Available online at www.sciencedirect.com

ScienceDirect

Journal of Urban Economics 64 (2008) 258–269

JOURNAL OF Urban Economics

www.elsevier.com/locate/jue

How do spatial and social proximity influence knowledge flows? Evidence from patent data

Ajay Agrawal [a,b,*], Devesh Kapur [c], John McHale [d]

[a] Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, ON, Canada, M5S 3E6
[b] NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, USA
[c] Center for the Advance Study of India, University of Pennsylvania, 3600 Market Street Suite 560, Philadelphia, PA 19104, USA
[d] Queen's School of Business, Queen's University, 143 Union Street, Kingston, ON, Canada, K7L 3N6

Received 15 March 2007; revised 2 January 2008

Available online 6 February 2008

Abstract

We examine how the spatial and social proximity of inventors affects access to knowledge, focusing especially on how the two forms of proximity interact. Employing patent citation data and using same-MSA and co-ethnicity as proxies for spatial and social proximity, respectively, we estimate a knowledge flow production function. Our results suggest that although spatial and social proximity both increase the probability of knowledge flows between individuals, the marginal benefit of geographic proximity is greater for inventors who are not socially close. We also report that the marginal benefit of being members of the same technical community of practice is greater in terms of access to knowledge for inventors who are not co-located. Overall, these results imply that spatial and social proximity are substitutes in their influence on access to knowledge. We discuss the implications of these findings in terms of the optimal dispersion of socially connected inventors.

JEL classification: Z13; O33; R12

Keywords: Productivity; Innovation; Knowledge flows; Agglomeration; Ethnicity; Diversity

1. Introduction

Paul Romer's widely cited endogenous growth theory casts knowledge rather than physical assets in a central role for producing economic growth (Romer, 1990). However, Romer's model is predicated on the notion that "anyone engaged in research has free access to the entire stock of knowledge." In reality, access to new knowledge is highly imperfect (Griliches, 1957). For example, prior empirical research has shown that, contrary to the notion of "free access," knowledge is more likely to flow between individuals who are located more closely together (Jaffe et al., 1993; Zucker et al., 1998). Yet geographic distance is just one of many forms of distance that can impede the transfer of knowledge. Conversely, one can find ways to be "near" sources of knowledge while being physically separated. For example, social or professional networks may lower the cost of accessing knowledge between members (Sorenson et al., 2006). Thus, to fully understand economic growth, we must consider not only the factors that influence the production of knowledge but also those that influence access to knowledge.

In this paper, we study knowledge access by examining whether spatial and social proximity are complements or substitutes in terms of enhancing knowledge flows between individuals. The social capital literature highlights both possibilities. Membership in multiple overlapping networks (one defined by spatial proximity and another defined by social proximity, for example) helps reinforce the deep bonds of trust that facilitate exchange of tacit knowledge (Coleman, 1988); yet an influential literature also stresses the importance of "structural holes" between networks and "weak ties" across networks in accessing non-redundant knowledge (Burt, 1992; Granovetter, 1973).

[*] Corresponding author at: University of Toronto, Rotman School of Management, 105 St. George Street, Toronto, ON, Canada. Fax: 001 416 978 5433.
E-mail address: ajay.agrawal@rotman.utoronto.ca (A. Agrawal).

0094-1190/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jue.2008.01.003

Prior research has shown that spatial proximity enhances access to knowledge between inventors (Jaffe et al., 1993; Thompson and Fox-Kean, 2005). A common explanation for this finding is that latent knowledge is more accessible by those who are located in the same geographic region as the inventor.[1] The inventor may be more willing to share knowledge with co-located individuals because they trust them more and/or because they perceive a greater likelihood of reciprocation. Similarly, social proximity through membership in a social network (e.g., ethnic, professional) may also enhance trust and the chance of reciprocation. For example, co-ethnic networks, such as Indians in the US, are often characterized as rich in social capital (Kalnins and Chung, 2006; Saxenian, 1999), lowering the cost of establishing trust between members. Even if members of a community do not know each other directly, they are more likely to be indirectly connected by knowing someone in common. In her study of Indian and Chinese engineers and entrepreneurs in Silicon Valley, Saxenian emphasizes the role of trust and reciprocity in ethnic communities that facilitate the sharing of knowledge and other resources among members, noting that: "The initial social connections often have a basis in shared educational experiences, technical backgrounds, language, culture, and history" (p. 37). Conceptually, we remain agnostic in this paper to the actual knowledge-sharing mechanisms facilitated through social proximity. In other words, we do not differentiate between knowledge sharing based on direct relationships (e.g., family, former work colleagues or school mates), indirect relationships (e.g., graduates of the same university but not the same cohort, common friends), reputation (e.g., status of particular universities or firms in the home country), or cultural cues (e.g., inferences based on knowledge about caste or culture-specific personality characteristics).[2]

How might we consider the influence of spatial and social proximity on knowledge flows simultaneously? One way is to construct and estimate a simple Knowledge Flow Production Function (KFPF). The intuition is that the likelihood of a knowledge flow between a given pair of inventors depends on the structure of relationships between those inventors—spatial, social, professional, etc. Furthermore, we pay particular attention to how different types of relationships interact.

We construct a KFPF in which only two factors mediate the probability of a knowledge flow between individuals, one spatial and one social. Specifically, we focus on whether individuals are co-located (in the same city) and/or are co-ethnic (of the same ethnic background). We then estimate this function using patent citation data and employing a matched sample method developed by Jaffe et al. (1993) and refined by Thompson and Fox-Kean (2005) to control for the underlying distribution of field-specific technological activity across geographic and ethnic space.

We find that both spatial and social proximity mediate knowledge flows. Co-location and co-ethnicity increase the probability of a knowledge flow between inventors; co-location increases the probability of a knowledge flow by 24% (assuming non-co-ethnic inventors) and co-ethnicity increases the probability by 14% (assuming non-co-located inventors). Furthermore, we find that co-location and co-ethnicity are substitutes rather than complements in the way their interaction influences knowledge flows; for example, co-location increases the probability of a knowledge flow by 24% for non-co-ethnic inventors but only by 2% for co-ethnic inventors. In other words, the marginal benefit of co-location is approximately 12 times larger for individuals who are not co-ethnic. Thus, in terms of facilitating access to knowledge, co-location appears to offer much greater benefits to individuals who are not otherwise socially connected.[3]

We also find that co-location increases knowledge flows to a greater extent between inventors working in different fields as compared to those working in the same field. Specifically, for inventors working in different fields, co-location increases the probability of a citation by approximately 30% compared to an increase of only 20% for inventors working in the same field. This result once again implies that social proximity, this time manifested through membership in a community of practice, substitutes for rather than complements spatial proximity with respect to the effect on knowledge flows.

Finally, we replace our binary measure of spatial proximity, co-location, with a continuous measure of distance measured in thousands of miles. Consistent with our prior findings, our results indicate that knowledge flows between inventors diminishes with distance. A 1000-mile increase in distance results in an approximately 2% reduction in the probability of a knowledge flow. Furthermore, the marginal impact of co-ethnicity increases with distance between inventors. While the marginal effect of co-ethnicity between inventors that are 1000 miles apart is to increase the likelihood of a knowledge flow by only 5%, the effect between inventors that are 3000 miles apart is more than double that (13%).

Our paper builds on recent work that has also stressed the role that ethnic networks play in facilitating knowledge exchange and other valuable economic interactions (Rauch, 2001).[4] In particular, Kerr (2005) reports results indicating that

[1] "Latent knowledge" is known by the inventor, is useful for the application or further development of the inventor's invention, but is not codified in patents or publications (Agrawal, 2006). The knowledge is not codified because either incentives to publish this knowledge are missing (e.g., knowledge embodied in failed experiments) or the knowledge is tacit such that it is particularly costly to codify.

[2] Although we allow for social proximity to affect knowledge flows through any mechanism in this paper, in the conclusion section we return to this topic and highlight the need to understand the specific mechanisms actually used to access knowledge in an ethnic community in order to draw general conclusions.

[3] We focus on co-ethnicity as one particular grouping for which membership raises the likelihood of sharing social capital. Of course, many such possible groupings exist. The social capital literature provides a useful framework for understanding knowledge-sharing networks more generally. This research has been impressively multidisciplinary, with important contributions by sociologists (Granovetter, 1973; Coleman, 1988; Burt, 1992), political scientists (Putnam, 2002), and economists (Knack and Keefer, 1997; Glaeser et al., 2002).

[4] A related literature focuses on the costs and benefits of ethnic diversity. Alesina and Ferrara (2005) provide a useful survey of how ethnic diversity affects economic performance. A major focus of this literature is on the damage done by ethnic conflict in heterogeneous societies (Easterly and Levine, 1997).

ethnic scientific communities play an important role in international technology diffusion. His findings suggest that a larger ethnic research community in the US improves technology diffusion to less advanced countries of the same ethnicity. In addition, Kalnins and Chung (2006) provide evidence from the US lodging industry that Gujarati immigrant entrepreneurs benefit from their ethnic group's social capital when already-successful members are co-located and in the same industry. Furthermore, these papers also suggest that their findings may extend beyond ethnicity to other social groupings.

Our paper proceeds as follows. In Section 2 we describe the US resident Indian diaspora, the socially connected network that is the basis of our empirical study. In Section 3 we define the KFPF and describe the methodology and data we employ for estimating the function's parameters. In Section 4 we present empirical results, and Section 5 discusses the welfare implications of our findings.

2. The US resident Indian diaspora

The US resident Indian diaspora is particularly suitable for the purposes of our study because, on average, the members:

(1) are reasonably identifiable (by last name),
(2) are highly active in technological innovation (we use patent citations to measure knowledge flows),
(3) identify strongly with their ethnicity (we use co-ethnicity as a proxy for social proximity), and
(4) are active across a broad range of geographies (many co-located and non-co-located observations in the sample).

Members of the US resident Indian diaspora are highly active in technological innovation, which is evident by their employment and patenting output. They are disproportionately concentrated in engineering (7%) and mathematical/computing professions (16%). As a comparison, roughly 1% of the native-born population works in each of these professions.[5,6] In addition, not only is the share of the Indian-American population working in technology significant, its role in the US innovation system has increased over time (Table 1).[7] The share of total USPTO-issued patents that have at least one Indian-named inventor has been rising steadily, approximately in line with the expanding Indian-born population.

Table 1
USPTO-issued patents by application year

	1976	1980	1985	1990	1995	2000
Total	71,040	72,129	78,646	108,684	156,777	164,340
One or more Indian inventor	651	788	1041	1934	4557	5334
Percentage Indian	0.9%	1.1%	1.3%	1.8%	2.9%	3.2%

Members of the US resident Indian diaspora identify strongly with their ethnicity, perhaps partly because many are of a recent vintage. Of the 2001 Indian-American population residing in the US, those born in the US were fewer than those born in India (0.7 million versus one million).[8] Furthermore, more than one third of the Indian-born came after 1996 and more than half after 1990.[9] Survey evidence underlines the strong ethnic identification: 53% visit India at least once every two years, 97% watch Indian TV channels several times a week, 94% view Indian Internet sites several times a week, 92% read an Indian newspaper or magazine several times a week, and 90% have an Indian meal several times a week.[10]

The Indian diaspora is active across multiple geographic areas. Table 2 provides a snapshot of the Metropolitan Statistical Area (MSA) locations of patenting activity by US- and Canada-resident Indian inventors. The table also shows the total level of patenting activity in each location and finally the share of patenting activity by Indian inventors in each MSA. For example, the San Francisco MSA received the largest number of patents by Indian inventors (and by all inventors). Indian inventors also received a relatively high share (11%) of the overall patents issued in that MSA.

These data illustrate that although inventive activity by the diaspora is geographically dispersed, it is not uniformly distributed (relative to the underlying distribution of overall patenting activity) but rather somewhat concentrated in particular cities such as San Francisco, New York, Chicago, and Austin. The Herfindahl index of concentration, calculated as $\sum_{i=1}^{N}(S_i^L)^2$, where S_i^L is the percentage share of patents issued in MSA i and N is the total number of MSAs, has a value of 667 for "Indian Patents" and 385 for "All Patents."

Finally, as we will discuss in the methodology section below, our identification strategy will need to address technological concentration by ethnicity. We offer descriptive data on this issue here. Table 3 shows the number of patents issued in each two-digit National Bureau of Economic Research (NBER)

At a more micro level, Borjas (1995) shows that ethnicity-based segregation at the level of neighborhoods slows down intergenerational wage convergence. But Alesina and Ferrara point out that some diverse societies are highly effective; work is continuing on the factors that make diversity an asset. Working in the tradition of Jacobs (1961), Ottaviano and Peri (2004) provide evidence of positive effects of diversity on the performance of US cities.

[5] Source: US Census Bureau and authors' calculations.

[6] In related work, Levin and Stephan (1999) and Stephan and Levin (2001) report that foreign-born and foreign-educated scientists and engineers (not necessarily from India) contribute disproportionately in terms of "exceptional contributions to US science" relative to what would be expected given their underlying distribution in the scientific labor force in the US.

[7] We describe our method for identifying Indian inventors in detail in the data section. Here we are measuring all inventors with Indian last names.

[8] Source: US Census Bureau, Current Population Survey, March Supplement, various years.

[9] The Indian-born population in the US numbered only 12,296 in the 1960 census. The population has grown dramatically in the last four decades, reaching 51,000 in 1970, 206,087 in 1980, 450,406 in 1990, and 1,022,552 in 2000. H-1B visas provided a major route of legal access to the US labor market in the 1990s for highly skilled individuals with job offers. Highly skilled Indians, especially those working in the computer industry, have been by far the largest beneficiaries of the H-1B visas. In fiscal year 2001, Indian-born individuals received almost half of all H-1Bs issued, 58% of which were in computer-related fields.

[10] Kapur (2004).

Table 2
Share of patents where at least one inventor is of Indian origin by location, US MSAs/CMSAs and Canadian CMAs (application year 1995)

MSA	Name	Indian patents	All patents	Indian share
7362	San Francisco Oakland San Jose, CA CMSA	2156	20396	10.6%
5602	New York Northern New Jersey Long Island, NY NJ CT	2017	17816	11.3%
1122	Boston Worcester Lawrence, MA NH ME CT CMSA	792	9660	8.2%
1602	Chicago Gary Kenosha, IL IN WI CMSA	770	7672	10.0%
4472	Los Angeles Riverside Orange County, CA CMSA	460	8862	5.2%
6162	Philadelphia Wilmington Atlantic City, PA NJ DE MD	429	5758	7.5%
640	Austin San Marcos, TX MSA	427	3147	13.6%
8872	Washington Baltimore, DC MD VA WV CMSA	386	4707	8.2%
6840	Rochester, NY MSA	286	3568	8.0%
1922	Dallas Forth Worth, TX CMSA	276	3887	7.1%
2162	Detroit Ann Arbor Flint, MI CMSA	253	5017	5.0%
7602	Seattle Tacoma Bremerton, WA CMSA	239	3720	6.4%
7320	San Diego, CA MSA	235	4312	5.4%
6640	Raleigh Durham Chapel Hill, NC MSA	220	2201	10.0%
6442	Portland Salem, OR WA CMSA	212	2211	9.6%
3362	Houston Galveston Brazoria, TX CMSA	202	3438	5.9%
5120	Minneapolis St. Paul, MN WI MSA	195	4967	3.9%
6280	Pittsburgh, PA MSA	174	1633	10.7%
1692	Cleveland Akron, OH CMSA	161	2703	6.0%
1080	Boise City, ID MSA	141	1009	14.0%
535	Toronto, ON, CMA (Canada)	140	1888	7.4%
520	Atlanta, GA MSA	135	2334	5.8%
7040	St. Louis, MO IL MSA	133	2088	6.4%
6200	Phoenix Mesa, AZ MSA	124	2198	5.6%
1642	Cincinnati Hamilton, OH KY IN CMSA	121	2460	4.9%
160	Albany Schenectady Troy, NY MSA	92	1362	6.8%
3480	Indianapolis, IN MSA	86	2144	4.0%
2082	Denver Boulder Greeley, CO CMSA	81	2634	3.1%
1840	Columbus, OH MSA	75	1193	6.3%
7160	Salt Lake City Ogden, UT MSA	62	1225	5.1%
3280	Hartford, CT MSA	33	1106	3.0%
4992	Miami Fort Lauderdale, FL CMSA	29	1202	2.4%
5082	Milwaukee Racine, WI CMSA	28	1323	2.1%
	Mean (MSAs listed above)	338	4238	7.0%
	Mean (all MSAs with non-zero patents)	45	613	4.1%

Note. Only locations with more than 1000 issued patents with application year 1995 are shown.

technology subcategory where at least one of the inventors is Indian.[11] The table also shows the total number of patents issued in each technology class and the share of patents where at least one of the inventors is Indian. Not surprisingly, computer hardware and software have the largest number of patents issued to Indians, who also have a relatively high share of the total number of patents issued in this class.

However, the table also shows that the impact of Indian inventors goes well beyond computers. Indeed, the highest Indian share is for organic compounds. Even so, Indian inventors are more technologically concentrated than overall inventors although the difference in concentration is less pronounced than for geographic concentration. The value of the Herfindahl index for technological concentration, for example, is 677 for patents with "One or More Indian Inventor" compared with 422 for "All Patents."

3. Estimating the knowledge flow production function

3.1. Definition of the KFPF

The KFPF measures the probability of a non-redundant knowledge flow to any inventor, i, from any other inventor, j (where $j \neq i$), based on well-defined structural relationships between the inventor pair (e.g., co-located, co-ethnic, co-specialist, etc.). We focus on the case where the existence of a given relationship is an all-or-nothing phenomenon (and thus can be measured by a dummy variable) but allow for a completely unrestricted set of interactions between the various types of relationships. We assume, however, that total knowledge flow from j to i is independent of both i's and j's relationships to other inventors and so abstract from issues of indirect access to knowledge through a network.[12]

[11] The three-digit patent classifications provided by the USPTO are mapped to 36 two-digit "subcategory codes" in Jaffe et al. (2002, pp. 452–454).

[12] See, for example, Burt (1992).

Table 3
Share of patents issued where one or more inventors are of Indian origin (application year 1995)

NBER subcategory	Description	One or more Indian inventors	All patents	Indian share
22	Computer Hardware & Software	528	9171	5.8%
31	Drugs	517	8873	5.8%
19	Miscellaneous-chemical	460	12205	3.8%
21	Communications	302	8532	3.5%
14	Organic compounds	301	4011	7.5%
15	Resins	242	5023	4.8%
46	Semiconductor devices	242	3776	6.4%
33	Biotechnology	235	5251	4.5%
69	Miscellaneous-others	188	10680	1.8%
45	Power Systems	114	4336	2.6%
24	Information Storage	113	3388	3.3%
12	Coating	97	2202	4.4%
52	Metal Working	90	3159	2.8%
41	Electrical Devices	79	3707	2.1%
43	Measuring & Testing	76	3665	2.1%
51	Mat. Proc. & Handling	76	5148	1.5%
32	Surgery & Med. Inst.	73	5444	1.3%
49	Miscellaneous-Elec.	64	3513	1.8%
42	Electrical Lighting	55	2154	2.6%
23	Computer Peripherials	54	2601	2.1%
59	Miscellaneous-Mechanical	54	5383	1.0%
54	Optics	44	3479	1.3%
53	Motors & Engines + Parts	35	3881	0.9%
44	Nuclear & X-rays	33	1559	2.1%
61	Agriculture, Husbandry, Food	32	2381	1.3%
55	Transportation	30	3450	0.9%
64	Earth Working & Wells	26	1303	2.0%
39	Miscellaneous-Drgs. & Med.	24	1010	2.4%
13	Gas	21	457	4.6%
11	Agriculture, Food, Textiles	13	802	1.6%
66	Heating	10	1104	0.9%
68	Receptacles	10	2299	0.4%
62	Amusement Devices	9	1473	0.6%
67	Pipes & Joints	9	912	1.0%
63	Apparel & Textile	7	1679	0.4%
65	Furniture, House Fixtures	6	2300	0.3%
	Total	4269	140311	3.0%

Letting R represent the total number of relationship types (e.g., co-located, co-member), K_{ij}, the probability of a knowledge flow from j to i is given by the general KFPF:

$$K_{ij} = \beta_0 + \sum_{s=1}^{S} \beta_s D_s. \tag{1}$$

The intercept in Eq. (1) is the probability of a knowledge flow when none of the relationships are present. S is the number of dummy variables required to represent all possible relationship types and all possible interactions between those relationship types.[13] Suppose, for example, three types of relationships are possible between an inventor pair. The number of dummy variables (S) required for a completely unrestricted model is then seven (three to capture the existence of each relationship, three to capture the interactions between each possible pair of relationships, and one to capture the interaction when all three of the relationships are present).

In this paper we focus on two types of relationships that potentially play an important role in facilitating knowledge flows between inventors: co-location and co-ethnicity. In this case, $R = 2$, $S = 3$, and the KFPF is given by:

$$\begin{aligned} K_{ij} &= \beta_0 + \beta_1(\textit{Co-Location}_{ij}) + \beta_2(\textit{Co-Ethnicity}_{ij}) \\ &\quad + \beta_3(\textit{Co-Location}_{ij} \times \textit{Co-Ethnicity}_{ij}), \quad j \neq i. \end{aligned} \tag{2}$$

The parameter on the interaction term determines whether co-location and co-ethnicity are complements or substitutes in the production of a knowledge flow. When β_3 is positive, the affect of co-location on the probability of a knowledge flow is greater for co-ethnic inventors; that is, co-location and co-ethnicity are complements. Conversely, co-location and co-ethnicity are substitutes when β_3 is negative. Later, we extend our analysis to consider distance between inventors rather than the binary measure of co-location and also explore whether communities of practice are complements to or substitutes for co-location.

3.2. Empirical methodology

Our objective is to identify the separate and joint effects of inventor co-location and inventor co-ethnicity on technological knowledge flows between inventors. The identification challenge is that inventive activity in particular technological areas is likely to be concentrated by location and ethnicity (Tables 2 and 3 show evidence of this). If this is true, we will observe a high incidence of citations among co-located and co-ethnic inventors even if co-location and co-ethnicity exert no causal influence on knowledge flows. Our identification strategy is to match each actual cited patent with a control patent that comes from the same technological class and time period as the actual cited patent. Assuming that the classes are sufficiently narrowly defined, the controls will have the same distribution across technologies as the actual citations, allowing us to control for incidental co-location and co-ethnicity effects.

With the controls selected, we estimate the effects of interest from the following simple regression using citations as a proxy for knowledge flows[14]:

[13] With R relationship types, the number of dummy variables needed, S, equals the number of possible combinations of relationship types from the set R, taken $r = 1, 2, \ldots, R$ at a time. This is given by the combinatorial formula: $S = \sum_{r=1}^{R} \frac{R!}{r!(R-r)!} = 2^R - 1$.

[14] Citations are not, however, straightforward to interpret. Patents cite other patents as "prior art," with citations serving to delineate the property rights conferred. Some citations are supplied by the applicant, others by the patent examiner (Alcacer and Gittelman, 2006; Hegde and Sampat, 2007), and some patents may be cited more frequently than others because they are more salient in terms of satisfying legal definitions of prior art rather than because they have greater technological significance. Cockburn et al. (2002) report, for example, that some examiners have "favorite" patents that they cite preferentially because they "teach the art" particularly well. Nonetheless, we are of the opinion that even examiner-added citations may reflect a knowledge flow. Jaffe et al. (2002) surveyed cited and citing inventors to explore the "meaning of patent citations"

$$P(\mathit{Citation}_{ij}) = \beta_0 + \beta_1 \mathit{Co\text{-}Location}_{ij} + \beta_2 \mathit{Co\text{-}Ethnicity}_{ij} + \beta_3(\mathit{Co\text{-}Location}_{ij} \times \mathit{Co\text{-}Ethnicity}_{ij}) + \varepsilon_{ij}, \quad i \neq j.$$

Citation, our dependent variable, is a dummy that takes a value of one when the observation relates to an actual citation and zero when it relates to a control. We index the citing inventor by i and the cited (or control) inventor by j. By construction, our sample contains an equal number of actual and control observations.[15] *Co-Location* is a dummy variable that takes a value of one when the original and cited (or control) inventors are located in the same MSA and zero otherwise. *Co-Ethnicity* is a dummy variable that takes a value of one when the cited (or control) inventor has an Indian surname (the original inventor always has an Indian surname, by construction). We exclude all self cites ($i = j$).[16]

To see how this regression allows us to identify the causal effects of interest, note that if the control matching procedure is effective and no causal link exists from co-location and co-ethnicity to citations, the coefficients on β_1, β_2, and β_3 should all be zero. Put differently, if we have well-matched controls and if no causal relationships are present, then information on co-location and co-ethnicity would not be helpful in predicting whether a given observation is an actual citation or a control.

What economic interpretation can be given to the coefficients? Suppose we observe a particular citation. For the cited patent, we can identify the entire set of patents from the same technological area and time period as the actual cited patent, what we call the control set. The coefficients allow us to calculate the increase in the probability of a citation relative to a random patent from the control set for various combinations of co-location and co-ethnicity between the inventors of the focal and cited patent. For example, suppose we are dealing with a citation where the focal and cited inventors are co-located but not co-ethnic. Suppose further that the estimated values of β_0 and β_1 are 0.50 and 0.25, respectively. These estimates imply that co-location is associated with a 50% increase in the probability of a citation relative to a random (non-co-ethnic) member of the control set.

The results allow us to test for statistically significant co-location effects (separately for both co-ethnic and non-co-ethnic inventors) and also for co-ethnicity effects (again separately for co-located and non-co-located inventors). A test of the significance of the interaction coefficient, β_3, provides a very direct way to determine whether co-location and co-ethnicity are significant complements or substitutes. For example, we would not be able to reject the null of complementarity if β_3 is statistically significant and positive.

The foregoing discussion underlines the key challenge associated with our method. A test of the null hypothesis of no co-location effect for non-co-ethnic inventors ($\beta_1 = 0$), for example, is actually a test of the joint hypothesis that we have effectively matched the controls and that no causal link is present from co-location to knowledge flows. A rejection of this null could follow from ineffective matching and/or the presence of a causal relationship. For this reason, we focus in detail in the next section on the method we use to make the control matches and discuss in the results section the likely robustness of particular findings to residual inadequacies in the matching procedure.

3.3. Data and sample construction

Data Source

We use the "front page" bibliographic data for patents published by the USPTO as the basis of the empirical work. These data contain the application and issue dates of each patent, the names and locations of the inventor(s), a technology classification, and a list of patents cited. We augment these data with a list of Indian names and the NBER Patent-Citations data file for additional fields, including the two-digit technology classification subcategory code.

We generate Indian name data from a list of 213,622 unique last names compiled by merging the phone directories of four of the six largest cities in India: Bangalore, Delhi, Mumbai (Bombay), and Hyderabad. Of these, 38,386 names appeared with a frequency of five or more. Of these, 13,418 matched a proprietary database of US consumers.[17] Finally, one of the authors and an outside expert coded each of these names as: (1) extremely likely to be Indian, (2) extremely unlikely to be Indian, or (3) could be either. The list of names used for this study includes only the 6885 last names that were coded as "extremely likely to be Indian."[18]

Unit of Analysis

Our unit of analysis is the inventor-patent-citation. Thus, a single patent that has two inventors and cites five prior patents

and find that approximately one-quarter of the survey responses correspond to a "fairly clear spillover," approximately one-half indicate no spillover, and the remaining quarter indicate some possibility of a spillover. Based on their survey data, the authors conclude that "these results are consistent with the notion that citations are a noisy signal of the presence of spillovers. This implies that aggregate citation flows can be used as proxies for knowledge-spillover intensity, for example, between categories of organizations or between geographic regions" (p. 400).

[15] Recall that we match each cited patent with a control patent. Although we disaggregate inventors when there are more than one on the focal patent, we treat multiple inventors as a set in the context of cited and control patents.

[16] We also conduct robustness checks where we remove examiner-added citations. The results become slightly stronger.

[17] InfoUSA prepared this database.

[18] We do not expect the frequency of false positives in our name data to be large. In a random phone survey ($N = 2256$), 97% of the individuals with last names from our sample list responded "yes" to the question: "Are you of Indian origin?" (Kapur, 2004). Nor do we expect the frequency of false negatives to be particularly large. Although we constructed our name set from the phone books of large metropolitan cities, the vast majority of Indian overseas migration to the United States is an urban phenomenon; the likelihood of an urban household in India having a family member in the US is more than an order of magnitude greater than a rural household. A different problem arises when people change their last name after migration. This is more likely with Indian women due to marriage. However, even among second-generation Asian-Americans, Indian-American women are least likely to marry outside the ethnic group (62.5% marry within the ethnic group, Le, 2004). Noise in our name data will bias our result downwards.

will generate ten unique observations. We employ this unit of analysis rather than simply patents since we are interested in the flow of knowledge between individuals rather than between inventions.

Control Patents

As noted above, the main methodological challenge in identifying the effects of co-ethnicity and co-location on knowledge flows is to control for the ethnic and locational clustering of inventive activity in particular technological areas at particular points in time. For example, we might observe Indian inventors in computer-related technologies residing in Silicon Valley citing a large number of other Indian inventors working in computer-related technologies and residing in Silicon Valley. This high level of co-ethnic and co-located cites could simply reflect the law of averages, as a relatively large fraction of inventors employed in Silicon Valley are of Indian origin and are working on computer-related technologies. Conversely, it could be because the combined effects of co-ethnicity and co-location are facilitating knowledge flows between inventors in this sector.

To address this issue, we build on a procedure developed by Jaffe et al. (1993) and refined by Thompson and Fox-Kean (2005) to identify a control patent for each observation.[19] We select a control patent for each observation that matches the cited patent on the following dimensions: (1) application year and (2) technology classification. While Jaffe et al. select controls from the set that matches the three-digit primary classification of the citing patent and Thompson and Fox-Kean enhance the methodology by selecting controls from the set that match on a single primary and secondary six-digit classification, we further refine the process and select from the set that matches on the highest possible number of six-digit classifications.[20]

In addition, we confirm that the control patent does not cite the original patent. If it does, we remove the patent from the set of potential controls and select the next best control patent. Finally, if no patents match the cited patent in at least the application year and the three-digit primary classification without being cited by the original patent, we remove the observation (original patent) from the data set.

We identify co-ethnic and co-localization effects as the extent to which the frequency of citations to co-ethnic or co-located inventors is over and above what we would expect given the ethnic and geographic distributions of inventive activity in the particular technological area of the cited patent.[21] The geographic or ethnic clustering of innovative activity in certain technology areas itself may be due to the lowered cost of establishing social relationships but also to something else such as thicker factor markets. Thus, focusing on knowledge flows that are more concentrated than the innovative activity in that particular field is a conservative approach.

Co-ethnicity Metrics

We examine the last name(s) of the inventor(s) on the cited patent associated with each observation. If at least one inventor has an Indian name, the inventor is designated as “of Indian origin” and we define the original and cited patents as “co-ethnic" (the former is Indian by construction). We do the same for control patents. Thus, although we disaggregate inventors on the focal patent if there are more than one, we consider multiple inventors as a set in the context of cited and control patents (i.e., a cited or control patent is Indian if any inventor in the set is Indian).

Co-location Metrics

We also examine the home address of the inventor for each observation. We assign inventors to an MSA based on their city and state information.[22] There are 268 US MSAs and consolidated metropolitan statistical areas (CMSAs) and 25 Canadian census metropolitan areas (CMAs), hereafter collectively referred to as “MSAs.”[23] We also create 63 “phantom MSAs” for individuals located in one of the 50 states or 13 provinces or territories that are in cities not assigned to one of the Census Bureau-defined MSAs. Once again we disaggregate inventors on the focal patent if there are more than one but consider multiple inventors as a set in the context of cited and control patents. Thus, we define the focal and cited (control) patents as co-located if the focal inventor is assigned to the same MSA as any inventor on the cited (control) patent.

[19] The Jaffe et al. and Thompson and Fox-Kean approaches involve the analysis of forward citations. To take advantage of the substantial growth in the Indian-born population in the US post-1990, our approach is to look backward to prior patents that are being cited by the patents granted to Indian inventors between 2001 and 2003. One can use either approach (backwards or forwards citations) to test for a disproportionate incidence of co-located or co-ethnic knowledge flows.

[20] We are able to find controls that match on more than one six-digit classification for approximately 60% of the observations in the sample (37% match on one six-digit classification and only 2% match on the three-digit primary classification). We only use observations for which we are able to find a control patent that matches on at least one six-digit classification. As a result, approximately 40% of our sample has controls that are as closely matched as those in Thompson and Fox-Kean and 60% of our sample has controls that are more closely matched.

[21] We also check whether “Indian patents” are cited more frequently than their non-Indian counterparts. Specifically, within cited patents with the same application year and three-digit classification code, we compare the total number of all citations received by Indian versus non-Indian patents. We do the same within control patents. The difference is not statistically significant.

[22] We use city and country information to assign Canadian inventors to a CMA.

[23] While MSAs and CMAs are similar in spirit, they are defined slightly differently. The Canadian criterion requires that the urban core have a population of at least 100,000 for a metropolitan area to exist. In contrast, for the period 1990–2000, the United States had two criteria to determine whether or not a metropolitan area existed: (1) where there is either a city of 50,000 or more inhabitants or (2) where there is a Census Bureau-defined urban area, i.e., a population of at least 50,000 and a total metropolitan population of at least 100,000 (75,000 in New England). Thus, the Canadian approach is the more restrictive of the two. We include Canada since this nation’s Indian-born population follows similar patterns to that of the US and our prior research on knowledge flows and social relationships included Canadian MSA data (Agrawal et al., 2006), facilitating comparison between the two studies. Also, the results presented remain almost identical when only US MSAs are included.

Distance Metrics

In addition to the binary co-location measure, we calculate the distance between inventors measured in miles. To construct this measure, we collect the coordinates (longitude and latitude) of the focal inventor at the city-state level. If the inventor did not reside in the US or Canada, we use the coordinates of the capital city of her country. If an inventor did reside in the US or Canada but coordinates could not be found for her city/region, we use the coordinates of the capital city of her state or province (the results are robust to dropping these observations). We use an identical approach for generating the coordinates of inventors on the cited and control patents. However, when cited or control patents have multiple inventors, we collect separate coordinates for each inventor, calculate distances from the focal inventor, and select the minimum as our value for distance. Finally, we use the great circle formula to calculate the distance between each pair of inventors.[24]

Sample Construction

We generate our sample by identifying all patents issued by the USPTO during the period 2001–2003, which totals 555,741. From this set, we identify those patents that have at least one inventor of Indian origin, which totals 19,612. On average, each of these patents has approximately 3.5 inventors and cites 16 prior patents. Since our unit of observation is the inventor-patent-citation, this results in 1,072,684 observations. Next, we remove those observations for which the inventor of the focal patent does not have an Indian name (although they co-invented with somebody who does) and those observations for which we are unable to identify a control for the cited patent. Finally, we remove actual-control pairs that do not match on the primary classification (even though they match on one or more six-digit classifications as described above); after following these procedures our sample includes 130,944 citing-cited pairs and an equal number of citing-control pairs for a total of 261,888 observations.

4. Results

Table 4 records the means of the variables that enter into our estimated KFPFs. The means are recorded separately for the actual and control citations. A comparison of the means shows that inventors of cited patents are more often co-located with focal inventors than are inventors of control patents. Similarly, the mean distance between cited and focal inventors is smaller than between control and focal inventors. Also, cited inventors have a higher incidence of co-ethnicity than do the control inventors.

The table also records the means of the interaction terms that will be a central focus of our regression analysis. The two-way interaction means provide a measure of the joint occurrence of three forms of relationships. In all cases, the joint occurrence of any pair of the three relationships—co-ethnicity, co-location, co-technology—is higher for the actual citations than for the controls. For example, the joint occurrence of co-location and co-ethnicity occurs for 1.5% of the actual citations compared with 1.1% for the controls, a difference that is statistically significant at the 1% level. However, these simple mean comparisons do not allow us to identify how the various forms of knowledge-flow facilitating relationships complement or substitute for one another. For that we need to estimate the parameters of the KFPF.

Table 5 records our estimated KFPFs where we use co-location as the single measure of geographic proximity. We explore two specifications. The first includes a co-location dummy, a co-ethnicity dummy, and the interaction between the two. The second adds two-way and three-way interactions with the co-technology dummy. We do not include a direct technology variable in the regressions because the technology classes of the actual citations and control citations are the same by construction. We estimate each specification using both Ordinary Least Squares (OLS) and Logit.[25] We concentrate on the OLS results in our discussion since both methods produce identical probabilities of an actual citation conditional on any given set of relationships.

The results show that both co-location and co-ethnicity significantly increase the probability that a "citation" is an actual citation rather than a control citation. Focusing on the first regression, co-location increases the probability that the observation is an actual citation by just under 12 percentage points, and co-ethnicity increases the probability of a citation by 7 percentage points. Thus, being co-located increases the probability of a patent being the one cited from a given control set by (0.1187/0.4853) or 24% relative to a non-co-ethnic inventor in the control set.[26] Using a similar calculation, being co-ethnic increases the probability of a patent being cited from a given control set by (0.0701/0.4343) or 14% relative to a non-co-located member of the control set. To the extent that our method of choosing the controls is effective (more on that below), these results are consistent with the hypotheses that co-location and co-ethnicity play strong causal roles in facilitating knowledge flows between inventors.

The most interesting finding is the large negative and statistically significant coefficient on the interaction term, $\hat{\beta}_3$. This result can be interpreted as evidence that co-location and co-ethnicity are substitutes in facilitating knowledge flows. In terms of the difference in marginal impact, co-location increases the probability of a knowledge flow by 24% for non-co-ethnic inventors but only by (0.0127/0.5554) or 2% for co-ethnic inventors. In other words, the marginal effect of co-

[24] $Distance = 3962.6 * a\cos((\sin(lat1) * \sin(lat2)) + (\cos(lat1) * \cos(lat2) * \cos(long1 - long2)))$.

[25] The reported standard errors are robust to the non-independence of observations drawn from clusters of observations based on the same citing patent. To see the potential for non-independence, take the example of two co-located Indian inventors on a given citing patent. A single citation made by this patent will generate four observations in our data set (two actual citations and two control citations). The value of the dependent variable (and thus the error term in the regression) will be the same for the two actual citations and also for the two control citations.

[26] Recall that the control set for a citation is the set of all patents occurring in the same technological field and time period as the actual cited patent.

Table 4
Descriptive statistics for regression variables

	Obs.	Actual citations		Control citations		Mean
		Mean	St. Dev.	Mean	St. Dev.	
Citation	261,888	1	0	0	0	1
Co-location	261,888	0.1260*	0.3318	0.0842*	0.2276	0.0418*
		(0.0009)		(0.0008)		(0.0012)
Distance/1000	261,888	2.2818*	2.2157	2.7336*	2.3323	−0.4518*
		(0.0061)		(0.0064)		(0.0089)
Co-ethnicity	261,888	0.0588*	0.2353	0.0465*	0.2105	0.0123*
		(0.0007)		(0.0006)		(0.0009)
Co-location	261,888	0.0150*	0.1215	0.0114*	0.1062	0.0036*
× Co-ethnicity		(0.0003)		(0.0003)		(0.0004)
Distance/1000	261,888	0.0649*	0.4134	0.0529*	0.3759	0.0120*
× Co-ethnicity		(0.0011)		(0.0010)		(0.0015)
Co-location	261,888	0.0758*	0.2646	0.0544*	0.2268	0.0214*
× Co-technology		(0.0007)		(0.0006)		(0.0010)
Distance/1000	261,888	1.2809*	1.9987	1.5033*	2.1823	−0.2224*
× Co-technology		(0.0055)		(0.006)		(0.0082)
Co-ethnicity	261,888	0.0351*	0.1839	0.0283*	0.1658	0.0068*
× Co-technology		(0.0005)		(0.0005)		(0.0007)
Co-location × Co-ethnicity	261,888	0.0094*	0.0965	0.0076*	0.0870	0.0018*
× Co-technology		(0.0003)		(0.0002)		(0.0004)
Distance/1000 × Co-ethnicity	261,888	0.0385*	0.3243	0.0314*	0.2933	0.0071*
× Co-technology		(0.0009)		(0.0008)		(0.0012)

Notes. 1. The value of the citation dummy is one for an actual citation and zero for a control citation. 2. Inventor and assignee self cites are excluded. 3. See text for description of how control citations are chosen. 4. Standard errors are in parentheses.

* Significance at the 1% level.

location is much smaller for inventors who are already connected through some other mechanism.

We offer two caveats with respect to the interpretation of these data. First, while Thompson and Fox-Kean (2005) demonstrate the benefits of refining the procedure for choosing controls (which we have further refined here), they also express concern that an adequate control selection procedure can ever be found. Although we have made significant efforts to select control patents that closely match cited patents in terms of technology class and year, concerns may remain that the controls are not matched closely enough. If the matches are not close enough such that innovative activity is concentrated by technology areas that are more finely defined than our controls, our co-ethnicity estimates may be biased upwards. In other words, β_2 will be biased if innovative activity is ethnically concentrated in technological areas more narrowly defined than those captured by the controls, perhaps for reasons other than localized knowledge flows.

We recognize this concern and therefore consider the co-ethnicity results ($\beta_2 > 0$) with caution. However, imperfect controls are less likely to bias the main result that co-ethnicity substitutes for co-location. Substitution is reflected in β_3, the negative and statistically significant coefficient on the interaction between co-location and co-ethnicity. Imperfect controls would only bias this estimate if the controls for citations that are co-ethnic and co-located are systematically better or worse than the average control. If the selected controls are systematically worse for co-ethnic and co-located inventors (one might imagine this is possible in a scenario where co-located and co-ethnic inventors are working in a very specialized technology area), this would bias our estimate upwards, in the opposite direction of our finding. However, in order for the bias to work in the same direction as our finding, the control patent would have to be systematically better for co-ethnic and co-located inventors. We are not aware of any conditions under which this would be true.

The third and fourth columns of Table 5 show the effects of allowing for co-technology interactions. The key question here is how being co-technologists affects the knowledge-flow facilitation role of co-location and co-ethnicity. The estimated KFPFs again show large co-location and co-ethnicity effects and a large negative interaction between the two. We now also find evidence of a substantial negative interaction between co-location and co-technology so that co-location matters less for knowledge flows when inventors are working the same technological area. For inventors working in different technology fields, co-location increases the probability of a citation by approximately 30% compared to an increase of 20% for inventors working in the same field.[27] The

[27] This is consistent with a prior finding that co-location results in a larger increase in the probability of a cross-field citation (from one technology field to another) than in the probability of a within-field citation (Agrawal et al., 2006). We attribute the lower marginal impact of co-location for within-field cites to a greater likelihood of alternative channels for establishing social relationships through communities of practice.

Table 5
Estimated knowledge flow production functions

Dependent Variable = Dummy for Actual Citation	OLS	Logit	OLS	Logit
Co-location	0.1187*	0.4810*	0.1467*	0.5997*
	(0.0049)	(0.0205)	(0.0064)	(0.0274)
Co-ethnicity	0.0701*	0.2813*	0.0718*	0.2881*
	(0.0064)	(0.0258)	(0.0094)	(0.0382)
Co-location × Co-ethnicity	−0.1060*	−0.4294*	−0.1064*	−0.4344*
	(0.0128)	(0.0520)	(0.0207)	(0.0856)
Co-location × Co-technology			−0.0454*	−0.1911*
			(0.0080)	(0.0336)
Co-ethnicity × Co-technology			−0.0028	−0.0115
			(0.0117)	(0.0473)
Co-location × Co-ethnicity			−0.0029	0.0168
× Co-technology			(0.0247)	(0.1016)
Constant	0.4853*	0.0588*	0.4853*	−0.0588*
	(0.0006)	(0.0025)	(0.0006)	(0.0025)
Observations	261,888	261,888	261,888	261,888
Number of citing patents	10,674	10,674	10,674	10,674
R^2	0.0054		0.0056	
Pseudo R^2		0.0039		0.0041

Notes. 1. Inventor and assignee self cites are excluded. 2. See text for description of how control citations are chosen. 3. Standard errors are in parentheses; standard errors are robust to citing-patent cluster effects.

* Significance at the 1% level.

coefficient on the interaction between co-ethnicity and co-technology is also negative but the size of the effect is small and statistically insignificant. This implies that the facilitating role of co-ethnicity is similar for inventors working in the same or in different technological areas.[28] Finally, the coefficient on the three-way interaction is statistically insignificant. This means that co-technology does not significantly influence the observed negative interaction between co-location and co-ethnicity and that co-ethnicity does not significantly influence the observed negative interaction between co-location and co-technology.[29]

Table 6
Estimated knowledge flow production functions with distance as geographic variable

Dependent Variable = Dummy for Actual Citation	OLS	Logit	OLS	Logit
Distance/1000	−0.0215*	−0.0866*	−0.0226*	−0.0910*
	(0.0006)	(0.0026)	(0.0007)	(0.0029)
Co-ethnicity	0.0116***	0.0466***	0.0251**	0.1015**
	(0.0068)	(0.0277)	(0.0105)	(0.0428)
Distance/1000 ×	0.0167*	0.0672*	0.0127**	0.0509**
Co-ethnicity	(0.0038)	(0.0155)	(0.0060)	(0.0242)
Distance/1000 ×			0.0020*	0.0081
Co-technology			(0.0006)	(0.0023)
Co-ethnicity ×			−0.0218***	−0.0884***
Co-technology			(0.0117)	(0.0122)
Distance/1000 × Co-ethnicity			−0.0061	0.0247
× Co-technology			(0.0071)	(0.0287)
Constant	0.5524*	0.2105*	0.5523*	0.2103*
	(0.0016)	(0.0065)	(0.0016)	(0.0065)
Observations	261,888	261,888	261,888	261,888
Number of citing patents	10,674	10,674	10,674	10,674
R^2	0.0101		0.0101	
Pseudo R^2		0.0073		0.0073

1. Inventor and assignee self cites are excluded. 2. See text for description of how control citations are chosen. 3. Standard errors are in parentheses; standard errors are robust to citing-patent cluster effects.

* Significance at the 1% level.
** Idem, 5%.
*** Idem, 10%.

Table 6 repeats the analysis using distance as an alternative measure of geographical proximity. The results are broadly consistent with Table 5. The coefficient on distance is negative and statistically significant. A 1000-mile increase in distance results in an approximately two percentage point reduction in the probability that the observation is an actual citation; once again we see that distance impedes knowledge flow. Consistent with our earlier finding of the substitutability of geographical proximity and co-ethnicity, we find a positive interaction between distance and the co-ethnicity dummy. This reinforces our earlier finding that co-ethnicity plays a more important role for more geographically separated inventors in terms of facilitating knowledge flows. As a specific example, the estimated coefficients of our KFPF indicate that while the marginal effect of co-ethnicity between inventors that are 1000 miles apart is to increase the likelihood of a knowledge flow by only 5%, the effect between inventors that are 3000 miles apart is to increase the likelihood by 13%.

The pattern of interactions between co-technology and the other forms of relationship are also broadly similar to those found in Table 5. Working in the same technological field is more facilitative of knowledge flows when inventors are further apart. Looked at another way, the advantage of geographical proximity tends to decline when inventors are working in the same technological field. We again find a negative interaction between co-ethnicity and co-technology, which is here signifi-

[28] Of course, since we are looking at patent citations, the technology of the cited patent must be relevant to the inventor on the focal patent. The question here is whether co-location and co-ethnicity are more or less important for citations that occur across rather that within technological fields.

[29] To examine how these effects vary across industries, we estimate separate KFPFs for each of the one-digit NBER technology classifications. The regressions use co-location and the geographic proximity variable and also include co-technology interactions. Although differences exist across the broad technology categories, the patterns of coefficients are overall quite similar to those reported for all technologies in the third column of Table 5. The coefficients on co-location range from a low of 0.1152 in Drugs & Medical to a high of 0.1906 in Mechanical (all statistically significant at the 1% level). The coefficients on co-ethnicity exhibit a wider range, from a low (and statistically insignificant) 0.032 in Electrical & Electronic to a high of 0.1558 in “Other.” Most interestingly, given the focus of this paper, the two-way interactions are generally negative in sign, although statistical significance varies substantially, reflecting in part the relatively small sizes of technology-specific sub-samples.

cant at the 10% level. The three-way interaction is once again statistically insignificant.[30]

Separately, we estimate the KFPF with just the distance and co-ethnicity variable (i.e., without the interaction). The estimated coefficients on distance/1000 and co-ethnicity are −0.0225643 and 0.0307164, respectively. Taking the ratio, these results imply that co-ethnicity has an equivalent effect on knowledge flows as being nearer by 690 miles.

Summing up these findings, we observe a general pattern of substitutability between the relationships that we have hypothesized affect knowledge flows. Geographical proximity matters most in the absence of social proximity that may otherwise facilitate access to knowledge. In terms of methodology, we have shown that the KFPF is a useful device for thinking about the drivers of knowledge flows and that one can empirically implement the function using patent citation data with careful attention to the choice of controls. We turn to the implications of these findings in the concluding section.

5. Conclusion

Our results show that, considered independently, co-location and co-ethnicity both enhance knowledge flows between inventors. However, in terms of their interaction, co-location and co-ethnicity substitute for rather than complement one another. What are the implications of our findings? In Agrawal et al. (2007), we use the KFPF to develop three simple models to determine the optimal distribution of ethnic inventors across the economy from the perspective of: (1) the city, (2) the national economy, and (3) individual ethnic inventors. In the context of the first model, where the city's simple objective is to maximize access to knowledge by its stock of local inventors, the estimated negative coefficient on the interaction between co-location and co-ethnicity is a sufficient condition for an ethnically diverse population of inventors to be optimal (rather than a homogeneous population).

In the context of the second model, where the country's objective is to distribute ethnic inventors across cities in order to maximize aggregate knowledge access over all cities, the estimated negative coefficient on the interaction between co-location and co-ethnicity is a sufficient condition for an equally distributed ethnic inventor mix to be optimal (rather than, say, all ethnic inventors being concentrated in one city). Finally, in the third model, where the individual ethnic inventor's objective is to maximize knowledge access for herself, the positive value of the sum of the estimated coefficient on co-location plus the estimated coefficient on the interaction is a sufficient condition for concentration—not dispersion—to be optimal (the only stable equilibrium).

Collectively, the models presented in Agrawal et al. (2007) both motivate the importance of the main empirical finding reported in this paper that co-location and co-ethnicity are substitutes while also illustrating the tension that arises since they are substitutes. The tension results from the model implication that dispersion is optimal from the perspective of the city and the national economy but concentration is optimal from the perspective of ethnic inventors themselves. Of course with free mobility of inventors, the actual dispersion across locations will be the result of numerous individual inventor decisions about where to live and work. As such, our empirical findings invite obvious policy responses, at least within the context of these simple models.

Overall, our paper points to the economic importance of social and spatial proximity in terms of mediating knowledge flow patterns. However, critical questions remain. Perhaps most urgent is the underlying mechanism that gives our key measure of social proximity—co-ethnicity—its economic salience. Do last names serve a cuing or reputational function for co-ethnics? Do co-ethnics benefit from lower cost access to latent knowledge (Agrawal, 2006) arising from social interactions predicated on common social circles, places of worship, or schools from which they graduated? Are these effects likely to be stronger or weaker for other channels of knowledge production, such as academic publishing? We need to understand the underlying mechanisms in order to draw general conclusions.

Our findings, along with others (Kapur and McHale, 2005; Nanda and Khanna, 2006), also point to the need to extend the scope of immigration models beyond just labor market effects to include the impact on knowledge flows and innovation. Moreover, our paper suggests that through a mix of location choice (relative to the location of related innovative activity) and recruitment decisions (in terms of social connections, or ethnic diversity in our specific case), firms may influence their innovation productivity. Indeed, the increased pace of recruitment of international talent in academia and private-sector labs as well as the rapid expansion of multinational R&D to international locations over the past quarter century suggests that firms may have already well recognized these important determinants of knowledge flow patterns.

Acknowledgments

We thank Thomas Astebro, Iain Cockburn, Bill Cooper, Jeff Furman, Zeynep Hansen, Ramana Nanda, Joanne Oxley, Bhaven Sampat, Brian Silverman, Jasjit Singh, Olav Sorenson, Daniel Trefler, and Arvids Ziedonis as well as seminar participants at NBER, MIT, ISNIE, Queen's University, University of Toronto, and the Canadian Economics Association who offered helpful comments. We also thank Alex Oettl who provided excellent research assistance. Errors and omissions are our own. The Social Sciences and Humanities Research Council of Canada (Grant No. 410-2004-1770) and Harvard University's Weatherhead Initiative grant funded this research. We gratefully acknowledge their support.

[30] We also run specifications with both the co-location dummy and distance variables included together. Not surprisingly, the results show that each matters, with the patterns for signs and significance broadly matching the estimated effects when we include co-location and distance separately. This suggests both a premium for pure proximity (i.e., co-location in an MSA) and also a negative gradient with the distance from an inventor's MSA.

References

Agrawal, A., 2006. Engaging the inventor: Exploring licensing strategies for university inventions and the role of latent knowledge. Strategic Management Journal 27 (1), 63–79.

Agrawal, A., Cockburn, I., McHale, J., 2006. Gone but not forgotten: Labor flows, knowledge spillovers, and enduring social capital. Journal of Economic Geography 6 (5), 571–591.

Agrawal, A., Kapur, D., McHale, J., 2007. Birds of a feather—Better together? Exploring the optimal spatial distribution of ethnic inventors. Working paper 12823, National Bureau of Economic Research.

Alcacer, J., Gittelman, M., 2006. How do I know what you know? Patent examiners and the generation of patent citations. Review of Economics and Statistics 88 (4), 774–779.

Alesina, A., Ferrara, E.L., 2005. Ethnic diversity and economic performance. Journal of Economic Literature 43 (3), 762–800.

Borjas, G.J., 1995. Ethnicity, neighborhoods, and human-capital externalities. American Economic Review 85 (3), 365–390.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard Univ. Press, Cambridge, MA.

Cockburn, I., Kortum, S., Stern, S., 2002. Are all patent examiners equal? The impact of examiner characteristics on patent statistics and litigation outcomes. Working paper 8980, National Bureau of Economic Research.

Coleman, J., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, S95–S120.

Easterly, W., Levine, R., 1997. Africa's growth tragedy: Policies and ethnic division. Quarterly Journal of Economics 112 (4), 1203–1250.

Glaeser, E.L., Laibson, D., Sacerdote, B.I., 2002. The economic approach to social capital. Economic Journal CXII, F437–F458.

Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology LXXIII, 1360–1380.

Griliches, Z., 1957. Hybrid corn: An exploration in the economics of technological change. Econometrica 25 (4), 501–522.

Hegde, D., Sampat, B., 2007. Applicant citations, examiner citations, and the private value of patents. Working paper, University of California at Berkeley.

Jacobs, J., 1961. The Death and Life of Great American Cities. Random House, New York, NY.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge flows as evidenced by patent citations. Quarterly Journal of Economics CVIII, 577–598.

Jaffe, A., Trajtenberg, M., Fogarty, M., 2002. The meaning of patent citations: Report on the NBER/Case Western Reserve Survey of Patentees. In: Jaffe, A., Trajtenberg, M. (Eds.), Patents, Citations, and Innovations: A Window on the Knowledge Economy. The MIT Press, pp. 379–402.

Kalnins, A., Chung, W., 2006. Social capital, geography, and survival: Gujarati immigrant entrepreneurs in the US lodging industry. Management Science 52 (2), 233–247.

Kapur, D., 2004. Survey of Indian Americans in the United States (SAIUS). Working paper, Harvard University.

Kapur, D., McHale, J., 2005. Sojourns and software: Internationally mobile human capital and high-tech industry development in India, Ireland, and Israel. In: Arora, A., Gambardella, A. (Eds.), From Underdogs to Tigers: The Rise and Growth of the Software Industry in Some Emerging Economies. Oxford Univ. Press, Oxford, UK.

Kerr, W., 2005. Ethnic scientific communities and international technology diffusion. Working paper, Harvard University.

Knack, S., Keefer, P., 1997. Does social capital have an economic payoff? A cross-country investigation. Quarterly Journal of Economics 112 (4), 1251–1288.

Le, C.N., 2004. Socioeconomic Statistics and Demographics. http://www.asian-nation.org/demographics.shtml. Accessed 22 July 2004, Asian-Nation: The Landscape of Asian America.

Levin, S., Stephan, P., 1999. Are the foreign-born a source of strength for US science. Science 285, 1213–1214.

Nanda, R., Khanna, T., 2006. Diasporas and domestic entrepreneurs: Evidence from the Indian software industry. Working paper, MIT.

Ottaviano, G., Peri, G., 2004. The economic value of cultural diversity. Working paper 10904, National Bureau of Economic Research.

Putnam, R.D., 2002. Bowling Alone. Simon & Schuster, New York, NY.

Rauch, J.E., 2001. Business and social networks in international trade. Journal of Economic Literature XXXIX, 1177–1203.

Romer, P., 1990. Endogenous technological change. Journal of Political Economy 98 (5).

Saxenian, A.L., 1999, "Silicon Valley's new immigrant entrepreneurs. The Public Policy Institute of California, San Francisco, CA.

Sorenson, O., Rivkin, J.W., Fleming, L., 2006. Complexity, networks, and knowledge flow. Research Policy 35, 994–1017.

Stephan, P., Levin, S., 2001. Exceptional contributions to US science by the foreign-born and foreign-educated. Population Research and Policy Review 20, 59–79.

Thompson, P., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review 95 (1), 450–460.

Zucker, L., Darby, M., Brewer, M., 1998. Intellectual capital and the birth of US biotechnology enterprises. American Economic Review 88 (1), 290–306.