International Review of Business Research Papers

Vol.5 N0. 2 March 2009 Pp. 420-429

Not Far From the Madding Crowd: The Role of Proximity in Biotechnology Innovation

Daniel K.N. Johnson*

Using patent citation data for the U.S., we test whether knowledge spillovers in biotechnology are sensitive to distance, and whether that sensitivity has changed over time. Controlling for self-citation by inventor, assignee and examiner, cohort-based regression analysis shows that physical distance is becoming less important for spillovers with time.

Field of research: Economics of innovation & technological change

1. Introduction

It has long been noted that firms within an industry often cluster geographically. Localization economies, which reduce the cost of inputs to firms in the local industry, have been studied in a variety of contexts (e.g. Henderson, 1986; Smith and Florida, 1994). For some industries, it is the nature of the knowledge itself that encourages co-location (Caballero and Jaffe, 1993; von Hippel, 1994). This paper examines knowledge flows within biotechnology using patent citations, controlling for self-citation by inventor, assignee, and examiner. We confirm the traditional evidence that inter-firm knowledge transfers decrease with distance, while adding the important caveat that the impact of physical distance has been diminishing over time. Thus, historically there is less reason for biotechnology firms to cluster geographically than there has been in the past. In section 2 of the paper, we review the relevant literature. Section 3 describes our data set and regression methodology. Section 4 presents our findings, while Section 5 concludes with implications for policy and further research.

---

* Daniel K.N. Johnson is an Associate Professor of Economics at Colorado College, 14 East Cache La Poudre Street, Colorado Springs, CO 80903, tel: (719) 389-6654, fax: (719) 389-6927, email: djohnson@ColoradoCollege.edu. Special thanks to Milena Mareva and David Popp for suggestions about data and programming. Financial support was provided by a National Science Foundation Award for the Integration of Research and Education, a Wellesley College Faculty Grant, a Colorado College Faculty Grant, a Mrachek Research Fellowship, a Mellon Research Block and the Chapman Fund.

2. Literature Review

The literature suggests that knowledge spillovers cluster geographically. The underlying supposition is that inventors are more aware of (or find more use for) inventions located close to them, and therefore build more heavily upon them. Empirical evidence stresses the important role of geography in the spillover of knowledge from one member of an innovation network to another (see Gelsing, 1992), and the importance of frequent personal contact, for example as Zucker et al. (1998) showed the important location of academic research superstars.Geographic proximity has been used to explain the location of R&D-intensive firms (Dorfman, 1988 and Carrincazeaux, 2001 are examples). However, the location of firms is not always a good predictor of the location of innovation. Feldman (1994) finds only a 0.42 correlation between innovation measures and value-added by region. This result is confirmed (Johnson and Brown, 2004) in an exploration of why the northeastern U.S. has dramatically declined as a share of the national patenting total. Other researchers have demonstrated a geographic pattern to European patent citations (Sjoholm, 1996; Maurseth and Verspagen, 1999).

Lundvall (1992) points out that the importance of geography should differ predictably by technology. While geography has little impact on stationary technologies (facing constant needs and opportunities), rapidly changing sectors should see geography as more important. That intuition was verified for a range of sectors (Jaffe et al., 1993; Jaffe and Trajtenberg, 1996; Almeida and Kogut, 1997). Since biotechnology knowledge becomes obsolescent very rapidly (Johnson and Santaniello, 2000), one might expect distance to matter greatly. However, most biotechnological information is not tacit, so will be relatively easy to communicate across long distances. In addition, biotechnology has emerged during a period when inter-regional communication has been increasingly affordable, so we might expect less localization of knowledge spillovers (Feldman, 1999).

3. Methodology

This paper uses biotechnology patent citations as a measure of knowledge spillovers. When a patent application is submitted for approval, it is accompanied by a list of citations to other patents. The intention is twofold: to build a convincing case that this application is novel, and to provide a legal record to protect patent rights in the future. The result is a paper trail of knowledge creation. Of course, patents do not perfectly reflect the creation of technology: some innovations are never patented and patents vary greatly in size and importance. However, Feldman (1994) showed that within the U.S., patents have high correlations with other measures of innovative activity. Citations themselves do not perfectly reflect the transfer of knowledge, as they may be inserted for a variety of other reasons. Jaffe et al. (2000) relates survey evidence showing that only ¼ of all patent citations correspond to a clear spillover of knowledge, with another ¼ having some possibility of a spillover. However, statistical tests indicate that overall citations can be interpreted as a noisy signal of spillovers. As a final definitional

challenge, "biotechnology" definitions differ between nations and over time (see Johnson and Santaniello, 2000). Therefore, we follow the most recent published biotechnology definitions of the U.S. Patent Office (USPTO, 1998), which include portions of eleven separate classes from the U.S. patent classification system[i].

Patent citations cluster for non-geographic reasons. For example, inventors may cite their own work, which would give a biased impression of the importance of geography. The same may be true of assignees, if employees of a firm are familiar with other patents held by the firm. Therefore we include self-citations in the analysis but control for them separately. We also identify an "examiner self-citation effect" to distinguish it from any geographic pattern we may observe. Using U.S. patent data from a combination of sources (Hall et al., 2001 among others), we recorded citations from all biotechnology patents granted between 1975 and 1994. Self-citation accounted for only one percent of all citations, with a further eight percent to the same assignee firm. Five percent were made to other patents sharing the same examiner, with considerable variation between examiners. For one examiner over seventy percent of the citations made to patents he examined hail from other patents he reviewed.

Our regression analysis extends the seminal work of Jaffe and his co-authors (Caballero and Jaffe, 1993; Jaffe and Trajtenberg, 1996). We add to the model by allowing coefficients to vary over time, and by directly controlling for the potential impact of technological similarity between regions and self-citation by inventors, assignees and examiners. The model recognizes that the likelihood that a patent k granted in year t will be cited by a subsequent patent, K, granted in year T, is at least in part a function of the attributes of patents k and K. Using exponential rates of decay and diffusion to model the flow of knowledge over time, that probability can be written as:

$$p(k,K) = \alpha(k,K)\delta(k,K)\exp[-\beta_1(\sum_{s=t_A}^{T} P_s)][1-\exp(-\beta_2(T-t)] \quad (1)$$

where α(k,K) represents the non-geographic attributes of patents k and K that affect the probability of citation, while δ(k,K) represents the relevant geographic attributes. β_1 represents the decay rate of knowledge, permitting older patents to be cited less frequently as they fall from the leading edge of their art. β_2 represents the rate of diffusion of knowledge, allowing the possibility that it takes time for new innovations to be recognized. Both exponential terms naturally depend (directly or indirectly) upon the time elapsed between granting of the cited and citing patents. In addition, there is ample evidence of “citation inflation” over time (Johnson and Popp, 2003), a fact which time-based dummy variables will capture.

We include six control variables (α parameters) and one geographic (δ) variable:

- whether or not patents k and K have the same inventor (α_{SV}),
- whether or not patents k and K have the same assignee (α_{SA}),
- whether or not patents k and K have the same examiner (α_{SE}),

- whether or not patents k and K have the same technology class (α_{SI}), pendency lag of cited patent k (α_{LAG}),
- year T of citing patent K, to account for citation inflation (α_T), and
- distance between state origins of patents k and K (δ_D).

While most of the variables are binary (including the year T, which is incorporated as a series of dummy year variables T_i), there are two notable exceptions. First, pendency lag is a continuous variable so patents were grouped into four categories: short (lag <=3 years), medium (3 < lag <=5 years), long (5 < lag <=10 years) and very long (lag>10 years). Second, (k,K) patent pairs are grouped into 100-kilometer cohorts ranging from high distance (over 2,300 kilometers between state capitals of k and K) to low distance (less than 100 kilometers), for twenty-four distance groupings. Sensitivity tests performed to include ten groupings (ranges of 250 kilometers each) produced similar results. For the sake of computability, we group the data into five-year time periods for α_T (1975-79, 1980-84, 1985-89, and 1990-94), to allow for some change in the key relationships over time.

Therefore, we postulate the functional forms

$$\alpha(k,K) = \alpha_{SV}^{D_{SV}} \alpha_{SE}^{D_{SE}} \alpha_{SA}^{D_{SA}} \alpha_{SI}^{D_{SI}} \alpha_{LAG}^{D_{LAG}} \alpha_T^{D_T} \quad (2)$$

$$\delta(k,K) = \delta^D \quad (3)$$

where δ is a single parameter raised to the power D, the number of units of physical distance between citing and cited locations. Thus, a distance of 200 km is associated with an estimated coefficient based on δ squared. Estimation of a more flexible functional form of δ(k,K), allowing each distance to enjoy its own independent δ coefficient, produces similar results. Using these parameters, the probability of a patent k granted in year t being cited by a patent K, granted in year T, can be estimated as:

$$p_{k,K} = \alpha_{SV}^{D_{SV}} \alpha_{SE}^{D_{SE}} \alpha_{SA}^{D_{SA}} \alpha_{SI}^{D_{SI}} \alpha_{LAG}^{D_{LAG}} \alpha_T^{D_T} \delta^D \exp[-\beta_1(\sum_{s=t_A}^{T} P_s)][1-\exp(-\beta_2(T-t))] + \varepsilon_{k,K} \quad (4)$$

A true geographic effect of clustering would be evidenced by a δ value less than unity (i.e. a lower probability of citation when the distance is high). Since most patents are never cited, we follow the work of Jaffe and his coauthors, grouping patents into "cohorts" of potential citations. Cohorts are constructed as mutually exclusive and exhaustive subsections of all possible citations. For example, one cohort may be defined as "all patents granted in 1976 and cited in 1978 by a patent that shares the same inventor, the same assignee, a different examiner, a different technology group, and hails from a state within 100 kilometers with a highly similar technology profile".
The expected number of citations to a cohort with specific values for the independent variables, hereafter abbreviated (X; t,T), is the likelihood of a single citation times the number of potentially citing (or cited) patents:

$$E[C_{X;t,T}] = (N_{X;t})(A_{X;T})(p_{k,K}) \tag{5}$$

where C is the number of citations to the cohort of patents described by the list of attached parameters, N is the number of (all) patents in that cohort available to be cited, and A is the number of potentially citing (biotechnology) patents granted in year T. This equation can now be rewritten using (2), (3), and (4) to give us:

$$\left(\frac{C_{X;t,T}}{(N_{X;t})(A_{X;T})}\right) = \alpha_{SV}^{D_{SV}} \alpha_{SE}^{D_{SE}} \alpha_{SA}^{D_{SA}} \alpha_{SI}^{D_{SI}} \alpha_{LAG}^{D_{LAG}} \alpha_{T}^{D_{T}} \delta^{D} \exp[-\beta_1(\sum_{s=t_A}^{T} P_s)][1-\exp(-\beta_2(T-t))] + \varepsilon_X \tag{6}$$

which can be estimated by non-linear least squares as long as the error term, ε_X, is well behaved. Because the data are grouped, we weight each observation by $\sqrt{(N_{X;t})(A_{X;T})}$ to avoid heteroskedasticity issues (Greene, 1993). As a comparison group, we omit from our estimation the coefficients reflecting simultaneous citations (where T=t), short cited patent lags, citations within the same state (distance = 0), and the last time dummy ($T_{1990\text{-}94}$).

Ten billion possible patent-to-patent citations are thus summarized into 5908 group-to-group observations of citations, each weighted by the number of citations they represent. Estimated coefficients hail from the power that group characteristics have in explaining frequent citations between certain groups, compared to infrequent citations between others. In particular, since we have included time as a component of our cohort definition, we can compare the different impact that distance (or any other variable) has wielded over time. We are statistically comparing cohorts which are identical in every way except for different timeframes, with the difference in citation frequency between those cohorts attributable to the different timeframe. It is equivalent to including a time dummy variable in a regression equation of individual observations, except that in this case each observation is a citation frequency within a cohort.

4. Findings / Discussion

A summary table of the salient data characteristics is presented below. Notice that there is an average 0.001 probability of citation between patents, with a slight downward trend over time. Although each patent cites more than prior patents did, there is an even more rapidly increasing population of patents in existence, making the probability fall over time. There are 5908 cohort observations in the full dataset: 2291 with a citing patent granted in 1975-79, 1671 in 1980-84, 1324 in 1985-89, and 622 in 1990-94.
Table 2 presents cohort-based regression estimates. Notice that significance is measured with a test of the null hypothesis that a given coefficient is unity. The results are broadly consistent with the literature (e.g. Johnson and Popp, 2003). For example, our regression yields time-specific constants that have diminished since 1980.

Unsurprisingly, both decay and diffusion are strikingly faster than elsewhere in the literature. There are positive and statistically significant self-citation effects by inventors, assignees and examiners, all as predicted. We obtain the counterintuitive result that increased physical distance makes citations more likely (δ>1), a statistically significant result. The reason is explained in the remaining columns of Table 2, where we report estimates of equation (7) separately for each time period. These regressions calibrate the decay and diffusion rates to the primary estimated values to explain the apparent anomalous coefficient.

Table 1: Summary of citation cohort data

Variable	Mean	St. Dev	Min	Max
Actual citations as share of potential citations				
1975-1979	0.002	0.036	0	1
1980-1984	0.002	0.022	0	1
1985-1989	0.001	0.009	0	1
1990-1994	0.001	0.015	0	1
	Freq of 0	Freq of 1	Min	Max
Same inventor				
1975-1979	1735	556	0	1
1980-1984	1267	404	0	1
1985-1989	1054	270	0	1
1990-1994	486	136	0	1
Same assignee				
1975-1979	1688	603	0	1
1980-1984	1238	433	0	1
1985-1989	1019	305	0	1
1990-1994	493	129	0	1
Same examiner				
1975-1979	1464	827	0	1
1980-1984	1057	614	0	1
1985-1989	901	423	0	1
1990-1994	450	172	0	1

Results are similar to the primary analysis, with one key distinction. The distance coefficient δ transitions from less than unity in 1975-79, to insignificantly different than unity in 1980-84, to significantly greater than unity in 1985-89 and 1990-94. In other words, the 1970's saw physical distance as a limiting factor to citations, with an additional hundred kilometers of distance dropping the probability of citation by five percent. In the later periods, distant citations became the norm, with longer citations more common than shorter ones. The results point unquestionably to the fact that physical distance has become less of a constraint with the passage of time. Perhaps the trend is due to the nature of the knowledge being created, but we suspect that it is more due to advances in communication, which allows easier transmission of information across great distances in the era of computerization, internet,

teleconferencing and cellular communication. In short, the principles underlying the inter-firm transfer of knowledge are changing in a striking fashion, making spillovers easier and longer than ever before.

Table 2: Regression results

Variable	1975-94		1975-79		1980-84		1985-89		1990-94	
Same inventor	0.20	(24.48)**	0.44	(21.65)**	4.71 $x10^{-3}$	(26.47)**	1.02 $x10^{-2}$	(6.68)**	6.16 $x10^{-8}$	(7.45)**
Same assignee	3.10	(17.28)**	82.96	(35.06)**	1.05	(0.10)	0.11	(82.28)**	1.54	(0.82)
Same examiner	5.32	(8.84)**	0.07	(23.62)**	57.86	(39.03)**	21.78	(44.37)**	5.74	(11.75)**
Same technology	2.28 $x10^{-7}$	(138.67)**	6.02 $x10^{-9}$	(50.59)**	4.95 $x10^{-9}$	(41.62)**	-9.67 $x10^{-5}$	(71.74)**	-6.07 $x10^{-9}$	(40.07)**
Pendency lags										
Medium	0.45	(2.79)**	0.34	(70.66)**	0.66	(19.55)**	0.26	(46.41)**	0.32	(59.59)**
Long	6.99 $x10^{-2}$	(190.20)**	0.04	(16.99)**	0.06	(19.54)**	0.06	(59.36)**	---	---
Very Long	1.41 $x10^{-2}$	(497.04)**	0.01	(11.03)**	6.30 $x10^{-3}$	(65.36)**	---	---	---	---
Constant										
1975-1979	1.28	(1.58)	---	---	---	---	---	---	---	---
1980-1984	3.38	(2.13)*	---	---	---	---	---	---	---	---
1985-1989	1.62	(2.23)*	---	---	---	---	---	---	---	---
Distance	1.01	(173.87)**	0.95	(13.82)**	1.01	(0.45)	1.05	(25.75)**	1.05	(7.78)**
Decay rate^	2.34 $x10^{-2}$	(14.32)**	2.34 $x10^{-2}$	---	2.34 $x10^{-2}$	---	2.34 $x10^{-2}$	---	2.34 $x10^{-2}$	---
Diffusion rate^	7.98 $x10^{-3}$	(2.99)**	7.98 $x10^{-3}$	---	7.98 $x10^{-3}$	---	7.98 $x10^{-3}$	---	7.98 $x10^{-3}$	---
Adjusted R^2	0.13		0.04		0.15		0.24		0.16	
Observations	5908		2291		1671		1324		622	

Notes: T-statistics are in brackets. ^ Decay and diffusion rates are estimated for the first column and calibrated for the last four columns. They report a standard t-test of $\beta=0$, unlike other coefficients which report a test of α, $\delta=1$. ** indicates 99% confidence, * 95% confidence.

5. Conclusion / Implications

We are left with a striking picture of the inter-firm transfer of biotechnological knowledge. Controlling for other factors, knowledge flows diminish with physical distance, but the constraining importance of distance has been receding with time. That is, knowledge is more likely to transfer over long distances now than it was twenty years ago. Paradoxically, it may now be more likely to jump long distances (e.g. from U.S. coast to U.S. coast), than it is to travel the shorter distances to neighboring states. The policy recommendations of this paper have therefore been heard before. In an age of more intense and distance-free communication, the conduits of knowledge transmission take on a new importance. Researchers and firms have obviously benefited tremendously from the movement to electronic patent searches and filings. In fact, that trend may have partially driven our results.

Long-distance knowledge transfers are increasingly the norm in biotechnology. The policy implications of this paper may be important not only for regions of the U.S. but for less developed nations as well. As the importance of physical distance has diminished over time, innovation has become possible at a wider array of locations, potentially drawing on a wider range of raw materials (such as agricultural germplasm) and ideas. This might imply a possibility for the deliberate fostering of non-traditional locations for biotechnology, with a prerequisite of vibrant communication with the research community elsewhere.

6. References

Caballero, R.J. and A.B. Jaffe, 1993, “How High are the Giants’ Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth,” in Olivier J. Blanchard and Stanley Fischer, eds., *NBER Macroeconomics Annual* 1993, MIT Press, Cambridge, MA.

Dorfman, N.S. 1988, "Route 128: The Development of a Regional High Technology Economy". The Massachusetts Miracle: High Technology and Economic Revitalization. D. Lampe, editor. MIT Press, Cambridge.

Feldman, M.P. 1994, *The Geography of Innovation*. Kluwer Academic Publishers, Dordrecht, Netherlands.

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

Gelsing, L. 1992, "Innovation and Development of Industrial Networks". National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. B. Lundvall, editor. Pinter Publishers, London.

Greene, W.H. 1993, *Econometric Analysis*, Macmillan Publishing Company, New York.

Hall, B.H., A.B. Jaffe and M.Trajtenberg, 2001, "The NBER Patent Citations Data File:

Lessons, Insights and Methodological Tools". *NBER Working Paper* #8498.
Henderson, J.V. 1986, "Efficiency of Resource Usage and City Size". *Journal of Urban Economics*. Vol 19, p.47-70.

Jaffe, A.B., M.Trajtenberg and R.Henderson, 1993, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations". *Quarterly Journal of Economics*. Vol 108, p.577-598.
Jaffe, A.B. and M.Trajtenberg, 1996, "Flows of Knowledge From Universities and Federal Labs: Modeling the Flow of Patent Citations Over Time and Across Institutional and Geographic Boundaries", *NBER Working Paper* #5712.
Jaffe, A.B., M.Trajtenberg and M.S.Fogarty, 2000, "Knowledge Spillovers and Patent Citations: Evidence from a Survey of Inventors". *American Economic Review*. Vol 90, p.215-218.
Johnson, D.K.N. and A.Brown, 2004, "How the West Has Won: Regional and Industrial Inversion in U.S. Patent Activity". *Economic Geography*, 80(3): 241-260.
Johnson, D.K.N. and D.Popp, 2003, "Forced Out of the Closet: The Impact of the American Inventors Protection Act on the Timing of Patent Disclosure". *Rand Journal of Economics*, 34(1): 96-112.
Johnson, D.K.N. and V.Santaniello, 2000, "Biotechnology Inventions: What Can We Learn from Patents?". Agriculture and Intellectual Property Rights: Economic, Institutional, and Implementation Issues in Biotechnology. V. Santaniello et al., editors. CABI Publishing, New York.
Lundvall, B.A. 1992, "User-Producer Relationships, National Systems of Innovation and Internationalisation". National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. B. Lundvall, editor. Pinter Publishers, London.
Maurseth, P.B. and B.Verspagen,1999, "Europe: one or several systems of innovation? An analysis based on patent citations". Economic Challenge to Europe: Adapting to Innovation-Based Growth. Edward Elgar Publishing, London.
Sjoholm, F. 1996, "International Transfer of Knowledge: The Role of International Trade and Geographic Proximity". *Weltwirtschaftliches Archiv*. Vol 132, p. 97-115.
Smith, D. and R.Florida,1994, "Agglomeration and Industry Location: An Econometric Analysis of Japanese-Affiliated Manufacturing Establishments in Automotive Related Industries". *Journal of Urban Economics*. Vol 36, p. 23-41.
USPTO Technology Assessment and Forecast Program,1998, "Technology Profile Report: Patent Examining Technology Center Groups 1630-1650, Biotechnology". United States Patent and Trademark Office report.
Von Hippel, E. 1994, "Sticky Information and the Locus of Problem-Solving". *Management Science*. Vol 40, p. 429-439.

---

[i] Specifically, the definition includes U.S. Patent Classes 47/1.1-47/1.4, 47/57.6-47758, 424/9.1-424/9.2, 424/9.34-424/9.81, 424/85.1-424/94.67, 424/130.1-424/283.1, 424/520-424/583, 424/800-424/832, 435/1.1-435/7.95, 435/40.5-435/261, 435/317.1-435/975, 436/500-436/829, 514/2-514/22, 514/44, 514/783, 530/300-530/427,

530/800-530/868, 536/1.11-536/23.74, 536/25.1-536/25.2, 800, 930, 935. We exclude class PLT (plant patents) due to data limitations on these documents.