Geographic Constraints on Knowledge Diffusion: Political Borders vs. Spatial Proximity

Jasjit Singh
INSEAD
1 Ayer Rajah Avenue
Singapore 138676
+65 6799 5341
jasjit.singh@insead.edu

Matt Marx
MIT Sloan School of Management
100 Main St., E62-478
Cambridge, MA 02142
+1 617 253 5539
mmarx@mit.edu

August 23, 2012

We thank INSEAD and the MIT Sloan School of Management for funding this research. We are grateful to Ajay Agrawal, Paul Almeida, James Costantini, Iain Cockburn, Pushan Dutt, Lee Fleming, Jeff Furman, Josh Lerner, Ilian Mihov, Peter Thompson, Brian Silverman and Olav Sorenson for feedback. We also thank seminar audiences at Boston University, INSEAD and Singapore Management University as well as participants at the Academy of Management 2010 and 2012 Meetings, Asia-Pacific 2011 Innovation Conference and NUS 2010 Conference on Research in Innovation and Entrepreneurship for comments. Any errors remain our own.

Abstract

Geographic localization of knowledge spillovers is a long-held tenet of economic geography. However, empirical research has examined this phenomenon by considering only one geographic unit (country, state or metropolitan area) at a time, and has also not accounted for spatial distance in such analyses. We disentangle different geographic effects by using a regression framework based on choice-based sampling in order to estimate the likelihood of citation between random patents. We find borders both at the country and state level to impose a constraint on knowledge diffusion over and above what geographic proximity in the form of metropolitan boundaries or geographic distance can explain. An identification methodology based on comparing inventor-added and examiner-added citation patterns points to an even stronger role for political borders than for spatial proximity per se. The state border effect, although robust on average, has been waning over the 30-year time period we examine. On the other hand, the country effect has in fact strengthened despite the commonly expected trend towards globalization and the seeming ease of cross-border communication with technological advancement.

Keywords: knowledge spillovers, borders, distance, economic geography, patent citations

JEL classification: O30, O33, R10, R12

1. INTRODUCTION

Establishing the micro-foundations of industrial agglomeration has become a key focus in economic geography. Moving beyond examining just exogenous locational characteristics, recent work has documented three endogenous mechanisms for why agglomeration takes place: benefits from labor pooling, efficiency gains from co-location of related industries, and localized knowledge spillovers (Ellison, Glaeser, and Kerr, 2010). Of these, knowledge spillovers have generated the most scholarly attention, perhaps because they are seen as critical for value creation and innovation in an increasingly knowledge-intensive economy. In this study, we take a closer look at the role played by various geographic elements in shaping such spillovers, including political borders and spatial proximity.

Even though several studies have documented localization of knowledge spillovers, the geographic levels most relevant for this phenomenon still remain unclear. A significant fraction of related empirical work has studied only country-level spillovers (Branstetter, 2001; Keller, 2002; Singh, 2007), with such studies sometimes being justification for assumptions used in theoretical models of economic growth (Romer, 1990; Grossman and Helpman, 1991). Other studies have taken borders at a less aggregate level – states – as the geographic unit of interest (Jaffe, 1989; Audretsch and Feldman, 1996; Almeida and Kogut, 1999; Rosenthal and Strange, 2001). The focus on either of these political borders is typically for one of two very different reasons. On the one hand, since it can be hard to obtain precise measures for geographic distance or co-location, some researchers have interpreted national or state borders as just a convenient proxy for proximity. On the other, others genuinely believe that national and state borders are important over and above proximity effects, for example due to institutional differences.

A glaring gap in the above literature remains that hardly any of the studies has rigorously tried to disentangle the effects operating at different geographic levels, which provides limited guidance regarding the exact geographic scope of knowledge spillovers. For example, although intra-country knowledge spillovers are found to be more intense than those across countries, this might simply reflect an aggregation of state- or metropolitan-level phenomena. Similarly, interpretation of state-level localization findings is unclear, as these might also be driven by effects operating more locally and are thus open to criticisms to the effect that "state boundaries are a very poor proxy for the geographical units within which knowledge ought to circulate" (Breschi and Lissoni, 2001: 982).

Perhaps motivated by such ambiguities, or by criticism like Krugman's remark that "states aren't really the right geographic units" in economic analysis (1991:43), recent research appears to have increased focus on exploring agglomeration at less coarsely defined geographic levels. For example, Rosenthal and Strange (2003) note that agglomeration attenuates sharply with distance, the strongest effects occurring within a ten-mile radius. Their work is underscored by the modeling of spatial clusters by Kerr and Kominers (2012) and a study of the Manhattan advertising industry by Arzaghi and

Henderson (2008) that shows entrants more likely to appear geographically proximate to incumbents due to spillover effects. This stream of work does not, however, typically examine state or national borders.

Despite a rich body of work examining different geographic levels, few have considered different levels of border and proximity effects *simultaneously* in order to unpack the contribution of each, leaving unresolved important questions such as whether localization effects demonstrated for larger geographic units (country or state) might merely be a manifestation of mechanisms merely on spatial proximity. It is unclear whether to interpret prior border-related findings simply as "distance is not dead" or as borders having an important and independent role on their own. Even studies that do examine multiple geographic levels, such as the path-breaking article by Jaffe, Trajtenberg and Henderson (1993), analyze different geographic units separately and study individual border effects without accounting for spatial proximity.

In addition to not unpacking various geographic levels, research on spillovers generally does not account for spatial distance, treating collocation within each geographic unit as just a measure for geographic proximity without attempting to disentangle border and distance effects. Identifying border effects truly associated with collocation within the same country or state, independent of distance, would require a simultaneous consideration of borders and distance. Although a few have used at least some distance-based measures, these have typically been too aggregate to disentangle all the geographic effects of interest. For example, although Keller (2002) employs data on distance between capital cities of countries, he does not consider different intra-country distances. Likewise, Peri (2005) considers distances between different pairs of states, but does not distinguish different city-to-city distances within a state. In this regard, there is a need to dig deeper into the geography of knowledge spillovers in a manner analogous to a body of work in the literature on international trade, which examines the role of geographic distance versus political borders at the country level (e.g., McCallum, 1995; Anderson and Wincoop, 2003) or state level (e.g., Wolf, 2000; Hillberry and Hummels, 2003, 2008).

Addressing this gap is important given the central role assumptions surrounding the geographic scope of agglomeration play in technological innovation, strategy, international economics and entrepreneurship. Our empirical approach builds upon the well-established tradition of using patent citations to measure diffusion of knowledge. As a further motivation, Figure 1 provides simple graphical evidence based regarding the geographic pattern of inter-firm citations to patents from U.S. inventors during 1975-2004.[1] Three observations are worth making. First, the likelihood of citation between random pairs of patents decreases with geographic distance. Second, the likelihood of citation is greater within country borders than across, greater within state borders than across, and greater within metropolitan area

[1] These charts were constructed using our dataset described later and employed in our regression analysis. Although our dataset was derived using stratified sampling from the population of patent pairs, we calculated the summary statistics by appropriately weighting each observation so that Figure 1 represents true population characteristics.

boundaries (measured using "CBSA" definitions explained later in the paper) than across. Third, the national and state border effects seem to be only partly explained by geographic proximity since there seems to be a border effect within each of the geographic distance buckets – a finding that continues to hold when further redefining or narrowing of the distance buckets. Figure 1 is, however, just based on summary statistics and does not account for a number of empirical issues explained later in the paper.

In more formal analysis reported later, we run a "horse race" among various geographic variables to isolate the level at which localization of knowledge spillovers operates most prominently. Specifically, we construct a dataset of patent pairs using choice-based sampling and then estimate a "citation function" that models the likelihood of citations between random patents. This framework departs from previous studies by making no ex ante assumptions about the correct geographic unit of analysis. Instead, it allows us to simultaneously account for collocation of the source and destination of knowledge within the same country, state or metropolitan area as well as account for fine-grained spatial distance.

Consistent with prior work, separate analyses we conduct at the national, state and metropolitan levels all exhibit spillover localization. Importantly, the findings hold even in regression models where these are considered simultaneously in order to account for the fact that considering individual units separately overstates their importance. We extend the analysis to models that first parametrically control for distance and then employ a set of non-parametric indicator variables. Much of the country- and state-level effects persist even though there are also independent effects for metropolitan areas as well as gradual decaying with distance, same-country localization again much stronger than within a state.

We view robust localization of knowledge flows within national borders as not a big surprise, given the well-documented linguistic, cultural, institutional and economic differences among countries (see, e.g., Coe, Helpman and Hoffmaister, 2009). However, time-trend analysis reveals a surprising strengthening of the same-country effect over time despite the accepted trend toward globalization and technological advances which supposedly smooth cross-border communication.

We find the state border effect even more puzzling and try to analyze it further. The finding turns out not to be driven by just one or two specific states (like California) or sectors (like computers or communication technologies). The result is seen even in a subsample comprised of patents close to state borders, indicating that the aggregate finding is also not driven by inadequately controlling for distance for cited patents in the interior. In fact, state borders are found to matter even in a very conservative test where metropolitan effects are completely isolated by considering only patents and (potential) citations within metropolitan areas that span state borders. We also analyze trends over time, and do find that – in contrast to the country border effect – the state border effect weakens considerably over time.

Finally, we address two challenges inherent in using patent citations to measure spillover localization. First, citation patterns are determined in part by technological relationships which cannot

be perfectly captured by any formal classification system (Thompson and Fox-Kean, 2005). Second, some citations are added by patent examiners, not inventors, and the extent to which the two represent spillovers likely differs (Alcacer and Gittelman, 2006; Thompson, 2006). To address these concerns, we first examine the robustness of our prior findings to using only inventor-added citations. We then employ an identification strategy (motivated by Thompson, 2006) that calculates true geographic effects as the difference between estimates from inventor versus examiner citations. Border effect findings remain robust even though this approach weakens the effect of proximity, further highlighting the importance of borders beyond pure geographic proximity in shaping knowledge diffusion patterns.

2. EMPIRICAL APPROACH

2.1. Constructing a patent-based dataset

We follow the well-established tradition of using citations between patents as an indicator of knowledge flows. Although citation-based measures are noisy in capturing true knowledge flows, surveys of inventors have established that citations—especially when employed in large samples—do capture knowledge flows meaningfully (Jaffe and Trajtenberg, 2002; Duguet and MacGarvie, 2005). Admittedly, even assuming that citations do correctly capture knowledge flows, it is not possible to decipher when a given citation represents a "spillover", i.e., a true externality for which the receiver does not fully pay. Nevertheless, we follow the prevalent view that using citations is reasonable because they at least partly represent spillovers and very often represent benefits the receiver gets in the form of "gains from trade" even in other cases where they represent purely market transactions.

Our dataset is based on United States Patent and Trademark Office (USPTO) patents with application years 1975 through 2009. In order to have at least a 5-year window to observe citations that a patent receives, we restrict our sample of cited patents to the period 1975-2004, with the set of potential patents citing these going all the way until 2009. Since recent literature (Alcacer and Gittelman, 2006; Thompson, 2006) has emphasized possible distinction between citations added by the inventors themselves versus patent examiners, we also keep track of this information when available (2001 onwards), and use it to complement our analyses using the full sample. Patent data also include inventors' city, state and country of residence. Since consistent state identification is available only for patents originating in the U.S., we restrict the cited patent sample (but not the citing patent sample) to U.S. inventors. Our calculation of geographic distances relies upon data from Lai, D'Amour and Fleming (2009) that map cities where inventors live to latitudes and longitudes.[2] We also map these cities to Core

[2] Our distance data are therefore restrictive in two ways. First, since we only observe a single latitude and longitude coordinate per city, we cannot calculate distances between inventors within a city or even be completely precise about distances between those in adjoining cities. Second, what the USPTO data contain is the city of residence of the inventor, which might sometimes not coincide with the city of where the inventor works. Patents do not always

Based Statistical Areas (CBSAs), which reflect metropolitan area commuting patterns and discard the small fraction of patents not falling within any CBSA.[3]

Before proceeding to construct a sample of patent pairs representing actual or potential citations, we restrict the cited patent sample to only patents whose geographic origin is unambiguously defined in order to avoid making arbitrary assumptions in trying to resolve locational ambiguity of a knowledge source. In other words, we exclude patents from geographically dispersed inventor teams, even though these might be an interesting (but different) topic to study. We also omit patents not assigned to any organization as well as those to non-firm sources (such as universities and government bodies) as the focus of this study is to examine inter-firm diffusion of knowledge. In the end, all the steps mentioned above yield a set of 631,586 potentially cited patents as sources of knowledge.

2.2. Constructing a matched sample of actual and potential citations

For each cited patent mentioned above, we collect data on all citations received during a 10-year window since its application and drop all within-firm citations. As a highly influential study by Jaffe, Trajtenberg and Henderson (1993 – hereafter, "JTH") points out, just calculating collation frequency within pairs of patents involved in realized citations would not suffice for establishing geographic localization of knowledge. Instead, what is needed is an appropriate control sample of potential (but unrealized) citations to establish a benchmark level of collocation expected given the existing geographic distribution of technological activity. To facilitate a comparison of our subsequent analysis with the JTH method, we therefore also start with their approach of matching each citing patent to a random control patent with the same three-digit technology class and application year as the original citing patent (but not from the same organization as the focal cited patent and also not actually citing it). Like JTH, we drop the small fraction of citations for which no match is found. This leads to a balanced sample of 4,007,217 realized citations (based on 631,586 cited patents) and exactly as many unrealized matched control citations.

The above JTH-style sample allows us to compare the extent of geographic collocation of the source and destination for the original citations versus control pairs, in turn using the country, state and metropolitan area as the geographic units of analysis in three *different* sets of calculations. While a useful

---

contain assignee address; even when present, this is often for the firm's headquarters. Thus inventor residential city is the best proxy for the invention's location; we control for possible commuting distances using metropolitan area.

[3] We found that about 15.3% of U.S. patents could not be matched to metropolitan area definitions (such as MSA, PMSA or CMSA) used in prior studies like Thompson (2006). We did not want to prevent dropping such a large fraction simply by relying upon the common approach of defining a "phantom metropolitan area" per state to handle the large number of exceptional cases, as doing so could confound metropolitan area effects with state effects. Therefore, we rely on a more comprehensive concordance between U.S. cities and the 2003 definitions of metropolitan areas from the U.S. Office of Management and Budget. Employing these so-called CBSAs has two benefits. First, they are more comparable in covering reasonable commuting distances for population centers across the U.S. Second, they allow mapping a larger fraction - over 96.3% instead of 84.7% - of U.S. patents to metropolitan areas. Our main results are, however, robust to using the older definition of metropolitan areas.

starting point, this approach is not well-suited to directly addressing our question: How much do national or state borders *per se* constrain knowledge flow, as opposed to the observed effects at these levels being manifestations of mechanisms that in fact operate at more local levels (such as city or CBSA) being driven purely by geographic distance? Our preferred approach for answering these questions is a regression framework that can simultaneously examine the effect of different geographic levels.

2.3. A regression framework for estimating citation likelihood

With collocation within a certain pre-defined geographic unit of analysis as the dependent variable in the JTH model, one cannot easily examine multiple geographic levels at the same time. One could try to ascertain the relative importance of different geographic levels by somehow comparing the findings across models; however, this would likely remain a statistically complex and unsatisfactory exercise. We instead rely on a regression framework that estimates likelihood of citation between two random patents, making the existence of a citation between a pair of patents the dependent variable and employing the entire set of geography-related variables *simultaneously* as explanatory variables in a single model.[4]

Our citation-level regression framework has the added advantage of flexibility in modeling technological relatedness between patents, allowing multiple levels of technological granularity to be considered at once. This addresses a challenge previous studies have faced in having to choose a specific technological granularity in constructing a JTH-style control sample. As Thompson and Fox-Kean (2005) and Henderson, Jaffe and Trajtenberg (2005) discuss, one faces a dilemma in using matching: the three-digit technology match commonly employed might be too crude to capture all relevant technological relationships, but using a finer classification could suffer from selection bias because a match would not be found for most of the sample. Both articles suggest a regression approach that simultaneously accounts for technological relatedness at multiple levels of granularity.[5]

A seemingly straightforward (yet incorrect) extension of the JTH methodology might be to employ a regression approach using a JTH-style matched sample in a (logit or probit) regression model, wherein the existence of a citation between a pair of patents is taken as the dichotomous dependent variable. However, this would imply that the matching procedure was in effect used to carry out sampling based on the dependent variable in the first place, since the JTH method draws a "zero" (unrealized citation) corresponding to each "one" (actual citation). This needs to be somehow corrected for in order to avoid biasing the estimates. Further, the potentially citing patents used in

[4] Our methodology builds upon studies such as Sorenson and Fleming (2004) and Singh (2005) that also model the citation likelihood between patents in a regression framework, though to study different research questions.

[5] This does not fully address the issue that no technological classification system – howsoever finely defined – can perfectly capture true technological relationships between patents. We address this concern later in the paper by extending our JTH-style as well as regression analysis using an approach motivated by Thompson (2006).

constructing the control pairs are drawn only from technology classes and years from which citations to the cited patent actually exist, ignoring the population of potentially citing patents from the remaining technology classes and years. As the technical appendix explains, this can further bias the results. Here, we describe a micro-level citation regression framework that ameliorates these issues.

Before discussing how we need to extend our matched sample to carry out patent-level regression analysis, it is useful for exposition to first imagine a sample of patent pairs (either realized or unrealized citations) constructed by pairing each of our initial set of potentially cited patents with a random draw of potentially citing patents. We could model the likelihood of a patent citation in this sample as a Bernoulli outcome *y* that equals 1 with a probability

$$\Pr(y = 1 \mid x = x_i) = \Lambda(x_i\beta) = \frac{1}{1 + e^{-x_i\beta}}$$

Here, i is an index for the sample of potential citations (i.e., patent pairs), $\boldsymbol{x}_i$ represents the vector of covariates and controls (described later), and $\boldsymbol{\beta}$ is the vector of parameters to be estimated.

Since the likelihood of a focal patent being cited by a random patent is extremely small, it is not practical to carry out the estimation based solely on the dataset constructed by using random sampling from the population of all potentially citing patents. Instead, consider employing a "choice-based" sample, wherein the sampled fraction γ of potentially citing patents that actually cite a focal patent is much larger than the fraction α of the patents that are not involved in a real citation to it. It is worth noting that a usual (unweighted) logistic estimation based on such a sample would lead to biased estimates, since the sampling rate here is different for different values of the dependent variable. One way to avoid the bias is to use the *weighted exogenous sampling maximum likelihood* (WESML) approach, which involves a modified logistic estimation based on first weighting each observation by the reciprocal of the ex ante probability of its inclusion in the sample (Manski and Lerman, 1977).[6]

The basic WESML approach as described above is based on employing a sample where the "zeroes" are drawn from the population of unrealized citations with the same ex ante likelihood. Recognizing that technological relatedness is a particularly strong driver of citation likelihood between patents, we can refine the choice-based sampling approach further to also get benefits from stratification on this explanatory variable. This implies allowing the parameter α to vary across different y=0 subpopulations (Manski and McFadden, 1981; Amemiya, 1985, Ch. 9).

Indeed, by carefully considering the respective subpopulations (defined by different technology classes and years of origin) from which we have effectively drawn our JTH-style control patents in the previous section, we can interpret our matched sample as above and appropriately

[6] See the Appendix for further detail. See also Greene (2003, Ch. 21) for a discussion of choice-base sampling.

calculate the weights to use with each control pair. However, as the technical appendix explains in more detail, this is not sufficient in itself. Using the WESML approach with the matched sample also requires extending the sample to ensure representation of potentially citing patents belonging to years and/or technology classes not represented in the original patent citations (and hence in the resulting matched sample). Doing so ensures that the strata considered are not only mutually exclusive but also exhaustive in representing the full population of potential citations. The above steps lead to our final sample of 13,728,582 patent pairs, which includes 4,007,217 actual citations (taking γ=1), 4,007,217 JTH-style matched pairs and 5,714,148 additional pairs from citing classes and years not represented in the matched sample. An example included in the technical appendix further illustrates the above sampling procedure as well as calculation of appropriate weights for all the control observations.

Rather than making specific assumptions about the temporal pattern of citations, we account for variation in citation likelihood with citation lag (i.e., years elapsed between the cited and citing patents) non-parametrically—that is, by including among the covariates the full set of indicator variables for different lags. We also include indicators for the cited patent's technological category and the citing patent's year of origin to account for systematic differences across sectors or over time.[7] Finally, since the citation probability might also be driven by other characteristics of the cited patent, we control for observables and cluster standard errors to account for unobserved ones.

3. EXTENDING THE TRADITIONAL MATCHING APPROACH

Before turning to our regression approach in the next section, we present some analysis that extends the more traditional JTH-style analysis. This should allow a reader familiar with prior literature to relate our study better to existing research in terms of both what kind of findings remain similar across the two approached and what new insights emerge specifically from using the regression approach.

3.1. Baseline analysis comparable with prior work

Following the empirical approach of JTH, we compare the incidence of geographic collocation of the potential knowledge sources as represented in actual citations as well as matched control pairs, in turn using the country, state, and metropolitan area as the geographic units of analysis. As the side-by-side comparison in Table 1 shows, our findings at each of the three units of analysis are quite comparable to those reported by JTH as well as a replication by Thompson and Fox-Kean (2005). The incidence of collocation for all three geographic units is statistically and economically greater between actual

[7] Our goal here is simply to control for citation lag and citing year effects without trying to identify one of these effects separately as in studies such as Rysman and Simcoe (2008). Given that perfect collinearity would result if citation lag and citing year effects are included as the usual sets of indicators, we omit one of the indicator variables.

citations and the corresponding matched control pairs: 74.7% vs. 57.6% at the country level; 13.4% vs. 6.2% at the state level; and 7.6% vs. 2.6% at the metropolitan level.[8]

3.2. Further investigation of the border effects

Again, it is difficult within the JTH framework to separate the extent to which localization spillovers are driven primarily by political borders, spatial proximity, or both. A full analysis will have to wait until we report findings from our regression approach below. However, we can carry out at least some informative analysis even within the JTH framework. In doing so, we focus in particular on the robustness and nature of the state border effect because, although localization at the country level might be less surprising given the well-documented linguistic, cultural, administrative and economic differences between countries (Coe, Helpman and Hoffmaister, 2009), the presence of a localization effect truly associated with state borders within a country like the U.S. would probably be puzzling.

Staying within the JTH framework, a first step in separating border and proximity effects is determining whether the state finding might be driven by observations that are also geographically distant from the state border. We therefore analyze diffusion of only knowledge originating near a state border to see if there is on average a state-border effect even for these in order to ensure that within-state localization reported above is not just a distance effect driven by cited patents in a state's interior. Specifically, columns (1)-(4) in Table 2 report findings from a JTH-style analysis using a subsample where the distance of a potentially cited patent's originating town or city to the closest state border is not more than 20 miles. Under the null hypothesis that state borders play no role in knowledge diffusion and that the previous findings were somehow driven by observations that are distant from the state borders, one would expect state-level localization to become difficult to observe for these observations. Comparing column (2) in Table (1) with column (2) in Table (2), we find that not to be the case. Even though state-level collocation in column (2) is substantially lower for citations in the near-border sample than the whole sample (7.1% in Table 2 vs. 13.4% in Table 1), the matched pair sample collocation incidence is also substantially lower in the near-border sample than the whole sample (2.7% in Table 2 vs. 6.2% in Table 1) so that the ratio reported calculated in column (4) is in fact higher in Table 2. In other words, taking account of geographic distribution of technological activity, we find no evidence that mere distance is driving the state effect reported earlier.

While the above analysis based on a subset of *cited* patents (representing the source of knowledge) originating near a state border increases confidence in the possibility that state borders do

[8] When Thompson and Fox-Kean subsequently employ nine-digit technology matching, they find that over two-thirds of their patents cannot be matched. Our approach is instead to stick to a three-digit initial match but control for a finer technological level through additional variables introduced directly into our regression model.

indeed have an independent effect, columns (5)-(8) refine this by restricting the set of potentially *citing* patents to those that originate within one of two states separated by the state border under consideration. For example, for a cited patent from Haverhill, Massachusetts (near the New Hampshire border) we would consider only (potential) citations from either Massachusetts or New Hampshire. Given that the citing patents in the matched pairs in our original sample could be from anywhere, this analysis relies on a new matched sample appropriate to the task purpose. Specifically, a control patent is now generated by matching the citing patent to a patent not just from the same 3-digit technology class and the same year but also originating from within the state dyad being considered.

The interpretation of the results reported in columns (5)-(8) is that, in a sample comprising only dyads of neighboring states, knowledge generated within 20 miles of a state border is still much more likely to be used within its state of origin than the neighboring state (after, as before, adjusting for geographic distribution of different technology classes). In other words, the finding of a state border remains qualitatively robust to using this alternate methodology. [9] Since we use a new sample that restricts actual and potential citations to be between neighboring states within the U.S., note that country border effects have been filtered out (so country-level analysis is no longer carried out) and that the reported numbers are also not comparable with the findings from columns (1)-(4).

One interesting feature of U.S. geography is that 34 of the 270 CBSAs include more than one state. For example, the CBSA containing Cincinnati, Ohio also extends into sections of Kentucky and Indiana. This allows us to test the border effect by examining whether in-state localization exists even for knowledge flows within such CBSAs. Specifically, columns (9)-(12) report the findings based on a subsample of the data in columns (5)-(8) where the observations only include cited patents originating in a multi-state CBSA. The observations are further restricted to citations coming from within the CBSA that are also matched to control citations also within the same CBSA. By construction, metropolitan effects have therefore been filtered out (so CBSA-level analysis is no longer carried out). Difference of means between incidences of geographic co-location within the state for actual citations versus corresponding controls remains statistically significant. Although their ratio is now much smaller, it should be noted that this is a very conservative test using a smaller, highly restrictive within-CBSA sample. Thus, just the fact that we find *any* state-border effect in this case is perhaps in itself quite remarkable. To a skeptic, this could be an indication instead that the state border effect is perhaps not as strong as it is made out to be in the earlier analysis. At this point, we are agnostic to an exact interpretation – preferring instead to address this debate in our regression framework where we

[9] In choosing the sample of cited patents near state borders, we have reported findings based on a cut-off of 20 miles as a compromise between being close to the border and having a reasonable sample size. In the spirit of Holmes (1998), we actually tried out progressively smaller windows starting from 50 miles and going all the way down to those within 5 miles of a state border. The findings remained robust in supporting of a state border effect.

are able to employ our full sample while still carefully accounting for spatial proximity in terms of both metropolitan co-location as well as geographic distance more generally.

3.3. Long-term time trends

Our sample size is orders of magnitude larger than those employed in previous studies, so we can carry out more detailed analyses reported in Table 3. Columns (1) through (4) segment our cited patents drawn from 1975-2004 into six five-year periods.[10] Localization of knowledge spillovers remains robust across all periods for all three geographic units. Further, we can examine the time trends by taking the ratio of collocation frequency for inventor pairs comprising actual citations vs. matched controls reported in column (4) as an indicator of the strength of the geographic effects. What is rather striking is that – despite much talk about globalization and decreasing relevance of geographic separation - the role of geography appears to have increased rather than decreased over time. Given that the JTH framework only analyzes each geographic unit in isolation, this analysis is however not able to disentangle whether the time trends are reflective primarily of underlying border effects, proximity effects or a combination of the two. We will therefore return to this issue later in the context of our preferred regression framework that accounts for all geographic effects simultaneously.

3.4. Inventor vs. examiner citations

Recent work has noted that many patent citations are included not by the inventors themselves but later by patent examiners (Alcacer and Gittelman, 2006). Therefore, it is useful to carry out analysis complementary to the above by examining just inventor-added citations, since these might arguably be more likely to reflect prior art that an inventor was aware of in coming up with the focal invention.[11] Columns (5) through (8) of Table 3 report the JTH-kind analysis based only on the subsample of citations added by inventors (and the corresponding controls). Since the inventor/examiner distinction is only available for citations post-2001, these calculations are reported only for the cited patent originating during one of the three five-year periods for which the citation window overlaps with availability of this information for a significant fraction of the citations. Comparing the extent of the localization effect calculated in column (8) versus column (4) reveals that a focus on just inventor-added citations significantly strengthens the geographic localization for all three geographic units of analysis. Unlike the results in column (4), the results in column (8) do not show any time trends –

[10] The sample size drops during the last five year period (2000-2004) because, while the earlier periods employ a full ten-year citation window, for this period we only observe citing patents through 2009.

[11] In addition to the fact that the inventor vs. examiner distinction is readily available only post-2001, a case can be made in favor of considering all citations rather than just inventor citations because inventors often deliberately omit reference even to relevant patents they know about due to strategic reasons (Lampe, 2011). From this point of view, analyzing just inventor citations is a more of a useful robustness check than necessarily always superior.

though that is largely reflective of the fact that the analysis cannot even be carried out for the first three periods due to unavailability of the inventor vs. examiner distinction for citing patents pre-2001.

Thompson (2006) exploits the inventor/examiner distinction to come up with a way to address a fundamental challenge with use of a JTH-style matching approach: since even the finest available technological classification might not capture some unobserved technological characteristics driving both patent citation patterns and geographical co-location, it is hard to make definitive statements about geographic co-location *leading to* increased knowledge diffusion. He suggests an identification strategy wherein one ought to take greater geographic localization for inventor-contributed citations only *relative to* that for citations added by examiners (who are "geography blind" and hence form an appropriate benchmark for comparison) as reliable evidence of localized knowledge spillovers. To be able to use this suggested benchmark in analyzing inventor citation findings from columns (5)-(8), we report analysis using just examiner-added citations (and corresponding matched controls) in columns (9)-(12). Comparing columns (8) and (12), the calculated ratio between collocation incidence for realized citations vs. matched patent pairs is found to be higher in all cases for inventor-added citations than for examiner-added citations, further establishing the robustness of the finding on geographic localization of knowledge spillovers. However, this still does not disentangle border vs. proximity effects through a simultaneous examination, for which we need to use our regression framework.

4. ANALYSIS USING OUR WESML REGRESSION FRAMEWORK

4.1. Simultaneous examination of multiple geographic levels

We now turn to the regression framework to simultaneously examine national and state borders after accounting for proximity effects related to metropolitan (i.e., CBSA) co-location and geographic distance. Table 4 summarizes the variables used in our analyses. Before trying to disentangle borders and proximity, however, it is instructive to get an overall sense of diffusion and geography. The analysis reported in column (1) of Table 5 is the simplest way of seeing this. The WESML regression estimates have an intuitive interpretation in terms of how an explanatory variable drives the likelihood of citation between random patents in the population, with the fact that citations are rare events making it possible to in fact directly interpret the logistic model coefficients as percentage effects on citation likelihood.[12] Column (1) implies that the likelihood of citation falls by 36% with a doubling of distance. (Again, distance is measured between inventor cities, not exact addresses.)

[12] In a logistic model, the marginal effect for a variable j is $\beta_j \Lambda'(\boldsymbol{x\beta})$, which turns out to equal $\beta_j \Lambda(\boldsymbol{x\beta})[1-\Lambda(\boldsymbol{x\beta})]$. In general, this would need to be calculated based either on the mean predicted probability or using the sample mean for $\Lambda(\boldsymbol{x\beta})$. But the fact that citations are rare events allows further simplification: since $\Lambda(\boldsymbol{x\beta})$ is much smaller than 1, $\beta_j \Lambda(\boldsymbol{x\beta})[1-\Lambda(\boldsymbol{x\beta})]$ is practically equivalent to $\beta_j \Lambda(\boldsymbol{x\beta})$. This means the coefficient estimate for β_j can be directly interpreted as the percentage change in citation probability with a unit change in variable j.

The analysis reported in column (2) also includes relevant control variables. Most importantly, the regression framework allows us to control for technological similarity and relatedness between patents using a series of associated variables rather than only relying on matching using three-digit technology class. The findings in column (2) imply that the likelihood of citation now falls by 27% with a doubling of distance – the difference between the column (1) and (2) estimates being driven mainly by the new technology controls. In line with the JTH argument, we find knowledge flows within the same or related technologies to be stronger than those across different technologies, as indicated by the positive and significant estimates for *same tech category*, *same tech subcategory*, *same tech class* and *relatedness of tech classes*. Noting that the Thompson and Fox-Kean (2005) critique regarding the inadequacy of three-digit technological controls, we have also included a control variable to capture overlap between the citing and cited patent along their secondary nine-digit technology subclasses (*overlap of tech subclasses*) and find that to have a strong effect as well.

Setting the geographic distance variable aside for now, columns (3), (4) and (5) successively introduce variables for co-location at three geographic levels: the country (*same country*), the state (*same state*) and the metropolitan area (*same CBSA*). The estimate for *same country* falls a little once *same state* is introduced in going from model (3) to model (4), and that the estimate for *same state* falls drastically once *same CBSA* is introduced in moving from model (4) to model (5). This highlights the benefit of using a regression approach to disentangle effects at the various geographic levels by simultaneously considering all three levels rather than relying just on separate analysis for each. In terms of magnitude, column (5) estimates imply a 77% greater likelihood of within-country knowledge flow than across national borders, a 41% greater likelihood for within-state flow than that across state borders, and 77% greater likelihood for within-CBSA flow than that across CBSA boundaries. [13]

Simultaneously considering multiple geographic units indicates that there is more to the national and state border effects than a mere aggregation of localization mechanisms operating at the metropolitan level. The estimates in column (5), however, do not rule out the possibility that such effects are not epiphenomenal with spatial distance since including the CBSA co-location variable does not account for distance-related effects that might be more gradual than CBSA collocation. To this point, the model in column (6) now also includes our distance variable from before. As expected, with geographic proximity now better controlled for through the combination of metropolitan co-location and distance, both border effects become smaller. Interestingly, the drop is much larger for the *same state* effect than the *same country* effect. (Strictly speaking, despite our footnote about

[13] If we carry out this analysis excluding the 9-digit technology control, geographic localization on all three dimensions turns out to be larger – with the difference being the greatest for metropolitan collocation. This is in line with intuition that geographic concentration of technological activity—which is what our technology-related control variables account for—is greater when viewed at a finer level of granularity for technology.

interpretation of coefficients as percentages, coefficients are not directly comparable across columns (5) and (6) due to different reference categories. But the above holds even after adjusting for that.)

To allow more flexibility in how distance constrains knowledge flows, column (7) repeats the analysis with a series of indicator variables for distance ranges. These fine-grained distance indicators are mutually exclusive, covering increasing distances starting in the sequence *distance 0 miles* (i.e., same city), *distance 0-10 miles*, and so on. The omitted category is distance greater than 6,000 miles. This non-parametric approach does not impose any functional-form assumption on how distance might affect the likelihood of citation, ensuring that the *same country* and *same state* variables more accurately measure border effects independent of geographic proximity. (Even more fine-grained indicators did not materially alter findings.) Not surprisingly, estimates for the distance indicators themselves reveal that knowledge flows are greatest when the source and recipients are collocated within the same city (i.e., distance = 0) and that the distance effect gradually falls (more or less monotonically) with distance. Once more, however, we find statistically and economically significant estimates for *same country* and *same state* even after we have accounted for geographic proximity using *same CBSA* and distance indicators. (Note that it is hard to directly compare *same country* and *same state* coefficients across columns (6) and (7) as the latter has a large number of new variables in the form of distance indicators.) This finding challenges an interpretation that localized knowledge diffusion reported by previous studies is merely a manifestation of intra-regional distances being on average smaller than cross-regional distances.[14]

4.2. Further investigation of the border effects

We now examine subsamples to figure out whether our findings are driven by particular kinds of patents. As already mentioned, one concern might be whether the state-level finding is driven by observations that are quite distant from the state border. Analogous to the near-border analysis presented for the JTH approach, we analyze diffusion of knowledge originating near state borders to see if there is on average a similar state-border effect even for these. Specifically, we look at the subset of potentially cited patents that lie within 20 miles of a state border. As column (8) in Table 5 indicates, the findings for the near-border cited patent sub-sample turn out to be very similar to those from the full sample (column (7)), including the same-state effect.

Next, we subset our sample by removing California as Silicon Valley has been often described as an outlier for diffusion (Almeida and Kogut, 1999). As the top state in terms of patenting activity

[14] In additional analysis, we tried models with indicators for *contiguous countries* and *contiguous states* to distinguish cases where the source and destination share a border. While we did find knowledge flow to be more intense between contiguous regions, we found that independent country and state border effects persist.

and one of the largest in terms of area, one might worry that our results depend on California in ways that state fixed effects do not capture. In column (9) of Table 5, both country and state localization are found to be robust to excluding California. To further investigate whether our findings are state-specific, in analysis not reported to conserve space, we also carried out analogous analyses for cited patent subsamples from the ten largest patenting states. The findings revealed that, in six of these ten cases, observed state-level localization of knowledge originating within the state borders could not be completely explained simply by geographic proximity effects in the form of metropolitan co-location and/or shorter geographic distances. In other words, the finding is not driven by just one or two specific states. In fact, California turned out to be one of the minority cases where state borders *do not* seem to have an effect independent of distance (but CBSA boundaries like those of the Silicon Valley still *do*), suggesting that – once one crosses out of areas like Silicon Valley – knowledge is no longer further constrained by the borders of California over and above effects related simply to distance.

Next, we turn to checking whether the results could similarly be driven by specific sectors. To start with, we exclude the one-digit NBER technology category *Computers & Communications* – a sector many scholars consider to be unique. As column (10) in Table 5 shows, the results are qualitatively unchanged. To further investigate if our findings are sector specific, we also carried out (but omit for space) separate analyses for cited patent subsamples from all six different one-digit NBER categories. The findings revealed that the findings are not driven by a specific sector. In fact, in five of the six cases, observed state-level localization of knowledge could not be completely explained simply by geographic proximity effects, the only exception being the category Hall, Jaffe and Trajtenberg (2001) label "Others". Similarly, repeating the analysis with two-digit NBER sub-categories reveals robust, independent state border effect for 30 of the 36 subsamples. Thus the state border finding appears not to be clearly driven by one or two specific states or sectors. We next investigate whether border effects are driven by particular time periods as opposed to being persistent.

4.3. Long-term time trends

Before disentangling long-term trends in border and proximity effects, it is useful to start with an overall sense of how the role of geography in knowledge diffusion has evolved over time. With this view, column (1) in Table 6 extends the analysis from column (2) in Table 5 by adding an interaction term *period * ln(distance + 1)* between the distance variable and the time period variable capturing the five-year period when the cited patent originated. (See Table 4 for detailed definition.) [15] Surprisingly,

[15] Recall that our analysis is carried out with a set of indicators for the time period of origin for the cited patent and for the time lag between the cited and citing patents. Since our citing as well as cited patents are of different vintage, our sample allows separately identifying cohort effects and citation lag effects in a way that previous studies with more restrictive samples (such as Thompson(2006)) were not able to.

and contrary to the widespread notion that the importance of distance has been eroding over time due to globalization and technological advancement, the decay in citation rate with distance seems to have *increased* over time, albeit the economical magnitude of this is not too large.

In column (2), we turn to disentangling time trends in the border vs. proximity effects, with the goal of figuring out whether the role of political borders has strengthened or weakened over time once proximity is accounted for. In addition to the distance variable, we re-introduce our other three geographic variables – *same country*, *same state* and *same cbsa*, but now also bring in their interaction effects with the time variable *period*. The trends turn out to differ across different variables: the effect of national borders seems to have increased over time while that for state borders and CBSA boundaries has decreased. Additional analyses in columns (3) and (4) add distance indicators and the full set of distance-period indicators respectively in order to more completely account for any distance-related effects and trends not captured above. The finding on the opposite time trends for country vs. state borders remain qualitatively robust, with the country effect still strengthening over time and the state effect weakening. However, the CBSA finding is more fickle, becoming statistically insignificant in column (3) and ultimately flipping sign to become positive (and statistically significant) in column (4). This might be due to the high correlation between the distance indicators and *same cbsa*.

As the model with the least functional form restrictions on distance, column (4) represents our specification of choice. Following Greene (2009), we interpret the results for the interaction terms in this non-linear model graphically by calculating the average predicted effect of a 0 to 1 transition for each of our variables - *same country*, *same state* and *same cbsa*. Specifically, by carrying out this exercise for the subsamples from different time periods, we plot the predicted effects for different periods in Figure 2. Examining the ratio of the predicted effect for the case where a specific variable (such as *same country*) is set to 1 vs. 0 helps comment on the economic magnitude of the trend. For example, the ratio between the cases with *same country* being 1 versus 0 *increases* from 1.42 in 1975-1980 (predicted probabilities of 5.0 in a million for *same country = 1* vs. 3.5 in a million for *same country = 0*) to 1.66 in 2000-2004 (4.4 in a million vs. 2.7 in a million). On the other hand, the ratio between the cases with *same state* being 1 versus 0 *decreases* from 1.38 in 1975-1980 (citation probabilities of 5.9 in a million for *same state = 1* vs. 4.3 in a million *same state = 0*) all the way down to 1.15 in 2000-2004 (predicted citation probabilities of 4.3 in a million vs. 3.8 in a million). We have offered a similar chart for CBSA for completeness; however, as our distance variable is based on the same single latitude and longitude value for all data from a city, we consider it too noisy to very reliably disentangle micro-level distance effects from CBSA effects. We therefore suggest extra caution in interpreting the findings regarding the CBSA effect, and largely treat that as a control variable for our purposes rather than discussing it extensively.

By employing interaction terms based on the *period* variable in the analysis above, we have been able to formally examine whether there are any long-term trends in knowledge diffusion patterns. However, for readers interested in more precise period-by-period findings that do not impose linear restrictions, in column (5) we report findings from interacting the geography variables with indicators corresponding to the six 5-year periods comprising the overall 30-year period 1975-2004 from which our cited patents sample is drawn. The omitted (reference) period for these interactions is the first period, namely, 1975-79. We observe that, relative to the border effects prevalent during 1975-79, the country border effects are stronger in four of the five subsequent periods (and statistically indistinguishable in the remaining one). On the other hand, relative to the same baseline period, the state border effects are weaker in three of the five subsequent periods (and statistically indistinguishable in the two remaining ones). Overall, these results are consistent with the time trends documented in the earlier analysis: country-level localization seems to have strengthened over time while state-level localization has diminished.[16]

4.4. Inventor vs. examiner citations

Having discussed some of the findings using our preferred regression approach that allows us to disentangle political border vs. geographic proximity effects in observed patterns of knowledge spillover localization, we now revisit the issue that many citations are generated not by inventors but by patent examiners. For easy interpretation, logistic regression estimates for the inventor versus examiner subsample are first separately reported in columns (1)-(3) and columns (4)-(6) respectively of Table 7. This is followed by the last three columns that examine the two subsamples together in a single multinomial logistic framework in order to allow more rigorous inference. As noted in section 3.4, unavailability of pre-2001 data on the inventor vs. examiner-added citation distinction restricts our analysis to patents receiving a meaningful number of post-2000 citations and thus reduces the number of observations considerably compared to Table 5. This restriction also makes it impossible for the inventor versus examiner distinction to shed further light on the long-term time trends.

We start with side-by-side analyses of the inventor-added citations subsample (which includes not only actual citations but also controls matched to those) in columns (1)-(3) and the examiner-added citations subsample (which also includes actual citations and corresponding controls) in columns (4)-(6) in Table 7. We begin by comparing columns (1) and (4) so as to assess the overall geographic effect. When not simultaneously accounting for political borders, the role of proximity appears to be

[16] One might wonder about the extent to which the temporal patterns in border effects could be an artifact of changes in sectoral composition in patenting activity. This turns out not to be important. In analysis not detailed here, we find that the increase in country-level localization as well as the drop in state-level localization is a more general temporal phenomenon than being driven simply by increasing dominance of specific sectors.

confirmed as the coefficient on *ln(distance = 1)* variable is almost twice as large for citations added by inventors than by examiners. However, simultaneously considering all our geographic units representing political borders and spatial proximity in the remaining columns questions whether a large part of the overall effect is truly comprised of an impact of proximity *per se*. The difference between the coefficients on *ln(distance + 1)* for columns (2) and (5) does not appear that large relative to the big gap between the two in columns (1) versus (4). Similarly, the coefficients on *same CBSA* are not too different between columns (2) and (5) and in fact become virtually indistinguishable between columns (3) and (6) as distance is accounted for non-parametrically in the form of our full set of indicator variables. This reinforces the concerns expressed by Thompson and Fox-Kean (2005) and Thompson (2006) that knowledge spillovers reported in earlier studies might to a significant extent have been a manifestation of the USPTO classification system (or, for that matter, *any* formal classification system) only imperfectly capturing true technological relationships across patents.

The previous finding on the influence of political borders, however, is not diluted as much by the inventor/examiner distinction. Comparing the estimates for *same country* in columns (2) and (3) with those in columns (5) and (6) respectively, examiner-added citations in fact show no country-level localization while the effect for inventor citations is economically and statistically highly significant. The state-level result also remains robust. However, although the *same state* coefficient is statistically insignificant for the examiner-added citation analysis in column (5), it turns significant for the preferred specification in column (6) once distance is accounted for in a non-parametric fashion. Still, the magnitude of the coefficient on *same state* in column (3) remains considerably larger than that in column (6). The relative weakness of the *same state* effect in this analysis might in part be due to the limited timeframe of the inventor-vs-examiner distinction, as we are able to observe patents only in the latter portion of our 30-year window. Recall from the earlier time trends analysis that the same-state effect was anyway weaker during this time period when considering all citations together. If we did have the inventor/examiner distinction data available for the earlier part of our sample, it is conceivable that the state effects might have been stronger.

Directly comparing estimates from non-linear regressions employing different subsamples (inventor vs. examiner citations) relies on our earlier observation that these estimates have a natural interpretation in percentage terms because citations are rare events. While intuitive, this approach leaves two open questions. First, given the different control groups for inventor and examiner subsamples, this direct comparison could be problematic. Secondly, it is not straightforward to test hypotheses regarding statistical distinguishability of estimates across different models. To address these concerns, we pool the two subsamples and run the analysis as a single (weighted) multinomial regression for the three mutually exclusive and exhaustive outcomes possible for any pair of random

patents: *Inventor Citation*, *Examiner Citation* and *No Citation*. The multinomial logit results reported in columns (7)-(9) take *No Citation* as the omitted (reference) category for ease of comparison with earlier column values, but we relied upon equivalent models taking *Examiner Citation* as the omitted category for testing hypotheses comparing coefficient values across inventor and examiner citations.

We are now able to formally test the extent to which the geography-related effects for both political borders and spatial proximity are robust to the inventor/examiner distinction. The findings remain qualitatively the same. In particular, most of the distinction between the coefficients on *ln(distance + 1)* between the *Inventor Citation* outcome and *Examiner Citation* outcome disappears in going from column (7) to column (8). In contrast, the coefficients for *same country* and *same state* remain much stronger for the *Inventor Citation* outcome than for *Examiner Citation* even in columns (8) or (9). The only distinction from before is that even the *same CBSA* effect is now significantly stronger for the *Inventor Citation* case than *Examiner Citation* case, although the extent of this difference is still at least somewhat smaller than for *same state* effect and much smaller than the *same country* effect. Thus our main qualitative finding – that the inventor versus examiner citation distinction dilutes the border effects less than the geographic proximity effects – continues to hold. Combining with the earlier sections, we therefore find that border effects are robust and not just an artifact of the geographic proximity of inventors or geographic distribution of technological activity.

5. Discussion, Caveats and Conclusion

The key contribution of this study is employing a novel regression framework based on choice-based sampling to *simultaneously* consider the impact of different geopolitical units in order to disentangle true border effects from geographic proximity effects. We also account for technological relatedness between the citing and cited patents at multiple levels of granularity, and further employ an identification approach inspired by Thompson (2006) to address concerns about unobserved aspects of technological relatedness. A robust finding of our study is that, on average, country and state borders serve as constraints on knowledge diffusion *even after accounting for geographic proximity* in the form of metropolitan co-location and geographic distance. We document that the findings are robust to examining only near-border samples in a variety of ways and are also not driven by just one or two specific states or sectors. In fact, application of the alternate identification strategy using the inventor/examiner distinction in citations only strengthens this finding regarding an independent effect of borders.

The finding that national borders have a strong effect might not be too surprising. The literature on international trade already suggests several border-related variables one could consider for digging deeper, such as linguistic, cultural, political and economic differences between countries. Indeed, in analysis not reported here, we found knowledge flows from the United States to other English-speaking

countries to be particularly strong even after accounting for the effect of geographic distance. A more general treatment of variables used in gravity-type models from international economics would, however, require a sample where not just the citing but also the cited patents are drawn from multiple countries.

What is perhaps surprising about even the country-level finding is that it has only grown stronger over a period which has seen the rise of information technology in general and the internet in particular. We take this as indication that inventors in the U.S. are disproportionately relying upon knowledge generated within the U.S. during a time when the fraction of patents originating overseas has been growing. However, this finding could also somehow be an artifact of our data. For example, absent the availability of inventor vs. examiner citation distinction for the full period, we cannot rule out a possibility that U.S. is simply getting more specialized in a way not captured by the formal technological classification system. If this is indeed true, we would observe U.S. patents as increasingly citing other domestic patents even in the absence of any true geographic trends associated with national borders.

Turning to the counter-intuitive finding on an independent state border effect, it is worth noting that a few studies (e.g., Holmes, 1998) have found state-level effects in related contexts before. In fact, a forthcoming paper by Belenzon and Schankerman (2012) documents state borders effects specifically in for knowledge diffusion. However, they examine only knowledge arising from universities, leading them to conclude that policies promoting within-state knowledge diffusion from state-funded public universities could be a driver of this finding. Our study reveals that the mechanisms driving state border effects like these might be more general, since they apply even to diffusion of knowledge arising in private companies. Connecting back to the growing literature emphasizing diffusion of knowledge through localized networks of people and organizations, future researchers should find it interesting to examine how different kinds of such formal or informal networks might be originating and operating at different geographic levels. It also seems worth investigating further exactly which of these mechanisms might have weakened relatively recently, leading us to observe that the state localization effect has – unlike the country effect - declined over time and become quite weak by the end of our sample period.

While further exploration of institutional and policies seems promising for future research, we cannot rule out that at least some of the effects we find will turn out not to be robust using alternate research designs. At a minimum, therefore, we view our study as an initial inquiry into border-related diffusion effects for flow of ideas, paralleling analogous studies looking to disentangling different border vs. proximity effects for flow of goods in the context in international trade (McCallum, 1995; Wolf, 2000; Anderson and Wincoop, 2003; Hillberry and Hummels, 2003, 2008). Further progress toward unpacking the geography of knowledge spillovers would also help refine existing theoretical models of innovation, entrepreneurship and growth, ultimately leading to more effective innovation-related policies.

References

Alcacer, J., and M. Gittelman. 2006. "Patent Citations as a Measure of Knowledge Flows: The Influence of Examiner Citations." *Review of Economics and Statistics* **88**(4) 774-779.

Almeida, P., and B. Kogut. 1999. "Localization of Knowledge and the Mobility of Engineers in Regional Networks." *Management Science* **45**(7) 905.

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, Mass.: Harvard University Press.

Anderson, J.E., and E.V. Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review* **93** 170-192.

Arzaghi, M., and J.V. Henderson. 2008. "Networking off Madison Avenue." *The Review of Economic Studies* **75** 1011-1038.

Audretsch, D., and M. Feldman. 1996. "R&D Spillovers and the Geography of Innovation and Production." *American Economic Review* **86**(3) 630-640.

Belenzon, S., and M. Schankerman. 2012. "Spreading the Word: Geography, Policy and University Knowledge Diffusion." *Review of Economics and Statistics*, Forthcoming.

Branstetter, L.G. 2001. "Are Knowledge Spillovers International or Intranational in Scope?" *Journal of International Economics* **53**(1) 53-79.

Breschi, S., and F. Lissoni. 2001. "Knowledge Spillovers and Local Innovation Systems: A Critical Survey." *Industrial and Corporate Change* **10**(4) 975-1005.

Coe, D.T., E. Helpman and A.W. Hoffmaister. 2009. "International R&D Spillovers and Institutions." *European Economic Review* **53**(7) 723-741

Duguet, E., and M. MacGarvie. 2005. "How Well Do Patent Citations Measure Knowledge Spillovers? Evidence from French Innovation Surveys." *Economics of Innovation and New Technology* **14**(5) 375.

Ellison, G., E. Glaeser, and W. Kerr. 2010. "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns." *American Economic Review* 100(June 2010):1195-1213.

Greene, W.H. 2003. *Econometric Analysis*, 5th ed. Upper Saddle River, N.J.: Prentice Hall.

Greene, William, "Testing Hypotheses About Interaction Terms in Nonlinear Models," Working paper, New York University (2009).

Grossman, G., and E. Helpman. 1991. *Innovation and Growth in the World Economy*. Cambridge, Mass.: MIT Press.

Henderson, R., A. Jaffe, and M. Trajtenberg. 2005. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment." *American Economic Review* **95**(1) 461-464.

Hillberry, R., and D. Hummels. 2003. "Intranational Home Bias: Some Explanations." *The Review of Economics and Statistics* **85**(4) 1089-1092.

Hillberry, R., and D. Hummels. 2008. "Trade Responses to Geographic Frictions: A Decomposition Using Micro-data." *European Economic Review* **52**(3) 527-550.

Holmes, T.J. 1998, "The Effect of State Policies on the Location of Manufacturing: Evidence from State Borders." *Journal of Political Economy* **106**(4) 667-705.

Jaffe, A.B. 1989. "Real Effects of Academic Research." *American Economic Review* **79**(5) 957.

Jaffe, A.B., and M. Trajtenberg. 2002. *Patents, Citations & Innovations: A Window on the Knowledge Economy*. Cambridge, Mass.: MIT Press.

Jaffe, A.B., M. Trajtenberg and R. Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* **434** 578-598.

Keller, W. 2002. “Geographic Localization of International Technology Diffusion.” *American Economic Review* **92**(1) 120-142.

Kerr, W. and S.D. Kominers. 2011. “Agglomerative Forces and Cluster Shapes.” Harvard Business School Working Paper 11-061.

Krugman, P. 1991. *Geography and Trade*. Leuven, Belgium: Leuven University Press.

Lai, R., A. D'Amour, and L. Fleming. 2009. “The Careers and Co-authorship Networks of U.S. Patent Holders Since 1975” Harvard Institute for Quantitative Social Science.

Lampe, R. 2011. Strategic Citation. *The Review of Financial Studies*, forthcoming.

Manski, C.F., and S.R. Lerman. 1977. “The Estimation of Choice Probabilities from Choice Based Samples.” *Econometrica* **45**(8) 1977-88.

Manski, C.F., and D. MacFadden. 1981. “Alternative Estimators and Sample Designs for Discrete Choice Analysis.” In C. Manski and D. McFadden, eds., *Structural analysis of discrete data with econometric applications*. Cambridge, Mass.: MIT Press.

McCallum, J. 1995. “National Borders Matter: Canada-U.S. Regional Trade Patterns.” *American Economic Review* **85**(3) 615-623.

Peri, G. 2005. “Determinants of Knowledge Flows and their Effect on Innovation.” *Review of Economics and Statistics* **87**(2) 308-322.

Romer, P.M. 1990. “Endogenous Technological Change.” *Journal of Political Economy* **98**(5 Part 2) S71-S102.

Rosenthal, S. and W. Strange. 2001. “The Determinants of Agglomeration.” *Journal of Urban Economics* **50** 191-229.

Rosenthal, S., and W. Strange. 2003. “Geography, Industrial Organization, and Agglomeration.” *The Review of Economics and Statistics* **85**(2) 377-393.

Rysman, M., and T. and Simcoe. 2008. Patents and the Performance of Voluntary Standard-Setting Organizations. *Management Science*. **54**(11) 1920-1934.

Singh, J. 2005. “Collaborative Networks as Determinants of Knowledge Diffusion Patterns.” *Management Science* **51**(5) 756-770.

Singh, J. 2007. “Asymmetry of Knowledge Spillovers between MNCs and Host Country Firms.” *Journal of International Business Studies* **38**(5) 764-786.

Sorenson, O., and L. Fleming. 2004. “Science and the Diffusion of Knowledge.” *Research Policy* **33**(10) 1615-1634.

Thompson, P. 2006. “Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-Added Citations.” *Review of Economics and Statistics* **88**(2) 383-389.

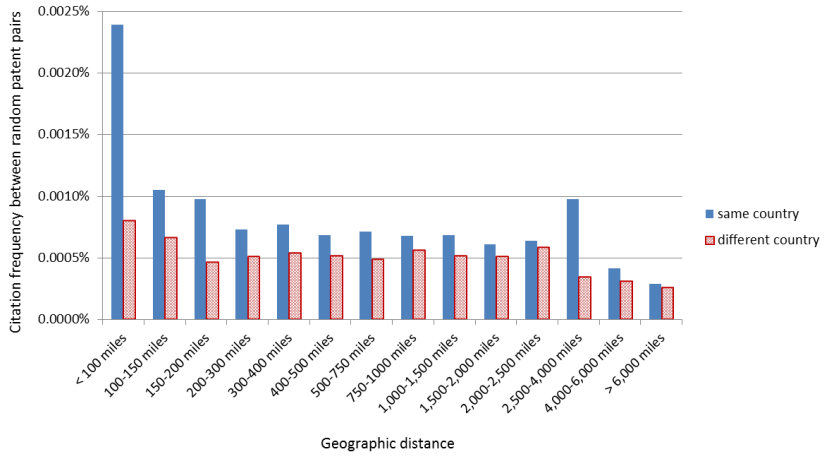
Thompson, P., and M. Fox-Kean. 2005. “Patent Citations and the Geography of Knowledge Spillovers: A Reassessment.” *American Economic Review* **95**(1) 450-460.

Trajtenberg, M. 2006. “The ‘Names Game’: Harnessing Inventors’ Patent Data for Economic Research.” NBER Working Paper 12479.

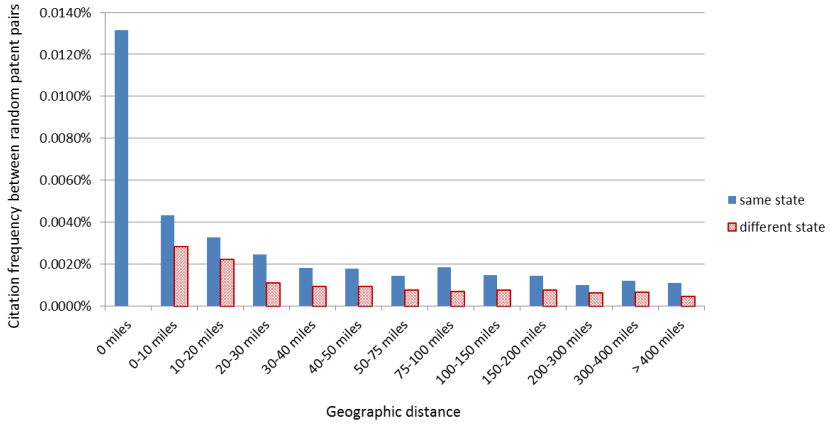
Wolf, H.C. 2000. “Intra-national Home Bias in Trade.” *Review of Economics and Statistics* **82**(4) 555-563.

Figure 1. Graphical depiction of the role of geography in patent citation likelihood

(i) Country borders vs. distance between inventor cities



(ii) State borders vs. distance between inventor cities



(iii) CBSA boundaries vs. distance between inventor cities

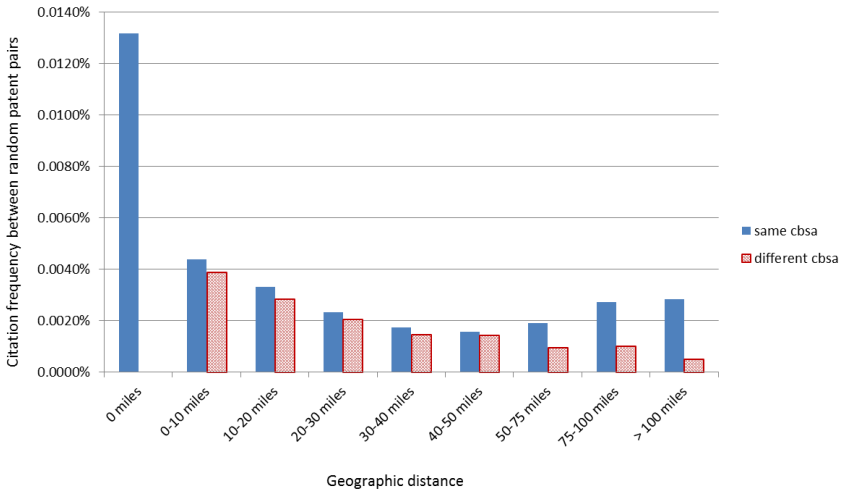
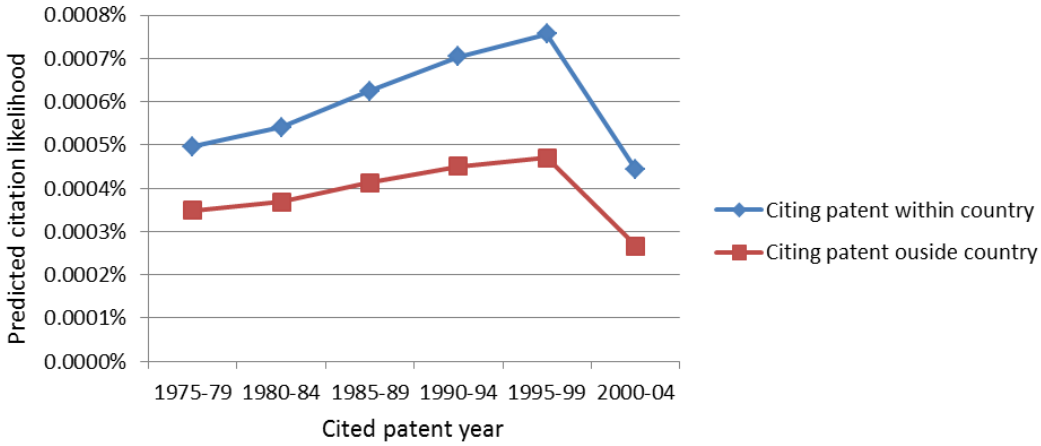
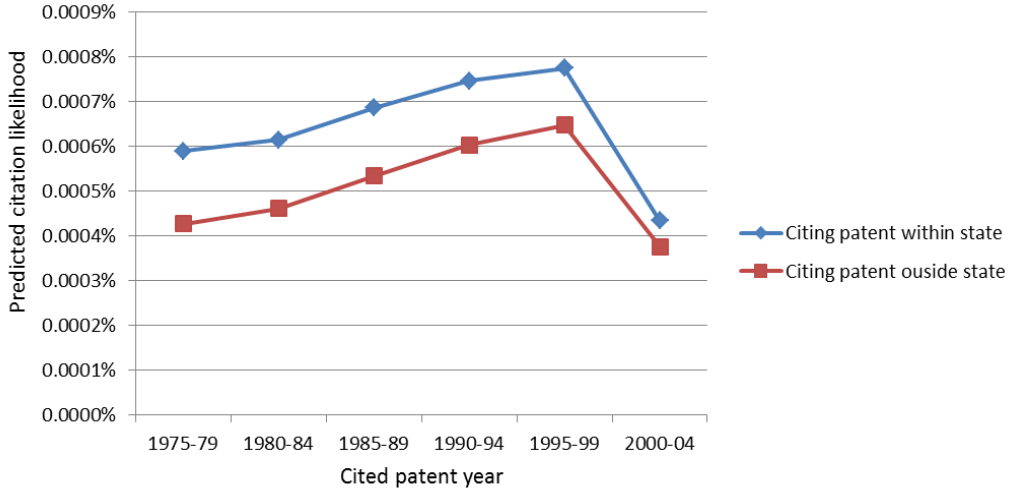


Figure 2. Predicted probabilities across different time periods

(i) Country border effect after accounting for other geographic levels



(ii) State border effect after accounting for other geographic levels



(iii) CBSA boundary effect after accounting for other geographic levels

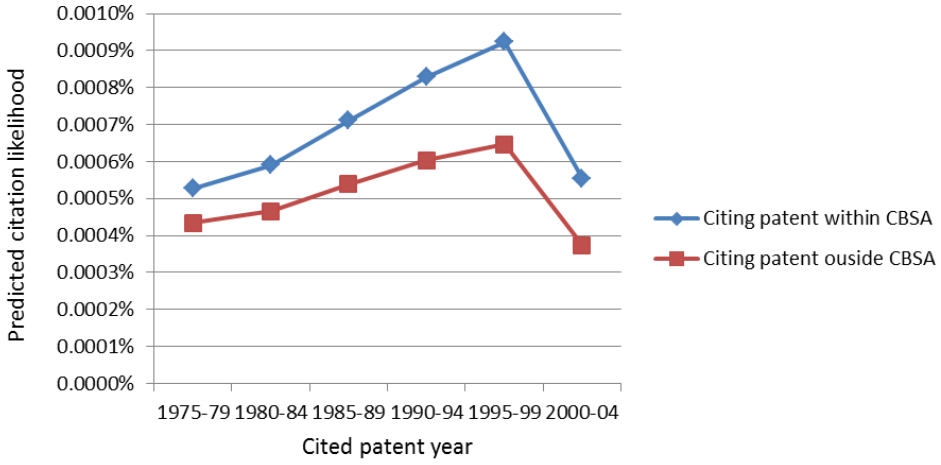


Table 1. Replicating findings from previous studies

	Our matched sample				Jaffe, Trajtenberg & Henderson sample				Thompson & Fox-Kean 3-digit sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Citations sample	Intraregion citations	Intraregion controls	Ratio (2)/(3)	Citations sample	Intraregion citations	Intraregion controls	Ratio (6)/(7)	Citations sample	Intraregion citations	Intraregion controls	Ratio (10)/(11)
Country-level analysis	4,007,217	74.7%	57.6%	1.30	7,759	68.0%	61.4%	1.11	7,627	68.6%	55.6%	1.23
State-level analysis	4,007,217	13.4%	6.2%	2.16	7,759	9.7%	5.1%	1.90	7,627	7.8%	5.0%	1.55
Metropolitan-level analysis	4,007,217	7.6%	2.6%	2.92	7,759	6.6%	1.7%	3.88	7,627	5.2%	3.5%	1.50

Notes: The Jaffe, Trajtenberg and Henderson (JTH) numbers reported here were calculated based on pooling of results for their different subsamples primarily using information available in their Table III in a manner similar to that reported by Thompson & Fox-Kean (TFK). The TFK sample statistics are for the first sample they construct by employing three-digit technology matching to be comparable to JTH. While TFK subsequently construct other samples using more fine-grained technology matching, we instead rely on regression models to similarly account for technology more finely. Using formal t-tests confirmed that difference of means between incidences of geographic co-location for actual citations versus corresponding controls were statistically significant in all cases, so the t-statistics have not been reported to conserve space.

Table 2. Further investigation of the state border effect

	Cited patent from near a state border				Cited patent from near a state border and citing patent from focal state dyad				Cited patent from near a state border and citing patent from focal state dyad as well as same CBSA as cited patent			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Citations sample	Intraregion citations	Intraregion controls	Ratio (2)/(3)	Citations sample	Intraregion citations	Intraregion controls	Ratio (6)/(7)	Citations sample	Intraregion citations	Intraregion controls	Ratio (10)/(11)
Country-level analysis	996,627	74.9%	58.4%	1.28								
State-level analysis	996,627	7.1%	2.7%	2.63	93,703	68.4%	55.7%	1.23	40,784	87.2%	82.8%	1.05
Metropolitan-level analysis	996,627	6.1%	2.2%	2.77	93,703	55.8%	38.0%	1.47				

Notes: To ensure that within-state localization reported above is not just a distance effect driven by cited patents in a state's interior, columns (1)-(4) carry out the JTH-style analysis using a subsample of our matched sample where the distance of the cited patent's originating town or city to the closest state border is not more than 20 miles. In columns (5)-(8), the set of actual citations is restricted to those arising either within the cited patents or in the closest neighboring state – with the set of control citations to use as a benchmark also being regenerated based on a matching with all potentially citing patents within these two states using their application year and a three-digit technology class. In columns (9)-(12), as an additional robustness check to distinguish the effect of metropolitan co-location from state borders, analysis has been further restricted to cited patents originating in CBSAs that cross state borders and having both actual as well as corresponding control citations arise within the CBSA (with one or both of them potentially still crossing the state border). Difference of means between incidences of geographic co-location for actual citations versus corresponding controls were statistically significant in all cases, so the t-statistics have not been reported to conserve space.

Table 3. Distinguishing different time periods as well as citations added by inventors vs. examiners

	Full matched sample				Inventor-added citation subsample				Examiner-added citation subsample			
	(1) Citations sample	(2) Intraregion citations	(3) Intraregion controls	(4) Ratio (2)/(3)	(5) Citations sample	(6) Intraregion citations	(7) Intraregion controls	(8) Ratio (6)/(7)	(9) Citations sample	(10) Intraregion citations	(11) Intraregion controls	(12) Ratio (10)/(11)
Country-level analysis												
1975-1979	262,657	66.7%	59.0%	1.13								
1979-1984	307,090	67.5%	56.5%	1.19								
1985-1989	504,546	73.4%	58.0%	1.27								
1990-1994	941,141	76.1%	57.3%	1.33	360,541	85.1%	57.5%	1.48	154,186	59.7%	55.1%	1.08
1995-1999	1,496,672	77.0%	58.2%	1.32	917,811	85.4%	59.0%	1.45	495,037	62.1%	56.6%	1.10
2000-2004	495,111	75.0%	55.9%	1.34	288,992	85.4%	57.1%	1.50	203,926	60.3%	54.2%	1.11
State-level analysis												
1975-1979	262,657	8.9%	4.6%	1.93								
1979-1984	307,090	9.4%	4.5%	2.09								
1985-1989	504,546	11.1%	4.9%	2.27								
1990-1994	941,141	13.4%	5.8%	2.31	360,541	15.7%	6.1%	2.57	154,186	9.5%	5.6%	1.70
1995-1999	1,496,672	14.7%	7.1%	2.07	917,811	16.8%	7.2%	2.33	495,037	10.8%	6.9%	1.57
2000-2004	495,111	16.3%	7.3%	2.23	288,992	19.3%	7.5%	2.57	203,926	12.1%	7.1%	1.70
Metropolitan-level analysis												
1975-1979	262,657	5.3%	2.1%	2.52								
1979-1984	307,090	5.6%	2.1%	2.67								
1985-1989	504,546	6.7%	2.1%	3.19								
1990-1994	941,141	8.0%	2.5%	3.20	360,541	9.4%	2.6%	3.62	154,186	5.3%	2.3%	2.30
1995-1999	1,496,672	7.9%	2.8%	2.82	917,811	9.1%	2.9%	3.14	495,037	5.6%	2.7%	2.07
2000-2004	495,111	9.4%	2.9%	3.24	288,992	11.4%	3.0%	3.80	203,926	6.7%	2.7%	2.48

Notes: Columns (1) through (4) employ exactly the same matched sample as the corresponding columns in the previous table except that the analysis has now been broken up into six five-year time periods based on the application year of the cited patent. The sample size drops during 2000-2004 because, while the first five periods employ the full ten-year citation window, the observed window is shorter for patents in this period given that we only observe citing patents until 2009. Columns (5) through (8) are based only on the subsample of citations added by inventors and their corresponding controls, and columns (9) through (12) are based only on the subsample of citations added by examiners and their corresponding controls. Since this distinction is only available for citing patents post-2001, this analysis is done only for the cited patent originating periods for which the citation window overlaps with availability of the inventor versus examiner distinction information for citations.

Table 4. Definitions of variables used during regression analysis

Political border variables	
same country	Indicator variable that is 1 if the citing and cited patents originate in the same country, i.e., the U.S. (given that our cited patent sample is drawn from the U.S. only)
same state	Indicator variable that is 1 if the two patents originate in the same state (within the U.S.)
Spatial proximity variables	
same cbsa	Indicator variable that is 1 if the citing and cited patents originate from inventors located in the same Core Based Statistical Area (CBSA) as per the 2003 definition of CBSAs by the U.S. Office of Management and Budget (CBSA definitions are meant to cover reasonable commuting distances and replace the prior MSA/PMSA/CMSA definitions for defining U.S. metropolitan areas in a more standardized fashion.)
distance	Distance, in miles, between the cities where the first inventors of the source and destination patents live (calculated as spherical distance between the latitude and longitude values for these cities)
Technological relatedness variables	
same tech category	Indicator variable that is 1 if the two patents belong to the same 1-digit NBER technology category
same tech subcategory	Indicator variable that is 1 if the two patents belong to the same 2-digit NBER technical subcategory
same tech class	Indicator variable that is 1 if the two patents belong to the same 3-digit USPTO primary technology class
relatedness of tech classes	Likelihood of citation (scaled by 100) between random patents with the same respective 3-digit primary technology classes that the focal cited and citing patents belong to
overlap of tech subclasses	Natural logarithm of one plus the number of overlapping 9-digit technology subclasses under which the patents are categorized
Patent-level variables	
references to other patents	Number of references the cited patent makes to other patents
references to non-patent materials	Number of references the cited patent makes to published materials other than patents
number of claims	Number of claims the cited patent makes
period	A sequential number representing which of our six five year time-periods the focal cited patent belongs to: 1975-79 being period 0, 1980-84 being period 1, 1985-89 being period 2, 1990-94 being period 3, 1995-99 being period 4 and 2000-04 being period 5.

Table 5. Simultaneous consideration of political borders and spatial proximity

	(1) Full Sample	(2) Full Sample	(3) Full Sample	(4) Full Sample	(5) Full Sample	(6) Full Sample	(7) Full Sample	(8) Near-Border Sample	(9) Excluding California	(10) Excl Comp & Comm
same country			0.863***	0.769***	0.766***	0.535***	0.451***	0.513***	0.447***	0.441***
			(0.006)	(0.006)	(0.006)	(0.011)	(0.016)	(0.032)	(0.021)	(0.024)
same state				0.750***	0.405***	0.109***	0.228***	0.253***	0.346***	0.230***
				(0.017)	(0.024)	(0.027)	(0.027)	(0.047)	(0.049)	(0.044)
same cbsa					0.769***	0.456***	0.337***	0.295***	0.433***	0.407***
					(0.030)	(0.029)	(0.032)	(0.071)	(0.054)	(0.045)
ln(distance + 1)	-0.364***	-0.271***				-0.137***				
	(0.001)	(0.003)				(0.005)				
distance 0 (i.e., same city)							1.665***	1.896***	1.661***	1.910***
							(0.067)	(0.207)	(0.103)	(0.095)
distance 0-10 miles							1.129***	1.342***	1.203***	1.236***
							(0.057)	(0.112)	(0.083)	(0.086)
distance 10-20 miles							0.990***	1.020***	0.923***	1.064***
							(0.049)	(0.088)	(0.073)	(0.070)
distance 20-30 miles							0.836***	0.517***	0.720***	0.919***
							(0.055)	(0.108)	(0.081)	(0.079)
distance 30-40 miles							0.552***	0.239*	0.386***	0.660***
							(0.071)	(0.125)	(0.116)	(0.100)
distance 40-50 miles							0.613***	0.433***	0.657***	0.802***
							(0.099)	(0.096)	(0.071)	(0.067)
distance 50-75 miles							0.595***	0.533***	0.583***	0.707***
							(0.040)	(0.066)	(0.050)	(0.052)
distance 75-100 miles							0.546***	0.529***	0.579***	0.665***
							(0.038)	(0.066)	(0.049)	(0.053)
distance 100-150 miles							0.599***	0.584***	0.614***	0.656***
							(0.033)	(0.053)	(0.040)	(0.048)
distance 150-200 miles							0.585***	0.553***	0.584***	0.670***
							(0.029)	(0.050)	(0.033)	(0.039)
distance 200-300 miles							0.479***	0.557***	0.491***	0.544***
							(0.029)	(0.043)	(0.035)	(0.042)
distance 300-400 miles							0.503***	0.520***	0.566***	0.567***
							(0.024)	(0.044)	(0.027)	(0.034)
distance 400-500 miles							0.479***	0.569***	0.508***	0.566***
							(0.024)	(0.048)	(0.029)	(0.036)
distance 500-750 miles							0.480***	0.494***	0.473***	0.519***
							(0.022)	(0.038)	(0.027)	(0.033)
distance 750-1000 miles							0.439***	0.483***	0.450***	0.484***
							(0.020)	(0.038)	(0.025)	(0.029)
distance 1000-1500 miles							0.419***	0.433***	0.441***	0.467***
							(0.020)	(0.038)	(0.025)	(0.030)
distance 1500-2000 miles							0.377***	0.400***	0.391***	0.405***
							(0.019)	(0.040)	(0.024)	(0.027)
distance 2000-2500 miles							0.368***	0.470***	0.417***	0.382***
							(0.019)	(0.038)	(0.025)	(0.028)
distance 2500-4000 miles							0.461***	0.487***	0.460***	0.507***
							(0.015)	(0.023)	(0.016)	(0.021)
distance 4000-6000 miles							0.112***	0.257***	0.154***	0.133***
							(0.010)	(0.027)	(0.011)	(0.012)
same tech category		1.103***	1.115***	1.111***	1.108***	1.106***	1.107***	1.088***	1.102***	0.893***
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.011)	(0.006)	(0.007)
same tech subcategory		1.298***	1.310***	1.300***	1.299***	1.297***	1.296***	1.298***	1.300***	1.460***
		(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.015)	(0.009)	(0.010)
same tech class		2.141***	2.154***	2.156***	2.145***	2.144***	2.144***	2.267***	2.215***	2.283***
		(0.016)	(0.014)	(0.016)	(0.015)	(0.015)	(0.014)	(0.025)	(0.017)	(0.016)
relatedness of tech classes		1.512***	1.604***	1.481***	1.518***	1.501***	1.502***	1.537***	1.650***	1.567***
		(0.129)	(0.104)	(0.127)	(0.112)	(0.116)	(0.104)	(0.193)	(0.138)	(0.124)
overlap of tech subclasses		1.687***	1.691***	1.686***	1.686***	1.684***	1.681***	1.716***	1.704***	1.845***
		(0.011)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.019)	(0.013)	(0.016)
ln(references to other patents + 1)		0.134***	0.135***	0.135***	0.136***	0.135***	0.135***	0.159***	0.149***	0.134***
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.012)	(0.006)	(0.007)
ln(references to non-patent materials + 1)		0.034***	0.034***	0.034***	0.033***	0.033***	0.033***	0.007	0.032***	0.039***
		(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.008)	(0.006)	(0.008)
ln(number of claims)		0.092***	0.092***	0.092***	0.093***	0.092***	0.092***	0.110***	0.095***	0.082***
		(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.010)	(0.006)	(0.006)
Period indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Citation lag indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-digit tech indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582	3,600,000	10,994,852	10,474,569
Pseudo-R2	0.0122	0.181	0.179	0.181	0.182	0.182	0.183	0.189	0.188	0.196
Wald chi2	124220	756991	785451	767431	759980	755320	763511	221768	616233	519519
Degrees of freedom	1	69	69	70	71	72	91	91	90	87

Notes: The unit of observation is pairs of patents representing actual or potential citations. The dependent variable is an indicator for whether or not the potentially citing patent actually cited the focal patent. A choice-based stratified sample is used, and a weighted logistic regression (WESML) approach is implemented using observation weights that reflect sampling frequency associated with different strata. The regression model also uses a constant term and indicator variables as indicated above, but these are not reported to conserved space and are available from the authors upon request. Robust standard errors are shown in parentheses, and are clustered on the cited patent. Asterisks indicate statistical significance (*** p<0.01, ** p<0.05, * p<0.1).

Table 6. Time trends in geographic knowledge diffusion patterns

	(1) Full Sample	(2) Full Sample	(3) Full Sample	(4) Full Sample	(5) Full Sample
same country		0.259***	0.140***	0.362***	0.249***
		(0.022)	(0.020)	(0.048)	(0.034)
same state		0.381***	0.367***	0.358***	0.332***
		(0.070)	(0.066)	(0.074)	(0.048)
same cbsa		0.616***	0.424***	0.125	-0.034
		(0.069)	(0.072)	(0.087)	(0.084)
ln(distance + 1)	-0.224***	-0.085***			
	(0.008)	(0.011)			
*period * same country*		0.089***	0.106***	0.021*	
		(0.007)	(0.004)	(0.011)	
*period * same state*		-0.089***	-0.048***	-0.035**	
		(0.017)	(0.016)	(0.015)	
*period * same cbsa*		-0.053***	-0.026	0.048**	
		(0.019)	(0.020)	(0.020)	
*period * ln(distance + 1)*	-0.011***	-0.017***			
	(0.002)	(0.003)			
*same country * period 1980-84*					0.061
					(0.080)
*same country * period 1985-89*					0.321***
					(0.051)
*same country * period 1990-94*					0.255***
					(0.047)
*same country * period 1995-99*					0.196***
					(0.043)
*same country * period 2000-04*					0.147***
					(0.053)
*same state * period 1980-84*					0.069
					(0.062)
*same state * period 1985-89*					-0.279**
					(0.133)
*same state * period 1990-94*					-0.011
					(0.060)
*same state * period 1995-99*					-0.171***
					(0.057)
*same state * period 2000-04*					-0.187***
					(0.072)
*same cbsa * period 1980-84*					0.113
					(0.114)
*same cbsa * period 1985-89*					0.513***
					(0.133)
*same cbsa * period 1990-94*					0.371***
					(0.100)
*same cbsa * period 1995-99*					0.412***
					(0.095)
*same cbsa * period 2000-04*					0.333***
					(0.119)
Distance-period indicators	No	No	No	Yes	Yes
Distance indicators	No	No	Yes	Yes	Yes
Period indicators	Yes	Yes	Yes	Yes	Yes
Citation lag indicators	Yes	Yes	Yes	Yes	Yes
Two-digit tech indicators	Yes	Yes	Yes	Yes	Yes
State indicators	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes
Number of observations	13,728,582	13,728,582	13,728,582	13,728,582	13,728,582
Pseudo-R2	0.181	0.183	0.183	0.183	0.183
Wald chi2	758828	754616	762307	778608	780425
Degrees of freedom	70	76	94	189	201

Notes: All notes from Table 5 apply here as well, except that regression coefficients for the control variables as well as for the distance-period and distance indicators (when applicable) are also omitted to further conserve space. As indicated, distance indicators are excluded in the first two models since a continuous distance variable has been directly included in those models.

Table 7. Inventor-added vs. examiner-added citations

	(1) Inventor Sample (Logit)	(2) Inventor Sample (Logit)	(3) Inventor Sample (Logit)	(4) Examiner Sample (Logit)	(5) Examiner Sample (Logit)	(6) Examiner Sample (Logit)	(7) Full Sample (Multinomial Logit)	(8) Full Sample (Multinomial Logit)	(9) Full Sample (Multinomial Logit)
Inventor Citation:									
same country		1.223***	0.753***					1.222***	0.754***
		(0.015)	(0.021)					(0.014)	(0.020)
same state		0.086***	0.246***					0.083***	0.243***
		(0.028)	(0.028)					(0.026)	(0.026)
same cbsa		0.417***	0.334***					0.411***	0.328***
		(0.035)	(0.041)					(0.032)	(0.037)
ln(distance + 1)	-0.352***	-0.166***					-0.353***	-0.168***	
	(0.004)	(0.007)					(0.004)	(0.007)	
Examiner Citation:									
same country					-0.034	-0.059		-0.024	-0.055**
					(0.027)	(0.038)		(0.015)	(0.023)
same state					-0.015	0.124**		0.002	0.118***
					(0.056)	(0.053)		(0.031)	(0.033)
same cbsa					0.408***	0.328***		0.316***	0.239***
					(0.077)	(0.077)		(0.044)	(0.049)
ln(distance + 1)				-0.173***	-0.147***		-0.163***	-0.140***	
				(0.006)	(0.013)		(0.004)	(0.008)	
Distance indicators	No	No	Yes	No	No	Yes	No	No	Yes
Period indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Citation lag indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-digit tech indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	4,651,156	4,651,156	4,651,156	3,377,722	3,377,722	3,377,722	5,828,778	5,828,778	5,828,778
Pseudo-R2	0.157	0.162	0.163	0.192	0.192	0.192	0.159	0.163	0.164
Wald chi2	254439	262810	267886	230040	232457	233936	466765	507039	511038
Degrees of freedom	66	69	88	66	69	88	132	138	176

Notes: All notes from Table 5 apply here as well, except that regression coefficients for the distance indicators as well as control variables are omitted. The first six columns employ weighted logistic regressions as before, but with only inventor-added citations and corresponding controls included in columns (1)-(3) and only examiner-added citations and corresponding controls included in columns (4)-(6). The last three columns employ weighted multinomial logistic regressions based on the combined sample. For multinomial logistic regressions, the likelihood of inventor-added as well as examiner-added citations is estimated using the no-citation case as the reference category. All analyses only include citing year 2001 onwards since inventor vs. examiner distinction is not available for earlier years. Given the citation window of at most 10 years, all cited patents originating pre-1991 therefore get dropped.

Appendix: Details of Our Sample Construction and Weights Calculation

A1. Basic Choice-Based Sampling

Choice-based sampling involves drawing a fraction (γ) of the "ones" and a smaller fraction (α) of "zeroes" from the population. The probability of a citation *conditional on a dyad being in the sample* follows from Bayes' rule:

$$\Lambda_i^{'} = \frac{\gamma \Lambda_i}{\gamma \Lambda_i + \alpha(1-\Lambda_i)} = \frac{\gamma}{\gamma + \alpha e^{-\beta X_i}} = \frac{1}{1+e^{-(\ln\left(\frac{\gamma}{\alpha}\right)+\beta X_i)}}$$

So the usual logistic estimation would lead to biased results (Greene, 2003). Since the functional form is still logistic, one way to correct the logit estimates is subtracting $ln(\gamma/\alpha)$ from the constant term. However, noting that such a correction is overly sensitive to the assumption of the logistic functional form being completely accurate, Manski and Lerman (1977) suggest instead the *weighted exogenous sampling maximum likelihood* (WESML) estimator obtained by maximizing the following weighted "pseudo-likelihood" function:

$$\ln L_w = \frac{1}{\gamma}\sum_{\{y_i=1\}} \ln(\Lambda_i) + \frac{1}{\alpha}\sum_{\{y_i=0\}} \ln(1-\Lambda_i) = -\sum_{i=1}^{n} w_i \ln(1+e^{(1-2y_i)x_i\beta})$$

where $w_i = (1/\gamma)\, y_i + (1/\alpha)(1-y_i)$. As Amemiya (1985, Section 9.5.2) demonstrates, consistency of WESML comes from the expected value of the weighted log likelihood turning out to be the same (except for a scaling factor) as the expected log likelihood for the same sample resulting through random (exogenous) sampling. WESML can be implemented using a logistic approach by "simulating" an exogenous sample by weighting each observation by the number of elements it represents from the population (i.e., by the reciprocal of the ex ante probability of inclusion of an observation in the sample). An appropriate estimator of the asymptotic covariance matrix is White's robust "sandwich" estimator. Strictly speaking, WESML is not statistically "efficient" (Imbens and Lancaster, 1996). Nevertheless, efficiency issue can be mitigated by employing sufficiently large samples.

A2. Combining Choice-Based Sampling with Stratification on Explanatory Variables

In basic choice-based sampling, the "zeroes" are all drawn from the $y = 0$ population with a uniform sampling rate (α). This approach can be generalized to obtain additional benefits from stratification on key explanatory variables—that is, allowing "α" to vary across different $y = 0$ subpopulations (Manski and McFadden, 1981; Amemiya, 1985, Ch 9). Let us define z as a label for different strata that takes values *1, 2, …, T*, and note that

$$\begin{aligned}\Pr(z = z_i \ and\, y = y_j \mid x = x_i) &= \Pr(z = z_i \mid x = x_i)\Pr(y = y_j \mid z = z_i \ and\, x = x_i) \\ &= \Pr(z = z_i \mid x = x_i)\Pr(y = y_j \mid x = x_i)\end{aligned}$$

The second equality comes by assuming that the vector **x** includes all information about z that affects outcome y—that is, **x** is a sufficient statistic for z. (In our settings, this means our controls sufficiently capture technology- and year-related effects on citation likelihood.) Defining the logistic outcome as $v = (z = z_i$ and $y = y_i)$ rather than just y, the log-likelihood function with exogenous (random) sample would be

$$\begin{aligned}\ln L &= \sum_{i=1}^{n} \ln\left[\Pr(z = z_i \ and\, y = y_i \mid x_i)\right] \\ &= \sum_{i=1}^{n} \left\{ y_i \ln\left[\Pr(z = z_i \mid x_i)\Lambda(x_i\beta)\right] + (1-y_i)\ln\left[\Pr(z = z_i \mid x_i)(1-\Lambda(x_i\beta))\right]\right\}\end{aligned}$$

This forms the basis for deriving the pseudo-likelihood function for choice-based sampling with stratification. As per the WESML method, each log-likelihood function term needs to be weighted by the inverse of the ex ante probability of that observation being included in the sample. These weights can still be computed as long as the sample as well as population counts for each stratum are known. Once we have the weights w_{tj} corresponding to $z = t$ ($t = 1, 2, \ldots, T$) and $y=j$ ($j = 0, 1$), the required pseudo-likelihood function is given by

$$\ln L_w = \sum_{i=1}^{n} \left\{ y_i w_{z_i 1} \ln\left[\Pr(z = z_i \mid x_i)\Lambda(x_i\beta)\right] + (1-y_i) w_{z_i 0} \ln\left[\Pr(z = z_i \mid x_i)(1-\Lambda(x_i\beta))\right] \right\}$$

$$= C - \sum_{i=1}^{n} w_i \ln\left(1 + e^{(1-2y_i)X\beta}\right)$$

where $w_i = y_i w_{z_i 1} + (1-y_i) w_{z_i 0}$ and $C = \sum_{i=1}^{n} w_i \ln\left[\Pr(z = z_i \mid x_i)\right]$.

Since C is independent of β, it can be ignored. Thus, a weighted logistic estimation can again be used, with the weights given by w_i. (Note that the weights now depend not just on y but also on the stratum z_i.)

A3. Applying WESML to (Extended) Matched Samples

This approach can be extended to matched samples such as the one we have constructed following JTH. For a given cited patent, since the matched patent is drawn randomly from the year and technology class of an actual citing patent, we can interpret each {citing year, citing class} combination as a different stratum and calculate the implied sampling rates based on the sample and population counts for each stratum to determine appropriate weights.

However, the matched sample is not representative of the population since the {citing year, citing class} combinations for which no actual citations ("ones") exist are ignored from the point of view of the potential citations ("zeroes"). To ensure the strata are mutually exclusive and exhaustive while still keeping their number manageable, we create (for each cited patent) a new observation by randomly selecting one potentially citing patent for each year (in the 10-year window) belonging to one of the technology classes from which no citation occurs (in that year). The weight for each of these is computed using the implied sampling rates for random draws from these subpopulations. An example should clarify the sample construction. One of our cited patents is 4205881, applied for in 1980 and in tech class 299. It receives two citations during 10 years: from 4441761 {year 1982, class 299} and 953915 {1989, 299}. Therefore patent pairs (4205881, 4441761) and (4205881, 4953915) represent actual citations ("ones") included with a weight of 1 (as we include all citations, i.e., set γ = 1). In JTH-based matching, citing patent 4441761 was matched to control patent 4402550 {year 1982, class 299}. In year 1982 and class 299, there were 92 potentially matching patents from which patent 4402550 was chosen through a random draw. So the observation (4205881, 4402550) was included as a control pair ("zero") with a weight of 92. Similarly, citing patent 4953915 mentioned above was matched to control patent 4974907 {1989, 299}. In year 1989 and class 299, there were 59 potential matches from which 4974907 was chosen. So the observation (4205881, 4974907) was included as a control pair ("zero") with a weight of 59. Finally, for each of the year 1981 through 1990, we selected a random potentially citing patent, constrained not to be from technology class 299 for the years 1982 and 1989 (as class 299 is already included in finer strata above just for these two years). The range of weights for these 10 observations ended up being between 61,578 and 99,371, depending on the number of eligible patents in the given citing year.