NBER WORKING PAPER SERIES

BIRDS OF A FEATHER - BETTER TOGETHER? EXPLORING THE OPTIMAL SPATIAL DISTRIBUTION OF ETHNIC INVENTORS

Ajay Agrawal
Devesh Kapur
John McHale

Working Paper 12823
http://www.nber.org/papers/w12823

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2007

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. This research was funded by the Social Sciences and Humanities Research Council of Canada (Grant No. 410-2004-1770) and by Harvard University's Weatherhead Initiative grant. Their support is gratefully acknowledged. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Birds of a Feather - Better Together? Exploring the Optimal Spatial Distribution of Ethnic Inventors

Ajay Agrawal, Devesh Kapur, and John McHale

NBER Working Paper No. 12823
January 2007
JEL No. O33,R12,Z13

ABSTRACT

We examine how the spatial and social proximity of inventors affects knowledge flows, focusing especially on how the two forms of proximity interact. We develop a knowledge flow production function (KFPF) as a flexible tool for modeling access to knowledge and show that the optimal spatial concentration of socially proximate inventors in a city or nation depends on whether spatial and social proximity are complements or substitutes in facilitating knowledge flows. We employ patent citation data, using same-MSA and co-ethnicity as proxies for spatial and social proximity, respectively, to estimate the key KFPF parameters. Although co-location and co-ethnicity both predict knowledge flows, the marginal benefit of co-location is significantly less for co-ethnic inventors. These results imply that dispersion of socially proximate individuals is optimal from the perspectives of the city and the economy. In contrast, for socially proximate individuals themselves, spatial concentration is preferred - and the only stable equilibrium.

Ajay Agrawal
Rotman School of Management
University of Toronto
105 St. George Street
Toronto, Ontario M5S 3E6
CANADA
and NBER
ajay.agrawal@rotman.utoronto.ca

John McHale
Queen's School of Business
Goodes Hall
143 Union Street
Kingston, Ontario
Canada K7L 3N6
jmchale@business.queensu.ca

Devesh Kapur
3600 Market Street, Suite 560
Centre for Advanced Study of India
University of Pennsylvania
dkapur@sas.upenn.edu

1 Introduction

What is the optimal spatial distribution of socially connected individuals from a knowledge flow perspective? Is it better to concentrate (all in one city) or disperse across the economy? That may depend on one's perspective. What is best for the city or national economy might not be so for the connected individuals themselves. The answer lies in the way in which knowledge flows are mediated by spatial and social proximity.

Why are we interested in the spatial distribution of socially connected individuals with respect to knowledge flows? 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)). Thus, to fully understand economic growth, we must consider not only the production of knowledge but also access to knowledge. Hence, to the extent the spatial distribution of socially connected individuals impacts knowledge access, this feature of the economy affects growth.

In terms of empirical evidence, one particularly salient imperfection with respect to the "free access" ideal is the apparent geographic localization of knowledge flows (Jaffe et al. (1993); Zucker et al. (1998)). That is, knowledge is more likely to flow between individuals who are located more closely together. Yet geographic distance is just one of many forms of distance that can impede the transfer of knowledge. Conversely, there are ways to be "near" sources of knowledge while being physically separated. For example, social or professional networks can lower the cost of accessing knowledge between members (Sorenson et al. (2006)).

How might we consider the influence of these factors simultaneously? In this paper, we use the device of a knowledge flow production function (KFPF) to model knowledge access. The intuition behind the function is that the likelihood of a knowledge flow between a given pair of inventors depends on the structure of relationships between those inventors - spatial, ethnic, professional, etc. We pay particular attention to how different types of relationships interact. We show that aggregating knowledge flows based on this function leads to interesting results on the factors that support knowledge access for various economic units (i.e., city and country).

We begin by constructing a simple 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 and/or co-ethnic.[1] Based on this KFPF, we model knowledge access to a city that is comprised of both ethnic and non-ethnic inventors. We show that, from the city's perspective, the optimal diversity in the city depends on whether spatial and social proximity are complements (reinforcing the other to further lower the cost of accessing knowledge) or substitutes (causing redundancy in the other).[2] Then, using the same KFPF, we model knowledge access to an economy (comprised of two cities) that has both ethnic and non-ethnic inventors. The optimal distribution of ethnic inventors across cities again depends on whether spatial and social proximity are substitutes or complements.

These models are used to examine the optimal allocation of the ethnic network from the vantage points of a city and of a national economy. However, with free mobility, the actual allocation will be the result of numerous individual inventor decisions about where to live and work. Therefore, we also model location decisions by inventors themselves. We show that although there are multiple equilibria, the equilibrium in which ethnic inventors are dispersed is only stable under certain conditions. We turn to empirical analysis to see whether these conditions hold in our data.

The KFPF is powerful in its generality and simplicity. Not only does it serve as the central building block for our three models, but we are also able to empirically estimate this function using patent citation data. 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 spaces, we find that co-location and co-ethnicity are substitutes rather than complements. In fact, our results suggest that the marginal benefit of co-location is four times larger for individuals who are not co-ethnic. In other words, in terms of facilitating access to knowledge, co-location appears to offer much greater benefits to individuals who are not otherwise socially connected.[3]

[1] Co-ethnic networks, such as Indians in the U.S., are often characterized as rich in social capital (Kalnins and Chung (2006); Saxenian (2002)).

[2] The social capital literature highlights both possibilities. Membership in multiple overlapping networks helps reinforce the deep bonds of trust that facilitate exchange of sensitive information; yet there is also an influential literature that stresses the importance of "weak ties" across networks in accessing non-redundant knowledge (Granovetter (1973); Burt (1992)).

[3] We focus on co-ethnicity as one particular grouping for which membership raises the likelihood of sharing social capital. Of course, there are many such possible groupings. 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

In the context of our models, the estimated coefficients imply that dispersion trumps concentration from the perspectives of both the city and the nation. However, from the perspective of ethnic inventors themselves, although a dispersed equilibrium exists, it is not stable; spatial concentration is the only stable equilibrium. The tension between what is optimal for the location versus what is optimal for individuals creates an interesting setting for studying migration patterns and related policies.

Our study 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 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 U.S. 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 also extend beyond ethnicity to other social groupings.

Our paper proceeds as follows. In section 2, we introduce the KFPF and construct the three models described above. Then, in section 3, we describe the U.S. resident Indian diaspora, the socially connected network that is the basis of our empirical study. Next, in section 4, we describe the methodology and data we employ for estimating the KFPF. In section 5, we present empirical results and discuss the implications of these for the three models. We conclude in section 6.

(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)). 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 La 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 U.S. cities.

2 Theory

2.1 The Knowledge Flow Production Function

The central building block of our three models is the KFPF. This function 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-specialists, 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.[5]

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 equation (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.[6] Suppose, for example, there are three types of possible relationships 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). The assumption that the probability of a knowledge flow between any given pair depends only on the relationships between that pair allows for straightforward aggregation to determine the total knowledge access for any individual inventor and also for any given collection of inventors (e.g., all inventors in a city or nation).

In this paper, we focus on two types of relationships that potentially play an important

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

[6] 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 $r = 1, 2, ..., R$ at a time. This is given by the combinatorial formula: $S = \sum_{r=1}^{R} \frac{R!}{r!(R-r)!} = 2^R - 1$

role in facilitating knowledge flows between inventors: co-location and co-ethnicity. Thus, in our case, $R = 2$, $S = 3$, and the KFPF is given simply by:

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

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. We now use the KFPF to build three simple knowledge flow models.

2.2 Optimal Dispersion: City-Level Model

Our first model examines how ethnic diversity influences total knowledge flows to a city. We assume the number of inventor slots in the city is fixed at N and our interest is the inventor allocation that maximizes the total knowledge flows received by the city. There are two types of inventors, which we label for convenience as minority type and majority type. There are $\overline{M}$ minority type and $\overline{Z}$ majority type inventors in the overall economy. A key assumption (common to all three models) is that the co-ethnicity effect is only present for minority type inventors.

The number of minority types in the city is M. Our first objective is to identify their share in the inventor workforce ($\frac{M^*}{N}$) that maximizes the total knowledge flow to the N inventors in the city. The total knowledge flows received by inventors in a given city (which we call city 1) equals the sum of the knowledge flows received by the inventors in that city:

$$K_1 = \sum_{i=1}^{N} \sum_{j=1}^{\overline{M}+\overline{Z}} K_{ij}, \ i \neq j \tag{3}$$

This total knowledge flow is usefully decomposed as the sum of eight components: 1) the flow within the local minority community; 2) the flow from the local majority community to the local minority community; 3) the flow from the non-local minority community to the local minority community; 4) the flow from the non-local majority community to the local

minority community; 5) the flow within the local majority community; 6) the flow from the local minority community to the local majority community; 7) the flow from the non-local majority community to the local majority community; and 8) the flow from the non-local minority community to the local majority community. This can be expressed mathematically by a straightforward albeit unwieldy expression:

$$\begin{aligned} K_1 = \quad & \big[M(M-1)(\beta_0+\beta_1+\beta_2+\beta_3)\big] + \big[M(N-M)(\beta_0+\beta_1)\big] + \big[M(\overline{M}-M)(\beta_0+\beta_2)\big] + \\ & \big[M(\overline{Z}-N+M)\beta_0\big] + \big[(N-M)(N-M-1)(\beta_0+\beta_1)\big] + \big[(N-M)(M)(\beta_0+\beta_1)\big] + \\ & \big[(N-M)(\overline{Z}-N+M)(\beta_0)\big] + \big[(N-M)(\overline{M}-M)\beta_0\big] \end{aligned} \tag{4}$$

The optimal share of minority type inventors in city 1 can be found by analyzing the first-order condition:

$$\frac{\partial K_1}{\partial M} = (\overline{M}-1)\beta_2 + (2M^*-1)\beta_3 = 0 \Rightarrow \frac{M^*}{N} = \frac{1}{2N} - \frac{\beta_2}{\beta_3}\left(\frac{\overline{M}-1}{2N}\right) \tag{5}$$

Proposition 1. *A diverse inventor mix is optimal for a city if and only if $\beta_3 < 0$ and the city is of sufficiently large size.*

Proof. If $\beta_3 < 0$, an examination of the second-order condition reveals that equation (5) identifies a maximum ($\frac{\partial^2 K_1}{\partial M^2} = 2\beta_3 < 0$). Inspection of equation (5) shows that this maximum requires a diverse inventor mix to maximize knowledge flows to inventors at city 1 for a large enough city (see Figure 1). The optimal share of minority type inventors is increasing in the strength of the (positive) co-ethnicity effect (β_2) and decreasing in the strength of the (negative) interaction effect (β_3).[7] If $\beta_3 > 0$, equation 5 identifies a minimum. In this case, it is optimal to fill all available slots with minority type inventors since, starting from any positive M, increasing M always leads to an increase in the total knowledge flow (see Figure 2).[8] Finally, if $\beta_3 = 0$, inspection of equation 5 reveals that the knowledge flow to city 1 is monotonically increasing in the share of minority inventors, again implying it is optimal to

[7] Intuitively, increasing the minority share involves a tradeoff between the benefit of increasing the number of inventors who are part of ethnicity-based knowledge exchanges and the cost of decreasing the value of the limited number of co-location slots. The first term in the first derivative can be viewed as the marginal benefit of increasing the size of this population by one person. The absolute value of the second term can be viewed as the marginal cost of diminishing the net benefit of co-location. The marginal benefit is constant and the marginal cost is increasing with M. As usual, the optimal level of M is the one that equates marginal benefit and marginal cost.

[8] We assume that the total number of minority-type inventors is greater than the total number of slots in city 1.

fill all available slots with minority type inventors. □

2.3 Optimal Dispersion: Economy-Level Model

Our second model examines how the distribution of ethnic inventors across cities influences total knowledge flows in the economy (which in our simple model consists of cities 1 and 2). We focus on the optimal share of minority type inventors allocated at city 1 ($\frac{M^{**}}{\overline{M}}$). Choosing this number to maximize the sum of knowledge flows at the two locations (K) now gives rise to the first-order condition:[9]

$$\frac{\partial K}{\partial M} = \beta_3\big(4M^{**} - 2\overline{M}\big) = 0 \Rightarrow \frac{M^{**}}{\overline{M}} = \frac{1}{2} \tag{6}$$

Proposition 2. *An equally distributed minority inventor mix is the unique optimum for the economy if and only if* $\beta_3 < 0$.

Proof. If $\beta_3 < 0$, an examination of the second-order condition reveals that equation (6) identifies a maximum ($\frac{\partial^2 K_1}{\partial M^2} = 4\beta_3 < 0$). Inspection of equation (6) shows that this maximum requires an equal distribution of inventors across the two locations (see Figure 3).[10]

If $\beta_3 > 0$, equation (6) identifies a minimum. In this case, it is optimal to concentrate all available minority type inventors at one location (see Figure 4).[11] Finally, if $\beta_3 = 0$, a unique optimum does not exist and the level of knowledge flows is independent of the distribution of minority-type inventors across cities. □

2.4 Optimal Dispersion: Inventor-Level Model

The two models above examine optimal dispersion of ethnic inventors from the vantage point of a city and of the national economy. But 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. Thus, we now turn to examine actual dispersion assuming that each inventor makes their location decision to maximize their own access to knowledge.

[9] To save space, we have not written down the knowledge access equation; it is an obvious extension of the equation for city 1 also including city 2.

[10] We assume that each city is large enough to take half the minority-type inventors.

[11] Note that it does not matter for the total knowledge flow which of the locations is chosen for this concentration. However, if only one city is large enough to accommodate the entire group, then that location should be chosen.

For simplicity, we continue to assume there are just two cities and make the additional assumption that the non-ethnic populations at locations 1 and 2 are Z_1 and Z_2, respectively (where $Z_1 \geq Z_2$). An individual minority-type inventor will be indifferent between the two locations when the knowledge accessed at each location is the same. Letting M^e denote the equilibrium number of minority-type inventors at city 1, we can use the KFPF to write the equilibrium condition as:

$$\begin{aligned} &M^e\big[\beta_0+\beta_1+\beta_2+\beta_3\big]+Z_1\big[\beta_0+\beta_1\big]+Z_2\beta_0+(\overline{M}-M^e)\big[\beta_0+\beta_2\big]= \\ &(\overline{M}-M^e)\big[\beta_0+\beta_1+\beta_2+\beta_3\big]+Z_2\big[\beta_0+\beta_1\big]+Z_1\beta_0+M^e\big[\beta_0+\beta_1\big] \end{aligned} \tag{7}$$

After some cancelation and re-arrangement, it will be helpful for our later discussion of stability to rewrite this condition as:

$$\frac{M^e}{\overline{M}}\big[\beta_1+\beta_3\big]+\frac{Z_1\beta_1}{\overline{M}}=\left(1-\frac{M^e}{\overline{M}}\right)\big[\beta_1+\beta_3\big]+\frac{Z_2\beta_1}{\overline{M}} \tag{8}$$

The equilibrium share at city 1 is then easily calculated as:

$$\frac{M^e}{\overline{M}}=\frac{1}{2}-\frac{(Z_1-Z_2)\beta_1}{2(\beta_1+\beta_3)\overline{M}} \tag{9}$$

The stability of the dispersed equilibrium described by equation 9 depends on the sign of $\beta_1+\beta_3$. This is most easily seen by graphing the two sides of equation 8. In Figure 5, we show the graph for the case where $\beta_1+\beta_3<0$. The intersection of the two curves identifies a stable equilibrium. To see that the equilibrium is stable, imagine that starting from the equilibrium some shock reduces the share of minority types at city 1. Minority-type inventors are then able to access more knowledge at city 1 than at city 2, which will induce knowledge-maximizing inventors to move, with the movement continuing until the initial equilibrium is restored.

In Figure 6, we show the graph for the case where $\beta_1+\beta_3>0$. Given the upward slope of the curves, the equilibrium identified by their intersection is now unstable. Starting again from this equilibrium, a shock that reduces the number of minority type inventors at city 1 will lead to a situation where more knowledge can be acquired at city 2 than at city 1, leading yet more inventors to leave city 1. This relocation process will continue until *no* inventors are at city 1.

On the other hand, if the initial shock leads to a larger share of inventors at city 1 than identified by equation 9, then the dynamics will lead *all* minority inventors to move to city 1. Thus, the equilibrium analysis shows that the individual members of the diaspora will concentrate when $\beta_1 + \beta_3 > 0$.

However, it is not necessarily the case that individual members will concentrate at the "right" location. Individuals will choose to move to the city that maximizes their access to knowledge, which is is conditioned on where others have located before them. If others have located in a smaller city, the individual may be faced with a collective action problem. From Figure 6, it is apparent that city 1 is the preferred location for concentration if $Z_1 > Z_2$ (which is assumed in the figure). But if, for historic reasons, the diaspora happens to concentrate in the smaller city 2, then that is where it will stay.

Finally, in the case where $\beta_1 + \beta_3 = 0$, the curves in Figures 5 and 6 become horizontal at their intercepts. In this case, there is exactly zero co-location effect for the minority-type inventors, which holds no matter what their initial spatial distribution. Minority types will all concentrate in the city with the larger number of non-minority types (city 1 in the depicted example). In contrast to the case where $\beta_1 + \beta_3 > 0$, there is no danger here of the dynamics of location choice leading to concentration at the "wrong" location.

Summing up this section, we have found that: 1) a negative interaction effect is sufficient for diversity to trump concentration for a city of a given size; 2) a negative interaction effect is sufficient for a geographically distributed diaspora to trump concentration for a national economy; and 3) the negative interaction effect needs to be greater than the (assumed) positive co-location effect for the dispersion equilibrium to be stable, assuming knowledge flow-maximizing, mobile inventors.[12]

These models provide insight into the perhaps surprisingly broad implications of the actual values of the KFPF coefficients. For example, a negative β_3 implies that dispersion of ethnic inventors is better than concentration from the perspectives of both the city and the overall economy. But what are the actual values of the KFPF coefficients? To answer this, we turn

[12]It is interesting to note that the necessary conditions for equilibrium stability shown in the third model are sensitive to the crowding out assumption. That is, in this model the number of majority-type inventors in each city is fixed so the total number of inventors in each city varies depending on the location choices of minority inventors. In other words, minority type inventors do not "crowd out" majority types but rather add to the existing population. Alternatively, if we model cities with perfect crowding out (a fixed number of slots in each city), the condition for a stable dispersed equilibrium is $\beta_3 < 0$. Intuitively, with perfect crowding out, a key force leading to concentration - the act of moving to a larger city makes that city even bigger and thus more attractive from the point of view of knowledge access - is no longer present.

to data and estimate the KFPF coefficients. We choose to focus on the U.S. resident Indian diaspora to construct our co-ethnic sample for a number of reasons. We explain these reasons next.

3 The U.S. Resident Indian Diaspora

The U.S. resident Indian diaspora is particularly suitable for the purposes of our study because, on average: 1) they are reasonably identifiable (by last name), 2) they are highly active in technological innovation (we use patent citations to measure knowledge flows), 3) they identify strongly with their ethnicity (we assume co-ethnicity to be a mechanisms for knowledge transfer), and 4) they are active across a broad range of geographies (many co-located and non-co-located observations in the sample).

The U.S. resident Indian diaspora is 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.[13,14] In addition, not only is the share of the Indian-American population working in technology significant, their role in the U.S. innovation system has increased over time (Table 1).[15] 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.

Members of the U.S. 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 U.S., those born in the U.S. were fewer than those born in India (0.7 million versus 1 million).[16] Furthermore, more than one third of the Indian-born came after 1996 and more than half after 1990.[17] Survey evidence underlines the strong ethnic identification: 53% visit

[13]Source: U.S. Census Bureau and authors' calculations.

[14]In related work, Stephan and Levin (2001) and Levin and Stephan (1999) 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 U.S.

[15]Our method for identifying Indian inventors is described in detail in the data section. Here we are measuring all inventors with Indian last names.

[16]Source: U.S. Census Bureau, Current Population Survey, March Supplement, various years.

[17]The Indian-born population in the U.S. 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 U.S. labor market in the 1990s for highly skilled individuals with job offers. Highly skilled Indians, especially those working in the computer

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.[18]

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 U.S.- 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). However, 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) technology subcategory where at least one of the inventors is Indian.[19] 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

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.

[18] Kapur (2004).

[19] 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.

index for technological concentration, for example, is 677 for patents with "One or More Indian Inventor" compared with 422 for "All Patents."

4 Empirical Methodology and Data

4.1 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 (Table 3 shows 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, the effects of interest are estimated from the following simple regression:

$$\begin{aligned} P(Citation_{ij}) = \quad & \beta_0 + \beta_1 Co-Location_{ij} + \beta_2 Co-Ethnicity_{ij} + \\ & \beta_3 (Co-Location_{ij} \times Co-Ethnicity_{ij}) + \varepsilon_{ij}, \\ & i \neq j. \end{aligned}$$

Citation is a dummy variable that takes a value of 1 when the observation relates to an actual citation and 0 if the observation relates to a control. The citing inventor is indexed by i and the cited (or control) inventor is indexed by j. By construction, there are an equal number of actual and control observations in our sample. *Co-location* is a dummy variable that takes a value of 1 when the original and cited (or control) inventors are located in the same MSA and 0 otherwise. *Co-Ethnicity* is a dummy variable that takes a value of 1 when the cited (or control) inventor has an Indian surname (the original inventor always has an Indian surname, by construction). All self cites ($i = j$) are excluded.[20]

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

To see how this regression allows us to identify the causal effects of interest, note that if the control matching procedure is effective and there is no causal link 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 citing and cited patent. For example, suppose we are dealing with a citation where the citing and cited inventors are co-located but not co-ethnic. Suppose further that the estimated values of β_0 and β_1 are 0.4 and 0.2, 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 there is no causal link from co-location to knowledge flows. A rejection of this null could follow from ineffective matching and/or the absence 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.

stronger.

4.2 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.[21] 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 6,885 last names that were coded as "extremely likely to be Indian."[22]

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 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,

[21] This database was prepared by InfoUSA.

[22] 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 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))). To the extent that there is noise in our name data, it will bias our result downwards.

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. Or 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.[23] 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 matches 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.[24]

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 there are no patents that match the cited patent in at least the application year and the three-digit primary classification without being cited by the original patent, the observation (original patent) is removed from the data set.

Co-ethnic and co-localization effects are identified 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.[25] The geographic or ethnic clustering of innovative activity in certain

[23]The Jaffe et al. and Thompson and Fox Kean approach involves the analysis of forward citations. To take advantage of the substantial growth in the Indian-born population in the U.S. 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. Either approach (backwards or forwards citations) can be used to test for a disproportionate incidence of co-located or co-ethnic knowledge flows.

[24]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 2% only 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.

[25]We also check whether "Indian patents" are cited more frequently than their non-Indian counterparts.

technology areas itself may be due to the lowered cost of establishing social relationships but also may be due to other local factors such as thicker factor markets. Thus, focusing on knowledge flows that are more concentrated than the innovative activity in that particular field may be considered a conservative approach.

Co-ethnicity Metrics

We examine the last name(s) of the inventors(s) on the cited patent associated with each observation. If a name matches, the inventor is designated as "of Indian origin" and we say the original and cited patents are "co-ethnic" (the former is of Indian origin by construction). We do the same for control patents.

Co-location Metrics

We also examine the home address of the inventor for each observation.[26] Inventors are assigned to an MSA based on their city and state information. There are 268 U.S. MSAs and consolidated metropolitan statistical areas (CMSAs) and 25 Canadian census metropolitan areas (CMAs), hereafter collectively referred to as "MSAs."[27] We also have created 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.[28] We say the original and cited patents are co-located if they both are assigned to the same MSA. Similarly, we say the original and control patent are co-located if they meet the same criterion.

Sample Construction

We generate our sample by identifying all patents issued by the USPTO during the period 2001-2003. There are 555,741 such patents. From this set, we identify those patents that have at least one inventor of Indian origin. There are 19,612 such patents. 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

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.

[26] City and country information is used for assigning Canadian inventors to a CMA.

[27] 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.

[28] We include Canada since this nation's Indian-born population follows similar patterns to that of the U.S. 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 U.S. MSAs are included.

remove those observations for which the inventor of the original patent does not have an Indian name (although they co-invented with somebody who does have an Indian name) and those observations for which we are unable to identify a control for the cited patent. Consequently, our sample includes 170,950 citing-cited pairs and an equal number of citing-control pairs for a total of 341,900 observations.

5 Results

The first and second columns of Table 4 record the estimated coefficients when the equation is estimated by OLS (with and without citing-patent fixed effects). The third and fourth columns record the results when a Logit specification is used for the same two samples. Since both specifications imply almost identical conditional probabilities for the occurrence of an actual citation, we limit our discussion to the OLS results. For both specifications without fixed effects, the reported standard errors are robust to the non-independence of observations drawn from clusters of observations based on the same citing patent.[29]

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 Column (1), co-location increases the probability of an actual citation by just over 12 percentage points, and co-ethnicity increases the probability of a citation by almost 7.5 percentage points. 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, $\widehat{\beta}_3$. As discussed in Section 2, this result can be interpreted as evidence that co-location and co-ethnicity are substitutes in facilitating knowledge flows. However, we can reject the null hypothesis that the co-location effect is offset by a negative interaction effect (i.e., $\widehat{\beta}_1 + \widehat{\beta}_3 = 0$) in favor of the alternative hypothesis, $\widehat{\beta}_1 + \widehat{\beta}_3 > 0$ (p-value = 0.002).

This finding can be considered in terms of the difference in marginal impact of co-location between inventors who are co-ethnic and those who are not. Co-location increases the prob-

[29] 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.

ability of a knowledge flow by 25% for non-co-ethnic inventors but only by 6% for co-ethnic inventors. In other words, the marginal effect of co-location is much smaller for inventors who are already connected through some other mechanism.[30]

Relating these results to the models in Section 2, the results are consistent with diversity being good for a location ($\widehat{\beta}_3 < 0$) - inventive activity at a location benefits from having inventors with a shared ethnicity, but not too many! The results are also consistent with the impact on national knowledge flows being maximized when inventors with a shared ethnicity are widely distributed rather than concentrated ($\widehat{\beta}_3 < 0$). However, from the individual inventor's perspective, the dispersed equilibrium is not stable ($\widehat{\beta}_1 + \widehat{\beta}_3 > 0$).

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, there may still be concerns 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 capture, 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 - 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,

[30] This is consistent with a prior finding that co-location results in a 74% increase in the probability of a cross-field citation (from one technology field to another) but only a 34% increase in the probability of a within-field citation (Agrawal et al. (2006)). The lower marginal impact of co-location for within-field cites is attributed to a greater likelihood of alternative channels for establishing social relationships through communities of practice.

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 able to think of any reasonable conditions under which this would be true.

6 Concluding Comments

We have examined the interaction between networks based on co-location and co-ethnicity for U.S. resident inventors of Indian origin. Our results show that co-location and co-ethnicity play significant roles in facilitating knowledge flows, but they appear to substitute for rather than complement one another. Our modeling shows that such substitutability is a sufficient condition for diversity to be optimal for a location and for a geographically distributed ethnic network to be optimal for the economy. However, the negative interaction effect must actually outweigh the co-location effect for a dispersed equilibrium to be stable; our results indicate this is not the case.

Overall, our paper points to the economic importance of ethnicity, geography, and knowledge. However, key questions remain unanswered. Perhaps most urgent is the underlying mechanism that gives ethnicity its economic importance. Do last names serve a cuing "reputational" function for co-ethnics? Do co-ethnics benefit from lower cost access to tacit knowledge 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 any general conclusions.

Our findings, along with others (Nanda and Khanna (2006), Kapur and McHale (2005)), 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.

References

Agrawal, Ajay, Iain Cockburn, and John McHale, "Gone But Not Forgotten: Labor Flows, Knowledge Spillovers, and Enduring Social Capital," *Journal of Economic Geography*, 2006, *6* (5), 571–591.

Alesina, Alberto and Eliana La Ferrara, "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 2005, *43* (3), 762–800.

Borjas, George J., "Ethnicity, Neighborhoods, and Human-Capital Externalities," *The American Economic Review*, 1995, *85* (3), 365–390.

Burt, Ronald S., *Structural Holes: The Social Structure of Competition*, Cambridge, MA: Harvard University Press, 1992.

Coleman, James, "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 1988, *94*, S95–S120.

Easterly, William and Ross Levine, "Africa's Growth Tragedy: Policies and Ethnic Division," *Quarterly Journal of Economics*, 1997, *112* (4), 1203–1250.

Glaeser, Edward L., David Laibson, and Bruce I. Sacerdote, "The Economic Approach to Social Capital," *Economic Journal*, 2002, *CXII*, F437–58.

Granovetter, Mark S., "The Strength of Weak Ties," *American Journal of Sociology*, 1973, *LXXIII*, 1360–1380.

Griliches, Zvi, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, October 1957, *25* (4), 501–522.

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

Jaffe, Adam B., Manuel Trajtenberg, and Michael Fogarty, *Patents, Citations, and Innovations: A Window on the Knowledge Economy*, Cambridge, MA: The MIT Press, 2002.

_ , _ , and Rebecca Henderson, "Geographic Localization of Knowledge Flows as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 1993, *CVIII*, 577–598.

Kalnins, Arturs and Wilbur Chung, "Social Capital, Geography, and Survival: Gujarati Immigrant Entrepreneurs in the U.S. Lodging Industry," *Management Science*, 2006, *52* (2), 233–247.

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

_ and John McHale, *Sojourns and Software: Internationally Mobile Human Capital and High Tech Industry Development in India, Ireland, and Israel in "From Underdogs to Tigers"*, Oxford, UK: Oxford University Press, 2005. forthcoming.

Kerr, William, "Ethnic Scientific Communities and International Technology Diffusion," 2005. Working paper, Harvard University.

Knack, Stephen and Philip Keefer, "Does Social Capital have an Economic Payoff? A Cross-Country Investigation," *Quarterly Journal of Economics*, 1997, *112* (4), 1251–1288.

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

Levin, Sharon and Paula Stephan, "Are the Foreign Born a Source of Strength for US Science," *Science*, 1999, *285*, 1213–1214.

Nanda, Ramana and Tarun Khanna, "Diasporas and Domestic Entrepreneurs: Evidence from the Indian Software Industry," 2006. Working paper, MIT.

Ottaviano, Gianmarco and Giovanni Peri, "The Economic Value of Cultural Diversity," 2004. NBER Working Paper 10904.

Putnam, Robert D., *Bowling Alone*, New York, NY: Simon & Schuster, 2002.

Rauch, James E., "Business and Social Networks in International Trade," *Journal of Economic Literature*, 2001, *XXXIX*, 1177–1203.

Romer, Paul, "Endogenous Technological Change," *Journal of Political Economy*, October 1990, *98* (5).

Saxenian, Anna Lee, *Local and Global Networks of Immigrant Professionals in Silicon Valley*, San Francisco, CA: The Public Policy Institute of California, 2002.

Sorenson, Olav, Jan W. Rivkin, and Lee Fleming, "Complexity, Networks, and Knowledge Flow," *Research Policy*, 2006, *35*, 994–1017.

Stephan, Paula and Sharon Levin, "Exceptional Contributions to US Science by the Foreign-Born and Foreign-Educated," *Population Research and Policy Review*, 2001, *20*, 59–79.

Thompson, Peter and Melanie Fox-Kean, "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 2005, *95* (1), 450–460.

Zucker, Lynne, Michael Darby, and Marilynn Brewer, "Intellectual Capital and the Birth of U. S. Biotechnology Enterprises," *American Economic Review*, March 1998, *88* (1), 290–306.

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
Percent Indian	0.9%	1.1%	1.3%	1.8%	2.9%	3.2%

Table 2: Share of Patents Where at Least One Inventor is of Indian Origin by Location, U.S. 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 1,000 issued patents with application year 1995 are shown.

Table 3: Share of Patents Issued Where One or More Inventors is of Indian Origin (Application year 1995)

NBER Subcategory	Description	One or More Indian Inventor	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%

Table 4: Co-Location, Co-Ethnicity, and the Probability of a Citation

	OLS		Logit	
Dependent Variable = Dummy for Actual Citation				
Co-Location	0.1208*	0.1421*	0.4897*	0.5583*
	-0.0036	-0.0033	-0.0193	-0.13
Co-Ethnicity	0.0748*	0.0772*	0.2893*	0.2999*
	-0.0045	-0.0046	-0.0235	-0.0181
Co-Location Co-Ethnicity	−0.0872*	−0.0904*	−0.3491*	−0.3538*
	-0.0089	-0.0095	-0.046	-0.0373
Constant	0.4861*	0.4824*	−0.0605*	
	-0.0004	-0.0009	-0.0023	
Fixed Effects	No	Yes	No	Yes
Observations	341900	341900	341900	341900
Number of Citing Patents	11248	11248	11248	11248
R^2 (pseudo R^2 for logit)	0.0058	0.0058	0.0042	

-Inventor and assignee self cites are excluded.
-See text for description of how control citations are chosen.
-* Indicates significance at 1% level.
-For the regressions without fixed effects, standard errors are robust to citing-patent cluster effects.
-Fixed effects regressions allow for citing-patent fixed effects.

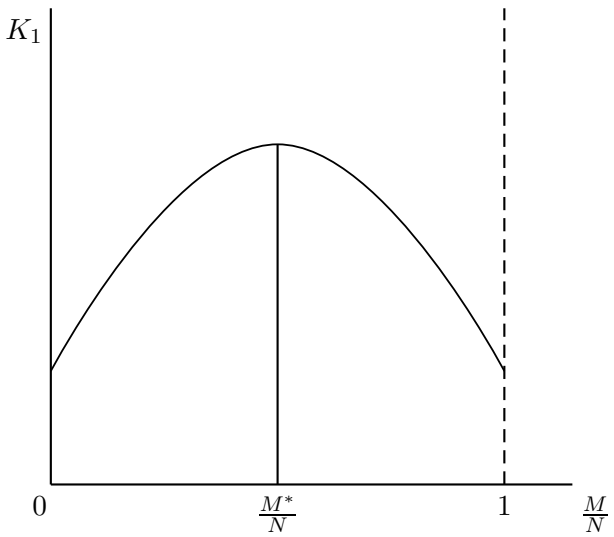

Figure 1: Optimal Diversity ($\beta_3 < 0$)

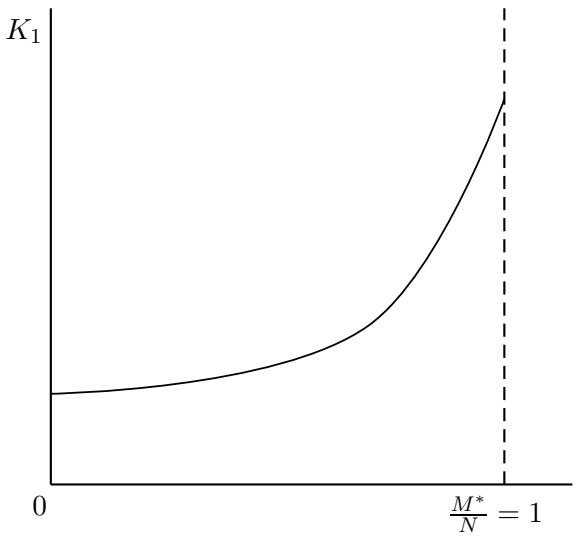

Figure 2: Optimal Diversity ($\beta_3 > 0$)

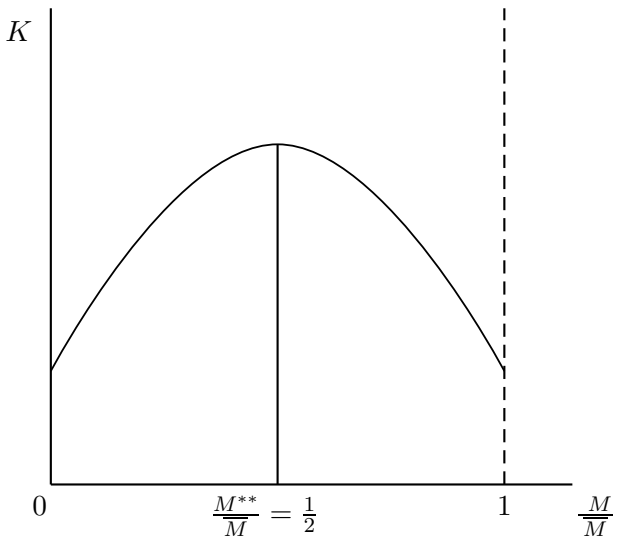

Figure 3: Optimal Distribution ($\beta_3 < 0$)

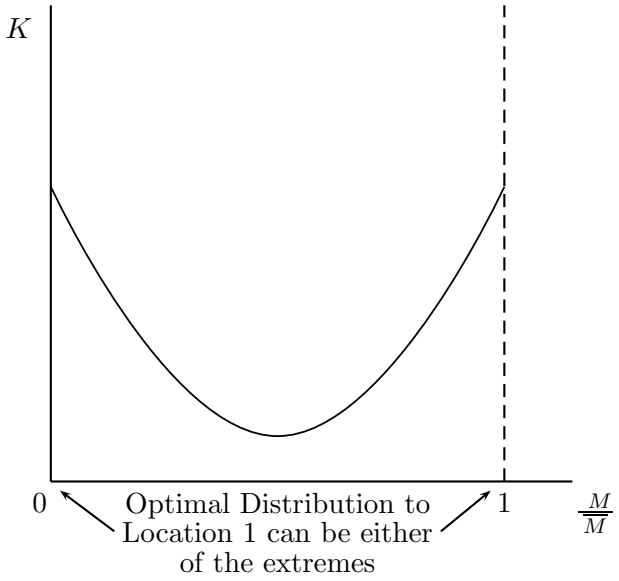

Figure 4: Optimal Distribution ($\beta_3 > 0$)

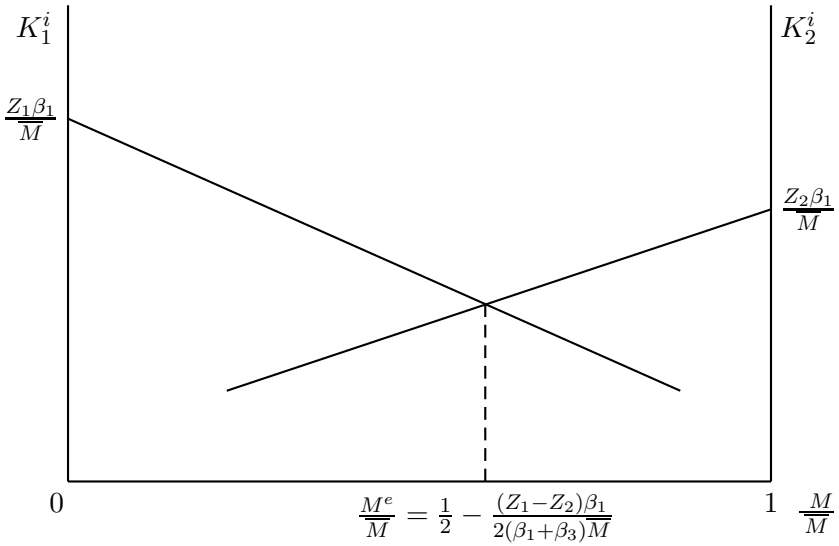

Figure 5: $(\beta_1 + \beta_3 < 0\ )$

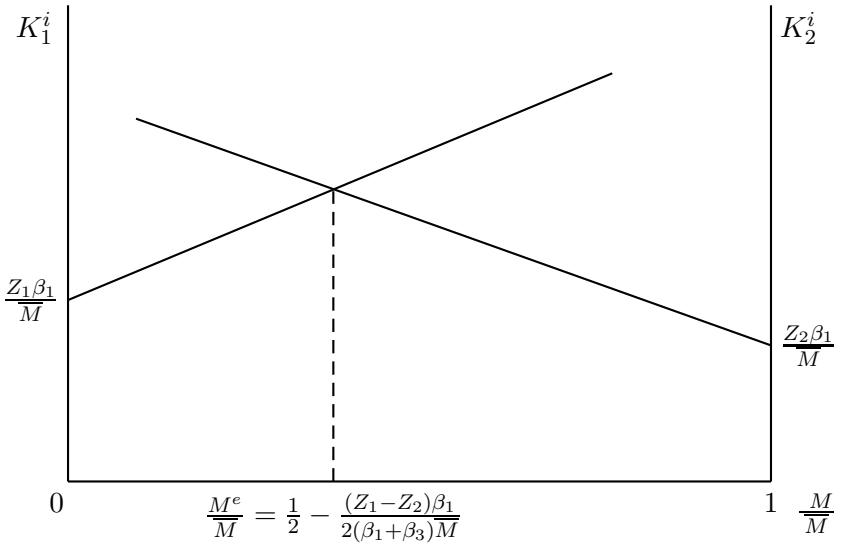

Figure 6: $(\beta_1 + \beta_3 > 0\ )$