Patent Citations and the Geography of Knowledge Spillovers:
Disentangling the Role of State Borders, Metropolitan Boundaries and Distance

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

Matt Marx
MIT Sloan School of Management
50 Memorial Drive, E52-561
Cambridge, MA 02142
+1 617 253 5539
mmarx@mit.edu

Lee Fleming
Harvard Business School
Morgan Hall 485
Boston, MA 02163
+1 617 495 6613
lfleming@hbs.edu

January 2010

We thank INSEAD, MIT Sloan School and Harvard Business School for funding this research. We are grateful to Ajay Agrawal, James Costantini, Pushan Dutt, Ilian Mihov and seminar participants at INSEAD for very helpful feedback. Any errors remain our own.

Abstract

We employ a regression framework based on choice-based sampling to estimate the probability of knowledge flow, measured using patent citations. This serves to extend research on the geography of knowledge spillovers, which has hitherto analyzed metropolitan, state or national effects only separately, through a simultaneous consideration of these geographic units. Fine-grained inventor location data is used to further disentangle the role of different geopolitical boundaries and distance. In addition to national border effects, we find a puzzling persistence of state-level spillover localization, a finding not explained merely as an aggregation of metropolitan effects, or outcome of spatial or social proximity. (100 words)

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

JEL classification: O30, O33, R10, R12

I. Introduction

In understanding the mechanisms underlying the geographic patterns of any economic phenomenon, it is helpful to disentangle the effects associated with different geopolitical boundaries from one another, and effects related directly with spatial distance. In the context of trade, empirical examination to distinguish between border and distance effects has been carried out both at the national level (e.g., McCallum, 1995; Anderson and Wincoop, 2003) and the state level (e.g., Wolf, 2000; Hillberry and Hummels, 2003, 2008). However, while geography is known to constrain not just the flow of trade but also the flow of knowledge (Jaffe, Henderson, and Trajtenberg, 1993; Thompson, 2006), few attempts have been made to simultaneously examine different geographic units in order to determine their relative importance in shaping the overall knowledge diffusion patterns. This is especially surprising given the overall importance attached to geographic localization of knowledge spillovers, a phenomenon of interest to researchers working in areas as diverse as technological innovation, economic geography, international economics and economic growth. The present study addresses this gap by unbundling the extent to which observed localization of knowledge flows is an outcome of: (i) discrete impediments associated with one or more geopolitical boundaries – national, state or metropolitan, and/or (ii) a decline in the extent of knowledge flows with distance.

To start with, existing research provides limited guidance regarding the relative importance of different geopolitical boundaries in shaping the geographic knowledge diffusion patterns. While intra-country knowledge flows have been reported to be more intense than those across countries (Branstetter, 2001; Keller, 2002), the extent to which national borders *per se* restrict knowledge flows has not been sufficiently examined. Similarly, given that intra-national political borders (such as states in the U.S.) are even less obviously a barrier to knowledge flow,

skepticism has been expressed regarding the extent to which state-level studies (Jaffe, 1989; Audretsch and Feldman, 1996) pick up effects that genuinely reflect fundamental state-level mechanisms at play. In fact, arguments are often made 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). Indeed, economic geographers argue that metropolitan boundaries should be the focus of analysis in examining such phenomena, echoing Krugman's remark that "states aren't really the right geographic units" (Krugman, 1991a:43). Despite concerns that observed effects of knowledge localization at the national or state level might be a manifestation of mechanisms operating more locally, any conjectures with regard to metropolitan co-location explaining the more aggregate findings still remain untested.

In addition to this limited understanding of the mechanisms potentially associated with different geopolitical boundaries, previous research has failed to distinguish these boundary effects from any effects driven just by distance. Hence, drawing an analogy from similar debates in the literature on trade (Anderson and Wincoop, 2003; Hillberry and Hummels, 2003, 2008), one might wonder about the extent to which effects seemingly associated with one or more geopolitical boundaries are actually a manifestation fundamentally related to geographic distance rather than any geopolitical boundary *per se*. In other words, it is unclear whether the constraints geography imposes on knowledge diffusion are primarily discrete effects associated with crossing one or more geopolitical boundaries, or in fact the result of a decline in the extent of knowledge diffusion with distance.

More insight into such issues pertaining to the geographic scope of knowledge flows would help refine existing theoretical models of innovation and growth, ultimately leading to more effective innovation-related policies. For example, localized knowledge spillovers at the

national level are assumed by many models of endogenous growth, whereby constraints on access to foreign knowledge can limit a lagging country's ability to catch up (Romer, 1990; Grossman and Helpman, 1991). Likewise, the extent to which knowledge spillovers may be localized even at a sub-national level, such as within states or provinces, can have important implications for how policies geared towards encouraging local R&D or facilitating knowledge diffusion (Peri, 2005). Finally, assumptions regarding the extent to which mechanisms underlying knowledge diffusion operate at the metropolitan level are an important component of the way economic geographers view the phenomenon of agglomeration of economic activity (Feldman and Audretsch, 1999; Glaeser, 1999; Fallick, Fleischman and Rebitzer, 2006).

In disentangling the various geographic effects mentioned above, our study contributes most directly to previous research employing patent citations as a measure of knowledge diffusion within and across regions (Jaffe, Henderson, and Trajtenberg, 1993; Jaffe and Trajtenberg, 2002; Thompson and Fox-Kean, 2005; Thompson, 2006). In this stream of literature, geographic localization of knowledge flows has typically been demonstrated through *separate* analyses at different geographic levels – country, state and metropolitan. However, we depart from this approach by making no *ex ante* assumptions about the appropriate geographic unit of analysis. Instead, we run a "horse race" among different geographic units (countries, states and metropolitan areas) to isolate the level at which mechanisms driving the localization effect operate most prominently. Since the pioneering matching-based methodology introduced by Jaffe, Henderson and Trajtenberg (1993) to detect localization of knowledge spillovers does not lend itself to running such a horse race, we implement a regression model based on choice-based sampling in order to directly estimate the probability of a patent citation as a function of an entire set of geography-related variables. This citation-level regression approach has the

additional benefit of allowing us to account for technological similarity or relatedness between patents at multiple levels of technological granularity. As elaborated below, this at least partly overcomes the challenge that matching-based studies have faced in having to choose a specific level of technological granularity at which to perform the match needed to construct a control sample of patents (Thompson and Fox-Kean, 2005; Henderson, Jaffe and Trajtenberg 2005).

We extend our regression approach to also account for spatial distance, allowing us to distinguish effects associated with geopolitical boundaries from the effect of distance. This is a significantly extension beyond the typical approach whereby geopolitical boundaries are interpreted as a proxy for distance itself, an approach that actually confounds the effects of geopolitical boundaries and distance. To the best of our knowledge, ours is the first study employing such fine-grained spatial data. While some related studies have employed distance-based measures, these have not been sufficiently fine-grained to be useful for disentangling the geographic effects of interest here. For example, while Keller (2002) employs data on distance between capital cities of countries, he does not differentiate between different intra-national distances. Similarly, although Peri (2005) refines this to consider differences in inter-state distances, even he does not distinguish different distances within a state. In contrast, we resort to a finer city-to-city distance measure to more precisely account for the effect of spatial distance in order to isolate knowledge diffusion effects truly attributable to geopolitical boundaries.

Consistent with previous studies, the independent analyses we conduct at the national, state and metropolitan levels exhibit evidence of localized knowledge diffusion. When all three are examined simultaneously, the country and state-level effects both turn out to be large and comparable in magnitude, while the metropolitan-level effect is somewhat smaller. Extending the analysis to explicitly account for spatial distance, we continue to find strong national and

state border effects, but the metropolitan effect practically disappears. As additional measures of geographic proximity, we also introduce indicator variables to capture whether the source and destination are in adjacent countries or states sharing a common border. While adjacency is found to increase knowledge flow intensity, this effect is still found to be smaller than intra-national and intra-state effects.

Further analysis is carried out to examine whether the border effects persists even when additional variables capturing the "social proximity" of the cited and citing inventor teams are introduced. Specifically, we account for self-citation not just at the level of the assignee organization but also at the level of an inventor moving between two teams of inventors. In addition, we at least partially account for direct or indirect collaborative ties between the cited and citing teams of inventors, inferred from co-authorship data for past patents. These analyses are motivated by previous research demonstrating that knowledge flows can be significantly affected by inventor mobility and the collaborative networks that result (Almeida and Kogut, 1999; Singh, 2005; Agrawal, Cockburn and McHale, 2006; Breschi and Lissoni, 2009). While these additional variables help explain some of the distance effect, they have little impact on the estimates for the national or state border effects.

We view our findings with regard to localization of knowledge flows associated with national borders (even after controlling for distance and other geographic units) as being in line with expectations, given the well-documented linguistic, cultural, administrative and economic differences between countries (see, e.g., Coe, Helpman and Hoffmaister, 2009). However, the persistence of a state border effect is puzzling, especially given the common perception that states are not a relevant unit of analysis for economic activity. As discussed in the conclusion section, this is a finding worth further investigation in subsequent studies.

II. Using patent citation data to examine geographic patterns of knowledge diffusion

Following the research tradition starting with the pioneering study by Jaffe, Trajtenberg and Henderson (1993), we employ citations between patents as an indicator of knowledge flow between inventors.[1] While citation-based measures can be noisy in capturing the underlying actual knowledge flows, direct surveys of inventors have established that citations do capture meaningful information, especially when employed in large samples (Duguet and MacGarvie, 2005; Jaffe and Trajtenberg, 2002).[2] In addition, the ability to derive information on the precise geographic location of inventors makes patent data particularly well suited to this study.

Our dataset combines raw data obtained from the United States Patent and Trademarks Office (USPTO) with additional fields made available by the National Bureau of Economic Research (NBER; see Jaffe and Trajtenberg 2002, Chapter 13). As part of a multi-year research effort, this combined dataset has been enhanced along four dimensions. First, for each assigned patent, the parent organization has been identified by carrying out an assignee name clean-up followed by a parent-subsidiary match.[3] Second, an elaborate inventor name matching procedure has been carried out to map all inventors to unique identifiers.[4] Third, worldwide locations of all

[1] Patent citations remain the only exception to Krugman's (1991a, p. 53) observation that "knowledge flows... are invisible; they leave no paper trail by which they may be measured or tracked." Admittedly, it is hard even with citation data to decipher when a knowledge flow really represents a "spillover", i.e., a true externality. Nevertheless, the literature has taken a view that studying knowledge flows using citation data is nevertheless interesting as they would at least partly represent spillovers, and that there might in fact be "benefits from trade" even from knowledge flows that are in reality market transactions rather than externalities.

[2] The issue of which citations to interpret as knowledge flows is still a subject of debate. For example, considering citations added by patent examiners rather than inventors might not be desirable when the inventor actually did not experience the implied knowledge flow, but is desirable if the inventor mistakenly or strategically omitted a citation (Alcacer and Gittleman, 2006; Lampe, 2009). Distinguishing between these is, however, very hard. While we would still have liked to repeat our analyses after excluding examiner-added citations as a robustness check, unavailability of machine-readable information regarding this for our sample period made this impractical.

[3] The NBER assignee data we started with is already an outcome of cleaning raw data from USPTO. We further cleaned these data and carried out a parent-subsidiary match using the NBER *Compustat* identifiers when available, Stopford's *Directory of Multinationals*, *Who Owns Who* directories and Internet sources (Singh, 2005).

[4] Our algorithms are based on those used by Singh (2008), and are similar to those implemented by Trajtenberg (2006) and Marx, Strumsky and Fleming (2009). While the details differ somewhat across these, the common idea is to match inventor records not just relying on the inventor name fields but also looking for a good enough match along some of the other fields such as technology classification, assignee, address, collaborator names and citations.

inventors have been mapped to latitudes and longitudes on the earth's surface, allowing the use of spherical geometry in calculating the precise distance between any two inventors.[5] Fourth, locations of U.S. inventors have been mapped to metropolitan areas (MSAs) where applicable.[6]

Our sample construction began with choosing cited patents originating in the U.S. during the period 1980-1986, considering only patents with inventors based in the same town or city so that their geographic origin is unambiguously defined.[7] For each patent, we determined the citations received during a 12-year window following the application year. As Jaffe, Trajtenberg and Henderson (1993) point out, simply examining the geographic co-location frequency of the inventors for the cited and citing patents does not suffice for making inference about geographic localization of knowledge flows. In particular, citations would appear excessively localized due to technological specialization of regions for reasons such as a local concentration of specialized labor or intermediate goods industries (Marshall, 1920; Krugman, 1991b). To adjust for this, Jaffe, Trajtenberg and Henderson (1993) propose a matching-based research design, wherein the prevalence of inventor co-location between the cited and citing patents is compared with that between the cited and "control" patents matched with the respective citing patents on their application date and technological characteristics.

To first replicate the findings of Jaffe, Trajtenberg and Henderson (1993), we start by constructing a matched sample of "control pairs" comprised of the cited patents and control

[5] This mapping, as described in Singh (2008), relied upon Geographic Names Information System of the U.S. Geological Survey, Geonet Names Server of the National Geospatial Intelligence Agency and other sources.
[6] The mapping relies upon a concordance between U.S. cities and metropolitan areas developed by Thompson (2006). These data actually include both "MSAs" and "CMSAs", although for brevity we simply use the term "MSA" for both. Previous studies sometimes also define an additional "phantom MSA" per state to handle cases where a location does not fall into an actual MSA, but doing so effectively confounds intra-MSA effects with intra-state effects. Since disentangling different geographic effects is a key goal of our study, we do not follow this practice.
[7] We also dropped the relatively rare cases where a cited patent's location could not be mapped to a precise latitude and longitude. In cases of imprecise data for the *citing* patent, the average latitude and longitude for all patents arising in the citing state (for U.S. inventors) or the country (for non-U.S. inventors) was used as an approximation.

patents having the same 3-digit technology class and application year as the respective citing patents in the citations sample. We then compute the extent of geographic co-location in the actual citations as well as the control pairs, separately using the country, state and MSA as the geographic unit of analysis in the three different sets of calculations. As the side-by-side comparison reported in Table 1 shows, our findings so far are comparable to those of Jaffe, Trajtenberg and Henderson (1993) in terms of evidence of geographic localization of knowledge diffusion irrespective of the geographic level of analysis employed.[8]

The matching approach, however, is not useful in directly confronting two key questions central to our study. First, how much do national or state borders *per se* constrain knowledge flows, as opposed to the corresponding effects reported by matching-based studies being a manifestation of mechanisms that in fact operate at a more local level (like MSA)? Second, do the geopolitical boundary effects really represent a discrete change associated with the boundaries themselves rather than being a manifestation of decay in knowledge diffusion with distance? The right approach to consider for answering these questions would be a regression framework that simultaneously examines the effect of different geographic boundaries, while also disentangling these boundary effects from the effect of spatial distance.

A regression approach could also help at least partially address a dilemma that Thompson and Fox-Kean (2005) have pointed out: the 3-digit technological match commonly employed may be too crude to capture relevant geographic distribution of technological activity, while a finer classification could suffer from a potential selection bias as a match for a large fraction of the sample would not be found. Suggesting a way forward, Henderson,

[8] The findings are also similar to those from the 3-digit matched sample used by Thompson and Fox-Kean (2005, Table 3). While that study goes on to refine the match by employing 9-digit technology subclasses instead, they run into a challenge that over two-thirds of the patents cannot be matched in this more fine-grained process. As explained below, our approach is to instead employ an extension of the 3-digit matched sample and control for a finer technological level through additional variables introduced directly into our regression framework.

Jaffe and Trajtenberg (2005) remark: "[I]t is certainly true that 'controlling for' technology in order to identify knowledge spillovers is very tricky... An example of a more structured approach to this issue that may help disentangle these forces can be found in Jaffe and Trajtenberg (2002, Ch. 7), where they estimate the probability of citations across countries, controlling for technological proximity." Our research design, which is an extension of a similar empirical framework employed by Singh (2005), follows the spirit of the above suggestion. Rather than selecting a specific granularity of technology, it accounts for technological relatedness at multiple levels simultaneously by using a regression approach to estimate the likelihood of citation between any two patents.[9]

III. A regression approach using choice-based sampling to estimate citation probability

We model the probability of patent *K* getting cited by patent *k* as a function P(*K*, *k*), assumed to take a logistic form. In a random sample of pairs of patents, the binary outcome *y* can be defined as being 1 if (and only if) a citation exists between the ordered pair. Formally, the Bernoulli outcome *y* for observation *i* is 1 with a probability

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

Here, $\mathbf{x_i}$ represents the vector of covariates and $\boldsymbol{\beta}$ of parameters to be estimated. However, since citations between random patents are extremely rare, it would not be practical to carry out the estimation based on a purely random sample drawn from a population of patent pairs.

[9] Jaffe and Trajtenberg (2002, Ch. 7) construct "cells" of patent pairs as the unit of observation, and define the number of realized citations in a cell as the dependent variable. However, they still control for the average technological distance between the citing and cited patents in a "cell" only based on the coarse 3-digit technology classification. So the Thompson and Fox-Kean (2005) critique still applies: e.g., cells with a greater fraction of pairs sharing a 9-digit technology will have a greater likelihood of citation as well as co-location. In contrast, we employ a finer unit of observation in our regressions: analyzing pairs of patents directly allows us to simultaneously account for relatedness at multiple - both coarse and fine - technological levels (1-digit, 2-digit, 3-digit and 9-digit).

Instead, it makes sense to employ "choice-based" sampling, wherein the fraction α of pairs drawn from the sub-population with $y = 0$ is several orders of magnitude smaller than the fraction γ of the pairs drawn from the sub-population with $y = 1$.[10] As the sampling rate now varies according to the value of the dependent variable, the usual logistic estimates would be biased. One way to avoid the bias is to use the *weighted exogenous sampling maximum likelihood* (WESML) approach, which weights each term in the log likelihood by the reciprocal of the *ex ante* probability of inclusion of an observation in the sample (Manski and Lerman, 1977).[11] Mathematically, this estimator maximizes the "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 $w_i = (1/\gamma)\,y_i + (1/\alpha)(1-y_i)$. The appropriate estimator of the asymptotic covariance matrix in this case has been shown to be White's robust "sandwich" estimator.[12]

The basic choice-based sampling assumes that the "zeroes" are all drawn from the overall $y = 0$ population at the same rate α. This can be generalized by allowing α to vary for different $y = 0$ sub-populations in order to combine benefits of choice-based sampling with those of stratification based on explanatory variables (Manski and McFadden, 1981; Amemiya, 1985, Ch 9). This allows interpreting each of the matched "controls" generated earlier as a random draw from the respective sub-population of "zeroes" for the given {cited patent, citing year, citing class} combination, allowing computation of an *ex ante* sampling

[10] For textbook discussion of choice-based sampling, see Amemiya (1985, Ch. 9) or Greene (2003, Ch. 21).
[11] If the logistic specification were accurate, one could still employ the usual (unweighted) logit by relying upon the coincidence that the random sampling and choice-based sampling maximum likelihood estimators coincide for all parameters except the intercept. However, unlike WESML, this approach is not robust to misspecification since the unweighted logit has no natural interpretation when the true response function is not logit (Xie and Manski, 1989).
[12] Although WESML is not an efficient estimator (Imbens and Lancaster, 1996), it continues to be widely employed as it is very intuitive as well as easy to implement. The efficiency issue can be mitigated by employing sufficiently large samples, an approach we follow here.

probability and hence a corresponding weight for the observation. However, the matched sample in itself does not suffice for estimation that one can meaningfully interpret as being applicable to the overall population, since representation also needs to be ensured for the {cited patent, citing year, citing class} combinations that were not included at all since no corresponding citations existed in the original sample. We therefore draw an additional control from the sub-population of potentially citing patents from all the non-represented citing classes for each {cited patent, citing year} combination, and again weight this control using the implied sampling rate. This leads to a final sample of 3,024,196 patent pairs (for 118,940 cited patents) for use in the regressions, including 798,458 actual citations, 798,458 matched pairs and 1,427,280 additional control pairs for the non-represented citing classes.[13]

Rather than making any specific assumptions about the temporal pattern of citations, our regression analysis accounts for variation in citation likelihood with time lag non-parametrically by employing a series of indicator variables for the lag (in years). We also include indicator variables for the cited patent's technological category and application year to account for systematic differences across sectors or over time. Finally, since the citation probability might also be driven by other characteristics of the cited patent, we control for such characteristics when possible and employ clustering in the standard error computation to account for unobserved ones. Table 2 summarizes the variables used in our analyses.[14]

[13] To ensure that our findings are not overly dependent on specific sampling process, we repeated all our regression analyses using a basic choice-based sample where all "zeroes" were drawn at random at the same rate α from the overall $y = 0$ population (and were accordingly assigned the same weight). All our main qualitative findings continued to hold. Not surprisingly, since this meant we no longer employed the technology-matched sample even as technological relatedness is a strong driver of citation, the standard errors were mostly somewhat larger.

[14] The distance variables are not defined in the relatively infrequent cases (less than 7%) where either the cited or the citing location could not be mapped to a precise latitude and longitude. To make sure that dropping these in the distance-related regressions did not bias our findings in any way, we repeated the analysis by using the average latitude and longitude for patents arising in the given state (for U.S. inventors) or country (for non-U.S. inventors) to calculate approximate distance in such cases. All the findings reported in the paper remained practically unchanged.

IV. Results

We begin by replicating prior findings for the three different geographic boundaries individually, with the results reported in the first three columns of Table 3. For comparability with matching-based analysis, this initial analysis accounts for technological similarity and relatedness only up to the 3-digit technology classification. Like the matching-based findings discussed earlier, we again observe the localization effect at all three levels: the country level (*same country* in col (1)), state level (*same state* in col (2)) and metropolitan level (*same msa* in col (3)). The regression estimates have an intuitive interpretation: they imply a 69% greater likelihood of a within-country knowledge flow than across national borders, 105% greater likelihood for within-state than across state borders, and 111% greater likelihood for within-MSA than across metropolitan boundaries.[15]

It is tempting to compare the results across the three columns and conclude that the localization effects at the MSA and the state level are comparable to each other and more intense than the effects at the country level. However, this can be misleading since adjudicating between the effects operating at different geographic levels requires simultaneous consideration of all three. Indeed, such an analysis – reported in col (4) – instead finds the estimated effects for *same country* (59%) and *same state* (59%) to be comparable, and that for *same msa* (49%) to be somewhat smaller. Further, noting that the Thompson and Fox-Kean (2005) critique regarding the inadequacy of 3-digit technological controls still applies here, col (5) introduces two additional controls for overlap between patents along the primary as well as additional 9-digit technology subclasses. Doing so causes the estimated

[15] In a logit model, the marginal effect of a variable j is $\beta_j \Lambda'(\boldsymbol{x\beta})$, which can be shown to equal $\beta_j \Lambda(\boldsymbol{x\beta})[1-\Lambda(\boldsymbol{x\beta})]$. In general, this needs to be calculated based either on the mean predicted probability or using the sample mean for $\Lambda(\boldsymbol{x\beta})$. But citations being rare events allows an even simpler interpretation: 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 the citation probability when the indicator variable j goes from 0 to 1.

effects of national borders (55% increase) and state borders (69% increase) to amplify even further compared to that of metropolitan boundaries (21% increase). At a minimum, these findings suggest that there is more to the national and state border effects than a mere aggregation of localization mechanisms operating simply at the metropolitan level.[16]

Our next task is to further disentangle the effect of different geopolitical boundaries examined above from that of spatial distance. Including a distance measure in the previous models would enable us to determine the extent to which within-region knowledge flows are driven primarily by geographic proximity rather than any mechanisms fundamentally associated with the corresponding geopolitical boundary. Before doing so, however, it is instructive to examine the effect of spatial distance in isolation. As reported in col (1) of Table 4, using spatial distance (*distance*) as the sole geographic variable also produces results consistent with localization of knowledge diffusion, implying a 23% fall in likelihood of citation with doubling of the distance between the source and destination inventors. Since we ultimately want to examine whether the border effects persist once distance has been fully accounted for, we now relax the specific functional form implicit in col (1) and employ a series of indicator variables for more flexibility in how distance might affect the likelihood of citation. As reported in col (2), the estimates still imply citation rate falling with distance.

We now consider together the indicator variables for distance and the geographic boundary variables employed earlier, allowing us to examine whether border effects really exist over and above the effect of distance. As a comparison of col (3) of Table 4 with the earlier findings from col (5) in Table 3 shows, the *same state* effect is now a little less strong

[16] As expected, the estimates for *same 1-digit technology*, *same 2-digit technology*, *same 3-digit technology*, *technological relatedness*, *same primary 9-digit technology* and *overlap of 9-digit technologies* indicate knowledge flow within the same or related technologies to be stronger than across technologies, and for *same assignee* shows within-assignee knowledge flow to be stronger than across assignees.

in magnitude relative to the *same country* effect. However, the more important observation remains that the estimates for *same country* and *same state* are both still quite large and robust, even as the *same msa* estimate now becomes indistinguishable from zero. In other words, though distance completely explains the metropolitan effect, the national and state border effects are to a large extent orthogonal to the effect of distance. This finding challenges the notion that the localized knowledge diffusion reported by previous studies is merely a manifestation of intra-regional distances typically being shorter than cross-regional distances.

Concern may arise that even employing a series of distance-related variables can only partially account for physical proximity. Therefore, following similar variables used in gravity models from the trade literature, col (4) in Table 4 introduces additional indicator variables *contiguous countries* and *contiguous states* to distinguish cases where two countries or states share a common border from cases where they do not. While we find knowledge flow to indeed be more intense among contiguous regions than non-contiguous regions, the knowledge flow between contiguous regions is still significantly smaller than within-region knowledge flow for both the country and the state level of analysis.

To dig deeper into possible mechanisms driving the national and state border effects, cols (5) and (6) extend the analysis to also account for the social connectedness of inventor teams. Specifically, motivated by the fact that inventor mobility tends to be geographically localized, col (5) introduces a new variable: *same inventor* (i.e., whether there exists an inventor who worked both in the original and the destination team). Extending the reasoning further and recognizing that direct and indirect collaborative ties across individuals – which can facilitate knowledge flow – tend to be geographically proximate as well, col (6) introduces additional variables for inventors in the cited and citing team being either *direct*

collaborators (i.e., someone in the destination team has in the past collaborated with someone in the original teams) or *indirect collaborators* (i.e., the teams are connected through a common third person with which someone from each team has previously collaborated).[17] While the relative effect of distance (especially *distance = 0 miles*) becomes significantly smaller as a result, the national border effect remains practically unchanged and the state border effect becomes only slightly smaller.[18] In other words, while the social proximity of inventor teams appears to indeed be a significant mediator of the distance effect, it does not help explain the persistence of border effects.

V. Conclusion

Our study makes an important contribution to the literature on the geography of knowledge flows. While the role of geopolitical boundaries and distance in constraining trade flows has been well studied, these geographic effects have not been disentangled in the context of knowledge spillovers: the localization of knowledge spillovers has previously been established only through separate analyses at the metropolitan, state or national level. Our empirical framework, based on a regression approach using choice-based sampling to estimate the probability of patent citation, allows for the simultaneous consideration of multiple geopolitical boundaries. In addition, while previous studies have typically interpreted geopolitical boundaries simply as a proxy for distance, we employ fine-grained inventor location

[17] It might appear counter-intuitive that direct ties have a smaller effect on citation probability than indirect ties. This turns out to be a result of the stringent technology controls we use. Not employing *technological relatedness*, *same primary 9-digit technology* and *overlap of 9-digit technologies* reverses this, as the relative magnitude of *direct collaborators* increases by a factor of three. This highlights a general point that good technology controls are crucial for avoiding over-emphasizing the role of collaborative ties, as collaboration is naturally more likely to occur among individuals working in similar or related technological areas (and hence more likely to cite each other).

[18] As an aside, it is interesting to note that *same inventor* has more explanatory power as a mediator for the distance effect (particularly *distance = 0 miles*) while *direct collaborators* and *indirect collaborators* have more explanatory power in explaining the state border effect (though still to a rather small extent).

data to explicitly account for spatial distance in order to separately identify effects directly associated with one or more geopolitical boundaries. Therefore, we are able to disentangle the extent to which previously reported localization findings reflect discrete effects truly associated with one or more geopolitical (i.e., national, state or metropolitan) boundaries as opposed to simply being a manifestation of aggregated city or metropolitan-level effects and/or a negative relationship between knowledge diffusion and distance. In accounting for technological similarity and relatedness between the citing and cited patents at multiple levels of granularity, our regression framework also gets around challenges that past matching-based studies have faced in having to choose a specific technological granularity.

Consistent with previous evidence on localized knowledge diffusion, we also find patent citation probability to be correlated with co-location, defined as being in the same country, state or metropolitan area. Intriguingly, when the three geopolitical boundaries are considered simultaneously, both country and state level effects remain robust, even as the metropolitan-level effect falls drastically. Even when the precise spatial distance is controlled for, the two border effects continue to remain strong even as the metropolitan-level effect practically disappears.

A finding that national borders matter even after controlling for distance seems to be in line with expectations. However, a state-level finding that is more than simply an aggregation of city or metropolitan level effects, remains robust even to accounting for spatial distance and contiguousness, and cannot be explained even when assignee self-citation, inventor self-citation and collaborative ties are accounted for, is particularly puzzling. Taken literally, this suggests that mechanisms fundamentally associated with not only country borders but also state borders seem to play an important role in shaping knowledge diffusion patterns, a finding that appears to

challenge previous studies of knowledge diffusion that have implicitly or explicitly assumed such borders to only be a proxy for spatial distance.

The present study could be extended in several ways. For example, a natural direction would be to look for mechanisms that explain the border effects. The literature on international trade suggests several obvious variables to consider at the national level, such as linguistic, cultural, administrative and economic distance between countries. Indeed, analysis not reported here found knowledge flows from the U.S. to other English-speaking countries to be particularly strong even after controlling for spatial distance. A more general treatment of gravity-type variables would, however, require a sample where not just the citing but also cited patents were drawn from multiple countries. This is beyond the scope of the present study, given our particular interest in more closely examining sub-national patterns of knowledge diffusion, particularly those associated with state and metropolitan boundaries for which we have associated data available only for the U.S.

At the state level, our study raises perhaps raises more questions than it answers. Future research should investigate why state borders seem to play a surprisingly robust role in shaping knowledge diffusion patterns. The finding is particularly unexpected for a country like the U.S., where state borders are regarded as historical constructs that impose few constraints on the flow of ideas. At a minimum, this calls for further empirical research paralleling the trade literature investigating how real and robust border effects really are in that context (e.g., McCallum, 1995; Wolf, 2000; Anderson and Wincoop, 2003; Hillberry and Hummels, 2003, 2008). We cannot rule out the possibility that, on further examination, some of the border effects we find would turn out to be not as robust as they first appear to be.

However, an intriguing possibility is that mechanisms operating fundamentally at the state level do play an important role in shaping the overall geographic patterns of knowledge diffusion. This would naturally call for a research agenda exploring what such mechanisms could be. For example, a potential factor could be a state's policy regarding the enforcement of employee non-compete agreements, a policy variable that has already been shown to have a significant effect on inter-firm mobility patterns of inventors (Marx, Strumsky and Fleming, 2009). In regression analysis not reported here, we found that states which do not enforce such agreements tend to see greater intra-state knowledge diffusion as well as greater knowledge inflow from other states. While only suggestive, the correlation is nevertheless intriguing enough to warrant further analysis. More generally, irrespective of whether the jury finally decides whether state borders inherently present an impediment to diffusion of knowledge or not, a closer examination of mechanisms shaping sub-national knowledge flow patterns would be fruitful in developing a better understanding of the geography of innovation and, ultimately, of economic growth.

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.

Agrawal, A., I. Cockburn, and J. McHale. 2006. "Gone but not Forgotten: Labor Flows, Knowledge Spillovers, and Enduring Social Capital." *Journal of Economic Geography* **6**(5) 571-591.

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

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

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

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

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.

Breschi, S. and F. Lissoni. 2009. "Mobility of Skilled Workers and Co-invention Networks: An Anatomy of Localized Knowledge Flows." *Journal of Economic Geography* **9**(4) 439-468.

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) 375-393.

Fallick, B., C. Fleischman, and J. Rebitzer. 2006. "Job-Hopping in Silicon Valley: Some Evidence Concerning the Micro-Foundations of a High Technology Cluster." *Review of Economics and Statistics.* **88**(3) 472-481.

Feldman, M.P. and D.B. Audretsch. 1999. "Innovation in Cities: Science-based Diversity, Specialization and Localized Competition." *European Economic Review* **43** 409-429.

Glaeser, E.L. 1999. "Learning in Cities." *Journal of Urban Economics* **46**(2) 254-277.

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

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

Imbens, G.W. and T. Lancaster. 1996. “Efficient Estimation and Stratified Sampling.” *Journal of Econometrics* **74** 289-318.

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

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.

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

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

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

Krugman, P. 1991b. “Increasing Returns and Economic Geography.” *Journal of Political Economy* **99**(3) 483-499.

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 MA, MIT Press.

Marshall, A. 1920. *Principles of Economics*. London, Macmillan.

Marx, M., D. Strumsky, and L. Fleming. 2009. “Mobility, Skills, and the Michigan Non-compete Experiment.” *Management Science* **55**(6) 875-889.

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.

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

Singh, J. 2008. “Distributed R&D, Cross-regional Knowledge Integration and Quality of Innovative Output.” *Research Policy* **37**(1) 77-96.

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

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.

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.

Table 1. Replicating findings from Jaffe, Trajtenberg and Henderson (1993)

	Our matched sample				Jaffe, Trajtenberg & Henderson sample			
	(1) Citations sample	(2) Intraregion citations	(3) Intraregion controls	(4) t-statistic	(5) Citations sample	(6) Intraregion citations	(7) Intraregion controls	(8) t-statistic
Country-level analysis								
Including assignee self-citations	798,458	73.3%	57.3%	215.5	8,914	71.2%		
Excluding assignee self-citations	697,116	69.9%	56.7%	169.3	7,759	68.0%	61.4%	8.6
State-level analysis								
Including assignee self-citations	798,458	18.0%	5.4%	253.9	8,914	17.7%		
Excluding assignee self-citations	697,116	9.4%	4.4%	118.9	7,759	9.7%	5.1%	11.0
Metropolitan-level analysis								
Including assignee self-citations	798,458	12.2%	2.8%	231.2	8,914	14.4%		
Excluding assignee self-citations	697,116	5.4%	2.1%	105.7	7,759	6.6%	1.7%	15.4

Notes: The Jaffe, Trajtenberg and Henderson (JTH) numbers reported here were calculated based on pooling of results for their different sub-samples primarily using information available in their Table III.

Table 2. Variable definitions for regression analysis

Country-level variables	
same country	Indicator variable that is 1 if the two patents originate from inventors located in the same country (in our sample U.S., since the cited patent sample is drawn only from the U.S.)
contiguous countries	Indicator variable that is 1 if the two patents originate from inventors located in the different countries with a common border
State-level variables	
same state	Indicator variable that is 1 if the two patents originate from inventors located in the same U.S. state
contiguous states	Indicator variable that is 1 if the two patents originate from inventors located in the different states with a common border
Metropolitan-level variable	
same msa	Indicator variable that is 1 if the citing and cited patents originate from inventors located in the same U.S. metropolitan area
Technology variables	
same 1-digit technology	Indicator variable that is 1 if the two patents belong to the same 1-digit NBER technology category
same 2-digit technology	Indicator variable that is 1 if the two patents belong to the same 2-digit NBER technical subcategory
same 3-digit technology	Indicator variable that is 1 if the two patents belong to the same 3-digit USPTO primary technology
technological relatedness	Citation propensity calculated as likelihood of citation (scaled by 100) between random patents drawn from the population with the same respective technology classes as the focal pair
same primary 9-digit technology	Indicator variable that is 1 if the two patents belong to the same 9-digit USPTO primary technology subclass
overlap of 9-digit technologies	Natural logarithm of one plus the number of overlapping (primary or other) 9-digit technology USPTO subclasses under which the two patents are categorized
Assignee variables	
same assignee	Indicator variable that is 1 if the two patents are owned by the same parent firm or organization
nonfirm assignee	The cited patent is assigned to an entity other than a firm (e.g., a universirty, research institute or government body)
Cited patent 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
Distance variable	
distance	Distance, in miles, between the cities where the first inventors of the source and destination patents live, calculated using spherical geometry based on the respective latitude and longitude
Inventor connectedness variables	
same inventor	Indicator variable that is 1 if at least one inventor is common for the two patents
direct collaborators	Indicator variable that is 1 if there is no common inventor between the two teams, but someone in the cited patent has in the past collaborated with someone from the citing patent
indirect collaborators	Indicator variable that is 1 if there is no common inventor or past collaboration, but there exists a third person who has collaborated with one of the cited as well as citing inventors in the past

Table 3. A "horse race" between different geopolitical boundaries

	(1)	(2)	(3)	(4)	(5)
same country	0.694***			0.594***	0.554***
	(0.012)			(0.011)	(0.026)
same state		1.045***		0.595***	0.688***
		(0.035)		(0.044)	(0.051)
same msa			1.113***	0.494***	0.209**
			(0.056)	(0.070)	(0.089)
same assignee	2.783***	2.415***	2.494***	2.261***	1.513***
	(0.060)	(0.061)	(0.062)	(0.062)	(0.11)
same 1-digit technology	1.057***	1.058***	1.057***	1.060***	1.063***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
same 2-digit technology	1.239***	1.242***	1.242***	1.236***	1.251***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)
same 3-digit technology	2.721***	2.677***	2.687***	2.684***	2.161***
	(0.029)	(0.029)	(0.030)	(0.029)	(0.029)
technological relatedness	4.090***	4.097***	4.089***	4.076***	3.356***
	(0.30)	(0.30)	(0.30)	(0.29)	(0.21)
same primary 9-digit technology					2.123***
					(0.11)
overlap of 9-digit technologies					1.744***
					(0.029)
nonfirm assignee	-0.0178	-0.00270	0.0214	0.0147	-0.0136
	(0.042)	(0.039)	(0.032)	(0.035)	(0.072)
references to other patents	0.0124***	0.0119***	0.0123***	0.0123***	0.0116***
	(0.0010)	(0.0011)	(0.0011)	(0.0011)	(0.0021)
references to non-patent materials	0.0820***	0.0608***	0.0643***	0.0676***	0.0428
	(0.021)	(0.022)	(0.022)	(0.022)	(0.038)
number of claims	0.00828***	0.00787***	0.00785***	0.00786***	0.00418***
	(0.00079)	(0.00077)	(0.00076)	(0.00079)	(0.0012)
Number of observations	3024196	3024196	3024196	3024196	3024196
Wald chi2	199974***	200095***	202663***	193711***	131128***
Degrees of freedom	29	29	29	31	33

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. Choice-based sampling was used, and weighted logit regression (WESML) approach was employed. Constant term and indicator variables for time lag, cited year and technology category not reported to conserved space. Robust standard errors in parentheses, clustered on the cited patent. Asterisks indicate statistical significance (*** p<0.01, ** p<0.05, * p<0.1).

Table 4. Also accounting for spatial and social proximity of inventor teams

	(1)	(2)	(3)	(4)	(5)	(6)
same country			0.455***	0.660***	0.658***	0.668***
			(0.079)	(0.11)	(0.10)	(0.097)
contiguous countries				0.462***	0.460***	0.465***
				(0.15)	(0.14)	(0.14)
same state			0.394***	0.542***	0.543***	0.505***
			(0.061)	(0.069)	(0.065)	(0.068)
contiguous states				0.231***	0.236***	0.229***
				(0.049)	(0.047)	(0.047)
same msa			-0.141	-0.135	-0.124	-0.120
			(0.10)	(0.10)	(0.096)	(0.097)
same assignee	1.370***	1.381***	1.338***	1.334***	1.195***	0.992***
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)
same inventor					1.539***	1.798***
					(0.23)	(0.23)
direct collaborators						0.665**
						(0.30)
indirect collaborators						1.242***
						(0.18)
ln(distance + 1)	-0.235***					
	(0.010)					
distance = 0 miles (i.e., same town or city)		2.284***	1.598***	1.243***	0.636***	0.618***
		(0.15)	(0.22)	(0.23)	(0.22)	(0.22)
distance >0 miles but <= 10 miles		1.547***	0.838***	0.483*	0.500**	0.384
		(0.24)	(0.26)	(0.27)	(0.25)	(0.28)
distance >10 miles but <= 20 miles		1.261***	0.569**	0.211	0.251	0.264
		(0.18)	(0.22)	(0.24)	(0.22)	(0.21)
distance >20 miles but <= 40 miles		1.416***	0.735***	0.366**	0.360**	0.324*
		(0.081)	(0.15)	(0.17)	(0.17)	(0.17)
distance >40 miles but <= 80 miles		1.112***	0.458***	0.0839	0.0890	0.0809
		(0.071)	(0.12)	(0.15)	(0.14)	(0.14)
distance >80 miles but <= 160 miles		0.947***	0.380***	0.0120	0.00959	0.0187
		(0.064)	(0.11)	(0.13)	(0.13)	(0.12)
distance >160 miles but <= 320 miles		0.731***	0.233**	-0.103	-0.0943	-0.104
		(0.054)	(0.097)	(0.12)	(0.11)	(0.11)
distance >320 miles but <= 640 miles		0.713***	0.257***	0.00871	0.0152	0.0125
		(0.033)	(0.084)	(0.11)	(0.11)	(0.10)
distance >640 miles but <= 1280 miles		0.502***	0.0570	-0.155	-0.142	-0.142
		(0.049)	(0.092)	(0.12)	(0.12)	(0.11)
distance >1280 miles but <= 2560 miles		0.595***	0.152*	-0.0586	-0.0560	-0.0643
		(0.033)	(0.082)	(0.11)	(0.11)	(0.10)
distance >2560 miles but <= 5120 miles		0.166***	0.109**	0.0745	0.0752	0.0779
		(0.047)	(0.048)	(0.050)	(0.050)	(0.051)
same 1-digit technology	1.065***	1.068***	1.067***	1.067***	1.072***	1.072***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
same 2-digit technology	1.249***	1.247***	1.246***	1.246***	1.244***	1.245***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
same 3-digit technology	2.132***	2.140***	2.140***	2.139***	2.146***	2.147***
	(0.029)	(0.029)	(0.029)	(0.029)	(0.030)	(0.029)
technological relatedness	3.330***	3.345***	3.337***	3.342***	3.368***	3.367***
	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)
same primary 9-digit technology	2.095***	2.085***	2.079***	2.079***	1.979***	1.922***
	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
overlap of 9-digit technologies	1.732***	1.733***	1.737***	1.737***	1.731***	1.734***
	(0.030)	(0.030)	(0.029)	(0.029)	(0.030)	(0.030)
nonfirm assignee	0.0100	0.0166	0.00504	0.000984	0.00401	0.0142
	(0.069)	(0.065)	(0.066)	(0.066)	(0.064)	(0.061)
references to other patents	0.0116***	0.0109***	0.0108***	0.0109***	0.0106***	0.0105***
	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0020)	(0.0020)
references to non-patent materials	0.0271	0.0170	0.0143	0.0142	0.00738	0.0119
	(0.039)	(0.037)	(0.037)	(0.037)	(0.038)	(0.037)
number of claims	0.00393***	0.00403***	0.00403***	0.00403***	0.00386***	0.00384***
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Number of observations	2818378	2818378	2818378	2818378	2818378	2818378
Wald chi2	120827***	127751***	128587***	128722***	132389***	136414***
Degrees of freedom	31	41	44	46	47	49

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. Choice-based sampling was used, and weighted logit regression (WESML) approach was employed. Constant term and indicator variables for time lag, cited year and technology category not reported to conserved space. Robust standard errors in parentheses, clustered on the cited patent. Asterisks indicate statistical significance (*** p<0.01, ** p<0.05, * p<0.1).