Spillover Geography Under the Microscope:  
Network-driven Transfers of Localized Knowledge  
Across Distance, Time and Political Boundaries

Robert C. Liu\(^1\)  
Department of Economics  
University of California, Los Angeles

Job Market Paper  
November 11, 2005

\(^1\)Computing and data resources for this paper were provided by the Center for International Science, Technology and Cultural Policy (CISTCP) of the UCLA School of Public Affairs; CISTCP is partially funded by the National Science Foundation and the University of California. I am most grateful to Michael Darby and Lynne Zucker for their encouragement and generous support of this project, as well as to Ken Sokoloff for valuable guidance. Hongbin Cai, Leah Brooks, Mariko Sakakibara and Dan Ackerberg also contributed substantive comments. The views expressed herein, as well as errors, are mine.
Abstract

The article describes a number of operational and methodological contributions in the search for geographically localized knowledge spillovers. The unique focus is on the role of individual inventors, and to examine, at this fundamental level of innovative activity, the impact of geographic distance, time, and political boundaries on transfers of knowledge. The analyses exploit patent citations which I identify as network-driven transfers—citations for which transmission mechanisms emphasize repeat contact with prior inventors. The matching of inventor locations with geographic coordinates also makes use of information hitherto unexploited in patent data and lead to fine-scaled results. Tests for localized spillovers, proximity effects, and distance and state regressions reveal systematic patterns: Localization attenuates over time generally, but increases for intrastate citations. After controlling for geographic distance, spillovers of new knowledge are also increasingly likely to remain within state boundaries. Network-driven transfers are correlated with proximity effects which become increasingly important over time.
1 Introduction

Knowledge moves in mysterious ways. Not so, if by mysterious we mean we are uncertain that it does; we know knowledge moves because abundant economy-wide evidence attests to it. The available evidence is both direct, in the form of incremental improvements in technology, and circumstantial. Knowledge spillover, for example, is commonly cited as a driving force of industrial agglomeration. That flows of knowledge capital are an important dynamic in modern economies is also beyond dispute. But flows of knowledge are often invisible and there is considerable complexity associated with its transmission between individuals and organizations. Individual innovators understand how unique advances came to be developed, but the outsider observes in the aggregate a web of mechanisms, institutional forces and conditions which together determine how new knowledge is transmitted. The goal of this paper is to untangle, at the basic level of innovators, unresolved issues concerning the attenuation of knowledge flows across geographic distance, time, and political boundaries. Addressing these issues remains vital as knowledge capital becomes an increasingly central component of productivity and growth.

Managers and policymakers understand the potential for spillovers in decisions regarding where to locate or concentrate R&D activity. Since Jaffe, Trajtenberg and Henderson (1993, hereafter JTH) and Audretsch and Feldman (1996) there has emerged a general consensus that spillovers are localized—that is they have an observable geographic component after controlling for the concentration of industry. Measuring spillovers, however, remains challenging\(^1\) and the evidence is inconclusive as to whether localized spillovers are important. Orlando (2004) recently finds that the R&D of geographically proximate firms has no im-

---

\(^1\)See Griliches 1992 for a survey of empirical issues.
pact on firm productivity within narrow industrial classifications. Thompson and Fox-Kean (2005a) address directly the JTH methodology and reverse their original findings of localization at the geographic levels of city and state.\textsuperscript{2} One contribution of the current paper is a demonstration of operational improvements which allow us to drop the typical discrete definitions of distance and to observe spillovers at many levels of geography, indeed from the immediate vicinity of originating innovations up to the country level, and any level in between. A look at localization effects at high geographic resolutions provides some context for earlier findings—the question becomes not whether spillovers effects exist (they do) but where, and perhaps when, the researcher should expect to observe them.

More importantly, I demonstrate key methodological advances which allow the detection of localization effects at a more fundamental level of innovative activity. When we say that subsequent innovations are geographically localized with respect to originating innovations, we infer that this occurs because individual inventors enjoy the benefits of proximity to or contact with originating inventors.\textsuperscript{3} To observe this phenomenon I use patent citations, a popular data source in which to observe spillovers from prior to subsequent innovations. A unique aspect of my use of citations, which are typically undifferentiated in empirical research, is the demonstration of a class of citations which I term \textit{network-driven transfers}. These are a subset of citations which have been linked to outside evidence of person-to-person contact between the citing and originating inventors. The data show that observed spillover effects are driven by particular types of knowledge flow. Network-driven transfers suggest an important productivity-enhancing aspect to inventor relationships—which appear sometimes as collaboration, other times as repeat citation—and reveal a social context.\textsuperscript{4}

\textsuperscript{2}I discuss JTH methodological issues in more detail later.

\textsuperscript{3}For these reasons I generally refer to patentees as “inventors” in the article. The former are persons who have filed applications ending in the successful grant of legal privileges, whereas the latter term reminds us of the crucial innovative role these individuals play.
to innovative activity.

The person-level approach leads to a more subtle test of localization which exploits location information at the level of individual inventors rather than the firm or patent as the unit of observation. I believe this distinction between firms and inventors to be important. Intuition based on one or the other as the point of reference can lead to very different predictions about how knowledge flows. Firms, for example, generally do not share knowledge which might be of value with competitors, but individuals may or may not be similarly motivated. Incentives and costs differ along other dimensions as well, but to my knowledge the underlying role of inventors in innovation’s black box has not been explored on a large scale. In terms of patents, I find that geographic distance does matter for both innovations and inventors. Localization effects attenuate with distance and with time in expected ways, but localization increases over time for intrastate citations. Political boundaries matter—spillovers are more likely to remain in-state over time after controlling for geographic distance. Network-driven transfers are correlated with inventor proximity effects which become increasingly important with time.

The paper is organized in the following manner: In Section 2, I provide a theoretical framework for empirical predictions and describe tests of localized spillover and proximity effects at the inventor level. Section 3 describes the data, including operational refinements by which to extract location information hitherto unexploited in patent data, and the procedure for identifying network-driven transfers in patent citations. I follow with a discussion of results in Section 4, including results from regressions of geographic distance and state boundaries over time. Section 5 concludes.
2 Observing Spillovers and Localized Knowledge

Theoretical Framework

What forces contribute to agglomeration and localization? Though commonly used together to describe geographic effects, the two terms are not interchangeable. The first refers to the clustering of industry or production, an entry effect. Localization, on the other hand, is associated with the accumulation of spillovers in a limited area and, thus, more of an output or productivity effect. Different starting points determine how the appropriate questions are framed: Asking why apple trees grow in orchards is different from asking why apples don’t fall far from the tree. To be sure each effect is endogenous to the other—more semiconductor firms in Silicon Valley leads to more learning and cumulative advances; more semiconductor innovations in one place leads more firms to locate to Silicon Valley, and so on. Other recognized causes of agglomeration include the pooling of factors like specialized labor or development of intermediate input markets. Likewise, localized spillovers depend not just on industry concentration but also the types of organizations involved, the nature of knowledge, and the policies governing their transfer. Apple trees cluster because that’s where apples fall but also because they are cheaper to care for that way, or the soil and climate are better. But why do apples not fall up or sideways, but instead fall down? The pull of the earth matters for apples—does geographic proximity to other innovators matter for spillovers?

As mentioned above, previous studies have produced mixed results on the importance of distance for spillovers. I am aware of none which examine this issue at the level of individual inventors; indeed if proximity effects exist at all, they exist at this basic level and,
hence, our choice of the unit of observation. The understanding that spillovers—and knowledge transfers generally, either compensated or uncompensated—are at a fundamental level composed of discrete flows of knowledge between individuals has an important exposition in the studies of Zucker and Darby on the biotechnology industry. In their work, firm entry and success is largely explained by human capital measures, in particular the geographic distribution of university “star” scientists (Zucker, Darby and Brewer 1998). Their evidence also suggests that knowledge transfers from scientists to firms are not spillovers in the strict sense but mediated by market-based contractual relationships (Zucker, Darby and Armstrong 1998). The picture which emerges is that of firms benefiting from localized flows of knowledge which themselves are composed of repeated interactions between firm scientists and university scientists. If proximity is important for these and other types of transfers, then successively nearer inventors would enjoy an increasingly greater benefit to their productivity.

Patent citation data are well-suited for examination of these effects. The individual-based approach of the current study precludes methodologies which depend on firm-based measures of knowledge capital such as R&D expenditures. A disadvantage with patents, however, is that not all granted patents are economically important. Patents are filed early in the inventive process, before the full value of innovations are known, but this also means they may more closely reflect the productive activity of individual inventors. Another potential downside of patents is that not all patent citations reflect actual transfers of knowledge between the cited and citing parties. Many citations have in fact been added later in an institutional filing and examination process in which inventors often play a limited role. In other work, I identify network-driven transfers or citations verified by outside evidence, from
patents and academic articles, indicating repeat contact between inventors (Liu, Zucker and Darby 2004). Such data diminish the influence of non-inventor citations and, more importantly, add predictive power in my person-based framework by acknowledging the impact of inventor pairs most likely to represent transfers of significant new knowledge. Nevertheless, if evidence exists of a geographic component in spillovers, we should expect to find it in patent citations.

The key assumption is that spillover localization and inventor proximity effects are important if the effects attenuate over distance. Attenuation can be understood, at either the innovation or inventor level, in terms of a cost to the transmission of knowledge with distance. Alternatively, such effects can be driven by the expectation of returns to innovative activity (Lamoreaux and Sokoloff 2000). Either approach would predict decreasing effects with distance or increasing localization effects at smaller distances. Political boundaries introduce a discontinuity along the distance dimension, as learning across state boundaries can incur a cost or incentives can enhance intrastate spillover. Actual spillovers likely involve both political boundaries and geographic distance.

Over time, conventional wisdom would suggest that diffusion and improvements in communications would lead to attenuation in both localization and proximity effects—the fading of localization in the JTH results is well-known. However, increasing technological complexity may make proximity to other inventors more important over time. A more sophisticated form of this argument is that much complex knowledge is tacit and embodied in individuals (Zucker, Darby and Armstrong). An increase in clustering over time, attributable either to spillovers or other agglomerative forces (Porter 2000), would also drive temporal effects in the other direction.
Tests of Localization

The framework requires a methodological design which tests at the inventor level and at the same time untangles localization from agglomerative effects. The former requirement rules out industry- (Audretsch and Feldman) and firm-based approaches (Hall and Mairesse 1995 and Orlando). The innovation-based methodology of JTH is the point of departure for my first set of tests, which extend the methodology to work at the inventor level.\textsuperscript{4} Their main contribution involves the selection of controls which capture the the spillover effect due to agglomeration alone. Using patent citation data, the procedure involves determining the rate at which a group of citing patents are collocated with their associated originating patents, either at the city, state or country level. The colocation rate is then determined in a similar fashion for a control sample selected—one control patent for each citing patent—based on the application year and patent class of citing patents. The matching rates for citing and control samples are then tested for equality by computing a \( t \)-statistic for the difference of two independent Bernoulli random variables. A significantly greater colocation rate for citing patents means that subsequent innovations tend to appear in the same geographic area, to a degree greater than the geographic distribution of industry or production would indicate. That the test statistic is constructed from easily computed sample statistics adds to its elegance, and the control group has a ready interpretation as the effect due to industry concentration.\textsuperscript{5} Note, however, that only the distribution of patent classes need be the same between citing and control samples. Control patents used in this way do not

\textsuperscript{4} JTH make use of inventor data but the analysis is performed at the level of the patent.

\textsuperscript{5} The interpretation of control patents need not be limited to effects arising from the distribution of industry or production. Lamoreaux and Sokoloff show that the assumption of a tight link between the location of innovation and production is sometimes incorrect. Nonetheless, a control patent is a random industry innovation whose location can be determined by a number a factors—spillovers, if localized, are localized after accounting for all of these factors.
“control” at the level of the observation. Results based on this approach serve to motivate and complement more systematic analyses that follow.

The new test is motivated by my interest in inventor behavior, but the key element which makes it possible is geographic distance. Why should one bother with alternatives to the typical city and state geographic units? First, there is considerable heterogeneity in the size and distribution of population concentrations across metropolitan areas and states. Though relevant for policy, states encompass a particularly wide range in terms of both population and industries represented. More importantly, for local transfers of knowledge, the geographic consideration for potential citing inventors (when such a consideration exists) will more often be geographic distance rather than the political boundary of a city. Political boundaries, however, likely have some impact; I examine the role of state boundaries in the analyses.

For the following tests of localization, rather than asking whether subsequent innovations are colocated with originating innovations, ask whether the inventors responsible for innovations are colocated with the originating inventors. A knowledge spillover can be understood to occur at either level, though a spillover leading to a new patent also implies a transfer at the inventor level. Patents typically have more than one inventor; the mean is approximately four inventors per patent over the last 30 years. A single patent citation thus involves multiple individuals across two patents. For $P$ originating inventors and $Q$ citing inventors, $PQ$ observations per citation can be examined for potential geographic proximity. If knowledge flows between individuals as we believe, then the distribution of these distances between patents linked by citation should be systematically different from distances computed between random patents not so linked.
To see why this might be so, consider the nearest two inventors as defined by the *near distance*, or the minimum in the set of $PQ$ distances, against the more geographically separated inventors defined by the *middle distance*, or median value in the set of $PQ$. Since, for a random control patent to which no knowledge has been transferred, the middle distance will be greater than near distance but we would not expect a pattern to their difference. On the other hand, in the case of a patent linked by citation to an originating patent, we would expect the difference between near and middle distances to change with distance. Being, say, 5 miles closer should not matter for two citing inventors who are both hundreds of miles from the source, but being 5 miles closer may well matter for two potential inventors located within the same metropolitan area as the source. The potential of proximity to enhance the probability of citation, and inventor productivity, should be observable at successively smaller distances. The effect should be nonexistent for random innovations which have nothing to do with originating patents other than occurring in the same industry and year as citing patents.

Now define the four Bernoulli random variables $X^m_1, X^n_1, X^m_0$ and $X^n_0$ with the respective parameters $p^m_1, p^n_1, p^m_0$ and $p^n_0$. At a fixed level of geography, $X^m_1$ is unity if the citing patent is colocated with the originating patent according to their *middle distance* as defined above. The two patents “match,” for example, at the 50-mile level if their inventors are located less than 50 miles apart. $X^n_1$ is defined similarly based on the *near distance* between inventors connected by citation. At any given level of geography $X^m_1 \leq X^n_1$ by construction. $X^m_0$ and $X^n_0$ indicate a geographic match in the same manner for inventors in a control patent with respect to an originating patent.
Now consider the functions of the sample means

\[ T^m_1 = \bar{X}^m_1 - \bar{X}^m_0 \quad \text{and} \quad T^n_1 = \bar{X}^n_1 - \bar{X}^n_0 \]

This is analogous to what is tested in JTH at the patent level, but with values determined using middle and near distance measures instead of by matching actual location observations to cities. The researcher can thus define geographic units of any size with the \( \bar{X} \)'s and \( T \)'s varying accordingly. Using distance in this way improves the resolution of results, but these remain tests of localized innovations; each test attributes one distance to citing patent and one to control patent and compares their distribution. An independence assumption gives the one-sided statistic

\[ t^i_1 = \frac{\hat{p}^i_1 - \hat{p}^i_0}{\sqrt{[\hat{p}^i_1(1 - \hat{p}^i_1) + \hat{p}^i_0(1 - \hat{p}^i_0)]/n}} \]  

which tests \( H_0 : p^i_0 \leq p^i_1 \) where \( i = m \) or \( n \) and the \( \hat{p} \)'s are the sample means. The test statistic is \( t \)-distributed.

I exploit the distance information in another way. Recall that, for each citing-originating patent pair, I observe all distances between inventors connected by citation. Over many patents this produces multiple observations on the distribution of citing distances which we presume to differ from the distribution of control distances. The distributions should be different if knowledge flows between individuals and if knowledge transmission attenuates with distance. Using middle and near distances together is one way to detect these more subtle effects, and others can be imagined.
Consider now

\[ T_2 = (\bar{X}_1^n - \bar{X}_0^n) - (\bar{X}_m^m - \bar{X}_0^m) = (\bar{X}_1^n - \bar{X}_1^m) - (\bar{X}_0^n - \bar{X}_0^m) \]

The first expression is equivalent to \( T_1^n - T_1^m \) and describes a test for an increase in localized innovations due to near inventors versus middle inventors. The second expression can be interpreted as that of testing for the differential impact of near inventors in citing patents versus control patents. Put another way, the statistic tests the null that there is no greater geographic “spread” between near and middle inventors in citations than in controls.

Any changes in the dispersion of co-inventors over time or across industries is captured in the control sample. Finding that near and middle inventors are not interchangeable for localization in this manner means that proximity matters for these individuals.

Before constructing the test statistic, however, we should be concerned about making an independence assumption across the expanded set of variables. Note the asymptotic variance of the earlier \( T_1 \) is in fact

\[ \left( \sigma_{X_1}^2 + \sigma_{X_0}^2 - 2\sigma_{X_1 X_0} \right)/n \]

where the last covariance term was assumed to be zero. But this is likely not the case—a citing patent is more likely to appear locally if its corresponding control patent does also. This is not a problem for the test, as assuming \( X_1 \) and \( X_0 \) are not positively correlated leads to larger sample variances in the test statistic and a more conservative test.

The asymptotic variance of \( T_2 \) includes the covariance terms

\[ 2(\sigma_{X_1^n X_0^n} - \sigma_{X_1^n X_1^n}) + 2(\sigma_{X_1^n X_0^m} - \sigma_{X_0^m X_0^m}) - 2\sigma_{X_1^n X_0^m} - 2\sigma_{X_1^n X_1^m} \]
which are arranged here to illustrate that, as with $T_1$, an independence (or uncorrelatedness) assumption produces a more conservative test. In the first group of terms, it is clear that $X_1^m$ will be more positively correlated with $X_1^n$ than with $X_0^n$ in the typical sample. Similarly, $X_0^m$ will be more positively correlated with $X_0^n$ than with $X_1^n$. The first two groups of terms are non-positive; the last two terms correspond to $T_1^m$ and $T_1^n$ covariance terms as above.

We can use the following simplified test statistic:

$$t_2 = \frac{(\hat{p}_1^m - \hat{p}_0^m) + (\hat{p}_1^n - \hat{p}_0^n)}{\sqrt{\hat{p}_1^m(1 - \hat{p}_1^m) + \hat{p}_1^n(1 - \hat{p}_1^n) + \hat{p}_0^m(1 - \hat{p}_0^m) + \hat{p}_0^n(1 - \hat{p}_0^n)}}/n \quad (2)$$

If a significant result in the first test (Equation 1) is evidence that innovations are localized—that is citing patents are colocated to a greater degree than random industry patents—a significant result in the second test (Equation 2) shows that the nearest inventors enjoy a disproportionate benefit from localized knowledge. I use the former to detect a localization effect and the latter to detect a proximity effect. Finding a proximity effect should not be surprising if observed spillovers actually represent many more repeated, unobserved transfers of knowledge. If geographic distance matters for such transfers this should be discernable in the data, to which I now turn.

3 Patent Data and Network-driven Transfers

Geographic Distance. The analyses require that we exchange distance measures for typical discrete definitions of inventor location. Computing distance between inventors relies on precise location information for multiple inventors in a patent. The practical problem of matching coordinates (latitude and longitude) to a large number of location observations
is no longer prohibitive given the existence of exhaustive geographical databases and tools like perl. For geographic sources I use a database of place names maintained by the U.S. Geological Survey and a dataset of USPS Zip Codes produced by the U.S. Census Bureau. Using string matching algorithms I am able to assign coordinates to approximately 98% of inventor observations in a large patent sample.\textsuperscript{6}

Computational issues involving the calculation of millions of distances are minor by comparison. For each originating-citing patent pair I compute the distance between each citing inventor and each originating inventor. For \( P \) originating inventors and \( Q \) citing inventors this produces a set of \( PQ \) distances. I use the two distance measures described earlier for the analyses in this paper—both the near distance, that is the distance between the two nearest inventors, and the middle distance, or the median value in the set of \( PQ \) distances. The same computations are performed for the corresponding control patent which leaves, for each originating-citing-control observation, four distance values. For reference, Table 1 lists typical distances between random individuals in select geographic units.

\textit{Control Patents.} Before describing the sample, a discussion of control selection is appropriate given recent interest in this issue. Thompson and Fox-Kean argue that U.S. patent classes, the basis for JTH control patents, are defined too broadly and result in controls which often have little to do with citing patents and tend not to be colocated with originating patents. This potentially underestimates the effect of agglomeration and biases toward findings of localization. They repeat the analysis using the U.S. subclass to show that disaggregated controls are colocated at higher rates. They find these rates to be not statistically different from the colocation rates of citing patents at all levels of geography

\textsuperscript{6}More details on data sources and data construction are provided in the Appendix.
Table 1: **Geographic Scale Reference.** Listed are distances for select geographic entities, for illustrative purposes. Miles are the mean distance between two random locations within the geographic unit, computed based on a uniformly-distributed disk of equal area. Definitions and area estimates are from the 2000 U.S. Census. An MSA is a Metropolitan Statistical Area; a large MSA can achieve designation as a Primary MSA (PMSA); a Consolidated MSA (CMSA) is a cluster of PMSAs.

<table>
<thead>
<tr>
<th>Miles</th>
<th>Geographic Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Mean Metropolitan Area (Central City)</td>
</tr>
<tr>
<td>24.6</td>
<td>Mean Metropolitan Area (MSA/PMSA)</td>
</tr>
<tr>
<td>48.5</td>
<td>Mean Metropolitan Area (CMSA)</td>
</tr>
<tr>
<td>40.2</td>
<td>Boston-Worcester-Lawrence CMSA</td>
</tr>
<tr>
<td>47.9</td>
<td>San Francisco-Oakland-San Jose CMSA</td>
</tr>
<tr>
<td>52.3</td>
<td>Washington-Baltimore CMSA</td>
</tr>
<tr>
<td>58.5</td>
<td>New York-Northern New Jersey-Long Island CMSA</td>
</tr>
<tr>
<td>96.0</td>
<td>Los Angeles-Riverside-Orange County CMSA</td>
</tr>
<tr>
<td>121.2</td>
<td>Median State (Iowa)</td>
</tr>
<tr>
<td>139.3</td>
<td>Mean State</td>
</tr>
</tbody>
</table>

except the country level. Both Thompson and Fox-Kean and the original authors in their comment (Henderson, Jaffe and Trajtenberg 2005) acknowledge a need for other measures of technological and industrial proximity.

To select controls I use an alternative measure of proximity available in U.S. patent data: the International Patent Classification (IPC). Lerner (1994) offers reasons for preferring IPCs in the current context. Quantitatively, he demonstrates that IPCs are correlated with value and can be used to measure patent scope or breadth. This follows from the fact that international classifications tend to be assigned by industry. The U.S. system, on the other hand, was conceived to facilitate the work of patent examiners, hence U.S. classifications more often reflect inventions’ structure or function. Though not the focus in this study, I nevertheless test six separate groups of control patents. For each citing patent, I select a random control from the same application year and patent class, where class is either the U.S. Patent Class (USPC) at two possible levels of aggregation, or the International Patent
Class (IPC) at four levels of aggregation.\textsuperscript{7}

In my sample, no control could be found for 0.19\% of citing patents at the USPC-1 level and 27\% at the USPC-2 level. The respective proportions for IPC-1 through IPC-4 were 0.30, 0.46, 1.3 and 6.4\% without an available control. Note that IPC-4 does better than USPC-2 (6.4\% versus 27\% discarded) even though they represent a comparable number of classifications (over 100,000). One reason for this is the selection algorithm which allows for each patent to belong to multiple IPCs but to only a single USPC. In other words, citing and control patents with multiple IPCs are allowed to share any one class in common. This different treatment is necessary because USPC indentifies a single primary class/subclass and separate secondary classes/subclasses, whereas multiple IPC classifications in a single patent are not distinguished in terms of their ordering.

Figures 1 and 2 depict colocation rates of citing and control patents in my sample together with results of JTH and Thompson and Fox-Kean. The final citation year and choice of controls in each case correspond with those of the other authors. Figures 3 and 4 depict colocation rates and test statistics for the localization effect for all six control groups. The level of aggregation matters—more aggregation yields controls colocated at lower levels and leads to stronger findings of localization at every level of geography. And yet the choice of controls does not matter because localization effects appear using\textit{ any} set of control patents. On the basis of above theoretical and practical considerations, I use controls selected from the IPC-4 level of aggregation (the IPC “group”) in the results to come. Other sample statistics follow.

\textsuperscript{7}These levels of aggregation (class and subclass in USPC; class, subclass, main group, and group in IPC) are built into the alphanumeric symbols. An additional top-level IPC section is not used. Further disaggregated sublevels are achieved via indentation schemes in both systems. The Lerner study uses the IPC subclass, or my IPC-2.
Sample Description. I begin with 20,000 originating patents randomly selected from those granted to U.S. inventors in 1976. These in turn were referenced by 178,435 citing patents through the end of 2004. I then identify citations that are “self-citations” by assignee, that is the same party (usually a firm or university) holds the rights to both patents. I eliminate these from the sample, not because they are not associated with spillovers—though most are neither uncompensated nor unintentional—but because citing and cited patents tend to be collocated. I additionally remove self-citations by inventor as these will similarly bias toward localization. Citing and originating patents for which there is at least one common inventor between them are eliminated. There are 14,609 self-citations in the sample; 8,827 are by assignee, 2,661 are by inventor, and 3,121 are by both assignee and inventor.

Of the remaining sample, I remove 6.4% of observations for which no IPC-4 control could be found and another 2.2% without a matched U.S. location in the originating patent. This leaves 149,726 originating-citing-control observations with each patent in an observation linked to one or more inventor locations.

Network-driven Transfers. Most empirical studies that make use of patent citations make no attempt to differentiate them—one citation is as good as another for determining which innovations are important, or where new knowledge goes. Many citations are also not economic spillovers in the sense that they are not uncompensated—representing knowledge exchanges, perhaps, rather than transfers. The cumulative nature of knowledge implies

---

8 The number of starting patents is somewhat arbitrary as the scope of the selection, cleaning, matching, and computation steps already required extensive automation. A balance was ultimately struck between runtime and the desire for sufficiently large subsamples. I also constructed samples for all originating patent years 1976–2004 in order to address questions regarding the changing nature of innovations over time, beyond the scope of the current paper.

9 See Appendix for further details.

10 One exception is Alcácer and Gittelmann (2004) on citations added by patent examiners. See also Jaffe, Trajtenberg and Fogarty (2000) on the meaning of patent citations.
the existence of mechanisms by which citing inventors learn prior knowledge. Sometimes this is limited to learning directly from what is disclosed in the cited patent. Likely to be more common are transfers involving other mechanisms, some of which also leave a paper trail. Liu, Zucker and Darby demonstrate at least four possibilities. Two types are collaborator citations or transfers involving inventors who have co-patented on other inventions or co-authored academic articles. The other two are repeat citations or those with evidence of similar inventor citations (in the same direction) or references in articles. Such knowledge transfers are network-driven because they are evocative of the social context in which innovations arise.

Network-driven transfers in the current study are those with any evidence (co-patenting, co-authoring, citing in patents, or referencing in articles) occurring within a 10-year window of the citing patent, that is 5 years before or after the year in which it was granted. This procedure separates the sample into two a priori groups of network-driven and other citations with 68,501 and 81,225 observations, respectively. The “other” group includes citations which are less likely to involve actual knowledge flow, such as citations added by lawyers or patent examiners.\textsuperscript{11}

\textsuperscript{11}I am prevented from further differentiation by present data and methodological hurdles; the theoretical issues also deserve special attention. More details on the Network-driven Transfer Data are in the Appendix.
4 Main Results

Localization and Proximity Effects

I provided a preview of the localization results in the previous section. Colocation rates for all citations and controls rise monotonically (by construction) with increasing levels of geography. Innovations are localized at all distances, that is, compared to random industry patents, spillovers accumulate locally to a greater degree no matter the scale of observation. The localization effect appears at mile one and increases with distance. The intuition here is that there is a scale to agglomeration and localization effects with the former attenuating more rapidly with distance.

The high levels of significance are due to sample size. A check using a separate, smaller (15%) sample constructed in an identical fashion produces similar colocation rates but markedly lower levels of significance. This is strong evidence that spillovers are indeed localized—the question is at what geographic scale the researcher should expect to observe the effect given their data. A finding of a proximity effect for the nearest inventors would strengthen this claim considerably. If distance is relevant to inventor productivity then it is certainly relevant to their output.

Near and Middle Inventors. I now consider the differential impact of near versus middle inventors to see if proximity matters for individuals. A positive finding of a proximity effect means that a disproportionate amount of localized spillovers accrue to inventors nearest the cited source. A select slice of the results are in Table 2. Figures 5 and 6 depict colocation.

\textsuperscript{12}The IPC-4 series in Figures 3 and 4 correspond to the main results. A different choice of controls does not alter findings.
Table 2: Results by Geographic Unit. Colocation rates and test statistics of originating patents in 1976 with citations received through grant year 2004; finer results are depicted visually in Figures 5 and 6. Shown are percent of citing ($\bar{X}_1^n$) and control ($\bar{X}_0^n$) patents colocated with originating patents; $t_1$ tests the equality of these rates for a localization effect. The $t_2$ statistic is associated with a test for a proximity effect. Observations: 149,726.

rates and test statistics computed at the fine scale of earlier figures. By construction, citations are colocated with originating patents earlier (going from smaller to larger distances) when determined with near inventors. Using near inventors, the rate is 4.9 and 3.1% for citations and controls at 25 miles (about the level of a metropolitan area or MSA), and 6.7 and 4.4% at 50 miles (the level of the largest metropolitan areas or CMSAs).

Again, innovations are localized everywhere. The test statistic for the proximity effect also exceeds thresholds for significance generally.\textsuperscript{13} The effect is highly significant below the 25-mile level with a peak at about 8 miles and decreases with distance. The statistic remains above the 5% significance threshold well into the interstate level (not depicted) and drops below it at about 750 miles. We should be careful in attributing meaning based on the level of a test statistic, but the shape of the statistic accords with intuition. Proximity to originating innovations confers more spillovers to the nearest inventors and there is (an evidently short) range beyond which the benefit diminishes quickly. The effect is strong.

\textsuperscript{13}By construction $t_2$ will never exceed $t_1$.
within this range, but there are diminishing returns to proximity as well. Being, say, 5 miles from the source is better than 10 miles, but 10 miles distant is much better than 15 miles.

Network-driven Transfers. The next set of results shows that the above effects are largely driven by identifiable components in the data. Figures 7 and 8 depict colocation rates and test statistics for the network-driven and “other” subsamples. There is a clear separation in the colocation rates and in the expected directions. At the 25-mile level, the difference between citation and control rates for all undifferentiated citations is 1.9%. Here, this difference is 3.8% for the network-driven sample and 0.9% for the other sample. At the 125-mile level the differences for undifferentiated, network-driven and other citations are 3.1, 5.2 and 1.4%, respectively.

The test statistic for localization is again above significance levels everywhere but at notably lower levels for the other sample. The effect is just detectable at the lowest distances. The proximity test for the network-driven sample is also significant at higher levels and exhibits identical features as results achieved with the full sample. The proximity effect diminishes rapidly between 8 and 50 miles (the limit of a large metropolis) after which it decreases at a slower rate, with the statistic dropping below the 5% level at 750 miles. By contrast, the effect is not detectable at all distances for the other sample.

These results add additional insight to the finding that proximity matters for potential citing inventors. They show that a large share of localized spillovers are due to cases where there is familiarity between inventors and, notably, when such contact is observable. About as common, or perhaps less common since the non-network-driven sample likely includes
many more “noisy” citations, are spillovers that go where the inventor relationship is not observable (as when a spillover is mediated by neighboring firms) or non-existent (as when learning is limited to knowledge disclosed in a patent). Understandably, proximity is not important for such transfers.

_Citing Cohorts through Time._ With Figures 9 and 10 we turn our attention to localization and proximity patterns over time. I obtain results after separating observations into four seven-year cohorts by the citing/control patent application year: (1) 1976–83, (2) 1984–90, (3) 1991–97, and (4) 1998–2004. The colocation rate of citing patents appears to drop between the first and second periods, as if innovations have begun to diffuse geographically. The rate for control patents, however, has not remained fixed and in fact is marked by an even larger decrease. When considered relative to random industry patents, the share of spillovers due to localization increases with each successive cohort. The test statistic for localization also makes this clear.

The proximity effect is not detectable for the earliest citing cohort at any geographic level. With each successive cohort the effect becomes stronger at all distances. The test statistic flirts with the 5% significance level between 25–32 miles for the second cohort and between 47–86 miles for the third; for citing patents beginning 1998 the effect disappears much later at the 750-mile level. The peak also occurs successively nearer to originating patents at 17, 9.0, 7.8 and 6.7 miles.

The increasing importance of proximity and localized spillovers imply that changes to the underlying mechanisms of knowledge transfer have served mostly to enhance these effects over time. The effects are also consistent with the observed clustering of industries, which
has persisted if not increased in recent decades (Porter). If such forces are at work in
driving agglomeration, then increased clustering would result in yet more proximity-driven
spillovers over time. The evidence suggests this may well be the case in spite of advances
(like the Internet) which would seem to diminish the importance of proximity. The data
allow for the investigation of proximity and localization effects in a more systematic way.

**Distance and State Regressions**

The results for network-driven transfers illustrate that localized spillovers and proximity
effects arise from underlying mechanisms of knowledge transfer. The scale and importance
of the effects, as well as the mechanisms involved, are different across industries. The effects
over time also deserve more systematic study. Until now I have relied on tests involving
sample colocation statistics. In regressions, the geographic location of the random industry
patent is allowed to control at the level of the observation for spillover location, as measured
by geographic distance from the source. The first set of specifications examine the localiza-
tion effect by determining effects on the citing or spillover distance. A positive correlation
with distance, or farther citations, is thus interpreted as a negative effect on localization.
The distances in the dependent variable \(CitingDisNear\) and control \(ControlDisNear\)
are the near distance in log miles. To examine the effect of time I include a lag variable
\(Lag\) whose units are log years between the citing/control patent and the originating patent.
I also examine the effects of network-driven transfers; \(NetworkDriven\) is unity for these
citations. Another factor affecting localization and proximity might be state boundaries.
I include a variable which is unity for intrastate spillovers \(CitingState\). To control for
industry heterogeneity I sort patents by eight industries based on the International Patent
Class.\footnote{The nesting of IPC levels facilitated this sorting, another advantage of international versus U.S. patent classes. Four industries are identified at the sector level (Necessities, Operations, Chemistry, Engineering); two at the class level (Electricity, Medical); and two at the subclass level (Biochemistry, Computing). Details are available from the author.} I include here two broad industries which are known to include geographically concentrated components: Biochemistry and Computing. OLS estimates are in Table 3.

In Column 1, the estimated coefficient on ControlDisNear is highly significant and positive. The Lag coefficient is also significant; that it is positive implies later citations are more geographically distant, after controlling for the location of random industry patents. This is evidence of fading localization overall. Column 2 adds NetworkDriven and its interaction with time. The negative coefficient implies that network-driven transfers, as before, are associated with more localized citations. The interaction coefficient is positive but not significant at the 5% level. CitingState in Column 3 controls for the differential impact of intrastate spillovers and citations that cross state boundaries. This does not have much of an impact on the other coefficients. The interaction of the state variable with time, however, is significant. The resulting lag effect for intrastate transfers is negative (0.037 + -0.068 = -0.31) and implies increasing localization with time. Replacing Lag with year dummies in Column 4 does not noticeably improve the fit or alter findings. Regarding industries, the effect of Biochemistry is among the more strongly localized, whereas Computing is among the least localized. This speaks to the mechanism of spillover in these industries and not the degree of agglomeration itself, say, in drugs or semiconductors. Interestingly, computer-related innovations appear to spill to greater distances than would be predicted given the distribution of computer patents.

Now I examine the same factors as related to a proximity effect. Table 4 contains estimates corresponding to the previous results but with a change to the dependent and control
Regressions of Innovation Distance ($Y = \text{CitingDisNear}$)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5.0 (0.036)*</td>
<td>5.1 (0.039)*</td>
<td>5.8 (0.028)*</td>
<td>5.9 (0.029)*</td>
</tr>
<tr>
<td>ControlDisNear</td>
<td>0.19 (0.0042)*</td>
<td>0.18 (0.0042)*</td>
<td>0.12 (0.0029)*</td>
<td>0.12 (0.0029)*</td>
</tr>
<tr>
<td>Lag</td>
<td>0.017 (0.0071)*</td>
<td>0.041 (0.0095)*</td>
<td>0.037 (0.0069)*</td>
<td></td>
</tr>
<tr>
<td>NetworkDriven</td>
<td>-0.35 (0.037)*</td>
<td>-0.18 (0.026)*</td>
<td>-0.15 (0.032)*</td>
<td></td>
</tr>
<tr>
<td>Lag * NetworkDriven</td>
<td>0.026 (0.015)</td>
<td>0.020 (0.010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CitingState</td>
<td>-3.1 (0.037)*</td>
<td>-3.2 (0.044)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lag * CitingState</td>
<td>-0.068 (0.015)*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operations</td>
<td>-0.13 (0.020)*</td>
<td>-0.10 (0.020)*</td>
<td>-0.047 (0.014)*</td>
<td>-0.045 (0.014)*</td>
</tr>
<tr>
<td>Chemistry</td>
<td>-0.29 (0.027)*</td>
<td>-0.21 (0.027)*</td>
<td>-0.20 (0.019)*</td>
<td>-0.20 (0.019)*</td>
</tr>
<tr>
<td>Engineering</td>
<td>-0.011 (0.022)</td>
<td>0.0049 (0.022)</td>
<td>0.065 (0.015)*</td>
<td>0.067 (0.015)*</td>
</tr>
<tr>
<td>Electricity</td>
<td>-0.066 (0.028)*</td>
<td>-0.041 (0.028)</td>
<td>0.028 (0.020)</td>
<td>0.029 (0.020)</td>
</tr>
<tr>
<td>Medical</td>
<td>0.048 (0.024)*</td>
<td>0.12 (0.025)*</td>
<td>0.10 (0.017)*</td>
<td>0.10 (0.017)*</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>-0.23 (0.044)*</td>
<td>-0.13 (0.044)*</td>
<td>-0.090 (0.031)*</td>
<td>-0.085 (0.031)*</td>
</tr>
<tr>
<td>Computing</td>
<td>0.048 (0.027)</td>
<td>0.086 (0.027)*</td>
<td>0.18 (0.019)*</td>
<td>0.19 (0.019)*</td>
</tr>
<tr>
<td>1978-79</td>
<td></td>
<td>0.0035 (0.027)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980-81</td>
<td></td>
<td>-0.035 (0.028)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1982-83</td>
<td></td>
<td>0.016 (0.030)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1984-85</td>
<td></td>
<td>0.080 (0.031)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1986-87</td>
<td></td>
<td>0.057 (0.030)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988-89</td>
<td></td>
<td>0.11 (0.030)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990-91</td>
<td></td>
<td>0.071 (0.030)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992-93</td>
<td></td>
<td>0.033 (0.030)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994-95</td>
<td></td>
<td>0.064 (0.030)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996-97</td>
<td></td>
<td>0.11 (0.030)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998-99</td>
<td></td>
<td>0.10 (0.031)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-01</td>
<td></td>
<td>0.052 (0.032)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002-04</td>
<td></td>
<td>0.084 (0.047)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observations</th>
<th>65,377</th>
<th>65,377</th>
<th>65,377</th>
<th>65,377</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.035574</td>
<td>0.043351</td>
<td>0.52638</td>
<td>0.52713</td>
</tr>
</tbody>
</table>

Notes: Interactions of year dummies with NetworkDriven and CitingState not reported in (4). Standard errors in parentheses; * indicates significance at 5% level or better.

Table 3: Localization Specifications. OLS estimates where distance is that between citing and originating patents in log miles. Control distance is constructed similarly. Lag is log years between citing/control and originating patents. All others are dummy variables. The omitted industry dummy is Necessities.
variables—instead of the absolute distance between citing and originating patents, distance here is the difference between near distance and middle distance. The new distance variables (CitingDisDiff and ControlDisDiff) demand a different interpretation for specifications which are otherwise the same. Greater distance values here imply a greater differential benefit to the nearest inventor. After controlling for the typical spread between inventors in random industry patents, a positive correlation with distance is interpreted as a positive effect on proximity, or as shifting spillover towards nearer inventors.

The positive lag coefficient in Column 1 implies that nearest citing inventors, those with the most potential to benefit from proximity to the source, become differentially nearer with time—an increasing proximity effect overall. The effect of network-driven transfers in Column 2 is also significant and in the expected positive direction. Its interaction with time is highly significant, and its magnitude compared to the coefficient for Lag implies that network-driven transfers account disproportionately for the increasing proximity effect over time. Column 3 results show that proximity is also enhanced over time for intrastate versus interstate spillovers.\textsuperscript{15} In terms of industries, proximity appears to be important for inventors involved with biochemicals and computers relative to inventors in other industries. The alternative treatment of lag effects in Column 4 again does not alter conclusions.

The results from distance regressions are generally consistent with those obtained using the sample statistic methodologies and yield further evidence for both localization and proximity effects. The evidence also suggests that proximity becomes more important with time. Notably, this change over time appears to be driven by mechanisms which increasingly require familiarity with originating inventors, and, consequently, less by mechanisms asso-

\textsuperscript{15}Here, CitingState is unity if both inventors (near and middle) share a state with the source.
Regressions of Inventor Distance ($Y = \text{CitingDisDiff}$)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.27 (0.027) *</td>
<td>0.36 (0.031) *</td>
<td>0.36 (0.031) *</td>
<td>0.39 (0.033) *</td>
</tr>
<tr>
<td>ControlDisDiff</td>
<td>0.42 (0.0037) *</td>
<td>0.41 (0.0037) *</td>
<td>0.41 (0.0037) *</td>
<td>0.41 (0.0037) *</td>
</tr>
<tr>
<td>Lag</td>
<td>0.14 (0.0078) *</td>
<td>0.040 (0.010) *</td>
<td>0.035 (0.011) *</td>
<td>0.41 (0.0037) *</td>
</tr>
<tr>
<td>NetworkDriven</td>
<td>0.085 (0.040) *</td>
<td>0.082 (0.040) *</td>
<td>0.24 (0.049) *</td>
<td>0.41 (0.0037) *</td>
</tr>
<tr>
<td>Lag * NetworkDriven</td>
<td>0.14 (0.016) *</td>
<td>0.13 (0.016) *</td>
<td>0.14 (0.016) *</td>
<td>0.14 (0.016) *</td>
</tr>
<tr>
<td>CitingState</td>
<td>0.043 (0.058)</td>
<td>0.095 (0.069)</td>
<td>0.043 (0.058)</td>
<td>0.095 (0.069)</td>
</tr>
<tr>
<td>Lag * CitingState</td>
<td>0.065 (0.023) *</td>
<td>0.12 (0.022) *</td>
<td>0.11 (0.022) *</td>
<td>0.12 (0.022) *</td>
</tr>
<tr>
<td>Operations</td>
<td>0.16 (0.022) *</td>
<td>0.12 (0.022) *</td>
<td>0.11 (0.022) *</td>
<td>0.12 (0.022) *</td>
</tr>
<tr>
<td>Chemistry</td>
<td>0.47 (0.029) *</td>
<td>0.37 (0.029) *</td>
<td>0.36 (0.029) *</td>
<td>0.37 (0.029) *</td>
</tr>
<tr>
<td>Engineering</td>
<td>0.14 (0.024) *</td>
<td>0.12 (0.024) *</td>
<td>0.11 (0.024) *</td>
<td>0.12 (0.024) *</td>
</tr>
<tr>
<td>Electricity</td>
<td>0.25 (0.031) *</td>
<td>0.22 (0.031) *</td>
<td>0.21 (0.030) *</td>
<td>0.22 (0.031) *</td>
</tr>
<tr>
<td>Medical</td>
<td>0.46 (0.027) *</td>
<td>0.35 (0.027) *</td>
<td>0.35 (0.027) *</td>
<td>0.36 (0.027) *</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>0.51 (0.047) *</td>
<td>0.38 (0.047) *</td>
<td>0.38 (0.047) *</td>
<td>0.39 (0.047) *</td>
</tr>
<tr>
<td>Computing</td>
<td>0.36 (0.030) *</td>
<td>0.31 (0.029) *</td>
<td>0.30 (0.029) *</td>
<td>0.30 (0.029) *</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th></th>
<th></th>
<th>Oct 93</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1978-79</td>
<td>-0.0032 (0.041)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980-81</td>
<td>-0.012 (0.043)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1982-83</td>
<td>0.038 (0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1984-85</td>
<td>0.0087 (0.047)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1986-87</td>
<td>0.084 (0.047)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988-89</td>
<td>-0.026 (0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990-91</td>
<td>0.0085 (0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992-93</td>
<td>0.062 (0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994-95</td>
<td>0.070 (0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996-97</td>
<td>0.082 (0.046)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998-99</td>
<td>0.070 (0.047)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-01</td>
<td>0.18 (0.050) *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002-04</td>
<td>0.098 (0.072)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Observations</th>
<th>64,495</th>
<th>64,495</th>
<th>64,495</th>
<th>64,495</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.19232</td>
<td>0.20447</td>
<td>0.20578</td>
<td>0.20673</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Interactions of year dummies with NetworkDriven and CitingState not reported in (4). Standard errors in parentheses; * indicates significance at 5% level or better.

Table 4: Proximity Specifications. OLS estimates where distance is the differential inventor distance between citing and originating patents in log miles. Control distance is constructed similarly. Lag is log years between citing/control and originating patents. All others are dummy variables. The omitted industry dummy is Necessities.
associated with one-off transfers. Conversely, it may well be that increasing proximity due to agglomeration is driving more network-driven transfers. What is clear is that these types of spillovers account for much of the effects and have come to represent a large share of knowledge transfers.

The evidence of increasing localization over time is limited to intrastate spillovers, whereas localization of interstate spillovers appears to fade with time. To further investigate the issue of state boundaries I perform a logit analysis where the dependent variable is the probability of intrastate spillover, that is when citing and originating patents share the same state. A positive correlation with the dependent variable is interpreted as an effect enhancing in-state spillover. I examine Lag and NetworkDriven effects as before. I also include a dummy which is unity if the control patent is also colocated by state; ControlState controls for fluctuations in interstate citations by industry and year. I also include specifications with distance as a control (CitingDisNear) to see if citations are more likely to remain in-state. The results are in Table 5.

The lag coefficient in Column 1 is significant and negative, implying that state localization fades with time in terms of the share of innovations. Network-driven transfers are positively related to intrastate citations, but their contribution to the lag is indeterminate. With distance fixed in Column 2, the network-driven effect disappears. This could mean that such networks function without regard to political boundaries. After controlling for distance, a network-driven transfer is no more or less likely to remain in-state. Note also that the lag coefficient has changed signs while remaining significant. With distance fixed, citations are now more likely to remain in the state over time. Computing innovations are highly localized at the state level whereas Biochemistry innovations are not. Columns 3 and 4
### Table 5: State Specifications

Logit estimates where dependent variable is unity if citing and originating patents are from the same state. *ControlState* is constructed similarly using control patent location. *Lag* is log years between citing/control and originating patents. *CitingDisNear* is distance between citing and originating patents in log miles. All others are dummy variables. The omitted industry dummy is *Necessities*.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constant</strong></td>
<td>-2.3 (0.059)*</td>
<td>6.4 (0.26)*</td>
<td>-2.3 (0.062)*</td>
<td>6.5 (0.30)*</td>
</tr>
<tr>
<td><strong>ControlState</strong></td>
<td>1.1 (0.030)*</td>
<td>1.2 (0.050)*</td>
<td>1.1 (0.030)*</td>
<td>1.2 (0.051)*</td>
</tr>
<tr>
<td><strong>Lag</strong></td>
<td>-0.042 (0.020)*</td>
<td>0.22 (0.10)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NetworkDriven</strong></td>
<td>0.46 (0.070)*</td>
<td>-0.070 (0.11 )</td>
<td>0.45 (0.085)*</td>
<td>0.075 (0.13 )</td>
</tr>
<tr>
<td><strong>Lag * NetworkDriven</strong></td>
<td>-0.0076 (0.028)</td>
<td>0.046 (0.044)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CitingDisNear</strong></td>
<td>-1.6 (0.049)*</td>
<td></td>
<td>-1.7 (0.057)*</td>
<td></td>
</tr>
<tr>
<td><strong>Lag * CitingDisNear</strong></td>
<td>-0.033 (0.019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Operations</strong></td>
<td>0.15 (0.042)*</td>
<td>0.13 (0.064)*</td>
<td>0.15 (0.042)*</td>
<td>0.13 (0.064)*</td>
</tr>
<tr>
<td><strong>Chemistry</strong></td>
<td>0.070 (0.054)</td>
<td>-0.47 (0.086)*</td>
<td>0.066 (0.054)</td>
<td>-0.47 (0.086)*</td>
</tr>
<tr>
<td><strong>Engineering</strong></td>
<td>0.15 (0.045)*</td>
<td>0.28 (0.069)*</td>
<td>0.15 (0.045)*</td>
<td>0.28 (0.069)*</td>
</tr>
<tr>
<td><strong>Electricity</strong></td>
<td>0.19 (0.056)*</td>
<td>0.14 (0.088)</td>
<td>0.18 (0.056)*</td>
<td>0.14 (0.089)</td>
</tr>
<tr>
<td><strong>Medical</strong></td>
<td>-0.074 (0.051)</td>
<td>-0.11 (0.080)</td>
<td>-0.065 (0.051)</td>
<td>-0.081 (0.081)</td>
</tr>
<tr>
<td><strong>Biochemistry</strong></td>
<td>0.12 (0.083)</td>
<td>-0.30 (0.14 )*</td>
<td>0.12 (0.083)</td>
<td>-0.29 (0.14 )*</td>
</tr>
<tr>
<td><strong>Computing</strong></td>
<td>0.22 (0.053)*</td>
<td>0.41 (0.086)*</td>
<td>0.21 (0.053)*</td>
<td>0.42 (0.086)*</td>
</tr>
<tr>
<td>1978-79</td>
<td>-0.024 (0.077)</td>
<td>0.39 (0.43 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980-81</td>
<td>-0.018 (0.080)</td>
<td>0.15 (0.43 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1982-83</td>
<td>0.084 (0.084)</td>
<td>0.081 (0.44 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1984-85</td>
<td>0.046 (0.087)</td>
<td>1.3 (0.50 )*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1986-87</td>
<td>-0.053 (0.089)</td>
<td>-0.22 (0.45 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988-89</td>
<td>-0.017 (0.087)</td>
<td>0.46 (0.45 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990-91</td>
<td>-0.087 (0.089)</td>
<td>0.26 (0.45 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992-93</td>
<td>-0.018 (0.088)</td>
<td>0.15 (0.43 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994-95</td>
<td>-0.14 (0.091)</td>
<td>0.72 (0.44 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996-97</td>
<td>-0.23 (0.093)*</td>
<td>0.66 (0.44 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998-99</td>
<td>-0.16 (0.094)</td>
<td>0.93 (0.45 )*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-01</td>
<td>-0.11 (0.097)</td>
<td>0.92 (0.46 )*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002-04</td>
<td>0.011 (0.14 )</td>
<td>1.5 (0.72 )*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>65,377</td>
<td>65,377</td>
<td>65,377</td>
<td>65,377</td>
</tr>
<tr>
<td><strong>Same State</strong></td>
<td>8,562</td>
<td>8,562</td>
<td>8,562</td>
<td>8,562</td>
</tr>
<tr>
<td><strong>Log Likelihood</strong></td>
<td>-24,614</td>
<td>-10,261</td>
<td>-24,581</td>
<td>-10,220</td>
</tr>
</tbody>
</table>

Notes: Interactions of year dummies with *NetworkDriven* and *CitingDisNear* not reported in (4). Standard errors in parentheses; * indicates significance at 5% level or better.
show comparable results using year dummies as before. I depict in Figure 11 the predicted probabilities using the results of Columns 1 and 2.

5 Conclusion

In empirical research patents are most often used as outcome measures. While economically meaningful, the patent as the unit of observation overshadows the fact that transfers of knowledge occur first at the level of the individual. Patent citations link pools of human capital as well as outcomes, and this article demonstrates a number of ways the patent data can be made to yield information on basic innovative activity. It is the author’s opinion that the current research adds some of the most convincing evidence to date of the existence and extent of localized spillovers.

The evidence for proximity effects is also dramatic. Inventors appear to benefit from other nearby inventors, as measured both by the share of localized spillovers and by the differential proximity of the nearest inventors in patent citations. Network-driven transfers, not surprisingly, are correlated with these effects. The findings offer insight on conditions affecting the probability of spillover, but the flow of causality between repeated, network-based contacts and increased productivity has yet to be established. Nevertheless, I believe the analysis highlights the potential of network-driven transfers for understanding the mechanisms involved.

A unique finding is that localization and proximity effects become more important over time, particularly for intrastate spillovers. An extension with further policy implications would make use of data on the nature of differences in innovation policies across states.
Increasing localization, both within and across state boundaries, is consistent with the view that increasing technological complexity or tacit knowledge makes location more important with time, in spite of advances which facilitate the transmission of information and knowledge. Alternatively, increasing localization over time may itself be a result of increasing agglomeration in a feedback loop. Data on location and geographic distance, in the context of a more structured approach, can play a role in disentangling these effects.
References


Appendix: Data Construction

Work involving the gathering and cleaning of U.S. patent data was performed on behalf of the Center for International Science, Technology and Cultural Policy (CISTCP). These data reside at CISTCP.\textsuperscript{16, 17}

Patents

The patent samples used in this study derive from a larger, comprehensive patent database constructed from multiple sources. The CISTCP data are similar to the widely-used NBER patent data\textsuperscript{18} in many respects and are developed from some of the same early sources, but include additional variables and are updated each week as new patents are issued. The PTO provides essentially complete patent data starting with patents granted in September 1996. Weekly “Grant Red Book” data are downloaded, machine-parsed and uploaded to the CISTCP patent database on a regular basis. This paper uses the following patent variables: grant year, application year, primary U.S. patent class and subclass, international patent class, citations, assignees, inventors (patentees), and, for each inventor, a location consisting of city and state and often the USPS ZIP Code.

Two types of self-citations are defined. The first, self-citation by assignee, occurs when

\textsuperscript{16}This work has been supported in part by grants to Lynne G. Zucker, Michael R. Darby, and their collaborators from the National Science Foundation (grants SES-0304727 and SES-0531146) and the University of California’s Industry-University Cooperative Research Program (grants PP99-02, P00-04, P01-02, and P03-01). Certain data included herein are derived from the Zucker-Darby Knowledge, Innovation, and Growth Project. ©Lynne G. Zucker and Michael R. Darby. All rights reserved.

\textsuperscript{17}Certain data included herein are derived from the Science Citation Index Expanded of the Institute for Scientific Information®, Inc. (ISI®), Philadelphia, Pennsylvania, USA: ©Institute for Scientific Information®, Inc. 200. All rights reserved.

\textsuperscript{18}The NBER Patent Citations Data File covers the period 1963–1999 and contains most of the important variables relevant to patent research, as well as useful patent measures constructed by the authors (Hall et al. 2001).
originating and citing patents share an identical assignee string. The matching algorithm allows for variations in capitalization, punctuation, and abbreviations of common terms such as “Co.” and “Inc.” Second, for a given citation, each inventor in the originating patent is machine-checked against each citing inventor; if there is at least one match by inventor name, the citation is defined to be a self-citation by inventor.

Self-citations were produced together with network-driven transfers for the complete patent data and are described more fully in Liu, Zucker and Darby. Available patent and academic article data are scoured exhaustively for evidence of repeat pairings corresponding with inventor pairings in citations. Employing multiple CPUs over many weeks, every inventor pair in a citation (multiple inventors means more than one pair) is checked against inventor and citation lists in patents and author and reference lists in articles; date, type and number of matches are compiled. The data source for journal articles is produced by the Institute for Scientific Information (ISI) and derive from the same source databases which underlie their online bibliometric search engine, the ISI Web of Science. These data include over 24 million articles from 1981–2004 and cover thousands of publications from all fields.

Geolocation

Geographic coordinates are assigned to inventors using data from the U.S. Geological Survey (USGS) and the U.S. Census Bureau. The former maintains the Geographic Names Information System (GNIS) which covers more than 180,000 populated places as of June 6, 2005. A table of Zip Codes and their coordinates was obtained from the Census Bureau

---

19 The NBER data include a numerical assignee identifier which could be used in a similar way.
20 http://geonames.usgs.gov
and represents Zip Codes defined as of November 1999.\textsuperscript{21}

An algorithm was developed to match the location variables in patents against the same variables in the geographic data. For determining a unique location, city and state take precedence over Zip Code because the former are available for all inventors. Furthermore, the USGS database includes historical place names, is kept current, and offers higher resolution—about five times more unique latitude-longitude observations are available for places than for Zip Codes.

For city names the algorithm accounts for format differences and abbreviations like “N.”, “St.” and “Mt.” If multiple cities share the same name and state, one is selected randomly using their population-weighted probabilities. To account for places which are unmatchable—places names with common variations and misspellings, and those undefined in the geographic database—the matching script consults a separate hand-coded file.\textsuperscript{22} For a large patent sample the procedure finds a matched location for about 98% of inventor observations.

\textsuperscript{21}http://www.census.gov/geo/www/tiger/zip1999.html. The Census Bureau actually uses redefined Zip Code Tabulation Areas (ZCTAs) in their ongoing activities.

\textsuperscript{22}Hand-coding is employed to a limited extent (28 cities) beyond which marginal returns diminish rapidly. The success rate before hand-coding is 96%.
Figures

Figures 1–10 depict, for various groups of citing and control patents, rates of colocation with originating patents by distance. Distance in Figures 1 and 2 is the *middle distance* or median value of computed distances between two patents. Distance in all other figures is *near distance*, as computed between the two nearest inventors, unless otherwise indicated. Test statistics are one-sided statistics constructed from sample colocation rates and are distributed as $t$. See text for details.

Figure 1: Comparison with JTH 1993, Colocated Patents
Figure 2: Comparison with Thompson and Fox-Kean 2005, Colocated Patents
Figure 3: USPC and IPC Controls, Colocated Patents
Figure 4: USPC and IPC Controls, Test Statistics
Figure 5: Near and Middle Inventors, Colocated Patents
Figure 6: Near and Middle Inventors, Test Statistics
Figure 7: Network-driven Transfers, Colocated Patents
Figure 8: Network-driven Transfers, Test Statistics
Figure 9: Citing Cohorts through Time, Colocated Patents
Figure 10: Citing Cohorts through Time, Test Statistics
Figure 11: Predicted Citations in Same State
Figure 1: **Comparison with JTH 1993, Colocated Patents.** Sample consists of originating patents from 1976 and citations received through 1989 (truncated to approximate JTH sample). Control patents share the top-level U.S. class (USPC-1). Observations: 70,813.

![Graph](image1)

Figure 2: **Comparison with Thompson and Fox-Kean 2005, Colocated Patents.** Sample consists of originating patents from 1976 and citations received through 2001 (truncated to approximate their sample). Control patents share the U.S. subclass (USPC-2). Observations: 102,354.

![Graph](image2)
Figure 3: **USPC and IPC Controls, Colocated Patents.** Sample consists of originating patents from 1976 and citations received through 2004. Observations: 159,484 (USPC-1), 116,751 (USPC-2), 159,393 (IPC-1), 159,083 (IPC-2), 157,772 (IPC-3), and 149,726 (IPC-4).

Figure 4: **USPC and IPC Controls, Test Statistics.** See Figure 3 above.
Figure 5: **Near and Middle Inventors, Colocated Patents.** Sample consists of originating patents from 1976 and citations received through 2004. Control patents share the IPC group (IPC-4). Observations: 149,726.

Figure 6: **Near and Middle Inventors, Test Statistics.** See Figure 5 above.
Figure 7: **Network-driven Transfers, Colocated Patents.** Sample consists of originating patents from 1976 and citations received through 2004. Control patents share the IPC group (IPC-4). Observations: 68,501 (Network-driven Transfers) and 81,225 (Other).

Figure 8: **Network-driven Transfers, Test Statistics.** See Figure 7 above.
Figure 9: **Citing Cohorts through Time, Colocated Patents.** Sample consists of originating patents from 1976 and citations received through 2004. Control patents share the IPC group (IPC-4). Observations: 45,872 (1), 35,029 (2), 41,165 (3), and 27,660 (4).

Figure 10: **Citing Cohorts through Time, Test Statistics.** See Figure 9 above.
Figure 11: **Predicted Citations in Same State.** Predicted values from Table 5 computed for various lags and distances and at the mean of the remaining variables. The Column 1 specification does not include distance; Column 2 estimates are evaluated at increments of 10 miles from 10–150 miles.