GEOGRAPHIC LOCALIZATION OF KNOWLEDGE SPILLOVERS AS EVIDENCED BY PATENT CITATIONS*

ADAM B. JAFFE
MANUEL TRAJTENBERG
REBECCA HENDERSON

We compare the geographic location of patent citations with that of the cited patents, as evidence of the extent to which knowledge spillovers are geographically localized. We find that citations to domestic patents are more likely to be domestic, and more likely to come from the same state and SMSA as the cited patents, compared with a "control frequency" reflecting the pre-existing concentration of related research activity. These effects are particularly significant at the local (SMSA) level. Localization fades over time, but only very slowly. There is no evidence that more "basic" inventions diffuse more rapidly than others.

The last decade has seen the development of a significant body of empirical research on R&D spillovers.[1] Generally speaking, this research has shown that the productivity of firms or industries is related to their R&D spending, and also to the R&D spending of other firms or other industries. In parallel, economic growth theorists have focused new attention on the role of knowledge capital in aggregate economic growth, with a prominent modeling role for knowledge spillovers (e.g., Romer [1986, 1990] and Grossman and Helpman [1991]).

We know very little, however, about where spillovers go. Is there any advantage to nearby firms, or even firms in the same country, or do spillovers waft into the ether, available for anyone around the globe to grab? The presumption that U. S. international competitiveness is affected by what goes on at federal laboratories and U. S. universities, and the belief that universities and other research centers can stimulate regional economic growth[2] are predicated on the existence of a geographic component to the

*We gratefully acknowledge support from the Ameritech Foundation, via the Ameritech Fellows program of the Center for Regional Economic Issues at Case Western Reserve University, and from the National Science Foundation through grant SES91-10516. We thank Neil Bania, Ricardo Caballero, Michael Fogarty, Zvi Griliches, Frank Lichtenberg, Francis Narin, seminar participants at NBER and Case Western Reserve University, and two anonymous referees for helpful comments. Any errors are the responsibility of the authors.

1. E.g., Jaffe [1986] and Bernstein and Nadiri [1988, 1989]. For a recent survey and evaluation of this literature, see Griliches [1991].

2. See, e.g., Minnesota Department of Trade and Economic Development [1988]; Dorfman [1988]; Feller [1989]; and Smilor, Kozmetsky, and Gibson [1988].

spillover mechanism. The existing spillover literature, however, is virtually silent on this point.[3]

In the growth literature it is typically assumed that knowledge spills over to other agents within the country, but not to other countries.[4] This implicit assumption begs the question of whether and to what extent knowledge externalities are localized. As emphasized recently by Krugman [1991], acknowledging the importance of spillovers and increasing returns requires renewed attention by economists to issues of economic geography. Krugman revives and explores the explanations given by Marshall [1920] as to why industries are concentrated in cities. Marshall identified three factors favoring geographic concentration of industries: (1) the pooling of demands for specialized labor; (2) the development of specialized intermediate goods industries; and (3) knowledge spillovers among the firms in an industry. Krugman believes that economists should focus on the first two of these, partially because he perceives that "[k]nowledge flows, by contrast, are invisible; they leave no paper trail by which they may be measured and tracked, and there is nothing to prevent the theorist from assuming anything about them that she likes" [Krugman, p. 53].

But knowledge flows do sometimes leave a paper trail, in the form of citations in patents. Because patents contain detailed geographic information about their inventors, we can examine where these trails actually lead. Subject to caveats discussed below relative to the relationship between citations and spillovers, this allows us to use citation patterns to test the extent of spillover localization. We examine citations received by patents assigned to universities, and also the citations received by a sample of domestic corporate patents. If knowledge spillovers are localized within countries, then citations of patents generated within the United States should come disproportionately from within the United States. To the extent that regional localization of spillovers is

3. Jaffe [1989] provides evidence that corporate patenting at the state level depends on university research spending, after controlling for corporate R&D. Mansfield [1991] surveyed industrial R&D employees about university research from which they benefited. He found that they most often identified major research universities, but that there was some tendency to cite local universities even if they were not the best in their field.

4. The existence of this implicit assumption was noted by Glaeser, Kallal, Scheinkman, and Shleifer [1991]: "After all, intellectual breakthroughs must cross hallways and streets more easily than oceans and continents." Grossman and Helpman [1991] consider international knowledge spillovers explicitly.

important, citations should come disproportionately from the same state or metropolitan area as the originating patent.[5]

The most difficult problem confronted by the effort to test for spillover-localization is the difficulty of separating spillovers from correlations that may be due to a pre-existing pattern of geographic concentration of technologically related activities. That is, if a large fraction of citations to Stanford patents comes from the Silicon valley, we would like to attribute this to localization of spillovers. A slightly different interpretation is that a lot of Stanford patents relate to semiconductors, and a disproportionate fraction of the people interested in semiconductors happen to be in the Silicon valley, suggesting that we would observe localization of citations even if proximity offers no advantage in receiving spillovers. Of course, the ability to receive spillovers is probably one reason for this pre-existing concentration of activity. If it were the *only* possible reason, then, under the null hypothesis of no spillover localization we should still see no localization of citations. As discussed above, however, there are other sources of agglomeration effects that could explain the geographic concentration of technologically related activities without resort to localization of knowledge spillovers. For this reason, we construct "control" samples of patents that are not citations but have the same temporal and technological distribution as the citations. We then calculate the geographic matching frequencies between the citations and originating patents, and between the controls and originating patents. Our test of localization is whether the citation matching frequency is significantly greater than the control matching frequency. Since the "control" matching frequency is, itself, likely to be partly the result of spillover-localization, we believe this to be a conservative test for the existence of localization.

The first section of the paper describes patents and patent citations, explains the construction of the control samples, and

5. Glaeser, Kallal, Scheinkman, and Shleifer [1991] characterize the "Marshall-Arrow-Romer" models as focusing on knowledge spillovers within the firms in a given industry. They examine the growth rate of industries in cities as a function of the concentration of industrial activity across cities, within-city industrial diversity, and within-city competition. They find that within-city diversity is positively associated with growth of industries in that city, while concentration of an industry within a city does not foster its growth. They interpret this contrast to mean that spillovers across industries are more important than spillovers within industries. As is discussed below, there is evidence from the R&D spillover literature to suggest that across-industry knowledge spillovers are, indeed, important. In this study, we do not consider the *industrial* identity of either generators or receivers of spillovers, though we do have some information on their *technological* similarity.

considers more carefully how citations might be used to infer spillovers. The second section presents the results of the tests of geographic localization. The following section examines whether the probability of geographic localization of any given citation can be explained by attributes of the originating or citing patents, or of relationships between them. A concluding section follows.

I. Experimental Design

A. Patents and Patent Citations[6]

A patent is a property right in the commercial use of a device.[7] For a patent to be granted, the invention must be nontrivial, meaning that it would not appear obvious to a skilled practitioner of the relevant technology, and it must be useful, meaning that it has potential commercial value. If a patent is granted,[8] a public document is created containing extensive information about the inventor, her employer, and the technological antecedents of the invention, all of which can be accessed in computerized form. Among this information are "references" or "citations." It is the patent examiner who determines what citations a patent must include. The citations serve the legal function of delimiting the scope of the property right conveyed by the patent. The granting of the patent is a legal statement that the idea embodied in the patent represents a novel and useful contribution over and above the previous state of knowledge, as represented by the citations. Thus, in principle, a citation of Patent X by Patent Y means that X represents a piece of previously existing knowledge upon which Y builds.

The examiner has several means of identifying potential citations. The applicant has a legal duty to disclose any knowledge of the prior art that she may have. In addition, the examiner is supposed to be an expert in the technological area and be able to

6. All of the data we use relate to patents granted by the U. S. patent office. About 40 percent of U. S. patents are currently granted to foreigners. Other countries, of course, also grant patents, leading to some ambiguity in the meaning of phrases like "U. S. patent" and "foreign patent." We shall use the phrase "U. S. patent" to mean a patent granted by the U. S. patent office, regardless of the residence of the inventor. We shall use the phrase "domestic patent" to refer to a patent granted (by the U. S. patent office) to an inventor residing in the United States. We shall use the phrase "foreign patent" to refer to a patent granted by the U. S. patent office to non-U. S. residents.

7. Ideas are not patentable; nor are algorithms or computer programs, though a chip with a particular program coded into it might be. The definition of a device was recently broadened to include genetically engineered organisms.

8. There is no public record of unsuccessful patent applications.

identify relevant prior art that the applicant misses or conceals. The framework for the search of the prior art is the patent classification system. Every patent is assigned to a nine-digit patent class (of which there are about 100,000) as well as an unlimited number of additional or "cross-referenced" classes. An examiner will typically begin the search of prior art using her knowledge of the relevant classes. For the purpose of identifying distinct technical areas, we utilize aggregations of subclasses to a three-digit level; at this level there are currently about 400 technical classes.[9]

For this study, we begin with two cohorts of "originating" patents, one consisting of 1975 patent applications and the other of 1980 applications. In each cohort we include all patents granted to U. S. universities and two samples of U. S. corporate patents[10] chosen to match the university patents by grant date and technological distribution. These sets of originating patents were chosen because we conjectured that the extent of geographic localization might differ depending on the nature of the originating institution. As discussed below, such differences turn out to be minor. The 1975 originating cohort contains about 950 patents that had received a total of about 4750 citations by the end of 1989. The 1980 originating cohort contains about 1450 that had received about 5200 citations by the same time.

B. Construction of "Control" Samples

The main idea of this paper is to compare the geographic location of the citations with the originating patent that they cite. But to make such a comparison meaningful, we have to consider how often we would expect them to match under some "null" hypothesis. That is, we need to compare the probability of a patent matching the originating patent by geographic area, *conditional* on its citing the originating patent, with the probability of a match *not conditioned on the existence of a citation link*. This noncitation-conditioned probability gives us a baseline or reference value against which to compare the proportions of citations that match.

9. Examples of three-digit patent classes are "Batteries, Thermoelectric and Photoelectric"; "Distillation: Apparatus"; "Robots"; seventeen distinct classes of "Organic Compounds"; and the ever-popular "Whips and Whip Apparatus."

10. The "top corporate" sample consists of patents granted to the 200 top-R&D-performing firms in the United States, as reported in S.E.C. 10-k forms and compiled by Compustat. The "other corporate" sample contains patents assigned to U. S. corporations that are not universities and not in the "top corporate" sample.

We call this baseline or reference probability the "control frequency."

Two considerations drove our choice for constructing the control frequency. First, the fraction of U. S. patents granted to foreigners has been climbing steadily during the period under study here. We do not want to conclude that citations are initially localized, but that this localization fades over time, simply because of this aggregate trend. Second, countries (and cities and states) differ in their areas of technological focus. Although such technological specialization is probably due, in part, to geographic localization of spillovers, we want to be conservative and test whether spillovers are localized *relative to what would be expected given the existing distribution of technological activity.*

To derive a control frequency that would be immune to contamination from either aggregate movements over time or localization based on the pre-existing concentration of technological activity, we went back to the patent data base and found a "control patent" to correspond to each of the citing patents. For each citing patent, we identified all patents in the *same patent class* with the *same application year* (excluding any other patents that cited the same originating patent). We then chose from that set a control patent whose grant date was as close as possible to that of the citing patent. This process yielded, for each set of citing patents, a corresponding control sample of equal size, whose distribution across time and technological areas is essentially identical to that of the citation data set. Each control patent is paired with a particular citing patent, allowing us to compare the geographic location of the control patent with that of the originating patent cited by its counterpart in the citing dataset. The frequency with which these control patents match geographically with the originating patent is an estimate of the frequency with which a randomly drawn patent that is not a citation, but has the same technological and temporal profile as the citation, matches geographically.

To put it slightly differently, when we calculate the frequency with which the citations match the geographic location of the cited patents, we are estimating the probability of geographic match for two patents, *conditional on there being a citation link and also conditional on the technological nature and timing of the citation.* When we calculate the frequency with which the "control" patents match geographically with the cited patents, we are estimating the probability of geographic match for two patents, *conditional only* *on the technological nature and timing of the citation.* If the citation match frequency is significantly higher, then that implies that citations are localized even after controlling for timing and technology.

C. Issues Relating to the Use of Citations to Infer Spillovers

With the construction of the control samples, we believe that we have designed a very clean test of the extent to which patent citations are geographically localized. Before going on, we must address the validity of drawing inferences about knowledge spill-overs from patent citations. For discussion purposes, we can classify the links that might exist between two inventions into one of three groups: spillovers accompanied by citations, citations that occur where there was no spillover, and spillovers that occur without generating a citation. Our experiment uses the first set, but clearly the other two are non-empty. The key question is whether and to what extent we expect that either of the latter two groups would be systematically more or less localized than the group we examine. Though there are a number of considerations, all difficult to quantify, we believe that on balance it is reasonable to draw inferences about spillovers from citations.

As a general consideration, it is important to keep in mind that any analogy between patent citations and academic article citations cannot be taken too far. Academics may cite a friend (or neighbor) just to be nice, since the price of doing so is infinitesimal, or even negative if a longer list of references is perceived as making the research look more thorough. An inventor who did the same in a patent application is, in effect, leaving money lying on the table: if those citations are included in the final patent the inventor has reduced the scope of her monopoly. Further, the patent examiner should not include such citations in the patent even if the inventor did put them there. Thus, it does not seem that "gratuitous" citations are a serious concern.

A deeper problem is created by "real" citations that are not spillovers. For example, suppose that a firm gets a patent on an invention and then contracts with another firm to make some part of it, or a machine necessary to make it, or any other aspect of the downstream development. It is possible that such a contractor might later get a patent on a related technology. To the extent that the flow of rents between these parties is governed by a complete contract, there could conceivably be no externality running from the original inventor to the contractor. If we now add to this hypothetical contract the assumption that such contracted development is relatively likely to be localized, we have the potential for the observed localization of citations to be greater than the true localization of knowledge spillovers.[11]

Although such "internalized spillovers" surely exist, it is likely that most citations that are not spillovers are of a different sort: citations (added by the examiner) to previous patents of which the citing inventor was unaware. Clearly, no spillover occurs in this case. Further, it seems likely that citations of this sort should not be any more geographically localized than the control patents. If many citations are in this category, it introduces "noise" into the citations as a measure of spillovers, and biases the results *away* from finding significant localization. Our a priori belief is that this category is much larger than the previous one, suggesting that spillovers are, on balance, probably more localized than citations, but readers with different beliefs should interpret our results accordingly.

Finally, there are an enormous number of spillovers with no citations, since only a small fraction of research output is ever patented. In particular, much of the results of very basic research cannot be patented. It is plausible that basic research generates the largest spillovers,[12] and also that basic research is communicated via mechanisms that are less likely to be localized, such as international journals. For this reason, it is probably appropriate to view our results as related to applied research, and to exercise care in extrapolating to the localization of spillovers from extremely basic research.

D. Geographic Assignment of Patents

The preceding discussion has presumed that the "location" of a patent is an unambiguous construct. The patent data contain the country of residence of each inventor, and the city and state of

11. As discussed below, we focus on tests of localization that exclude citing patents that are owned by the same organization as the originating patent, precisely because such "self-citations" do not represent an externality. Citations by other organizations that have an economic relationship with the original inventor could be viewed as similar to self-citations. There is, however, a significant difference: we expect that, in general, the contract between the two parties will be quite incomplete, making it more likely than not that the citing organization could capture some rents from the original invention and hence benefit from at least a partial spillover.

12. This question is analyzed in detail in Trajtenberg, Henderson, and Jaffe [1992].

residence for U. S. inventors.[13] Use of this information is complicated by the fact that patents can have multiple inventors who can live in different places. The following procedure was followed:

1. For U. S. *inventors,* city/state combinations were placed in counties using a commercially available city directory; each U. S. inventor was then assigned to an SMSA[14] based on state and county. For this purpose an additional "phantom" SMSA was created in each state, encompassing all counties in the state outside of defined SMSAs. Approximately 98 percent of inventors were successfully assigned to SMSAs.

2. Assignments of each *patent* to a country, a state, and an SMSA were then made based on pluralities of inventors. So, for example, a patent with one inventor living in Bethesda, MD, one in Alexandria, VA, and one in rural Virginia would be assigned VA for its state and Washington, DC, for its SMSA. Ties were assigned arbitrarily, except that ties between true SMSAs and phantom SMSAs were resolved for the true one and ties between United States and foreign were resolved in favor of foreign.[15]

II. Results on Extent of Localization

As a prelude to the geographic analysis, Table I and Figure I present some descriptive data about the citations and their relationship to the originating patents. Table I shows that about 80–90 percent of the 1975 patents and 70–80 percent of the 1980 patents had received at least one citation by the end of 1989, with the higher proportion in each case applying to the university patents. Mean citations received (including zeros) were four–six for 1975 and three–four for 1980, again with the higher numbers corresponding to the university patents.[16] The average lag between the originating application year and the application year of the citing

13. Published data on the geographic distribution of U. S. patents (including those cited in the popular press) are based on the location of the organization or individual to which the patent is "assigned" by the inventor (usually her employer). For analysis of localization, the location of the assignee is not a desirable datum. There is an ambiguity in such data relating to the way employees of multinational corporations make their assignments. An employee of "Honda" in the United States could assign her patent to "Honda U.S.A., Inc." or she could assign it to the parent company in Japan. In the former case, the patent office would call it a domestic patent, and in the latter case a foreign patent; similarly for IBM Switzerland.

14. These assignments were made based on the 1981 SMSA definitions. In areas where Consolidated Metropolitan Statistical Areas were defined in 1981, these were used; elsewhere Metropolitan Statistical Areas were used. Hence we use the generic term "SMSA."

15. At the country level, 98 percent of patents were assigned unanimously. At the state level, 90 percent were assigned unanimously; an additional 4 percent had more than half of inventors in a single state. At the SMSA level, 86 percent were assigned unanimously, and an additional 6 percent had a clear majority. Overall, 4.5 percent of patents were assigned to "phantom" SMSAs.

16. Our companion paper [Trajtenberg, Henderson, and Jaffe, 1992] explores in detail the use of citation intensity and related measures for measuring the basicness and appropriability of inventions.

TABLE I
DESCRIPTIVE STATISTICS

Originating dataset	Percent receiving citations	Total no. of citations	Mean citations received	Average citation lag[a,b]	Percent self-citations[b]	Percent same patent class[b,c]
1975						
University	88.6	1933	6.12	6.53	5.6	54.3
Top corporate	84.2	1476	4.70	7.17	18.6	55.7
Other corporate	82.3	1341	4.22	7.82	9.1	57.5
1980						
University	79.9	2093	4.34	4.36	8.9	56.3
Top corporate	79.9	1701	3.54	4.41	24.6	58.3
Other corporate	74.1	1424	2.95	4.46	12.6	57.2

a. Application year of citing patent minus application year of originating patent.
b. For those patents receiving any citations.
c. Comparison is at the three-digit level (see text).

patent is 6.5 to 8 years for the 1975 cohort, and a little over 4 years for the 1980 cohort.

The inference that a citation indicates a possible knowledge spillover is much less clear in the case where the citing patent is owned by the same organization as the originating patent. For this reason, we distinguish what we call "self-citations." A self-citation is defined as a citing patent assigned by its inventors to the same party as the originating patent, which is, by construction, either a university or a domestic corporation. Not surprisingly, the self-citation rate differs for the different sources of originating patents, with universities having the lowest and top corporations the highest rates.[17] Finally, Table I shows that 55 to 60 percent of citations have a primary patent class that is the same as the primary patent class of the originating patent, indicating that the originating and citing patents are *technologically* close to one another.

Figure I provides additional detail on the distribution of lags between originating and citing patents, again defined as the difference in application years. The figure shows that citations are few in the early years,[18] and reach a plateau after about three years.

17. The apparent increase in self-citation rates between 1975 and 1980 is probably spurious; self-citations tend to come earlier than other citations. See Trajtenberg, Henderson, and Jaffe [1992] for more on this issue.

18. Patents are typically granted one to three years after application; thus, a citation lag of zero or one implies that the citing patent may well have been applied for before the originating patent was actually granted. Pending applications are not public, so in this case the citation would almost surely have been identified by the examiner.

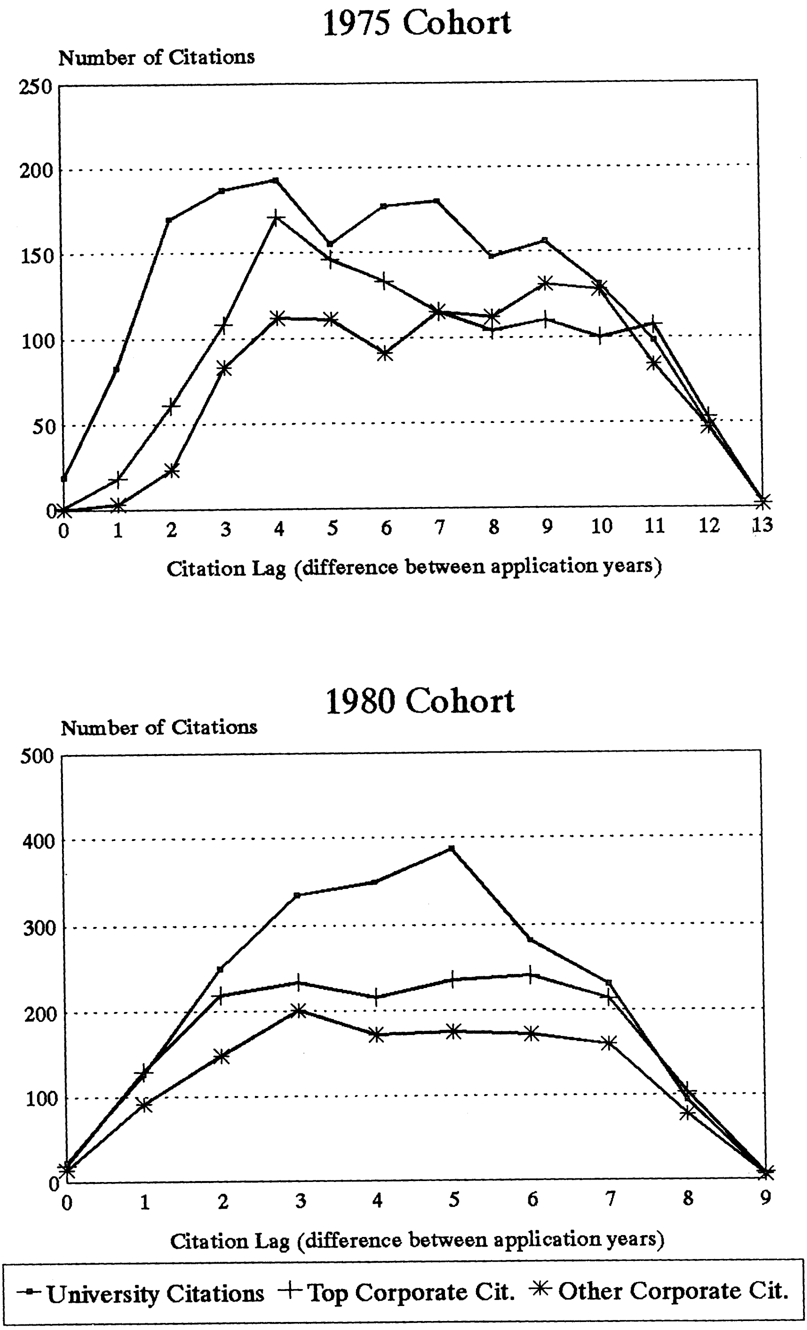
1975 Cohort
Number of Citations
250
200
150
100
50
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Citation Lag (difference between application years)
1980 Cohort
Number of Citations
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9
Citation Lag (difference between application years)
University Citations
Top Corporate Cit.
Other Corporate Cit.

FIGURE I

TABLE II
SMSA DISTRIBUTIONS FOR SOME DATASETS

Location	1975 University originating	1975 Top corporate originating	Citations to 1975 university	Citations to 1975 top corporate	Controls for citations to 1975 university	All citations to patents from in NY SMSA
Foreign	—	—	31.8	31.4	35.8	31.2
Boston	15.0	3.1	7.5	4.6	5.1	4.0
Los Angeles/ Anaheim	7.0	4.8	9.0	5.7	6.1	3.9
San Francisco/ Oakland	5.1	1.4	3.8	3.7	6.1	3.5
Madison, WI	4.2	—	1.6	—	0.5	0.6
Philadelphia/ Wilmington	4.2	9.3	5.4	8.2	4.5	9.1
Rural Iowa	3.8	—	1.6	0.6	0.2	—
San Jose	3.5	2.8	4.0	3.4	—	1.9
New York/ NJ/CT	3.2	13.5	9.7	11.7	13.7	28.5
Salt Lake City	3.2	—	2.1	—	0.5	0.4
Detroit/ Ann Arbor	2.6	2.4	2.6	1.7	1.7	1.2
Minneapolis/ St. Paul	1.3	5.2	2.8	2.9	1.9	2.1
Chicago	1.9	4.2	3.9	5.7	5.6	4.2
Albany	0.6	3.1	1.9	2.1	1.3	0.8

All figures are percentages. SMSA percentages for citations and controls are relative to *domestic* total.

It is not possible to tell for sure from these data when (if ever) that plateau tails off; the apparent tail-off in both panels of the figure is due at least in part to the 1989 observational cutoff.[19] For 1975 the higher citation rate for university patents is particularly pronounced in the early years; this pattern is not apparent in the 1980 cohort.

Before getting to our formal test of localization, an examination of Table II is useful to get a sense for the extent of geographic concentration in these data. It shows the fraction of patents coming from abroad and from a selection of major U. S. SMSAs for

19. The dropoff in both panels corresponds approximately to application year 1987 (1975 + 12 and 1980 + 7). Typically, a significant fraction of applications have not been granted within two years, so when we looked in 1989 this fraction of 1987 applications were not yet granted.

several of the datasets. Not surprisingly, a measurable fraction of university patents comes from Madison, WI; this is not true for corporate patents. A measurable (though smaller) fraction of the *citations* of university patents comes from Madison, and this fraction is larger than that for the controls. Indeed, the controls for the university citations look generally "more like" corporate patents than do the citations, suggesting that localization may be present. Other qualitative evidence of localization is apparent in the table, including the high percentage of NY SMSA citations that come from the NY SMSA.

The basic test of localization is presented in Table III. For each geographic area and each originating dataset, it presents the proportion of citations that geographically matched the originating patent. These proportions are shown both with and without self-citations. The matching proportions for the control samples are then shown, as well as a t-statistic testing the equality of the control proportions and the citation proportions (excluding self-citations).[20]

We focus first on the 1975 results on the left of the table. Starting with the country match, we find that citations *including self-citations* are domestic about 6 or 7 percent more often than the controls. Excluding self-citations eliminates this difference for the top corporate citations and cuts it roughly in half for the others. The remaining difference between the citations excluding self-cites and the controls is only marginally significant statistically.

Looking at the 1975 results for states, we find that citations of university patents come from the same state about 10 percent of the time; this rises to 15 percent for other corporate and 19 percent for top corporate. Excluding self-citations, however, makes a big difference. The university and top corporate proportions are cut to 6–7 percent, and the other corporate to just over 10. For the university and other corporate cohorts, the matching frequencies excluding self-citations are significantly greater than the matching control proportions.

20. Let p_c be the probability that a citation comes from the same geographic unit as the originating patent; let p_o be the corresponding probability for a randomly drawn patent in the same patent class (control). We test $H_o{:}p_c = p_o$ versus $H_a{:}p_c > p_o$ using the test statistic:

$$t = \frac{\hat{p}_c - \hat{p}_o}{\sqrt{[\hat{p}_c(1 - \hat{p}_c) + \hat{p}_o(1 - \hat{p}_o)]/n}},$$

where $\hat{p}_c$ and $\hat{p}_o$ are the sample proportion estimates of p_c and p_o. This statistic tests for the difference between two independently drawn binomial proportions; it is distributed as t.

TABLE III
GEOGRAPHIC MATCHING FRACTIONS

	1975 Originating cohort			1980 Originating cohort		
	University	Top corporate	Other corporate	University	Top corporate	Other corporate
Number of citations	1759	1235	1050	2046	1614	1210
		Matching by country				
Overall citation matching percentage	68.3	68.7	71.7	71.4	74.6	73.0
Citations excluding self-cites	66.5	62.9	69.5	69.3	68.9	70.4
Controls	62.8	63.1	66.3	58.5	60.0	59.6
t-statistic	2.28	−0.1	1.61	7.24	5.31	5.59
		Matching by state				
Overall citation matching percentage	10.4	18.9	15.4	16.3	27.3	18.4
Citations excluding self-cites	6.0	6.8	10.7	10.5	13.6	11.3
Controls	2.9	6.8	6.4	4.1	7.0	5.2
t-statistic	4.55	0.09	3.50	7.90	6.28	5.51
		Matching by SMSA				
Overall citation matching percentage	8.6	16.9	13.3	12.6	21.9	14.3
Citations excluding self-cites	4.3	4.5	8.7	6.9	8.8	7.0
Controls	1.0	1.3	1.2	1.1	3.6	2.3
t-statistic	6.43	4.80	8.24	9.57	6.28	5.52

Number of citations is less than in Table I because of missing geographic data for some patents. The t-statistic tests equality of the citation proportion excluding self-cites and the control proportion. See text for details.

At the SMSA level, 9 to 17 percent of total citations are localized. This again drops significantly when self-citations are excluded, but 4.3 percent of university citations, 4.5 percent of top corporate citations, and 8.7 percent of other corporate citations are localized excluding self-cites. This compares with control matching proportions of about 1 percent, and these differences are highly significant.

The results for citations of 1980 patents (right side of Table III) are even stronger and more significant. For every dataset, for every geographic level, the citations are quantitatively and statistically significantly more localized than the controls. The general increase in the proportion of U. S. patents taken by foreigners is reflected in a decline of 3 to 6 percent in the control percentages matching by country. The citation matching percentages actually rise, however, particularly for top corporate citations. It is impossible to tell from this comparison whether this represents a real change, or whether it is the result of the 1980 citations having shorter average citation lags. Since this gets to the issue of explaining which citations are localized, we postpone discussion until the next section.

Before moving on, the results on the extent of localization can be summarized as follows. For citations observed by 1989 of 1980 patents, there is a clear pattern of localization at the country, state, and SMSA levels. Citations are five to ten times as likely to come from the same SMSA as control patents; two to six times as likely excluding self-citations. They are three to four times as likely to come from the same state as the originating patent; roughly twice as likely excluding self-cites. Whereas about 60 percent of control patents are domestic, 70 to 75 percent of citations and 69 to 70 percent of citations excluding self-cites are domestic. Once self-cites are excluded, universities and firms have about the same domestic citation fraction; at the state and SMSA level there is weak evidence that university citations are less localized. For citations of 1975 patents, the same pattern, but weaker, emerges for citations of university and other corporate patents. For top corporate there is no evidence of localization at the state or country levels, though the SMSA fraction is significantly localized. Thus, we find significant evidence that citations are even more localized than one would expect based on the pre-existing concentration of technological activity, particularly in the early years after the originating patent.

III. Factors Affecting the Probability of Localization

The contrast between the 1975 and 1980 results suggests that localization of early citations is more likely than localization of later ones. This accords with intuition, since whatever advantages are created by geographic proximity for learning about the work of others should fade as the work is used and disseminated. Another hypothesis that is implicit in the previous discussion is that citations that represent research that is technologically similar to the originating research are more likely to be localized, because the individuals pursuing these related research lines may be localized. In addition, attributes of the originating invention or the institution that produced it may affect the probability that its spillovers are localized.

To explore these issues, we pooled the citations (excluding self-cites) to university and corporate patents for each cohort, and ran a probit estimation with geographic match/no match between the originating and citing patents as the dependent variable. As independent variables we included the log of the citation lag (set to zero for lags of zero), dummy variables for top corporate and other corporate originating patents, interactions of the lag and these dummies, and a dummy variable equal to unity if the citation has the same primary class as the originating patent. We also included a dummy variable that is unity if the *control* patent corresponding to this citation matches geographically with the originating patent, to control for the general increase over time in the fraction of U. S. patents granted to foreigners.

We also included two variables relating to the originating patent suggested by our work on basicness and appropriability of inventions [Trajtenberg, Henderson, and Jaffe, 1992]. The first, "generality" is one minus the Herfindahl index across patent classes of the citations received.[21] It attempts to capture the extent to which the technological "children" of an originating patent are diverse in terms of their own *technological* location. Thus, an originating patent with generality approaching one has citations that are very widely dispersed across patent classes; generality of zero corresponds to all citations in a single class. We argue elsewhere that generality is one aspect of the "basicness" of an invention. One might hypothesize that basic research results are less likely to be localized, because their spread is more likely to be through communication mechanisms (e.g., journals) that are not localized. The other variable characterizing the originating invention is the fraction of the originating patent's citations that were

21. Let C_{ik} be the number of citations received by patent i from subsequent patents whose primary patent class is k, and let C_i be the total number of citations received by patent i. The measure of generality is then

$$G_i = 1 - \sum_k \left(\frac{C_{ik}}{C_i}\right)^2.$$

TABLE IV
GEOGRAPHIC PROBIT RESULTS

	Country match		State match		SMSA match	
	1975	1980	1975	1980	1975	1980
Dummy for control	0.139	0.085	0.396	0.300	*	0.283
Sample match	(0.045)	(0.041)	(0.124)	(0.102)		(0.172)
Log of citation	−0.078	0.094	−0.264	0.198	−0.123	0.037
lag	(0.049)	(0.056)	(0.073)	(0.079)	(0.057)	(0.086)
Dummy for top	−0.114	−0.010	−0.383	0.013	−0.234	−0.208
corporate	(0.168)	(0.127)	(0.249)	(0.177)	(0.288)	(0.200)
Dummy for other	0.069	0.053	−0.214	−0.007	0.325	−0.042
corporate	(0.209)	(0.134)	(0.277)	(0.189)	(0.291)	(0.207)
Log-lag	0.046	−0.016	0.226	0.007	0.102	0.156
*top corp. dummy	(0.091)	(0.086)	(0.138)	(0.115)	(0.156)	(0.131)
Log-lag	0.008	−0.026	0.307	0.036	0.037	0.039
*other corp. dummy	(0.108)	(0.091)	(0.147)	(0.124)	(0.155)	(0.138)
Dummy for matching	−0.085	0.069	−0.013	0.034	−0.057	−0.016
patent class	(0.050)	(0.045)	(0.073)	(0.058)	(0.080)	(0.068)
Generality of origin	0.092	0.177	0.026	−0.140	0.013	−0.298
patent	(0.091)	(0.088)	(0.136)	(0.111)	(0.150)	(0.130)
Origin fraction	−0.813	0.162	0.815	0.883	1.174	0.828
Self-citations	(0.180)	(0.124)	(0.246)	(0.134)	(0.237)	(0.154)
# of observations	3581	4217	3573	4215	3566	3972
# of matches	2363	2925	256	490	197	298
Log likelihood	−2269	−2559	−894	−1459	−736	−1022

*The number of observations for which the control patent matched at the SMSA level was so small that this parameter could not be estimated.
Standard errors are in parentheses. All equations also included five technological field dummies.

self-cites. We take a high proportion of self-cites as evidence of relatively successful efforts by the original inventor to appropriate the invention. We expect that the nonself-citations to such a patent are more likely to be confined to suppliers, customers, or other firms that the inventing firm has a relationship with, and may therefore tend to be localized.

Finally, the extent of localization depends fundamentally on the mechanisms by which information flows, and these mechanisms may be different in different technical fields. For this reason, we also included dummy variables for broad technological fields.[22]

The results are presented in Table IV. Because of the presence

22. (1) Drugs and Medical Technology; (2) Chemicals and Chemical Processes Excluding Drugs; (3) Electronics, Optics, and Nuclear Technologies; (4) Mechanical Arts; and (5) All Other.

of the interaction terms between the lag and the corporate dummies, the coefficient on the lag itself corresponds to the fading of localization of citations of university patents. There is evidence in the 1975 results of such fading. This effect is statistically significant at the state and SMSA levels; its quantitative significance is discussed further below. For the citations of corporate patents, the interaction terms measure the difference between their fading rates and those of university citations. These terms are generally not statistically significant. In only one case (other corporate, 1975) could we reject the hypothesis of equality of fading rates at traditional confidence levels. There is, however, weak evidence that the corporate citations do not fade as rapidly as those of university patents, at least at the state and SMSA levels. The coefficients on the corporate dummies themselves capture differences in the predicted probability of localization for citations with lags of zero or one year. These are all insignificant, and there is no clear pattern.

The matching-patent-class and generality variables do not work well. The effects are generally insignificant, and show no consistent pattern. The effect of the self-citation fraction, however, is strong and puzzling. At the state and local level, there is a very significant effect in the predicted direction: citations of patents with a high self-citation fraction are more likely to be localized. This is *not* just saying that self-citations are localized, since they are excluded; it is the other citations that are more localized. At the country level, at least in 1975, this effect is reversed and is significant. Taking all results together, it suggests that for patents with a lot of self-citations, the nonself-citations are more likely to be foreign, but those that are domestic are more likely to be in the same state and SMSA as the originating patent.

The 1980 results are disappointing. The coefficient on the time lag term switches sign, though it is generally insignificant. One possibility is that these citations span too short a time period to capture the lag effect well. To test this possibility, we reran the estimation in Table IV on the 1975 citations, excluding all that were granted after 1984. This analysis tells us what we would have believed about citations of 1975 patents if we had looked for them only as long as we have looked for the citations of the 1980 patents. The results (not reported here) looked "more like" the 1980 results than the original 1975 results did. In particular, the coefficient on the lag term was insignificant, and was positive at the SMSA level.

TABLE V
PREDICTED LOCALIZATION PERCENTAGES OVER TIME (BASED ON 1975 PROBIT RESULTS FOR CITATIONS OF UNIVERSITY PATENTS)

	Predicted percentage for:		
Lag	Same country	Same state	Same SMSA
0 or 1 year	67.1	9.7	4.8
5 years	65.5	6.5	4.0
10 years	64.6	5.3	3.7
25 years	63.5	4.0	3.3

Thus, it may be that the "perverse" results for the 1980 sample would go away if we had later citations to include.

A probit coefficient does not have an economically meaningful magnitude, because of the need to standardize the variance of the underlying error distribution. However, we can calculate what the coefficients imply about changes in the predicted probabilities. This is done in Table V, using the 1975 lag coefficient.[23] Table V was constructed by calculating the predicted localization probability using the results of Table IV, evaluating the citation lag at different values, and evaluating the other independent variables at the mean of the data. It shows that the estimates correspond to a reduction in the localization fraction after, for example, ten years, from 67.1 percent to 64.6 percent at the country level, 9.7 percent to 5.3 percent at the state level, and 4.8 percent to 3.7 percent at the SMSA level.

IV. DISCUSSION AND CONCLUSION

Despite the invisibility of knowledge spillovers, they do leave a paper trail in the form of citations. We find evidence that these trails, at least, are geographically localized. The results, particularly for the 1980 cohort, suggest that these effects are quite large and quite significant statistically. Because of our interest in true externalities, we have focused on citations excluding self-cites. For some purposes, however, this is probably overly conservative. From the point of view of the Regional Development Administra-

23. As discussed above, this is the point estimate of the lag coefficient for citations of university patents. The point estimates are different for the corporate originating patents, but we have not performed separate calculations for each dataset.

tor, it may not matter whether the subsequent development that flows from an invention is performed by the inventing firm, as long as it is performed in her state or city. Our results are also conservative because we attribute none of the localization present in the control samples to spillovers, despite the likelihood that spillovers are, indeed, one of the major reasons for the pre-existing concentration of research activity.

We also find evidence that geographic localization fades over time. The 1980 citations, which have shorter average citation lags, are systematically more localized than the 1975 citations. By using a probit analysis, we produced estimates of the rate of fading. These estimates seem to suggest a rate of fading that is both smaller than one would expect, and smaller than would be necessary to explain the difference between the 1975 and 1980 overall matching fractions. One possibility is that the difficulty of measuring the rate of fading is due to the "contamination" of citations by the patent examiner. As noted in footnote 18, it is particularly likely that citations with very short lags were added by the examiner. If we believe that such citations are less likely to represent spillovers and less likely to be localized, then this would tend to bias toward zero our measure of the effect of time on localization.[24]

We find less evidence of the effect of technological area on the localization process. Citations in the same class are no more likely to be localized. Overall, there is not really any evidence in these data that the probability of coming from a given geographic location conditional on patent class is different from the unconditional probability. This may be due to the arbitrary use of the "primary" patent class, to the exclusion of the "cross-referenced" classes. There is no legal difference in significance between the primary and cross-referenced classes, and in many cases the examiners do not place any significance on which class is designated primary. In future work we hope to explore whether using the full range of information contained in the cross-referenced classes provides a better technological characterization of the patents.

In this context it is worth noting that part of what is going on is probably that knowledge spillovers are not confined to closely related regions of technology space. Approximately 40 percent of

24. Attempts to test for this possibility by including a dummy variable for age 0 or 1, as well as other explorations of possible nonlinearity in the match-lag relationship, were inconclusive.

citations do not come from the same primary patent class; even at the level of the five broad "technological fields" listed in footnote 21, 12 to 25 of percent of citations are across fields. This is consistent with Jaffe [1986], which found that a significant fraction of the total "flow" of spillovers affecting firms' own research productivity comes from firms outside of the receiving firm's immediate technological neighborhood.

We find surprisingly little evidence of differences in localization between the citations of university and corporate patents. The largest difference is that corporate patents are more often self-cited, and self-cites are more often localized. The probit results do not allow rejection of the hypothesis that the initial localization rates for nonself-citations are indistinguishable for the different groups. They do provide some weak evidence that this initial localization is more likely to fade for the university patents, at least at the state and local levels.

In order to provide a true foundation for public policy and economic theorizing, we would ultimately like to be able to say more about the mechanisms of knowledge transfer, and about something resembling social rates of return at different levels of geographic aggregation. The limitations of patent and citation data make it difficult to go much farther with such questions within this research approach. Ex post, the vast majority of patents are seen to generate negligible private (and probably social) returns. In future work we plan to identify a small number of patents that are extremely highly cited. It is likely that such patents are both technologically and economically important [Trajtenberg, 1990]. Case studies of such patents and their citations could prove highly informative about both the mechanisms of knowledge transfer, and the extent to which citations do indeed correspond to externalities in an economic sense.

HARVARD UNIVERSITY
TEL-AVIV UNIVERSITY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

REFERENCES

Bernstein, J., and M. Nadiri, "Interindustry R&D Spillovers, Rates of Return, and Production in High-Tech Industries," *American Economic Review Papers and Proceedings*, LXXVIII (May 1988), 429–34.

Bernstein, J., and M. Nadiri, "Research and Development and Intra-industry Spillovers: An Empirical Application of Dynamic Duality," *Review of Economic Studies*, LVI (April 1989), 249–67.

Dorfman, N., "Route 128: The Development of a Regional High-Technology

Economy," D. Lampe, ed., *The Massachusetts Miracle: High Technology and Economic Revitalization* (Cambridge, MA: MIT Press, 1988).
Feller, Irwin, "R&D Theories and State Advanced Technology Programs," Paper prepared for the American Association for the Advancement of Science Annual Meeting, January 1989.
Glaeser, Edward, H. D. Kallal, J. A. Scheinkman, and A. Shleifer, "Growth in Cities," NBER Working Paper No. 3787, July 1991.
Griliches, Z., "The Search for R&D Spillovers," NBER Working Paper No. 3768, July 1991.
Grossman, G., and E. Helpman, *Innovation and Growth in the Global Economy* (Cambridge, MA: M.I.T. Press, 1991).
Jaffe, Adam, "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value," *American Economic Review,* LXXVI (1986), 984–1001.
——, "Real Effects of Academic Research," *American Economic Review,* LXXIX (1989), 957–70.
Krugman, P., *Geography and Trade* (Cambridge: M.I.T. Press, 1991).
Mansfield, E., "Sources and Characteristics of Academic Research Underlying Industrial Innovations," mimeo, University of Pennsylvania, 1991.
Marshall, A., *Principles of Economics* (London: Macmillan, 1920).
Minnesota Department of Trade and Economic Development, Office of Science and Technology, *State Technology Programs in the United States,* 1988.
Romer, P., "Increasing Returns and Long-Run Growth," *Journal of Political Economy,* XCIV (1986), 1002–37.
——, "Endogenous Technological Change," *Journal of Political Economy,* XCVIII (1990), S71–S102.
Smilor, R., G. Kozmetsky, and D. Gibson, *Creating the Technopolis: Linking Technology, Commercialization and Economic Development* (Cambridge, MA: Ballinger Publishing Co., 1988).
Trajtenberg, M., "A Penny for Your Quotes: Patent Citations and the Value of Innovations," *Rand Journal of Economics,* XXI (1990), 172–87.
Trajtenberg, M., R. Henderson, and A. Jaffe: "Ivory Tower Versus Corporate Lab: An Empirical Study of Basicness and Appropriability," NBER Working Paper No. 4146, 1992.