

SCIENCE AND THE DIFFUSION OF KNOWLEDGE*

Olav Sorenson

Anderson Graduate School of Management
University of California, Los Angeles
110 Westwood Plaza, Box 951481
Los Angeles, CA 90095-1481
osorenso@anderson.ucla.edu

Lee Fleming

Harvard Business School
Harvard University
Morgan T97
Boston, Ma. 02163
lfleming@hbs.edu

April 12, 2001

Word count: ~ 5,400

PRELIMINARY DRAFT

* Both authors contributed equally to this work. Funding from the UCLA Academic Senate and 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. Conversations with Michael Darby and Jim March contributed to the development of the paper. All remaining faults remain our own.

SCIENCE AND THE DIFFUSION OF KNOWLEDGE

Scientists, social scientists and politicians frequently credit basic science with stimulating technological innovation, and with it economic growth. To support this idea, researchers have shown that patents based on university research receive more citations – a measure of patent importance – than those developed outside of academia. That research and much of the rhetoric it supports implicitly assumes that the application of scientific methods enables the invention of higher quality technologies. Nevertheless, another possibility exists. The norm of ‘communism’ and the related practice of publication may speed the diffusion of information developed in the scientific community. By examining patent data, this paper seeks to determine to what extent do quality differences versus communication explain the citation premium accorded to university and science-based patents. The analyses suggest that heightened communication explains a substantial amount of the difference, a result with important implications for both future research and public policy.

1. Introduction

The idea that the development of science has driven the rapid economic growth of the Western world enjoys wide acceptance today. Though identifying the precise point at which this notion entered the marketplace of ideas proves difficult, the fact that both de Toqueville (1848) and Marx (1844) pointed to science as a key progenitor of the technological progress that stimulates economic growth suggests that this claim has held sway in intellectual circles for at least 150 years.¹ The widespread belief in this relationship generates important social consequences. Early sociologists noted that this claim to material usefulness affords science its status in modern society (Merton, 1938, 1942; Weber, 1946). Economists have drawn on this widely accepted relationship to justify the public funding of scientific research (e.g., Bush, 1945; Kuznets, 1959; Malakoff, 2000). And international organizations, such as UNESCO and the World Bank, use this logic to support the promotion of these expenditures worldwide (Schofer, Ramirez and Meyer, 2000).

In the attempt to validate this claim, researchers usually examine one of three types of evidence.² The earliest research on this topic analyzed the impact of public and private research and development expenditures on productivity increases in the United States, showing that these expenditures do indeed appear to stimulate growth (Mansfield 1972; Rosenberg, 1976; Schmitz, 1989; Adams 1990). Nevertheless, cross-national studies fail to confirm this effect, often finding a negative relationship between indicators of scientific research and economic growth (Williams, 1964; Shenhav and Kamens, 1991; Schofer, Ramirez and Meyer, 2000). A more recent strain of research – arguing that scientific practices can stimulate productivity even at the firm level – demonstrates that companies that adopt a scientific orientation outperform those that do not (Henderson and Cockburn, 1994; Gambardella, 1995; Zucker and Darby, 1996; Zucker, Darby and

¹ The precise dating eludes us, but an examination of the classics of political economy points to the early 19th century. Although the idea appears absent from the works of Adam Smith (1776) and Ricardo (1821), one can find it in de Toqueville, Marx and the later work of Alfred Marshall (1920). Gieryn (1983) also mentions it as an element of a 1866 discourse involving John Tyndall.

² Other approaches offered, though not followed, include Sveikauskas' (1981) correlation of scientific employment to economic growth and Jaffe's (1989) investigation of the relationship between university research expenditures and local patenting rates.

Brewer, 1998).³ However, Stern (1999) points out and provides strong evidence that firms adopting a scientific orientation might accrue these benefits from compensating differentials in wages (i.e. talented workers accept lower wages in exchange for the ability to publish their research), rather than the productivity enhancements from science. The most direct line of investigation shows that patents that originate out of university research receive more citations in the future, a measure of patent importance, suggesting that university research yields more important inventions (Jaffe and Trajtenberg, 1996; Henderson, Jaffe, and Trajtenberg 1998).

These accounts typically assume – whether implicitly or explicitly – that science benefits society because its methods offer a superior process for generating high-quality knowledge that can enable new technologies. Nevertheless, this highly favorable view of the scientific method conflicts with a substantial literature in the sociology of science that points to the difficulty the scientific community faces in defining its own boundaries (e.g., Collins, 1982; Gieryn, 1983). It also contrasts markedly with descriptions of the highly contentious and often subjective debates that frequently surround the interpretation of new evidence.⁴ An alternative, perhaps somewhat more sociological, explanation exists. In addition to its methods, science also includes a set of social conventions (Merton, 1942). One in particular strikes us as important: the norm of publication (Bernal, 1939). Absent publication, much of the knowledge gained from scientific research must pass through interpersonal networks. To the extent that these social networks localize in both physical and social space, transmission will most frequently occur to those in close proximity to the inventors. Though these interpersonal networks allow for the diffusion of knowledge through space, publication elevates the diffusion process beyond the limited contacts available in social networks, thereby accelerating the diffusion of knowledge. Though this diffusion might prove functional in the sense that it reduces the duplication of research effort, it suggests that the norm of publication rather than the scientific method of developing knowledge provides this benefit.

³ Interestingly, all of these studies investigate firms within a single industry, biotechnology. Thus, the more general impact of commercial enterprises adopting scientific practices remains unknown.

⁴ A host of examples exist. ADD Recently, Collins (2000) reviews the rejection of high visibility gravitational radiation research.

To gain purchase on this issue, this study compares the citation patterns of inventions that draw on the scientific literature to those that do not. We attempt to identify the importance of communication in two ways: First, we compare the future patent citation rates across three groups of patents. The first group consists of patents that cite the scientific literature, while the second includes patents that cite published sources that do not meet the standards of scientific research (e.g., marketing reports and popular journals). The third group comprises all patents that do not cite published sources. If science generates higher quality knowledge, then the first group of patents should receive more citations than the other two groups. On the other hand, if communication accelerates the diffusion of information, then both groups of patents that cite published sources should receive more citations than the group that does not. Second, we compare the spatial diffusion of future patents that cite these various published sources to those that do not reference such literature. If publication expands the rate at which information diffuses through space, then patents that reference published sources should experience an expansion in the geographic range of the patents that cite them.

The results suggest that communication plays a large role in explaining the citation premium attributed to patents based on scientific research. Patents derived from science-based research do not show a significantly higher likelihood of being cited by all future patents. Nevertheless, patents that reference published materials – whether scholarly or not – receive citations at a much faster rate from distant inventors, suggesting that publication accelerates the flow of information through space. Thus, the citation premium accorded to science-based research appears to stem almost entirely from the fact that the faster dissemination of knowledge increases the number of inventors aware of these inventions. Though this rapid diffusion of information might benefit society by reducing the degree of inefficient duplicated effort, the scientific method, per se, does not appear to lead to higher quality inventions.

In addition to the theoretical issues, this question also addresses important policy issues. For example, if much of the benefit of science accrues from its norms of openness, then

public funding of research should include stipulations of quick and public disclosure of results to take advantage of the potential societal benefits. Moreover, if the technological benefits of science to society derive mainly from communication, then faster review and publication might increase the pace of technological change. For example, faster publishing and web-based dispersal of information may well increase the rate of patenting and invention.

2. The Role of Science in Technological Advance

In considering the role that science plays in technological advance, one must first distinguish how science differs from non-science. Two types of criteria separate scientific and non-scientific activity in the literature: the method of knowledge generation and the professional norms. Philosophy tends to focus on the logic of the scientific method and how that method may engender the efficient production of knowledge. Meanwhile, sociological accounts typically focus on the norms of the scientific community. Let us review each of these accounts.⁵

2.1 Science as a Method

Much of the work in the philosophy of science focuses on the logic of theory development in scientific research. By focusing on these methods, science seeks both to differentiate itself from other pursuits, such as metaphysics, and to legitimate its activities. Thus, Comte (1853) distinguishes science from other activities by its reliance on observation and logic to generate theory. Nevertheless, this description fails to differentiate science from other forms of experiential learning. A more precise modern statement of the value of the scientific method comes from Popper (1965) and the positivist school, who claim that the generation of falsifiable statements forms a more logically sound basis for building knowledge.

⁵ Though both portray scientists as belonging to a nearly homogenous population, scientists must admittedly contend with overlapping and conflicting roles in multiple communities (Moore, 1996). Likewise, the factors that demarcate science can vary across disciplines and often appear in the practices of 'non-scientists' (Gieryn, 1983). Nevertheless, while scientists likely differ in interesting ways from one another, we focus on those factors that *typically* differentiate those labeled as scientists from non-scientists.

Though more recent research raises questions regarding the validity of the positivist model of science, even the modifications to this theory typically argue that science seeks a close match to empirical observation. Thus, Kuhn (1958) notes that theories progress in two stages: normal science and scientific revolution. During normal science, research progresses along the path described in the positivist model. Through the course of this research, anomalies that fail to fit the dominant theory inevitably arise. Instead of simply declaring the theory wrong, scientists instead attempt to extend the theory incrementally until these theory fixes become unwieldy. At that point, the field reaches a stage of crisis and seeks out a new dominant paradigm. Nevertheless, even in Kuhn's account, science offers a superior means of generating knowledge. By developing falsifiable theories, scientists can dramatically reduce the size of the space they must sample in their experiments.

The norms of science enforce this methodology. Students being trained in the sciences receive extensive indoctrination into the importance of using the 'scientific method'. As students, they take entire courses devoted to the proper methods for developing and testing theories. Following those courses, most programs place them in apprentice-like positions through research assistantships and post-doctoral appointments in which faculty can closely monitor the activities of these pupils and ensure that they conform to accepted practices. Even after leaving the training stage of their careers, the peer review process affords the community a mechanism for continuing to enforce a set of research standards.

To the extent that these methods support the more effective accretion of knowledge, one might expect the adoption of scientific methods to stimulate technological development. If one views the process of invention as a search of a large and multi-dimensional technological space, then the development of theories provides an efficient mechanism for reducing the proportion of the space that one must sample (Fleming and Sorenson, 2000). When sampling the space proves costly, either in terms of time or money, techniques that reduce the need to sample will yield economic benefits. Thus, one might expect science to offer a more efficient search process if its methods offer a superior mechanism for developing knowledge.

Work in the sociology of science, however, raises questions regarding the extent to which science actually does rely on accurate descriptions of reality as an objective function. This methodologically relative view of science also resonates with a wide range of sociological studies that highlight the role of social processes in science. Thus, we also consider a less instrumental perspective on the role of science.

2.2 Science as an Institution

Robert K. Merton encouraged sociologists to consider science as an institution. Through their training and both positive and negative career incentives, scientists internalize a set of values that guide their activities. Though Merton (1942) identified several importance norms in the scientific community – including universalism, communism, disinterestedness and organized skepticism – ‘communism’ seems most important to the question of knowledge diffusion.

The norm of ‘communism’ plays an important role in scientific research. ‘Communism’ refers to the idea that individual scientists believe that their property rights over their ideas extends only to the credit associated with finding it first (Merton, 1942). Though this norm might appear to reduce the incentives for innovation, since all recognition flows to the first discoverer it still provides powerful incentives for primacy (Merton, 1957). Additionally, because public dissemination of findings establishes primacy, the combination of the communal nature of the knowledge invented and the desire for recognition both reinforce the desire to publish research results (Merton, 1942).

This norm also serves a functional purpose. Bernal (1939) notes that the growth of science coincided with the rejection of the idea of secrecy. The rapid diffusion of knowledge through publication and other media allows researchers to avoid the duplication of effort and advance their research beyond the already known. 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."⁶

In addition to published journals, the scientific community has also generated a wide range 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. For example, sociologists meet every year at a central location to encourage interaction with other. In the physical and biological sciences, scholars tend to focus on meetings that congregate researchers with much narrower ranges of interests. By forming networks that bridge geographic space, these organizations expand the range of information diffusion (Sorenson and Stuart, 2001). However, while these organizations stimulate interaction among researchers with related backgrounds, disciplinary silos may actually impede interaction and the movement of ideas across academic departments⁷ and limit fraternization between scientists and non-scientists (Fleming and Sorenson, 2001).

The training that scientists receive provides them with a common language for communicating ideas. On the one hand, this language may ease the codification and transmission of complex ideas. Nevertheless, like other professions this language excludes outsiders – a fact that Merton (1938) refers to as the "cult of unintelligibility." These specialized vocabularies likely impede the flow of ideas outside the community boundaries (Merton, 1938) and from scientific to engineering communities (Allen, 1977). They may even delimit the transmission of information across specialties within disciplines (Durkheim, 1893). However, as Abbot (1988) argues, these private vocabularies allow professions to maintain their status by excluding outsiders, so even if common language could transmit the same ideas scientists might have a vested interest in maintaining these specialized languages.

⁶ Interestingly, Merton (1942) notes that this phrase dates to before the 12th century, so this cumulative ideology does not represent a new one.

⁷ The strength of this delimitation appears to vary dramatically from field to field. For example,

3. Empirical Strategy

To understand better the effect of science on technological development, this study focuses on the differences between patents that cite non-patent materials and those that do not. Patents provide a window on the generation and diffusion of knowledge. Similar to studies of citation patterns across academic papers, the citations that appear on patents offer a paper trail for tracking the diffusion of information. Patent citations offer a more objective measure of linkages, however, than citation patterns in publications. Unlike authors, inventors have an incentive to minimize their prior art citations. Patents represent a property right over the commercial developments based on some parcel of intellectual property. As such, patent applicants proceed with strong incentives to avoid actions that could delimit the range of their claim. Both citations to other patents and citations to prior publications can potentially reduce the scope of their claims and limit the effectiveness of any future legal action defending it. Thus, patent applicants will seek to minimize these citations.

Nevertheless, the patent review process places a limit on the failure to cite ‘true’ prior art. Based upon personal expertise and automated searches, patents examiners add missing citations to applications in the review process (Carr, 1995). References to non-patent materials more likely come from the patent applicants themselves (Tijssen, 2001). Although patent examiners often add citations to prior patents, their mandate does not extend their search to the vast array of published materials on which one could potentially draw.

3.1 Citation Rates

The most basic evaluation of the impact of science involves looking simply at the citation rates of patents that draw on science relative to those that do not. A substantial body of research based on patents uses the number of citations a focal patent receives from future patents as an indicator of the focal patent’s economic and technical importance (e.g., Jaffe, Trajtenberg and Henderson, 1993). Though this measure undoubtedly offers a noisy proxy for these outcomes, it does allow the comparison of patents across classes and several studies have confirmed this general relationship (Trajtenberg, 1990; Albert, et

al., 1991; Hall, Jaffe and Trajtenberg, 2000). Transporting this measure to our question leads to the following expectation: *If science offers a better method for developing knowledge, then patents based on its methods should receive more citations than those that do not employ the scientific method.* Thus, we begin by looking simply at these citation rates, using non-patent citations as an indicator of whether or not the invention draws from scientific research.

Patents cite a wide range of sources besides previous patents; most notably, they make reference to various scientific, technical, and corporate literatures. 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 (e.g., Tijssen, 2001). Little work investigates the effects of scientific and non-scientific citations on large samples. To enable such an analysis, a trained researcher categorized each of the 16,728 sources referenced on our sample of 17,264 patents.⁸ She sorted the references into seven mutually exclusive and exhaustive categories:

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 focus on the patents citing 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

⁸ 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. However, 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, we feel confident that the classification scheme identifies distinct and cognitively meaningful categories.

those necessary to receive publication in a peer-reviewed journal. One example of this type of citation appears in patent 4922432, 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 path design appears many years earlier in an engineering textbook (Mead and Conway, 1980); hence, the paper would likely fail the peer review process in a journal.

Technical Report: Most of the items in this category refer to research institute publications; for example, the Battelle Institute, “Final Report on High-Performance Fibers II, An International Evaluation to Group Member Companies” by Donald C. Slivka et al., 1987 appears on patent 4935180. Though many of these reports may describe the results of scientific research, the classification of them remains problematic since the institutions self-publish these reports, thereby avoiding peer review.

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 that appears 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 to simply 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: The publications in this category typically refer 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*. Practically, it contains popular magazines and journals, such as *Concrete and Cement Age*, *Guns and Ammo* and *Harvard Business Review*, which do not ascribe to the norms of scientific research.

Table 1 shows the distribution of these various references in the sample. Patents that cite the scientific literature exhibit a citation premium, supporting earlier research that finds a citation premium among patents that arise out of university research (Jaffe and Trajtenberg, 1996; Henderson, Jaffe, and Trajtenberg 1998). On average, these patents receive 4.79 citations in the five years following their granting versus the 3.54 citations received by patents that do not reference published sources, a premium of 35%. Nevertheless, every other form of publication also appears to accord the patents that reference it a citation premium. Although some of these categories, such as conference proceedings also confound publication with scientific method, at least two types of publications – the corporate non-technical and non-index journals – make no claim to scientific methods. Patents that reference non-technical corporate publications receive 4.61 citations on average, a 30% premium, while those that refer to non-index journals receive 5.4 citations – 53% more than patents that do not cite any publications and 13% more than patents citing journals in the scientific index.

These results raise doubt that the methods used in scientific research can account for the larger number of citations these patents receive. If those methods did explain the citation premium, then patents drawing on scientific methods should receive more future citations than both patents that do not cite published sources and those that reference non-scientific publications. Nevertheless, non-scientific publication, such as corporate advertisements and popular magazines, appear to enjoy citation premiums that equal or exceed those received by patents based on scientific research. Though this finding seems more consistent with the diffusion explanation of the role of science, it does not provide direct

evidence of the diffusion process. Therefore, we continued by analyzing the geographic distribution of citation linkages.

3.2 Citation Distribution

Citations tend to localize in space because the networks that tie inventors together connect most densely when these actors lie in close geographic proximity to one another. A variety of studies, beginning with Park (1926) demonstrate that social actors more frequently form ties when they reside near to one another. Bossard (1932) found that the likelihood of marriage between two individuals decreased rapidly as the distance between them lengthened, a finding confirmed in both marital and friendship ties in a series of subsequent investigations. Studies of the workplace indicate that the frequency of interaction between co-workers depends on the proximity of their offices to one another (e.g., Allen, 1977). More recently, Sorenson and Stuart (2001) demonstrate that location plays a strong role in determining which entrepreneurs venture capitalists decide to fund. The importance of geographic distance arises from two factors: First, the likelihood that any two individuals will meet in the course of their day-to-day activities – thereby having the opportunity to form a tie – declines rapidly with distance (Hawley, 1971). Second, even when a tie does occur, the costs of maintaining that tie likely increase, as actors must bridge increasingly wide expanses to interact (Zipf, 1949). Thus, individuals budgeting scarce resources will discontinue these distant ties more readily.

By removing the flow of information from the constraints of social networks, publication should accelerate the diffusion of ideas in geographic space. When private information travels through interpersonal ties, it cannot escape the spatial limitations of the network. Hence, studies repeatedly show circumscribed transmission of tacit or confidential data. For example, Hedstrom (1994) finds that the geographic configuration of communication networks in the Swedish population can explain the spatial contagion in these organizations' founding. Baker (1984) shows in his study of a commodities exchange trading floor that the structure of interaction influences the price volatility of securities. Nevertheless, when information becomes publicly available, its transmission should transcend these network limitations, flowing much more freely to loosely- or un-

connected individuals. *If publication in scientific journals accelerates the spatial diffusion of information, these patents should receive citations from more geographically distant future patents.*

To investigate the diffusion of ties, we model the probability that a future patent cites a given focal patent as a function of distance, publication and a variety of control variables - essentially an estimation of tie formation. Although many studies of tie formation analyze every possible dyad and use logistic regression to estimate the effects of a covariate vector on the likelihood of a tie (e.g., Podolny, 1994; Gulati, 1995), this strategy creates two problems: Methodologically, it fails to account correctly for non-independence across cases, as each actor enters the analysis many, many times.⁹ The large number of repeat occurrences of each actor can lead to systematic underestimation of the standard errors for actor attributes that do not change from dyad to dyad. Pragmatically, this strategy presents a second obstacle; the observation of all possible dyads can prove burdensome computationally, especially for large networks. For example, consideration of all potential dyads in our data would require us to create a matrix with more than eleven billion cells.

Sampling randomly from the set of potential patent dyads offers one potential solution to these issues. Nevertheless, this approach falls short of the ideal because it ignores the fact that the realized ties provide most of the information for the estimation of the factors that affect tie likelihood (Coslett, 1981; Imbens, 1992; Lancaster and Imbens, 1996). Thus, we include all 70,271 citations that actually appear in our sample of 17,264 patents. For comparison, we then create a matched sample of potential cites that did not occur, pairing each of the 17,264 patents in the sample with four patents chosen randomly from the patents assigned between July 1990 and June 1996. Though this generated a data set of 139,487 dyads, our analyses restricted the estimated sample to the 74,266 cases where the cited patent holder resides in the U.S. To address the fact that focal patents enter the data more than once, we report robust standard errors estimated without the assumption of independence across observations on the same patent.

⁹ Doreian (1981), Krackhardt (1988), and Mizruchi (1989) discuss a variety of approaches for addressing the non-independence problem.

The use of a matched sample introduces one new problem. Logistic regression can yield biased estimates when the proportion of positive outcomes in the sample does not match the proportion of positive outcomes in the population. In particular, uncorrected logistic regression using a matched sample tends to produce underestimates of the factors that predict a positive outcome (King and Zeng, 2000). Large samples do not necessarily alleviate this problem. Following Sorenson and Stuart (2001), we adjust the coefficient estimates using the method proposed by King and Zeng (2000) for the logistic regression of rare events to correct for this potential bias.¹⁰

The results of these analyses appear in table 3. The analysis focuses on journals that appear in the scientific index because these patents most clearly stem to some degree from the efforts of scientists. If science offers a superior method for generating knowledge, then patents that draw on these methods should enjoy higher citation rates from all future patents. On the other hand, if science increases the velocity of information diffusion through its norms of openness, then publication should expand the geographic scope of citing patents. The first model simply shows that patents that reference articles from journals in the Scientific Index have a higher probability of receiving a citation from any future patent (the variable, *scientific index* represents a simple count of the number of these publications referenced). As one would expect in a diffusion process, model 2 demonstrates that while the odds of receiving a citation

¹⁰ The traditional logistic regression model considers the dichotomous outcome variable to be a Bernoulli probability function that takes a value 1 with the probability π :

$$\pi_i = \frac{1}{1 + e^{-X_i\beta}}$$

where X is a vector of covariates and β is a vector of parameters. Researchers typically use maximum likelihood methods to estimate β . King and Zeng (2000) prove that the following weighted least squares expression estimates the bias in β generated by oversampling rare events:

$$\text{bias}(\hat{\beta}) = (X'WX)^{-1} X'W\xi,$$

where $\xi = 0.5 \cdot Q_{ii} [(1 + w_1)\hat{\pi}_i - w_1]$, 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 (events) in the sample relative to the fraction in the population. Essentially, one regresses the independent variables on the residuals using W as the weighting factor. Tomz (1999) implements this method in the *relogit* Stata procedure.

typically decline rapidly with the distance¹¹ between the future patent applicant and the focal patent holder, publication mitigates this effect, each publication reducing the spatial effects by 7%.

Model 3 demonstrates that the effects remain robust even after controlling for a variety of other potential confounding effects. *Activity control* accounts for differences in citation patterns across patent classes, by including the number of prior art citations made by the focal patent. The control positively impacts the probability of a tie. When the future patent and the focal patent both belong to the *same class* (a dummy variable indicating common membership in a primary class), the likelihood of a citation goes up dramatically as these patents reside in close proximity in technological space. Patents that span multiple classes may receive more citations simply because they apply to a broader range of future technologies, similar to an academic article that cites a broad variety of literatures. Nevertheless, the model fails to show such an effect for *number of classes*. Technologies also vary in their closeness to the technological frontier. *Recent technical area* – the average of the patent numbers of the focal patent’s prior art (higher numbers indicate more recent technology) – controls for this difference across technologies. Young fields receive future citations at a faster rate. *Foreign assignee* shows that information does not easily spread beyond political borders as this dummy variable indicates that foreign assignees cite U.S. patent holders at an even lower rate than distant American inventors (e.g., those in Hawaii). Finally, the *time* between the granting of the focal patent and the application of the future patent in days (logged) appears to increase the likelihood of a citation occurring and reduce the importance of distance, as one would expect as social networks diffuse information geographically (Sorenson and Stuart, 2001). Though the inclusion of these controls greatly improves the fit of the model, it

¹¹ We calculate distance by assigning the longitude and latitude at the center of the zip code in which the patent assignees reside to each patent. Using spherical geometry, the distance between the two points, i and j , is:

$$d_{ij} = C \left[\arccos \left(\sin(lat_i) \sin(lat_j) + \cos(lat_i) \cos(lat_j) \cos(|long_i - long_j|) \right) \right],$$

where latitude (lat) and longitude ($long$) are measured in radians. C converts the result into linear units of measure, $C = 3437$ corresponds to miles. We take the inverse of geographic distance because zero represents a logical bound to which we can then assign foreign patents whose distance lies far, but at an unknown exact point, from the U.S. based focal patents.

does not diminish the interaction between *scientific index* and *distance*. Patents that draw on the scientific literature enjoy a broader geographic dispersion of future citations.

Table 4 shows that similar patterns of results hold for publications that do not ascribe to the norms of science. For this comparison, we selected the two types of publications that appear most distant from the scientific community: non-index journals and non-technical corporate publications (mostly advertisements). If the act of publication itself engenders the diffusion of information, then these publications, though devoid of other trappings of the scientific community, should show a similar pattern of results. Indeed, they do. Models 5 and 6 present parallel specifications to models 2 and 3 for non-index journals, while models 7 and 8 do so for non-technical corporate publications.

4. Discussion

The results support the idea that science-based patents receive more citations because the act of publication allows their ideas to diffuse more rapidly. Nevertheless, some other alternative accounts warrant consideration.

One difficulty in investigating the distribution of citations geographically comes from separating the flow of knowledge from the distribution of inventive activity. If citations appear highly localized, two factors could explain that distribution. On the one hand, interpersonal communication might distribute important information through space. However, citations could also localize simply because everyone researching the problem resides in a concentrated geographic area. Jaffe, Trajtenberg, and Henderson (1993) controlled for this problem by restricting their analysis to patents within a particular technology class. They measured the likelihood of a local citation to a patent, compared to a temporally proximal patent in the same technology class. Though this sampling method can yield biased estimates of the predictors of citation likelihood, it does control for the distribution of activity. To address this possibility, we estimated a second set of models predicting the distance between citing and cited patent including fixed-effects by class to capture differences in the distribution of activity (see Appendix). These models

support the argument that patents that cite published sources receive cites from further away.

Another explanation offered for the citation premium for university patents argues that academics only patent their most valuable research. Presumably then, the quality of knowledge produced by both universities and the private sector does not differ substantially. Researchers working in academia simply maintain a higher threshold over which an idea must pass before they consider it worthy of patenting. In support of this idea, as the number of patents issued to universities has increased, the citation premium appears to have declined (Henderson, Jaffe and Trajtenberg, 1998; Hicks, et al., 2001). Nonetheless, this explanation applies to the number of citations that one would expect a science-based patent to receive rather than the geographic distribution of those citations. Thus, it cannot explain the pattern of results seen here.

The fact that publication speeds the flow of knowledge through space poses an interesting question to a 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 (Rosenberg 1990, Cohen and Levinthal 1990, Henderson and Cockburn, 1994; Zucker and Darby, 1996; Zucker, Darby and Brewer, 1998). If the technological knowledge they generate spills over to rivals more rapidly, the promotion of communal norms has costs as well as benefits. These results then appear to complement recent work by Stern (1999), which suggests that firms benefit by promoting science not because its methods inherently offer value but rather because talented employees desire the ability to publish. Hence, they will accept lower compensation from firms that allow them to stay active in the scientific community.

The results also provide insight into previous results on the role of universities and knowledge spillovers in technological change. By isolating the role of publication, this research offers a new perspective on the mechanisms of technological diffusion. For example, Henderson, Jaffe, and Trajtenberg (1998) demonstrated that university patents

receive more citations (although the citation premium has been decreasing in recent years). Our results demonstrate that this positive effect has likely been a correlate, however, of the fact that university patents draw upon published science.

References

- Abbot, Andrew. 1988. *The System of Professions*. Chicago: University of Chicago Press.
- Albert, M., D. Avery, Francis Narin and P. McAllister. 1991. "Direct Validation of Citation Counts as Indicators of Industrially Important Patents," *Research Policy* 20: 251-259.
- Allen, Thomas. 1977. *Managing the Flow of Technology*, Cambridge, MA: MIT Press.
- Almeida, P. and B. Kogut. 1999. "Localization of Knowledge and the Mobility of Engineers in Regional Networks," *Management Science* 45:7:905-917.
- Audretsch, D. and M. Feldman. 1996. "R&D Spillovers and the Geography of Innovation and Production," *American Economic Review*, vol. 86:3:630-640.
- Adams, James D. 1990. "Fundamental Stocks of Knowledge and Productivity Growth." *Journal of Political Economy* 98: 673-702.
- Bazerman, Charles. 1988. *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison, WI: University of Wisconsin Press.
- Bernal, John D. 1939. *The Social Function of Science*. New York: Macmillan.
- Bossard, James S. 1932. "Residential Propinquity as a Factor in Marriage Selection," *American Journal of Sociology*, 38: 219-224.
- Bush, V. 1945. *Science: The Endless Frontier*. Washington, DC: US Government Printing Office.
- Carr, F. 1995. *Patents Handbook: A Guide for inventors and Researchers to Searching Patent Documents and Preparing and Making an Application*. Jefferson, N.C.: McFarland,
- Cohen, Wesley and Daniel Levinthal 1990, "Absorptive Capacity: A New Perspective on Learning and Innovation," *Administrative Science Quarterly*, 35:128-152.
- Cole, Stephen and Jonathan R. Cole. 1968. "Visibility and the Structural Bases of Awareness of Scientific Research." *American Sociological Review* 33: 397-413.
- Collins, H.M. 1982. "Knowledge, Norms and Rules in the Sociology of Science." *Social Studies of Science* 12: 299-309.
- Collins, H.M. 2000. "Surviving Closure: Post Rejection Adaptation and Plurality in Science." *American Sociological Review* 65: 824-845.

Compte, Auguste. 1853. *The Positive Philosophy of Auguste Comte* (Harriet Martineau, trans.). London: Chapman.

Coslett, Stephen R. 1981. "Maximum Likelihood Estimator for Choice-based Samples." *Econometrica* 49: 1289-1316.

Doreian, Patrick. 1981. "Estimating Linear Models with Spatially Distributed Data." Pp. 359-388 in Samuel Leinhardt (ed.), *Sociological Methodology 1981*. San Francisco: Jossey-Bass.

Durkheim, Emile. [1893] 1984. *The Division of Labor in Society* (W.D. Halls, trans.). New York: Free Press.

Fleming, Lee and Olav Sorenson. 2001. "Technology as a Complex Adaptive System: Evidence from Patent Data," *Research Policy*, forthcoming.

Friedkin, Noah E. 1978. "University Social Structure and Social Networks Among Scientists." *American Journal of Sociology* 83: 1444-65.

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

Gieryn, Thomas F. 1983. "Boundary Work and the Demarcation of Science from Non-Science: Strains and Interests in the Professional Ideologies of Scientists." *American Sociological Review* 48: 781-95.

Gulati, Ranjay. 1995. "Social Structure and Alliance Formation Patterns: A Longitudinal Analysis." *Administrative Science Quarterly* 40: 619-652.

Hall, Bronwyn, Adam B. Jaffe and Manuel Trajtenberg. 2000. "Market Value and Patent Citations: A First Look." Working paper 7741, National Bureau of Economic Research.

Harhoff, Dietmar, Francis Narin, Frederic M. Scherer and Katrin Vopel. 1999. "Citation Frequency and the Value of Patented Inventions." *Review of Economics and Statistics* 81: 511-15.

Hawley, Amos H. 1971. *Urban Society*. New York: Ronald.

Henderson, Rebecca and Iain Cockburn. 1994. "Measuring Competence? Exploring Firm Effects in Drug Discovery." *Strategic Management Journal* 15(Winter special issue): 63-84.

Henderson, Rebecca, Adam B. Jaffe and Manuel Trajtenberg 1998, "Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-1988," *The Review of Economics and Statistics*, Feb 1998; Vol. 80, Iss. 1; pg. 119.

Hicks, Diana, Tony Breitzman, Dominic Olivastro and Kimberly Hamilton. 2001. "The Changing Composition of Innovation Activity in the U.S. – a Portrait Based on Patent Analysis." *Research Policy*: forthcoming.

Imbens, Guido. 1992. "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-based Sampling," *Econometrica*, 60: 1187-1214.

Jaffe, Adam B. 1989. "Real Effects of Academic Research." *American Economic Review* 79: 957-70.

Jaffe, Adam B. and Manuel Trajtenberg. 1996. "Flows of Knowledge from Universities and Federal Labs: Modeling the Flow of Patent Citations over Time and across Institutional and Geographic Boundaries."

Jaffe, Adam B., Manuel Trajtenberg and Rebecca Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* XX: 577-98.

King, Gary and Langche Zeng. 2000. "Logistic Regression in Rare Events Data." *Political Analysis*: in press.

Krackhardt, David. 1988. "Predicting with Networks: Nonparametric Multiple Regression Analyses of Dyadic Data." *Social Networks* 10: 359-382.

Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions* (2nd Ed.). Chicago: University of Chicago Press.

Kuznets, Simon S. 1959. *Six Lectures on Economic Growth*. New York: Free Press.

Lancaster, Tony and Guido Imbens. 1996. "Efficient Estimation and Stratified Sampling." *Journal of Econometrics* 74: 289-318.

Malakoff, D. 2000. "Does Science Drive the Productivity Train?" *Science* 289: 1274-6.

Mansfield, E. 1972. "Contribution of R&D to Economic Growth in the United States," *Science* 175: 477-486.

Marshall, Alfred. 1920. *Principles of Economics* (8th Ed.). London: Macmillan.

Marx, Karl. [1844] 1975. "Economic and Philosophical Manuscripts." Pp. 279-400 in *Early Writings*, Rodney Livingstone and Gregor Benton (trans.), New York: Vintage Books.

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

Merton, Robert K. 1938. "Science and the Social Order." *Philosophy of Science* 5: 321-27.

Merton, Robert K. 1942. "Science and Technology in a Democratic Order." *Journal of Legal and Political Sociology* 1: 115-26.

Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* 22: 635-59.

Mizruchi, Mark S. 1989. "Similarity of Political Behavior Among Large American Corporations." *American Journal of Sociology* 95: 401-424.

Moore, Kelly. 1996. "Organizing Integrity: American Science and the Creation of Public Interest Organizations." *American Journal of Sociology* 101: 1592-1627.

Nelson, R. 1986. "R&D, Innovation, and Public Policy," *American Economic Review* 76:2:186-189.

Podolny, Joel M. 1994. "Market Uncertainty and the Social Character of Economic Exchange." *Administrative Science Quarterly* 39: 458-483.

Popper, Karl R. 1965. *The Logic of Scientific Discovery*. New York: Harper & Row.

Ricardo, David. [1821] 1911. *The Principles of Political Economy and Taxation* (3rd Ed.). London: J.M. Dent & Sons.

Rosenberg, Nathan. 1976. *Perspectives on Technology*. Cambridge: Cambridge University Press.

Rosenberg, Nathan. 1990, "Why Do Firms Do Basic Research (with Their Own Money)?" *Research Policy* 19: 165-174.

Schmitz, James A., Jr. 1989. "Imitation, Entrepreneurship, and Long-Run Growth." *Journal of Political Economy* 97: 721-39.

Schofer, Evan, Francisco O. Ramirez and John W. Meyer. 2000. "The Effects of Science on National Economic Development, 1970 to 1990." *American Sociological Review* 65: 866-887.

Schumpeter, Joseph A. 1942, *Capitalism, Socialism, and Democracy*. New York: Harper and Row Publishers.

Shevhav, Yehouda and David Kamens. 1991. "The 'Costs' of Institutional Isomorphism in Non-Western Countries." *Social Studies of Science* 21: 427-545.

Smith, Adam. [1776] 1976. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Clarendon Press.

Sorenson, Olav and Toby E. Stuart. 2001. "Syndication Networks and the Spatial Distribution of Venture Capital Investments." *American Journal of Sociology* 106: forthcoming.

Stern, Scott. 1999. "Do Scientists Pay to Be Scientists?" Working paper 7410, National Bureau of Economic Research.

Sveikauskas, Leo. 1981. "Technological Inputs and Multifactor Productivity Growth." *Review of Economics and Statistics* 63: 275-82.

Tijssen, R. 2001, "Global and Domestic Utilization of Industrial Relevant Science: Patent Citation Analysis of Science-Technology Interactions and Knowledge Flows," *Research Policy*, 30: 35-54.

Tomz, Michael. 1999. **relogit** (Stata ado file). Available at <http://gking.harvard.edu/stats.shtml>.

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

Trajtenberg, Manuel. 1990. "A Penny for Your Quotes: Patent Citations and the Value of Innovations." *Rand Journal of Economics* 21: 172-87.

Weber, Max. 1946. "Science as a Vocation." Pp. 129-56 in *From Max Weber: Essays in Sociology*, Hans H. Gerth and C. Wright Mills (eds.), New York: Oxford University Press.

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

Zucker, Lynn and Michael Darby. 1996. "Star Scientists and Institutional Transformation: Patterns of Invention and Innovation in the Formation of the Biotechnology Industry." *Proceedings of the National Academy of Sciences* 93: 709-16.

Zucker, Lynn, Michael Darby and M. Brewer. 1998. "Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises." *American Economic Review* 88: 290-306.

Table 1: Publication types, frequency and future citations[†]

Publication type	Total patents	Percent of sample	Average citations
Scientific index journal	3,118	18.1%	4.79•
Conference proceedings	483	2.8%	6.12•
Technical report	334	1.9%	5.29•
Corporate publication – technical orientation	337	2.0%	6.00•
Book	922	5.3%	4.24•
Corporate publication – non-technical	711	4.1%	4.61•
Non-index journal	298	1.7%	5.40•
No references to publications	12,769	74.0%	3.54

[†] 17,264 patents, the total across all categories exceeds this because many patents cite more than one type of non-patent prior art. • indicates a mean significantly different from the patents without references to publications at the $p < .01$ level.

Table 2: Co-occurrence matrix[‡]

Publication type	2	3	4	5	6	7
1. Scientific index	345• (4.0)	194• (3.2)	177• (2.9)	538• (3.2)	131 (1.0)	142 (2.6)
2. Conference proceedings		75• (8.0)	61• (6.5)	116• (4.5)	35 (1.8)	38• (4.6)
3. Technical report			42• (6.4)	99• (5.6)	30 (2.2)	28• (4.9)
4. Corporate technical				79• (4.4)	71• (5.1)	37• (6.4)
5. Book					59 (1.6)	60• (3.8)
6. Corporate non-technical						54• (4.4)
7. Non-index						

[‡] • indicates a cell count significantly different from random coincidence at the $p < .01$ level.

Table 3: Rare events logit models of the likelihood of a focal patent receiving a citation from a future patent*

	Model 1	Model 2	Model 3	Model 4
Scientific index	.017** (.003)	.017** (.003)	.015 (.012)	.017 (.012)
Scientific index / distance		-.371** (.083)	-.756** (.078)	-40.82** (1.46)
1 / distance		5.08** (.178)	10.53** (2.27)	29.17** (2.38)
Activity control			.372** (.117)	.489** (.117)
Same class			5.20** (.101)	5.10** (.101)
Number of classes			.008 (.002)	.010 (.012)
Recent technological area			.415** (.087)	.349** (.087)
Foreign assignee			-.730** (.087)	-.732** (.087)
Time (grant to cite)			.755** (.060)	.729** (.057)
Time / distance			-.419** (.180)	-1.86** (.189)
Scientific index X time / distance				3.10** (.116)
Constant	-9.75** (.013)	-9.87** (.013)	-22.24** (.868)	-21.97** (.116)
Log-likelihood	-52157.1	-49257.6	-31180.9	-31179.4

* 74,266 cases, 51.7% represent ties (vs. .00059% in population) • $p \leq .05$ ** $p \leq .01$

Table 4: Rare events logit models of the likelihood of a focal patent receiving a citation from a future patent*

	Model 5	Model 6	Model 7	Model 8
Non-index journal	.149** (.046)	.292** (.061)		
Non-index / distance	-12.21** (.540)	-40.45** (.416)		
Corporate non-technical			.045** (.013)	-.028 (.039)
Corporate non-technical / distance			-3.55** (.409)	-2.75** (.376)
1 / distance	5.26** (.165)	62.40** (2.27)	5.10** (.167)	9.52** (2.27)
Activity control		.483** (.116)		.431** (.116)
Same class		5.15** (.101)		5.22** (.102)
Number of classes		-.012 (.012)		.003 (.012)
Recent technological area		.083 (.089)		.308** (.090)
Foreign assignee		-.741** (.087)		-.733** (.087)
Time (grant to cite)		.765** (.061)		.753** (.061)
Time / distance		-4.47** (.180)		-.366* (.181)
Constant	-9.86** (.013)	-21.04** (.875)	-9.86** (.013)	-21.84** (.877)
Log-likelihood	-50383.0	-31180.0	-50372.1	-31168.2

* 74,266 cases, 51.7% represent ties (vs. .00059% in population) • $p \leq .05$ ** $p \leq .01$

Supplemental Appendix: Distance Models

Another approach to examining the diffusion of information selects only those citations that actually occur and estimates the distance between citing and cited patents. The models below show this method.

This method suffers from some potential forms of selection bias, however, it does allow the estimation of interaction terms that become somewhat messy in the models predicting tie likelihood. For example, these models show that the expanded diffusion decreases over time, as one would expect as network-based diffusion substitutes for the dissemination of information through published sources. The models also suggest that publication plays an even more important role in diffusing information for recently developed technologies.

The models estimate the logged distance between citing and cited patents. The tables provide estimates for both dummy and count measures of citations to publications. The findings appear robust across these specifications. In addition, models 3 through 8 include fixed effects for each patent class. These fixed effects should account for differences in the geographic distribution of research activity across classes.

Table App. 1: Models of distance between patents and the prior art they cite

	Model 1 (counts)	Model 2 (dummies)	Model 3 (counts w/ FE)	Model 4 (dummies w/ FE)
Activity control	.353** (.037)	.310** (.037)	.051 (.097)	.060 (.097)
Number of classes	.107** (.015)	.109** (.015)	.103** (.015)	.101** (.015)
Time (application to grant)	.000 (.046)	-.001 (.046)	-.098* (.045)	-.119** (.046)
Time (grant to cite)	.674** (.025)	.671** (.025)	.524** (.023)	.524** (.022)
Recent technical area	-.917** (.030)	-.872** (.031)	-.573** (.031)	-.553** (.031)
Scientific index	.066** (.005)	.162** (.037)	.058** (.005)	.187** (.037)
Non-index journal	.202** (.045)	.551** (.094)	.128** (.043)	.223** (.089)
Corporate non-technical	.242** (.017)	1.36** (.064)	.157** (.016)	1.02** (.061)
Constant	5.88** (.124)	5.86** (.124)	5.21** (.181)	5.10** (.182)
Class fixed effects			YES (364; Sig.)	YES (364; Sig.)
R-squared	.048	.055	.046	.049

Table App. 2: Models of distance between patents and the prior art they cite

	Model 5 (counts w/ FE)	Model 6 (dummies w/ FE)	Model 7 (counts w/ FE)	Model 8 (dummies w/ FE)
Activity control	.062 (.097)	.069 (.097)	.049 (.097)	.059 (.097)
Number of classes	.107** (.015)	.102** (.015)	.103** (.015)	.100** (.015)
Time (application to grant)	-.088 (.046)	-.099* (.046)	-.097* (.045)	-.118** (.045)
Time (grant to cite)	.525** (.023)	.523** (.023)	.561** (.024)	.574** (.027)
Recent technical area	-.628** (.032)	-.648** (.033)	-.573** (.031)	-.552** (.031)
Scientific index	.049** (.006)	.113** (.039)	.058** (.005)	.183** (.037)
Non-index journal	.131** (.043)	.295** (.090)	.126** (.043)	.225** (.089)
Corporate non-technical	.176** (.017)	1.08** (.062)	.164** (.016)	1.03** (.061)
Scientific index X recent tech. area	.073** (.015)	.472** (.081)		
Non-index X recent tech. area	.206* (.100)	.319* (.142)		
Corporate non-technical X recent tech. area	.093** (.024)	.360** (.103)		
Scientific index X time (grant to cite)			-.267** (.074)	-.126* (.055)
Non-index X time (grant to cite)			-.918 (.670)	-.191 (.137)
Corporate non-technical X time (grant to cite)			-.672* (.276)	-.282** (.098)
Constant	5.41** (.124)	5.47** (.124)	5.18** (.182)	5.06** (.182)
Class fixed effects	YES (364, Sig.)	YES (364; Sig.)	YES (364; Sig.)	YES (364; Sig.)
R-squared	.048	.052	.046	.050