

Sorenson, Olav; Singh, Jasjit; Fleming, Lee

**Working Paper**

## Science, social networks and spillovers

Papers on Economics and Evolution, No. 0512

**Provided in Cooperation with:**

Max Planck Institute of Economics

*Suggested Citation:* Sorenson, Olav; Singh, Jasjit; Fleming, Lee (2005) : Science, social networks and spillovers, Papers on Economics and Evolution, No. 0512, Max Planck Institute of Economics, Jena

This Version is available at:

<https://hdl.handle.net/10419/31797>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# PAPERS on Economics & Evolution



MAX-PLANCK-GESELLSCHAFT

# 0512

**Science, social networks and spillovers**

**by**

**Olav Sorenson  
Jasjit Singh  
Lee Fleming**

The *Papers on Economics and Evolution* are edited by the Evolutionary Economics Group, MPI Jena. For editorial correspondence, please contact: [evopapers@econ.mpg.de](mailto:evopapers@econ.mpg.de)

ISSN 1430-4716

© by the author

Max Planck Institute of Economics  
Evolutionary Economics Group  
Kahlaische Str. 10  
07745 Jena, Germany  
Fax: ++49-3641-686868

**SCIENCE, SOCIAL NETWORKS AND SPILLOVERS**

**OLAV SORENSON**

LONDON BUSINESS SCHOOL

REGENT'S PARK

LONDON NW1 4SA, UNITED KINGDOM

[osorenson@london.edu](mailto:osorenson@london.edu)

**JASJIT SINGH**

INSEAD

1 AYER RAJAH AVENUE

SINGAPORE 138676

[jasjit.singh@insead.edu](mailto:jasjit.singh@insead.edu)

**LEE FLEMING**

HARVARD BUSINESS SCHOOL

SOLDIERS FIELD

BOSTON MA 02163, USA

[lee.fleming@hbs.edu](mailto:lee.fleming@hbs.edu)

Presented at the “Regional innovativeness – mechanisms of knowledge flows and accumulation” workshop at the Max Planck Institute for Research into Economic Systems, Jena; we thank Andres Meder for his comments.

**ABSTRACT**

Previous empirical research has established that science appears to stimulate the widespread diffusion of knowledge. The exact mechanism through which science catalyzes knowledge flow, however, remains somewhat ambiguous. This paper investigates whether the observed knowledge diffusion associated with science-based innovation genuinely stems from the norm of openness and incentives for publication, or whether it arises as an artifact of scientists having more dispersed social networks that facilitate the dissemination of tacit knowledge. Our findings support the former possibility: We use patent citation patterns to track knowledge flows, and find that science-based innovations diffuse more widely even after controlling for the underlying social networks of researchers as measured using data on prior collaborations.

## INTRODUCTION

Scientists, social scientists and politicians have attributed much of the acceleration in economic growth over the past two centuries to the advancement of science (Marx, 1944; Bush, 1945; Kuznets, 1959). Despite this widespread belief in the value of science, however, we know relatively little about which mechanisms may or may not contribute to this linkage. One mechanism that has received a substantial amount of attention is the norm of openness. Whereas commercially motivated inventors typically attempt to keep their findings secret in the hopes of benefiting as much and as long as possible from the fruits of their labor, both the reward system and the values in the community of scientists compel them to disseminate knowledge gained to others as quickly as possible (Merton, 1942; Dasgupta and David, 1994). As a result, public science presumably benefits society by generating a high level of knowledge spillovers, which in turn increase the efficiency of research by minimizing the duplication of effort (Nelson, 1959; Arrow 1962) and may also stimulate innovation and economic growth (Marshall, 1922; Romer, 1986; Aghion and Howitt, 1992).

Evidence in support of more efficient knowledge diffusion through science has come primarily in three forms. On the one hand, a vibrant stream of historical research has sought to link the rise of Western nations to the rapid accumulation of knowledge associated with the Enlightenment (e.g., Rostow, 1975; Mokyr, 2002). Though an important strand of research, these broad historical accounts offer only limited traction for distinguishing an increase in spillovers from other potential mechanisms linking science to economic growth (though Bernal, 1939, and David, 2004, argue that the norms regarding openness in science arose at almost precisely the time of the Enlightenment). A second strand of investigation meanwhile has explored the motivations of scientists, finding that these individuals do indeed appear to want to disseminate

their discoveries widely to gain recognition. Recent research, for example, has demonstrated that this motivating factor exists even for scientists working in profit-maximizing firms (Murray, 2003; Stern, 2004). And at a more micro level, another group of researchers has examined patent and publication data to understand the linkages between scientific research (often identified narrowly as that occurring in a university setting) and regional rates of innovation and invention (e.g., Jaffe, 1989; Autant-Bernard, 2001). Among these, Sorenson and Fleming (2004) provide the most direct and systematic evidence that science promotes the diffusion of knowledge. Specifically, they find that patents that reference non-patent prior art (i.e. published materials) – whether peer-reviewed or not – receive citations at a higher rate from more distant patents in both geographic and technical space.<sup>1</sup>

Though consistent with a view that publication dramatically extends the spatial reach of spillovers, the evidence to date does not conclusively establish publication as *the* primary mechanism through which science fosters diffusion. An alternative possibility exists: those engaged in science and in developing technology related to it might simply have wider ranging social networks than traditional inventors. Social scientists have long recognized the importance of boundary-spanning individuals in diffusing knowledge (e.g., Allen, 1977; Tushman, 1977), and recently, several papers have rigorously demonstrated that technological knowledge diffuses primarily through social relations. For example, both Breschi and Lissoni (2004) and Singh (2005) find that collaboration networks – social networks formed when researchers work together on the same invention – account for much of the variance in citation patterns within and across regions. In other words, the underlying patterns of direct and indirect social relations drive the observed localization of knowledge flows; spillovers remain largely local because inventors,

---

<sup>1</sup> Consistent evidence also appears in the location choices of German technology start-ups (Audretsch, Lehmann and Warning, 2004), where firms exploiting published knowledge exhibit less sensitivity to the location of universities than those that require access to tacit knowledge.

like most people, primarily interact with others that live and work in close proximity to them. But scientists may travel more extensively and maintain more distant networks than other individuals. Audretsch and Stephan (1996), for example, report that more than 70% of the links between academic scientists and the biotechnology firms with which they work cross regional boundaries. We also know that foreign-born individuals account for disproportionately large shares of the scientists in the United States as well as other countries (Levin and Stephan, 1998). If this mobility translates into more geographically dispersed social networks, then the more rapid diffusion of science-based knowledge might simply reflect the more extensive reach of scientists' contacts, rather than the importance of publication to stimulating spillovers. It is therefore important to determine whether farther-reaching social networks or publication itself accounts for the widespread diffusion of science-based knowledge.

We extend work by Sorenson and Fleming (2004) to distinguish between these competing possibilities, investigating whether differences in the spatial dispersion of social networks can account for the greater spatial diffusion of citations to patents referencing published materials. In particular, we test whether the likelihood of a patent receiving a citation varies as a function of whether or not the focal patent builds on scientific research while simultaneously controlling for the social distance between the two (groups of) inventors. Our information on the structure of social networks comes from the collaborations of inventors across patents (cf. Breschi and Lissoni, 2004; Singh, 2005). Although this approach limits our ability to test the importance of social networks (because we only consider the effects of one particular type of relation), we nonetheless believe it provides useful information because the technical knowledge important to fomenting invention more likely passes through these professional collaborations than other, more purely social, relations.

Our results strongly affirm the importance of publication – and by extension the norm of openness in science – to the diffusion of knowledge. Though we do not observe any potential payments for the transmission of this knowledge, and hence cannot say with certainty that this increased diffusion involves spillovers, it seems unlikely that scientists can effectively exclude others from, and therefore credibly demand payments for, access to their published research. We find that publication has the greatest marginal benefit to knowledge diffusion when interpersonal ties do not link the source and destination (teams of) inventors. For patents that do *not* reference scientific articles, the likelihood that the knowledge resulting from the source researcher(s) diffuses to other inventors increases greatly when a network path – for example, prior collaborators, collaborators of prior collaborators, etc. – connects the two parties. On the other hand, for patents that do reference scientific articles, having such network relations offers almost no additional advantage to the probability of knowledge diffusion. In other words, consistent with sociologists' expectations of the relationship between social network-based and broadcast-based diffusion, interpersonal social relations and publication appear to act as substitutes in the diffusion of knowledge. Science accelerates spillovers by removing the dispersion of knowledge from the relatively restricted range of social networks and opening it to all capable of absorbing the codified version.

### SCIENCE AND SPILLOVERS

Two types of criteria separate scientific activity from non-scientific activity in the literature. The first, with its locus in the philosophy of science, focuses on the logic of the scientific method (e.g. experimental design), as well as how and why that method might produce more accurate theories about the nature of the world. Meanwhile, a second set, originating with Merton (1942)



and derived primarily from the sociology of science, focuses on science as an institution – understanding the career incentives facing scientists and the norms and values promulgated by the academic community. Both of the mechanisms that we consider as potential explanations for the more rapid diffusion of knowledge developed by scientists relate to science as an institution.

Though Merton and other sociologists have identified several central norms in the scientific community, with regard to the question of spillovers, one in particular seems most relevant: the idea of ‘communism’. ‘Communism’ refers to the notion that individual scientists do not expect to gain from their discoveries beyond the rewards associated with the credit associated with finding them first (Merton, 1942). Other scientists have free access to the ideas generated by their predecessors, as long as they acknowledge those prior scientists’ contributions (e.g., through a citation).

One might expect this norm to minimize the incentives for innovation. To the contrary, however, scientists receive strong, though mostly indirect, incentives to innovate because the community invariably rewards scientists on the basis of the number and the importance of the discoveries they have made (Merton, 1957). These rewards come in a variety of forms – recognition through citations, prizes, and the naming of species, theories and elements, and resources through research grants, endowed chairs, university-funded laboratories and graduate students. In essence, these rewards attach a private good (an incentive) to the public good of new and valuable information (Dasgupta and David, 1994).<sup>2</sup>

---

<sup>2</sup> Gustin (1973) questions the importance of these incentives given that a large number of scientists continue to publish even late in their careers after it has become clear that their odds of garnering such accolades has diminished to essentially zero. The motivations for publishing therefore may well depend more on the internalization of the norms of the community than on these incentives.

*Spillovers through publication*

Taken together, the norm of communism and the incentives surrounding first discovery also engender an intense desire among scientists to publish new findings and ideas as quickly as possible (Merton, 1942; Dasgupta and David, 1994). Although other forms of communication, such as presenting at conferences and seminars, also conform to the norm of communism, publication offers a particularly effective means of disseminating one's discoveries to the scientific community. Quick publication also allows scientists to establish and defend their claims to primacy on a discovery. Publication thus plays an important dual role.

A concomitant advantage of this drive to publish is the more rapid diffusion of new knowledge. To facilitate the efficient transmission of ideas, scientists have developed highly specialized vocabularies and grammars to codify complex information. These languages become a type of shorthand for efficiently transmitting information by allowing single words or phrases to represent a large number of interconnected ideas (Cowan and Foray, 1997), just as a cookbook saves space by indicating, for example, that a cook should 'julienne' a beet instead of describing the entire process of cutting the vegetable into thin, matchstick like strips. Scientists learn these languages, as well as how to implement the processes to which they refer through university education and by interacting with others working on similar topics. By embedding a great deal of information within these terms, articles can therefore transmit vast quantities of information efficiently, allowing journals to broadcast information that once required interpersonal communication (Senker, 1995).

Once the results of a discovery take written form, they can flow far and wide. Publishers distribute printed journals both to individuals and libraries around the world where they become available to researchers everywhere. Even when journals had to move over the ground through

horse-drawn carriages and across the water in ships, news of important discoveries could reach every corner of the planet in just a few weeks. And, with the advent of the Internet, we have reached the point where the dissemination of codified knowledge to others has become more or less instantaneous and costless – with the exception of the costs associated with codifying the knowledge in the first place (Brökel, 2005). Moreover, since written records represent archives that individuals can access at any time, publication also eliminates the need for individuals to meet in time; published knowledge does not die with its discoverer (Cowan and Foray, 1997).

To the extent that other researchers can access the information embodied in these publications, they can potentially benefit by building on these findings without having to expend the resources to rediscover them (Bernal, 1939; Nelson, 1959; Arrow, 1962). From the perspective of for-profit firms, however, the kind of innovation (basic science) that the scientific community rewards, and consequently publishes, might nonetheless differ from the kind of innovation (with more commercial application) a firm might optimally desire (Gittelman and Kogut, 2003; Murray, 2003).

#### *Spillovers through networks*

Science might also generate spillovers through at least one other mechanism: wider ranging social networks. In addition to journals, the scientific community has also created a wide array of organizations to facilitate the flow of communication: conferences, departments, academies, etc. These organizations help researchers to form and maintain distant linkages across employers and locations, thereby producing social networks capable of transmitting information across organizational and regional boundaries. Though trade associations and peer networks may form similar connections among non-scientists, relative to scientific organizations, these organizations

tend to be small in size (Zuckerman and Sgourev, 2005), and hence offer more limited opportunities for forming valuable connections.

Greater mobility might also extend the networks of scientists to more distant regions. Though no research appears to have compared the movement of scientists to other types of technical workers that might engage in invention, some evidence suggests that scientists might exhibit unusually high levels of geographic mobility. Levin and Stephan (1998), for example, report that an uncommonly high proportion of scientists migrated to the United States from other countries. Modern science has also developed institutional practices, such as the post-doc, that explicitly encourage young researchers to experience and form ties to distant labs (cf. Melin, 2004). Histories of science, moreover, suggest that these patterns of mobility of individuals moving from one institution and region to another have long existed (e.g., Gribbin, 2004). To the extent that individuals maintain contact with those with whom they interacted in their prior locations, this mobility will foster denser networks of linkages across institutions and regions.

Though more geographically dispersed social networks could also explain differences in the spatial reach of spillovers generated by science, publications and social networks differ markedly in their implications both for what type of information could pass through these channels and for who could access it. On the one hand, social networks offer greater bandwidth, in the sense that they can transport both codifiable (though potentially not yet codified) and tacit knowledge (Cowan, David and Foray, 2000; Brökel, 2005). On the other hand, they operate over a more circumscribed set of potential recipients: Whereas anyone capable of reading and understanding the codified knowledge printed within them can access a publication, only those with direct or indirect connections to sources can draw on knowledge traveling through social

networks. We exploit this difference in estimating the degree to which each of these factors might account for the greater observed range of the diffusion of scientific knowledge.

### EMPIRICAL STRATEGY

Patents provide a window of observation on the diffusion of knowledge through their references to other patents and published materials. Patents reference two kinds of ‘prior art’: (i) those earlier patents that the new invention extends, and (ii) other materials (e.g., scientific and technical publications) containing ideas on which the inventor built. Our empirical strategy presumes that references in patents to both other patents and non-patent prior art signify cases where an invention builds, to some extent, on the information embodied in these sources.

Inventors would prefer to minimize these references. As patents confer a property right, both references to other patents and citations to prior publications can reduce the scope of a patent’s claims and consequently reduce the effectiveness of any future attempt to defend it legally. Even in cases where the patent applicant also owns a prior patent on which the new invention builds, she may wish to avoid naming that patent as prior art for at least two reasons. On the one hand, if the new patent received approval without citing the earlier patent, it might effectively extend the term of the assignee’s property right over that earlier invention. On the other hand, even if the inventor applied for both patents concurrently, receiving two patents over an overlapping domain of intellectual space could provide useful redundancy in enforcing property rights. The patent review process, however, mitigates the incentives to not cite; using personal expertise and automated searches, patent examiners in the review process insert relevant citations, where missing, to applications before granting a patent.<sup>3</sup>

---

<sup>3</sup> Alcazer and Gittelman (2004) report that a large share of citations result from this review process, suggesting that patent examiners provide an important check on inventors’ incentives to not report relevant prior art.

Since patent offices release publicly the information reported in the applications of the patents they grant, the patent itself almost certainly accelerates the diffusion of the knowledge embodied in it. One might then assume that our ability to find a marginal effect to publication would depend on a propensity for inventors to keep secret some portion of the underlying knowledge and/or on the failure, or inability, of potential inventors to monitor all new patent awards. One should note, however, that we do not assume that these non-patent references provide other inventors with information on the focal inventions themselves; rather, we believe that both patents build on some underlying piece of (potentially published) knowledge. In this sense, the disclosure aspect of patents should not greatly influence our analysis as we seek to use patents more as a means of tracking bodies of latent knowledge than to follow the diffusion of awareness of the focal invention. Publication widens the range of potential inventors aware of this underlying information, thereby expanding the pool of individuals capable of building on a particular piece of knowledge.

Figure 1 provides a diagrammatic illustration of this assumption. The squares denote patents, while the oval represents a piece of non-patent (potentially published) knowledge. In situations where patent 1 extends published work, other inventors can more easily access this knowledge. We thus anticipate that publication increases the probability that others build on the same knowledge, and hence reference patent 1 (as in the case of patent 2).

Assuming that the inventor of patent 2 has an awareness of the literature referenced by patent 1 might strike some as a rather heroic assumption. To assess the validity of this claim, Sorenson and Fleming (2004) surveyed a sample of patent holders. More than half (62%) of the patent holders reported an awareness of at least some of the *specific* references listed on the

patents they cited and 71% indicated an awareness of either the cited articles or other similar articles. The empirical evidence therefore appears to support this assumption.

INSERT FIGURE 1 ABOUT HERE

### DATA AND MEASURES

Our analysis drew on a sample of utility patents constructed for earlier papers (for details, see Fleming and Sorenson, 2004). In relation to the research question considered here, this sample has several useful features: It covers a short window of time (May-June 1990) to limit the extent to which time-varying heterogeneity in the level of activity across fields influences our findings. It includes sufficient information on forward (future) citations to estimate the factors affecting diffusion. And, most importantly for this paper, a trained researcher has already assigned to categories the 16,728 non-patent references in our sample.

Our analysis focuses on the effects of two variables. The first of these is a measure of whether or not an invention draws on scientific research. The trained researcher classified the non-patent references into seven mutually exclusive and exhaustive categories: *Scientific Index* journals, conference proceedings, technical reports, technical corporate publications, non-technical corporate publications, books and non-index journals. Sorenson and Fleming (2004) provide a detailed description of the rules governing this classification system and summary statistics concerning these non-patent references. In identifying science-based patents, our analysis relies on references to *Scientific Index* journals (i.e., information on whether or not any specific patent cites a publication listed in the *Scientific Index*, a database of peer-reviewed

scientific journals).<sup>4</sup> This category includes both the high prestige journals, such as *Science* and *Nature*, and a multitude of more and less well-known journals.

Some prior research suggests that patent examiners appear to have much less influence on the assignment of non-patent prior art. In a small sample, Tijssen (2001), for example, finds that the majority of these references came from the inventors. One might then wonder why patent applications include these references (since they limit the scope of the patent, and therefore reduce the value of the property right conferred). In a series of interviews with patent holders, Sorenson and Fleming (2004) found that inventors had two primary explanations.<sup>5</sup> First, many viewed the citation of relevant material – whether a patent or a paper – as the “right thing to do.” Second, they also believed that these non-patent references might confer legitimacy on their applications, and consequently increase the likelihood of them being granted.

The second central measure is the strength of social connection between two inventors. Large-scale, systematic data on interpersonal relations has generally been unavailable for studies of information diffusion. In the case of inventors, however, the patent data allow us to track a subset of each individual’s overall social network: those relations arising from collaboration on inventions. Singh (2005) uses collaboration information for patents registered with the U.S. Patent Office (USPTO) to construct a longitudinal database of interpersonal relations among all inventors listed on U.S. patents from 1975 to 1996. This database allows him to construct a ‘social proximity graph’ involving all inventions with more than one inventor, which in turn provides a measure of the ‘social distance’ between any two (teams of) inventors. For example,

---

<sup>4</sup> Citations to *Scientific Index* journals cluster with other types of non-patent references related to scientific research, such as conference proceedings, technical reports, technical corporate publications and books (Sorenson and Fleming, 2004). Any less restrictive definition of scientific research, therefore, would correlate highly with the one we use. We nonetheless consider most appropriate the use of only citations to *Scientific Index* journals because these include some minimum levels of quality and public availability.

<sup>5</sup> Self-citations do not appear to account for these non-patent references; Sorenson and Fleming (2004) report that only 3% of inventors also appear as authors on the non-patent references that their patents cite.



if two inventors have collaborated on a prior invention, then they would have a path length of one on this graph. A collaborator of a collaborator is a path length of two, and so forth.

In our estimations, we cluster dyads into three groups depending on the path length: A *short network path* indicates that it would take three or fewer steps to link the inventors in a patent dyad, while a *long network path* implies that more than three (but a finite number of) steps exist between the two. We use pairs of inventors that cannot reach each other through the collaboration network as the baseline category for estimation.

### REGRESSION METHODOLOGY

To analyze the diffusion of citations in geographic and social space, we estimated the probability that a future patent cites a given focal patent as a function of our variables of interest and a variety of control variables. Dyads of focal patents and future (potential) citing patents thus become the unit of analysis. To avoid the potential problems associated with estimating non-independent cases, we assembled the data for this analysis using a case-control sample design (alternatively one could reduce the effects of network autocorrelation through estimation; e.g., Krackhardt, 1988); in other words, we paired a set of future patents that cited our focal patents (cases) with a second set that did not (controls). Our sample consists of all 60,999 citations that actually exist for our sample of 17,264 patents. In addition, we coupled each of the 17,264 focal patents with four future patents that did not cite it (chosen at random from all patents granted between July 1990 and June 1996).<sup>6</sup> This process produced a data set of 130,055 dyads, but data limitations on some of the control variables forced us to restrict our analysis to the 75,278 cases

---

<sup>6</sup> We chose four patents for the ‘control’ group so that the sample would include roughly equal proportions of realized and unrealized dyads. To address the fact that focal patents enter the data more than once, we report robust standard errors. Though one could potentially ‘match’ the control sample to the case sample on one or more dimensions, we chose not to do so because matching precludes one from estimating the effects of any dimension on which one matches.

where the inventors of the focal patent listed addresses within the United States. The inventors for the (potential) citing patent could, however, reside anywhere in the world. Our analysis therefore is a study of the global diffusion of knowledge originating in the United States, though we see no reason why these results should not generalize to knowledge originating anywhere in the world.

To account properly for the effects of the sampling procedure, our estimations employed a rare events logistic regression methodology.<sup>7</sup> Logistic regression yields biased estimates when the proportion of positive outcomes (citations in this case) in the sample does not match the proportion in the population. In particular, our matching algorithm generated a sample with a much higher proportion of citations than found in the population as a whole. To adjust for this fact, we use the *weighted exogenous sampling maximum likelihood* (WESML) estimator. This procedure modifies the usual maximum likelihood function used for logistic regression by weighting each term in the likelihood function by the number of population elements it represents (i.e. by the inverse of the sampling probability for the actual and potential citations respectively).

Our analysis also includes several control variables. First and foremost, since we conceptualize the process as one of diffusion, we need controls for distance. All patents report the home addresses of the inventors on the front page of the patent application. We therefore generated a geographic distance measure by matching inventors' 3-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.

---

<sup>7</sup> See King and Zeng (2001) for a survey of the state of the art in this methodology. Both Sorenson and Fleming (2004) and Singh (2005) have applied (slightly different forms of) rare events logistic regression in the context of patent citations.

Postal Service.<sup>8</sup> We calculated the distance in miles between all patent dyads using spherical geometry; taking the natural log of this value accounted for the fact that the relevance of each additional mile declines with distance. Since we do not have exact distance information for non-U.S. locations, we simply set the *log (geographic distance)* variable to zero for pairs involving such observations and instead capture the average effect of knowledge diffusion across national borders using a dummy variable *foreign*.

Distance also exists in a technological sense. To create a measure of the distance between two patents in terms of technological space, we calculated the overlap in subclass assignments between each focal patent, *i*, and each (potential) citing patent, *j*:

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

where *s* is a vector of subclass assignments, with each cell being a binary indicator variable of membership in a subclass (i.e. one denoting assignment to the subclass). The measure ranges from zero to one, with larger values representing more distant technologies. We also included a *same class* indicator variable for dyads belonging to the same primary technology class.

In addition to the distance measures, the citation probability models include several additional controls. *Self-cite* signifies dyads where both patents belong to the same assignee. The *number of prior art citations* counts the references on the focal patent to earlier patents. A *technology activity control* captures differences in the average level of activity in different technological domains by averaging the typically number of citations received by a patent with the same set of subclass assignments as the focal patent (see Fleming, 2001, for a complete description and

---

<sup>8</sup> The USPTO reports 5-digit zip information, but we opted to use cleaned data from CHI, an information provider, to reduce measurement error. CHI has telephoned every patent holder to verify inventors' addresses; however, it only maintains this information at the 3-digit level. Where the patent lists more than one inventor, we used the location corresponding to the address of the first inventor. Models where we randomly selected a location from the listed inventors, however, produced equivalent results.

discussion of the logic behind this measure). And fixed effects for the time lag (in years) between the cited and (potential) citing patent capture systematic differences in the probability of citation that result from having different time lags between the pairs of patents as well as systematic cross-year differences in citation probability. The definitions for all of our variables appear in Table 1. Table 2 reports summary statistics for these variables, while Table 3 provides a correlation matrix.

## RESULTS

Table 4 reports the results from the WESML regressions. As already mentioned, each observation in the sample used consists of a focal patent and a (potential) citing patent. The dependent variable is the indicator variable *citation*, which takes a value of one when the focal patent receives a citation, and is zero otherwise. Column (1) reports the results from only including the dummy *Science Index*. Consistent with previous research, we find that science-based patents have a greater probability of being cited. In column (2), we introduce all of our control variables except those related to the network path length between inventors. Interestingly, we can no longer reject the hypothesis that the coefficient on *Science Index* does not differ statistically from zero, though this conclusion stems entirely from an increase in the standard error for the *Science Index* coefficient (and the result is somewhat fragile – under a wide range of specifications the coefficient remains positive and significant).

Among the control variables, the probability of knowledge flow between two inventing teams diminishes with distance, as one would expect. This result holds true both for domestic patents (potentially citing patents with inventors located in the U.S.) as well as foreign patents (patents arising from inventors located outside the U.S.). As one might expect, being located

outside the United States has a much larger negative effect on the probability of citation. On average, being located outside the U.S. has an equivalent effect to the two inventors being separated by 3,800 miles ( $=\exp^{-5.448/-0.661}$ ) – in other words a substantially larger effect than the distance between Boston and San Diego (~2,600 miles). Knowledge flow also rises with technological similarity, both in terms of greater subclass overlap and for patents belonging to the same primary technology class. Finally, consistent with previous literature, we find the probability of patent citation increases when the focal patent and the (potential) citing patent belong to the same assignee.

In column (3), we now introduce two new variables - *short network path* and *long network path* – to capture cases where the source and destination inventors belong to the same connected component of the collaboration network with a minimum path length of less than or equal to three, or of greater than three, respectively. Recall that teams not connected through the collaboration network form the baseline category. Consistent with Singh (2005), we find that the probability of knowledge diffusion increases with proximity (that is, the estimated coefficients follow the ordering: *short network path* > *long network path* > 0). Finally, column (4) reports the estimates of whether social connections indeed matter less for the diffusion of science-based innovations, the argument posited by Sorenson and Fleming (2004) as a possible reason for why science-based innovations diffuse more widely. Consistent with their claim, the value of social networks to diffusion declines for inventions building on published knowledge, as indicated by the negative and significant coefficients for *Science Index X short network path* and *Science Index X long network path*.

To understand better the meaning of these coefficients, we calculated the change in the relative risk of a citation associated with publication and the existence of a network connection

(compared to a focal patent, unconnected in the collaboration network to the potential citing patent, building on unpublished knowledge). The results of our calculations appear in Table 5. When the focal patent does not build on published knowledge, a social connection between the inventor of the focal patent and the inventor of the potentially citing patent increases the probability of a citation by 3.5 to 11 times, depending on the closeness of their connection. The marginal benefits to these social connections decline markedly, however, when the focal patent draws on published knowledge. In fact, given the size of the standard errors, we cannot reject the possibility that linkage in the collaboration network confers *no* advantage to the access of published knowledge.

#### DISCUSSION AND CONCLUSION

Our paper opened with a call for more detailed investigation of the mechanisms underlying the greater dispersion of scientific knowledge. In particular, we sought to distinguish between two institutional mechanisms that could explain this phenomenon. On the one hand, the norm of openness and the incentives to publish research findings might lead scientists to codify knowledge to a greater degree, thereby broadcasting new knowledge to all capable of receiving it. On the other hand, the greater mobility of scientists and the existence of organizations to promote inter-organizational and –regional ties might engender the development of more extensive social relations among scientists, facilitating a more rapid diffusion of knowledge through these networks. We exploit a novel data set to account for direct and indirect collaboration ties between researchers and find that, even after controlling for these social relations, innovations that build on published scientific research appear to diffuse far more rapidly than the ones that do not. Moreover, our results suggest that *social connections to the*

*inventor provide no marginal benefit* to other inventors seeking to build on a prior invention that itself draws on publicly available (published) knowledge.

Not only do our results point to publication, rather than more extended social networks, as the mechanism underlying the more rapid diffusion of knowledge developed by science, but also they help to dispel another alternative explanation for the value of science: that the scientific method produces higher quality innovations. Much of the literature on the value of science, particularly to firms, has either explicitly or implicitly assumed that the process of scientific research generates knowledge of greater generality and value; one might therefore expect this knowledge to diffuse more widely. This alternative explanation, however, cannot account for why social connections would matter greatly to the diffusion of unpublished knowledge but matter little at all to the spread of published knowledge. The fact that the two act as substitutes suggests that publication and social networks play similar roles in diffusing knowledge. An account regarding quality differences between the innovations of science and non-science suggests that the effect of building on science should be orthogonal (or perhaps even complementary) to social connectedness.

To the extent that the codification and publication of knowledge accelerates its flow, policymakers should look to additional mechanisms for encouraging this process. The more rapid diffusion of knowledge benefits society in at least three ways. Most obviously, the public availability of research results reduces the likelihood that multiple firms and/or individuals engage in redundant research, thereby increasing the efficiency of R&D investments (Bernal, 1939; Nelson, 1959; Arrow, 1962). If the process of invention involves the recombination of elements into novel configurations, then the widespread dissemination of knowledge might further increase the pace of innovation by expanding the number of elements, and hence the

combinatorial potential, available to any given inventor (Weitzman, 1998). Finally, the public dissemination of knowledge may also increase the efficiency of *production* based on that knowledge. In the absence of such public availability, only a limited number of firms can compete on the provision of the goods or services related to the innovation. Diffusion of the underlying knowledge enables the entry of additional firms and likely transforms the basis of competition from one of access to the knowledge to a regime of the most efficient production and effective implementation of the innovation. Policies and institutions therefore that either speed the dissemination of knowledge (e.g. Internet-based working paper archives) or increase the proportion of knowledge that becomes codified in the public domain (e.g. stipulations requiring the publication of government-funded research findings) can improve societal welfare.

Our results also have implications for our understanding of past research. The fact that social networks do not confer an advantage in accessing published knowledge calls into question the pervasiveness of ‘tacit’ knowledge found in the literature. The existence of tacit knowledge, and the need to access it through face-to-face contact, has been used to interpret a wide range of results, such as the importance of personnel movements to knowledge transfer and the tendency for firms to agglomerate into industrial clusters. Our results, however, suggest that science allows people with the appropriate training to interpret and use published knowledge without having to rely upon (localized) social networks for access. Science, in other words, appears to facilitate the codification of knowledge. Our findings therefore support the idea that much uncoded knowledge may simply remain so because the costs of codification exceed the benefits for the holders of it (Nelson and Winter, 1982; Cowan, David and Foray, 2000; Brökel, 2005). Factors that either reduce the costs of codification (such as the existence of specialized



vocabulary) or increase the incentives surrounding dissemination thus can potentially benefit society by expanding the stock of codified knowledge.

These findings nonetheless raise questions as to why private firms would invest in basic scientific research. One possibility is that “performance of basic research may be thought of as a ticket to admission to an information network” (Rosenberg, 1990, p. 170). Science enables the codification of knowledge partially through the specialized languages that scientists develop. Understanding one of these languages not only requires training in the relevant field but also constant contact with the community as the language evolves to accommodate new discoveries that require additional baseline information for efficient codification. On the other hand, our results make it clear that firms that publish their findings will see the benefits of these discoveries erode faster. One might therefore see our results as more consistent with Stern (2004), who contends that firms benefit from doing science not because engaging in science improves the innovation process but because highly skilled employees have a preference for directing their own research programs and publishing their findings and therefore will accept lower wages from firms willing to give them such freedom.

A similar ambiguity remains on whether of not developing countries and regions should stimulate the growth of indigenous scientific communities. On the one hand, developing countries and regions may need to invest in training scientists to first become aware of and make sense out of the cumulative codified knowledge that exists in the rest of the country and world. But such a conclusion appears inconsistent with the absence of a positive macro-level relationship between investments in science and economic growth found in many studies (Shenhav and Kamens, 1991; Schofer, Ramirez and Meyer, 2000; Fritsch and Slavtchev, 2005). As noted above, the public dissemination of information also changes the basis of competition.

In this case, even if developing regions contribute to the base of scientific knowledge, their local industries may still lack the capabilities and complementary assets necessary to compete with firms from more developed regions in the provision of these goods and services.

In both cases of firm and developing region investments in science, the interpretation of results and prescriptions for policy turn critically on the degree to which access to published knowledge depends on the absorptive capacity of the potential recipient. Hence, a better understanding of whether potential recipients differ in their abilities to access published knowledge remains an important, and open, question. Others have made similar arguments, suggesting that ‘tacitness’ is best viewed as a dyadic concept concerning the degree of shared context between a sender and receiver (Nelson and Winter, 1982; Cowan, David and Foray, 2000), but existing research has by and large treated the degree of tacitness as a fundamental (and typically immutable) property of any particular piece of knowledge. Continued progress on these policy issues therefore requires research into whether or not – and what types of – investments in science enable access to the stock of codified (published) knowledge. Standing on the shoulders of giants may first involve a fair bit of climbing to get there.

#### REFERENCES

- Aghion, Philippe, and Peter Howitt. 1992. “A model of growth through creative destruction.” *Econometrica*, 60: 323-351
- Alcacer, Juan, and Michelle Gittelman. 2004. “How do I know what you know? Patent examiners and the generation of patent citations.” Working paper, New York University
- Allen, Thomas J. 1977. *Managing the Flow of Technology*. MIT Press, Boston, MA
- Arrow, Kenneth. 1962. “Economic welfare and the allocation of resources for invention.” Pp. 609-625 in R.R. Nelson (Ed.), *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Princeton, NJ: Princeton University Press

- Audretsch, David B., Erik E. Lehmann, and Susanne Warning. 2004. "University spillovers: Does the kind of science matter?" *Industry and Innovation*, 11: 193-206
- Audretsch, David B., and Paula Stephan. 1996. "Company-scientist locational links: The case of biotechnology." *American Economic Review*, 86: 641-652
- Autant-Bernard, Corinne. 2001. "Science and knowledge flows: Evidence from the French case." *Research Policy*, 30: 1069-1078
- Bernal, John D. 1939. *The Social Function of Science*. New York: Macmillan
- Breschi, Stefano, and Franco Lissoni. 2004. "Mobility and social networks: Localised knowledge spillovers revisited." In *The Role of Labour Mobility and Informal Networks for Knowledge Transfer*, Dirk Fornahl, Christian Zellner, and David Audretsch (eds.), Berlin: Springer Verlag
- Brökel, Tom. 2005. "The spatial determinants of knowledge transfer and its economic implications." Presented at the Regional Innovativeness – Mechanisms of Knowledge Flows and Accumulation Workshop, Jena, Germany
- Bush, Vannevar 1945. *Science: The Endless Frontier*. Washington, DC: US Government Printing Office
- Cowan, Robin, Paul A. David and Dominique Foray, 2000. "The explicit economics of knowledge codification and tacitness" *Industrial and Corporate Change* 9: 211-253
- Cowan, Robin, and Dominique Foray. 1997. "The changing economics of codification and the diffusion of knowledge." *Industrial and Corporate Change* 6: 595-622
- Dasgupta, Partha, and Paul David. 1994. "Towards a new economics of science." *Research Policy* 23: 487-521
- David, Paul A. 2004. "Understanding the emergence of 'open science' institutions: Functionalist economics in historical context." *Industrial and Corporate Change*, 13: 571-589
- Fleming, Lee. 2001. "Recombinant uncertainty in technological search." *Management Science* 47: 117-132
- Fleming, Lee, and Olav Sorenson. 2004. "Science as a map in technological search." *Strategic Management Journal* 25: 909-928
- Fritsch, Michael, and Viktor Slavtchev. 2005. "The role of regional knowledge sources for innovation." Presented at the Regional Innovativeness – Mechanisms of Knowledge Flows and Accumulation Workshop, Jena, Germany

- Gittelman, Michelle. and Bruce Kogut. 2003. "Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns." *Management Science* 49(4): 366-382
- Gribben, John. 2004. *The Scientists: A History of Science Told Through the Lives of Its Greatest Inventors*. Random House
- Gustin, Bernard H. 1973. "Charisma, recognition and the motivation of scientists." *American Journal of Sociology*, 78: 1119-1134
- Jaffe, Adam B. 1989. "The real effects of academic research." *American Economic Review*, 79: 957-970
- Krackhardt, David. 1988. "Predicting with networks: Non-parametric multiple regression analyses of dyadic data." *Social Networks*, 10: 359-382
- Kuznets, Simon S. 1959. *Six Lectures on Economic Growth*. New York: Free Press
- Levin, Sharon, and Paula E. Stephan. 1998. "Are the foreign born a source of strength for U.S. science?" *Science*, 285: 1213-1214
- Marshall, Alfred. 1922. *Principles of Economics*, 8<sup>th</sup> Edition, London: McMillan
- Marx, Karl. [1844] 1975. "Economic and philosophical manuscripts." Pp. 279-400 in *Early Writings*, Rodney Livingstone and Gregor Benton (trans.), New York: Vintage
- Melin, G. 2004. "Postdoc abroad: Inherited scientific contacts or establishment or new networks?" *Research Evaluation*, 13: 95-102
- Merton, Robert K. 1942. "Science and technology in a democratic order." *Journal of Legal and Political Sociology* 1: 115-126
- Merton, Robert K. 1957. "Priorities in scientific discovery: A chapter in the sociology of science." *American Sociological Review* 22: 635-659
- Mokyr, Joel. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton: Princeton University
- Murray, Fiona. 2003. "Innovation as co-evolution of scientific and technological networks." *Research Policy* 31: 1389-1403
- Nelson, Richard R. 1959. "The simple economics of basic scientific research." *Journal of Political Economy* 67: 297-306
- Nelson, Richard and Sidney Winter (1982). *An Evolutionary Theory of Economic Change*. Harvard University Press: Cambridge, MA.

- Romer, Paul M. 1986. "Increasing returns and long-run growth." *Journal of Political Economy*, 94: 1002-1037
- Rosenberg, Nathan. 1990, "Why do firms do basic research (with their own money)?" *Research Policy* 19: 165-174
- Rostow, Walt W. 1975. *How It All Began: Origins of the Modern Economy*. New York: McGraw Hill
- 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
- Senker, Jacqueline. 1995. "Tacit knowledge and models of innovation." *Industrial and Corporate Change* 4: 425-447
- Shenhav, Yahouda, and David Kamens. 1991. "The 'costs' of institutional isomorphism in non-Western countries." *Social Studies of Science*, 21: 427-525
- Singh, Jasjit. 2005. "Collaborative networks as determinants of knowledge diffusion patterns." *Management Science*, 51: 756-770
- Sorenson, Olav, and Lee Fleming. 2004. "Science and the diffusion of knowledge." *Research Policy*, 33: 1615-1634
- Stern, Stern. 2004. "Do scientists pay to be scientists?" *Management Science* 50: 835-853.
- Tijssen, Robert 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
- Tushman, Michael. 1977. "Special boundary roles in the innovation process." *Administrative Science Quarterly*, 22: 587-605
- Weitzman, Martin L. 1998. "Recombinant growth." *Quarterly Journal of Economics*, 113: 331-360
- Zuckerman, Ezra S., and Stoyan Sgourev. 2005. "Peer capitalism: Parallel relationships in the U.S. economy." *American Journal of Sociology*, forthcoming.

**Table 1: Definition of variables**

<b>Citation</b>	The binary dependent variable, which is 1 for actual citations and 0 for the unrealized potential citations in the sample
<b>Science index</b>	An indicator variable that is 1 and only if the cited patent has at least one citation to a scientific journal
<b>Short network path</b>	An indicator variable that is 1 if and only if the citing and cited patents are related through a chain of collaborative links with length not greater than 3
<b>Long network path</b>	An indicator variable that is 1 if and only if the citing and cited patents are related through a finite chain of collaborative links that exceeds a length of 3
<b>Log (geographic distance)</b>	$\ln(\text{distance in miles})$ between the cited patent and potentially citing patent if both are US-based and 0 in case the potentially citing patent is not US-based
<b>Foreign</b>	An indicator variable that is 1 if and only if the potentially citing patent is from outside the U.S.
<b>Subclass overlap</b>	A measure of how close two patents are technologically, as measure by the subclasses listed for the patent
<b>Same class</b>	An indicator variable that is 1 if and only if the citing and the cited patent share the same primary technological class
<b>Self-cite</b>	An indicator variable that is 1 if and only if the citing and cited patent belong to the same assignee
<b>Number of prior art citations</b>	Number of prior art citations made by the cited patent
<b>Technology activity control</b>	A control variable for the expected number of citations given the technologies a patent involves
<b>Time lag</b>	Number of years between the application dates of the citing and the cited patent

**Table 2: Summary statistics**

	<b>Observations</b>	<b>Mean</b>	<b>Standard Deviation</b>	<b>Minimum</b>	<b>Maximum</b>
<b>Citation</b>	75,297	0.518	0.500	0	1
<b>Science index</b>	75,297	0.226	0.418	0	1
<b>Short network path</b>	72,785	0.073	0.260	0	1
<b>Long network path</b>	72,785	0.216	0.411	0	1
<b>Log (geographic distance)</b>	75,297	3.816	3.228	0.000	8.390
<b>Foreign</b>	75,297	0.339	0.474	0	1
<b>Subclass overlap</b>	75,277	0.444	0.853	0	19
<b>Same class</b>	75,297	0.264	0.441	0	1
<b>Self-cite</b>	75,297	0.121	0.326	0	1
<b>Number of prior art citations</b>	75,287	9.867	8.913	0	110
<b>Technology activity control</b>	75,287	1.257	0.425	0.329	3.025
<b>Time lag</b>	75,278	3.071	1.773	0	6

**Table 3: Correlations matrix**

	Citation	Science index	Short network path	Long network path	Log (geographic distance)	Foreign	Subclass overlap	Same class	Self-cite	Number of prior art citations	Technology activity control
<b>Citation</b>	1.000										
<b>Science index</b>	0.068	1.000									
<b>Short network path</b>	0.268	0.059	1.000								
<b>Long network path</b>	0.051	0.091	-0.147	1.000							
<b>Log (geographic distance)</b>	0.089	-0.016	-0.179	0.049	1.000						
<b>Foreign</b>	-0.243	-0.006	-0.186	-0.017	-0.845	1.000					
<b>Subclass overlap</b>	0.502	0.010	0.243	-0.010	-0.008	-0.121	1.000				
<b>Same class</b>	0.566	0.027	0.187	-0.001	0.027	-0.130	0.434	1.000			
<b>Self-cite</b>	0.042	-0.040	0.117	-0.056	-0.011	-0.042	0.048	0.037	1.000		
<b>Number of prior art citations</b>	0.064	0.063	0.039	0.006	0.015	-0.034	0.055	0.043	0.020	1.000	
<b>Technology activity control</b>	0.187	0.159	-0.003	0.103	0.028	-0.025	-0.020	0.122	-0.064	0.012	1.000
<b>Time lag</b>	0.141	0.013	-0.014	0.106	0.070	-0.064	-0.007	-0.085	-0.006	0.007	0.044



**Table 4: WESML Regression analysis for probability of patent citations**

	(1)	(2)	(3)	(4)
<b>Science index</b>	0.338** (0.033)	0.377 (0.291)	0.202 (0.280)	1.455** (0.143)
<b>Science index X Short network path</b>				-2.948* (1.155)
<b>Science index X Long network path</b>				-2.556** (0.588)
<b>Short network path</b>			1.423* (0.634)	2.442** (0.652)
<b>Long network path</b>			0.440* (0.188)	1.287** (0.154)
<b>Log (geographic distance)</b>		-0.661** (0.048)	-0.572** (0.042)	-0.490** (0.034)
<b>Foreign</b>		-5.448** (0.344)	-4.995** (0.395)	-4.335** (0.296)
<b>Subclass overlap</b>		5.209** (0.266)	5.207** (0.261)	5.518** (0.239)
<b>Same class</b>		3.978** (0.284)	3.786** (0.299)	3.875** (0.276)
<b>Self-cite</b>		0.669* (0.320)	0.731* (0.305)	0.754* (0.316)
<b>Number of prior art citations</b>		-0.008 (0.017)	-0.008 (0.017)	-0.003 (0.015)
<b>Technology activity control</b>		0.264 (0.243)	0.107 (0.234)	0.138 (0.232)
<b>Fixed effects for time lag</b>	Included	Included	Included	Included
<b>Number of Observations</b>	75,278	75,267	72,773	72,773

Robust standard errors in parentheses  
\* significant at 5%; \*\* significant at 1%

**Table 5: Relative risks of the probability of a citation**

Relative risk of citation	Science Index = 1	Science Index = 0
Short path	2.58	11.50
Long path	1.20	3.62
No path	4.28	1.00

**FIGURE I**

**MODEL OF KNOWLEDGE FLOWS**

