Third, we generated a sample using the same methodology as Jaffe et al. (1993) and estimated the probability of within region citations; the results again revealed that references to published materials reduced the degree of localization (by this method, a journal reference decreased the probability of a citation within the same three-digit zip by 12% and within the same state by 14% – both effects significant at $p < .001$). Though technological fields vary in their geographic distribution of activity, these differences do not appear to explain our results.

Our findings speak to a range of policy implications. At the firm level, for example, the fact that publication extends the flow of knowledge offers interesting insight into the burgeoning literature on the growing use of science in for-profit organizations. These studies frequently argue that firms that adopt the norms of science experience superior performance compared to those that do not (e.g., Henderson and Cockburn, 1994). Although many of these studies intimate that science improves research quality, our results point instead to two alternative explanations: on the one hand, our findings appear to complement recent work by Stern (2004), which argues that firms benefit by promoting science not because it improves the R&D process, but rather because talented employees desire the ability to publish (allowing firms to pay lower wages). Another possible explanation stems from the fact that publication reduces the constraints of location. Firms must typically locate near to critical information resources because access to them comes through spatially constrained social networks (Sorenson and Stuart, 2001). By drawing information from published sources, however, firms that do science-based research may gain flexibility in location, allowing them to avoid labor market competition by locating facilities far from rivals. Like Stern (2004), this possibility points to reductions in labor costs, as well as potentially in spillovers of private knowledge, as the advantages of adopting a scientific orientation.

From a public policy point of view, our results affirm the importance of public disclosure. As noted above, much of the societal benefit to science-based research in our analysis appears to stem from the dissemination of information. The rapid diffusion of knowledge can improve the efficiency of research investments by reducing the degree of effort duplication. To the extent that this reduction in overlapping research activities stimulates innovation rates (and potentially economic growth), public funding sources should consider requiring timely publication of the results of the research projects they support to maximize societal benefits. More broadly, the importance of openness also points to the potential value of promoting other mechanisms that facilitate the availability of information. Data centers that provide access to the raw data underlying analyses, for instance, may reduce the costs of replicating and building upon prior results (King, 1995). Similarly, resource centers containing the physical artifacts of research might further stimulate the diffusion of knowledge in cases where a portion of that knowledge eludes codification. Furman and Stern (2004), for example, find that the donation of materials to biological resource centers dramatically increases (more than doubles) the pace of research examining these organisms. Such policies, moreover, may become increasingly important as scientists and universities become more and more interested in garnering commercial gains, in addition to (or even in lieu of) academic accolades, from research.

The importance of openness also points to a potential dark side to the increasingly strong linkage between industry and science. Over the past two decades, for-profit firms have increasingly substituted for public institutions as a source of financial backing for academic research (now estimated at 20–25%, Behrens and Gray, 2001). Though some have hailed this trend as positive, bringing the discipline of the market to the allocation of research funds and encouraging academics to consider more applied technological problems, it seems likely that industrial sponsors may also seek to restrict access to the research that they support in the hopes of appropriating the potential rewards – maximizing firm profitability, but not necessarily public welfare (witness the current debate over the disclosure of negative clinical trial results for pharmaceuticals). Future work therefore should consider this issue carefully as the importance of openness may justify legislation or shifting the balance back toward greater public sponsorship of academic research.

Although the belief that science promotes economic growth plays an important role both in the allocation of resources to science and to its status in society, we lack a clear understanding of both the accuracy of this belief and the mechanisms underlying the relationship. Do scientists actually generate more useful knowledge? This paper takes one step toward resolving this ambiguity by demonstrating that science appears to benefit technological innovation to a large extent by expanding the flow of information in geographic and social space. This acceleration of the diffusion process occurs because incentives within the scientific community encourage scientists to publish their research, thereby making the information publicly available. In the absence of publication, knowledge must diffuse through the relatively more viscous medium of interpersonal social networks. Science thus appears to stimulate innovation by expanding the spatial reach of spillovers.

Acknowledgements

Funding from the Harvard Business School supported this research. Corey Billington and Ellen King of Hewlett-Packard graciously provided the patent data, and Sue Borges contributed superb research assistance. Feedback from the participants of the NBER Productivity Seminar, the Harvard Workshop on Applied Statistics and the HBS Technology and Operations Management Seminar, and particularly comments from Michael Darby, Rebecca Henderson, Gary King, Jim March, Santiago Mingo, Woody Powell, Scott Stern, Art Stinchcombe, Peter Thompson, and three anonymous reviewers proved useful in the development of the paper. All remaining faults remain our own.

References

Abbott, A., 1988. The System of Professions: An Essay on the Division of Labor. University of Chicago Press, Chicago.

Adams, J.D., 1990. Fundamental stocks of knowledge and productivity growth. Journal of Political Economy 98, 673–702.

Allen, T., 1977. Managing the Flow of Technology. MIT Press, Cambridge, MA.

Arrow, K., 1962. Economic welfare and the allocation of resources for invention. In: Nelson, R.R. (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton University Press, Princeton, NJ, pp. 609–625.

Bazerman, C., 1988. Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. University of Wisconsin Press, Madison, WI.

Behrens, T.R., Gray, D.O, 2001. Unintended consequences of cooperative research: impact of industry sponsorship on climate for academic freedom and other graduate student outcome. Research Policy 30, 179–199.

Bernal, J.D., 1939. The Social Function of Science. MacMillan, New York.

Bossard, J.S., 1932. Residential propinquity as a factor in marriage selection. American Journal of Sociology 38, 219–224.

Cameron, A.C., Trivedi, P.K., 1986. Econometric models based on count data: comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1, 29–53.

Carr, F.K., 1995. Patents Handbook: A Guide for Inventors and Researchers to Searching Patent Documents and Preparing and Making an Application. Jefferson, NC: McFarland.

Cowan, R., Foray, D., 1997. The changing economics of codification and the diffusion of knowledge. Industrial and Corporate Change 6, 595–622.

Dasgupta, P., David, P., 1994. Towards a new economics of science. Research Policy 23, 487–521.

Eamon, W., 1985. From the secrets of nature to public knowledge: origins of the concept of openness in science. Minerva 23, 321–347.

Festinger, L., Schacter, S., Back, K.W., 1950. Social Pressure in Informal Groups. Harper, New York.

Fleming, L., 2001. Recombinant uncertainty in technological search. Management Science 47, 117–132.

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strategic Management Journal 25, 909–928.

Furman, J., Stern S., 2004. Climbing atop the shoulders of giants: the impact of institutions on cumulative research. Working paper, Boston University.

Gambardella, A., 1995. Science and Innovation: The U.S. Pharmaceutical Industry During the 1980s. Cambridge University Press, Cambridge.

Hawley, A.H., 1971. Urban Society. Ronald, New York.

Henderson, R., Cockburn, I., 1994. Measuring competence? Exploring firm effects in drug discovery. Strategic Management Journal 15 (winter special issue), 63–84.

Henderson, R., Jaffe, A.B., Trajtenberg, M., 1998. Universities as a source of commercial technology: a detailed analysis of university patenting, 1965-1988. Review of Economics and Statistics 80, 119–127.

Jaffe, A.B., Trajtenberg, M., 1996. Flows of knowledge from universities and federal labs: modeling the flow of patent citations over time and across institutional and geographic boundaries. Proceedings of the National Academy of Sciences 93, 12671–12677.

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

King, G., 1995. Replication, replication. PS: Political Science and Politics 28, 443–449.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political Analysis 9, 137–163.

Kitch, E.W., 1977. The nature and function of the patent system. Journal of Law and Economics 20, 265–290.

Lanjouw, J.O., Schankerman, M., 2004. Patent quality and research productivity: measuring innovation with multiple indicators. Economic Journal 114, 441–465.

Lazarsfeld, P.F., Merton, R.K., 1954. Friendship as a social process: a substantive and methodological analysis. In: Berger, M., Abel, T., Page, C.H. (Eds.), Freedom and Control in Modern Society. Van Nostrand, New York, pp. 18–66.

Mansfield, E., 1972. Contribution of R&D to economic growth in the United States. Science 175, 477–486.

Marx, K., [1844] 1975. Economic and philosophical manuscripts. In: Early Writings, Rodney Livingstone and Gregor Benton (trans.), New York Vintage Books, pp. 279–400.

Mead, C., Conway. L., 1980. Introduction to VSLI Systems. Addison-Wesley Reading, MA.

Merton, R.K., 1942. Science and technology in a democratic order. Journal of Legal and Political Sociology 1, 115–126.

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

Mowery, D.C., Ziedonis, A.A., 2002. Academic patent quality and quantity before and after the Bayh-Dole act in the United States. Research Policy 31, 339–418.

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

Nelson, R.R., 1959. The simple economics of basic scientific research. Journal of Political Economy 67, 297–306.

Nelson, R.R., 1982. The role of knowledge in R&D efficiency. Quarterly Journal of Economics 97, 453–470.

Owen-Smith, J., 2001. Managing laboratory work through skepticism: processes of evaluation and control. American Sociological Review 66, 427–452.

Owen-Smith, J., Powell, W.W., 2004. Knowledge networks as channels and conduits: the effects of spillovers in the Boston biotechnology community. Organization Science 15 (1), 5–21.

Podolny, J.M., Stuart, T.E., 1995. A role-based ecology of technological change. American Journal of Sociology 100, 1224–1260.

Reitzig, M., 2004. Technical quality, market potential and the value of inventions: what do patent indicators really measure? Working paper, Copenhagen Business School.

Rogers, E.M., 1995. Diffusion of Innovations, fourth ed. Free Press, New York.

Rosenberg, N., 1990. Why do firms do basic research (with their own money)? Research Policy 19, 165–174.

Schofer, E., Ramirez, F.O., Meyer, J.W., 2000. The effects of science on national economic development 1970 to 1990. American Sociological Review 65, 866–887.

Senker, J., 1995. Tacit knowledge and models of innovation. Industrial and Corporate Change 4, 425–447.

Shenhav, Y., Kamens, D., 1991. The ‘costs’ of institutional isomorphism in non-western countries. Social Studies of Science 21, 427–545.

Singh, J., 2004. Collaborative networks as determinants of knowledge diffusion patterns. Working paper, Harvard Business School.

Sorenson, O., Stuart, T.E., 2001. Syndication networks and the spatial distribution of venture capital investments. American Journal Sociology 106, 1546–1588.

Stephan, P., 1996. The economics of science. Journal of Economic Literature 34, 1199–1235.

Stern, S., 2004. Do scientists pay to be scientists? Management Science 50, 835–853.

Stouffer, S.A., 1950. Intervening opportunities: a theory relating mobility and distance. American Sociological Review 5, 845–867.

Sveikauskas, L., 1981. Technological inputs and multifactor productivity growth. Review of Economics and Statistics 63, 275–282.

Tijssen, R.J.W., 2001. Global and domestic utilization of industrial relevant science: patent citation analysis of science-technology interactions and knowledge flows. Research Policy 30, 35–54.

de Toqueville, A., [1848] 1966. Democracy in America (George Lawrence, trans.). New York Harper & Row.

Trajtenberg, M., 1990. A penny for your quotes: patent citations and the value of innovations. Rand Journal of Economics 21, 172–187.

Zipf, G.K., 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley Reading, MA.