Journal of Urban Economics 69 (2011) 43–55

Contents lists available at ScienceDirect

Journal of Urban Economics

www.elsevier.com/locate/jue

Brain drain or brain bank? The impact of skilled emigration on poor-country innovation

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

[a] Rotman School of Management, University of Toronto and NBER, Toronto, ON, Canada M5S 3E6
[b] Center for the Advanced Study of India, University of Pennsylvania, Philadelphia, PA 19104, USA
[c] National University of Ireland, Galway, Ireland
[d] College of Management, Georgia Institute of Technology, Atlanta, GA 30308, USA

ARTICLE INFO

Article history:
Received 12 December 2008
Revised 1 June 2010
Available online 3 July 2010

JEL classification:
O33
R12
Z13

Keywords:
Knowledge flows
Emigration
Ethnicity
Diaspora

ABSTRACT

The development prospects of a poor country or region depend in part on its capacity for innovation. In turn, the productivity of its innovators, whom are often concentrated around urban centers, depends on their access to technological knowledge. The emigration of highly skilled individuals weakens local knowledge networks (brain drain) but may also help remaining innovators access valuable knowledge accumulated abroad (brain bank). We develop a model in which the size of the optimal innovator Diaspora depends on the competing strengths of co-location and Diaspora effects for accessing knowledge. Then, using patent citation data associated with inventions from India, we estimate the key co-location and Diaspora parameters. The net effect of innovator emigration is to harm domestic knowledge access, on average. However, knowledge access conferred by the Diaspora is particularly valuable in the production of India's most important inventions as measured by citations received. Thus, our findings imply that the optimal emigration level may depend, at least partly, on the relative value resulting from the most cited compared to average inventions.

1. Introduction

The development impact of skilled migration from poor countries has long been a contentious issue. Scholars are even far from a consensus on the narrower question: What is the impact on innovation when a poor country loses a large fraction of its science and engineering workforce through emigration?

One school of thought argues that such talent is often wasted at home. Migration to more supportive environments raises global innovation, and some gains flow back to the poor country through the imports of products with improved technology or lower cost (Kuhn and McAusland, 2006). Furthermore, gains may flow back to the developing country via returnees with enhanced skills, personal connections, and ideas for innovation (Saxenian, 2005).

Another school of thought focuses on the importance of domestic technology innovators. Despite their typically considerable distance from the technology frontier, domestic innovators could be important for various reasons: (1) international technology diffusion may be slow due to the localization of knowledge spillovers;[1] (2) rich-country innovation may not properly address the needs of poorer countries;[2] and (3) domestic knowledge production may be necessary to create the capacity to absorb foreign technology.[3]

However, the most important form of innovation for a poor country is likely the adoption of technologies developed elsewhere (World Bank, 2008). In other words, the greatest opportunities for growth in a poor country lie in moving towards the international frontier rather than in pushing that frontier forward. Highly skilled domestic innovators are likely to be central to this catch-up process.

The availability of new datasets showing high and generally increasing poor- to rich-country emigration rates for tertiary-educated workers has heightened concern about the "brain drain" (Docquier and Marfouk, 2005; Dumont and Lemaitre, 2005). These rates measure the absence of tertiary-educated nationals from the economy. In many cases, inventors have acquired their education abroad, and so the rates are not actually measures of the outflow of individuals trained domestically (the usual connotation of the term "brain drain"). These rates are extremely high for many small,

* Corresponding author. Address: University of Toronto, Rotman School of Management, 105 St. George Street, Toronto, ON, Canada M5S 3E6. Fax: +1 416 978 5433.
E-mail address: ajay.agrawal@rotman.utoronto.ca (A. Agrawal).

[1] For example, Keller (2002) presents evidence on international technology diffusion. Also, Jaffe et al. (1993, hereafter "JTH") document the localization of knowledge spillovers.

[2] For example, Basu and Weil (1998) present a model in which the appropriate technology is specific to a country's available inputs.

[3] For example, Cohen and Levinthal (1989) argue that R&D has the indirect benefit of increasing a firm's capacity to absorb technology being developed elsewhere. In addition, Caselli and Coleman (2001) show that importing technology embodied in computers is positively related to domestic human capital stocks.

0094-1190/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jue.2010.06.003

poor countries. For example, Docquier and Marfouk (2005) estimate that 41% of those with a tertiary education and born in a Caribbean country now live in an OECD country.

Although tertiary emigration rates tend to be considerably lower for larger developing countries, emigration rates for the most educated and talented are much higher (Kapur and McHale, 2005). To take the example of India, researchers estimate the overall tertiary emigration rate to be about 4%, while the emigration rates from the elite Indian Institutes of Technology (IITs) are substantially higher. For example, an analysis of the brain drain from the graduates in the 1970s of one of the country's top engineering universities, IIT-Mumbai, reveals that 31% of its graduates settled abroad, while the estimated migration rate of engineers more generally was only 7.3% (Sukhatme and Mahadevan, 1987). Recent alumni data in the case of IIT-Kharagpur, another top university, find 4007 registered alumni in India, 3480 in the US, and another 739 spread over 59 countries.[4]

At the same time, substantial flows of financial remittances highlight the many *benefits* to the country of origin from international migration, extending not just to money but also to the flows of ideas and technologies from its Diaspora. The latter raises the possibility that the migration of skilled human capital from poor countries may not just be a negative "brain drain;" it could also have more a positive effect as a "brain bank," accumulating knowledge abroad and facilitating its transfer back to domestic inventors (Kerr, 2008).

In this paper, we develop and estimate a model in which the *access* of domestic innovators to knowledge drives innovation. This approach combines Paul Romer's classic model of innovation and growth, where the existence of new ideas that might be built upon is the basis of innovation, and "anyone engaged in research has free access to the entire stock of knowledge" (Romer, 1990, p. S83), with the notion that knowledge spillovers are spatially localized (Jaffe et al., 1993; Rosenthal and Strange, 2001) and thus less costly to access for individuals located closer to the inventor. For a poor country, the degree of access to the existing stock of knowledge is likely of particular importance, warranting the emphasis on proximity to other inventors.[5] To this end, we will explore both spatial and ethnic proximity.

The main building block of our model is the Knowledge Flow Production Function (KFPF). For any domestic innovator, the KFPF gives the probability of receiving knowledge from any other innovator based on structural aspects of their relationship. We focus in particular on whether innovators are co-located in the domestic economy, share a Diaspora connection (co-ethnic), or are unconnected by location *or* nationality. We assume the outputs of domestic innovators depend on their overall access to knowledge from domestic, Diaspora, and foreign sources. The total innovation output of the national economy is then simply the sum of the innovation outputs of domestic inventors.

Hence the central tradeoff in the model: The emigration of a domestic innovator leads to a direct reduction in domestic innovator stock and weakens the network of co-located innovators but can also lead to new access to foreign-produced knowledge through the Diaspora. The latter effect will be stronger where enduring connections to the Diaspora exist and where emigrant innovators increase their knowledge stock by moving to environments with better resources, colleagues, and incentives to innovate.

These conflicting effects lead to the idea of the *optimal Diaspora*—the emigrant stock that maximizes national knowledge access. We show that the optimal Diaspora depends on the relative size of the co-location and Diaspora effects. We also examine extensions to the model that allow for circulation between the home economy and the Diaspora, non-random selection of emigrants and returnees, and heterogeneous KFPFs based on the importance of the innovation.

The empirical challenge is to identify the co-location and Diaspora effects in the KFPF. To accomplish this, we construct a novel sample from patent data linked with Indian last name data and then build on a widely used method that employs patent citations as a proxy for knowledge flows between inventors and "matched citations" to control for the underlying distribution of inventive activity across geographic and ethnic space. This allows us to isolate the causal impacts of location and Diaspora connections on the probability of a knowledge flow.

Our empirical focus is on the knowledge access of frontier innovators in a poor country. This focus allows us to take advantage of the rare instance of a "paper trail" for national and international knowledge flows afforded by the recording of citations on a patent (Jaffe et al., 1993). We stress again that frontier innovation will typically be of second-order importance for growth in poor countries. However, to the extent that networks for knowledge access operate similarly for frontier- and implementation-based innovation, the findings on the drivers of knowledge flows at the frontier should provide a valuable clue to the relative importance of local versus Diaspora knowledge networks and thus the likely impact of skilled emigration on poor-country knowledge access and innovation.

Our paper complements the large urban economics literature on the geography of agglomeration economies (see Rosenthal and Strange (2004), for a survey). An important recent finding in this literature is that such economies decline with distance (Rosenthal and Strange, 2003, 2008; Anderrson et al., 2004, 2009; Arzaghi and Henderson, 2008). Scholars have found the attenuation effect to be most pronounced for agglomeration economies due to localized knowledge spillovers (Rosenthal and Strange, 2001). Our focus is somewhat different in that we concentrate on spillovers from individuals who have become geographically separated but potentially retain a connection through ethnicity-based networks.[6] Agrawal et al. (2006) provide evidence that knowledge flows from mobile inventors go disproportionately to their prior locations, which suggests enduring connections. Agrawal et al. (2008) show that both co-location and co-ethnicity support knowledge flows, with a negative interaction between the two. It is possible, then, that the attenuating effects of distance will be weaker for our sample of Indian-origin inventors.

We organize the rest of the paper as follows. In the next section, we model an optimal innovator Diaspora. In Section 3, we describe our empirical strategy for identifying the causal effects of co-location and Diaspora membership on knowledge flows. In Section 4, we describe our patent citation and Indian-name data, presenting our results in Section 5. In Section 6, we discuss the implications of our findings.

2. The optimal Diaspora

2.1. Permanent migration

We first develop a simple model of an optimal innovator Diaspora, abstracting initially from the possibility of return, innovator

[4] See http://www.iitfoundation.org/directory/stats/. Accessed September 3, 2004.

[5] Klenow and Rodriguez-Clare (2004) argue that international technology spillovers explain some of the basic facts about cross-country income levels and growth rates. Using a calibrated endogenous growth model, they show that relatively small barriers to international technology diffusion—knowledge access, in the language of our model—can lead to large cross-country differences in income levels.

[6] Another source of enduring connections is the movement of workers between firms. Workers moving from firm to firm within a location – "job hopping" – can support knowledge exchange and productivity (see Fallick et al. (2006) and Freedman (2008) for recent applications).

heterogeneity, and differences in the KFPF related to the value of innovations. Our focus is on knowledge production in a relatively poor country, which we call India without loss of generality. The essential idea is that the productivity of India-residing innovators depends on their access to knowledge. This access in turn depends on their relationships to other innovators and also on the productivity of those innovators. We allow connectivity to be affected by co-location and co-nationality and also for the possibility that innovators are more productive abroad because of better incentive structures and resources. The emigration of an innovator results in a direct loss to the stock of Indian innovators, thinning domestic knowledge networks, but could actually increase total knowledge access if the diasporic linkages and productivity gains are large enough. The model's goal is to identify the size of the Diaspora that maximizes the access to knowledge of India-residing innovators.[7]

The KFPF captures the probability of a knowledge flow between any pair of innovators (at least one of whom is a resident of India) based on certain structural relationships between those innovators. We express the probability of a knowledge flow to a particular Indian innovator, i, from another innovator, j, as:

$$K_{ij} = f + \alpha_{ij}\gamma f + \beta_{ij}\delta f, \tag{1}$$

where f is the (base-case) probability of a knowledge flow if the other innovator is neither a resident of India nor a member of the Indian Diaspora, α_{ij} is a dummy variable that takes the value of 1 if innovator j is also a resident of India, γ is the proportionate knowledge-flow premium from being co-located, β_{ij} is a dummy variable that takes the value of 1 if j is a member of the Indian Diaspora, and δ is the proportionate premium for being in the Diaspora. Note that the value of γ reflects the combined effects of co-location and the (possibly negative) relative productivity effect of doing science in India, whereas the value of δ reflects the effect of the Diaspora connection and any productivity gap that might exist between members of the Diaspora and foreigners. Denoting the total number of Indian innovators (both India-based and emigrant) as N, the total size of the Indian scientific Diaspora as D, and the total number of foreign innovators as Z, we express the total (expected) knowledge flow to i with this knowledge access equation:

$$K_i = Zf + (N - D - 1)(1 + \gamma)f + D(1 + \delta)f. \tag{2}$$

We then find the aggregate knowledge access of India-residing innovators by multiplying both sides of (2) by the total number of such innovators:

$$\begin{aligned} K &= (N - D)K_i \\ &= (N - D)Zf + (N - D)(N - D - 1)(1 + \gamma)f + (N - D)D(1 + \delta)f. \end{aligned} \tag{3}$$

We assume innovation depends on both the access to knowledge and the absorptive capacity to turn that knowledge into valuable economic output. In this paper, we focus only on knowledge access and assume it is positively associated with output: $I_i = I(K_i)$ $\frac{\partial I_i}{\partial K_i} > 0$. Of course, the knowledge access to innovation will be country specific and depend, *inter alia*, on the available capital stock, the presence of complementary human capital, and security of property rights.

[7] The model allows for a trade-off between the costs of weakened local knowledge networks and the benefits of access to more distant knowledge. Recent work in urban economics has highlighted other potential trade-offs associated with labor pooling. Combes and Duranton (2006) develop a model in which labor pooling has two opposing effects: It allows greater access to knowledge produced by other firms, but the potential for one's own workers to be poached forces firms to pay higher wages to retain their workforce. Gerlach et al. (2009) develop a model with the same deglomerative force but in which the agglomerative force comes from asymmetric R&D investments that produce a diversified portfolio of technologies at the industry level.

We find the Diaspora size, D^*, that maximizes national knowledge access (and thus innovation) from the first-order condition:

$$\frac{\partial K}{\partial D} = 2D^*(\gamma - \delta) - Z - N(1 + 2\gamma - \delta) + (1 + \gamma) = 0. \tag{4}$$

Rearranging Eq. (4), we obtain an expression for the optimal Diaspora as a fraction of the total stock of Indian innovators:

$$\frac{D^*}{N} = \left(\frac{1 + 2\gamma - \delta}{2(\gamma - \delta)}\right) + \left(\frac{1}{2(\gamma - \delta)}\right)\left(\frac{Z}{N}\right) - \left(\frac{1 + \gamma}{2(\gamma - \delta)}\right)\left(\frac{1}{N}\right). \tag{5}$$

Eqs. (3)–(5) allow us to characterize the conditions under which a Diaspora is beneficial for knowledge access and innovation. We do this in two steps. First, an examination of Eqs. (3) and (4) reveals that, for this first-order condition to identify a maximum, we require from the second-order condition that δ is greater than γ:

$$\frac{\partial^2 K}{\partial D^2} = 2(\gamma - \delta) < 0 \quad \Rightarrow \quad \delta > \gamma. \tag{6}$$

Otherwise, the national knowledge access will decline monotonically with the size of the Diaspora (see the first equality in Eq. (4)). We first assume that this condition does not hold. A positive Diaspora is never beneficial in this case. We can give this necessary condition a more intuitive explanation. Suppose in the extreme that the potential emigrants contribute nothing directly to domestic innovation while at home. Their only contribution comes indirectly from the knowledge that flows from them to other domestic innovators. Whether their absence helps or harms, in that case, depends simply on whether domestic innovators access more knowledge from them when at home or abroad, i.e., on the relative magnitudes of δ and γ.

Second, we use Eq. (7) to identify the necessary and sufficient condition for a strictly positive Diaspora to be beneficial:

$$\frac{D^*}{N} > 0 \quad \Longleftrightarrow \quad \delta > 1 + 2\gamma + \frac{Z}{N} - \frac{1 + \gamma}{N}. \tag{7}$$

This condition is quite stringent. Even in the extreme case where N is sufficiently large enough that we can ignore the last two terms and where there is no co-location premium (i.e., $\gamma = 0$), the Diaspora premium must be greater than 100% for a Diaspora to be beneficial for the total knowledge flow to India-residing innovators.[8]

2.2. Circulatory migration

The model with permanent migration abstracts from one potentially important element: the return of emigrant innovators. Such returnees are likely to have developed connections with foreign innovators while away, connections that may endure on their return to facilitate ongoing knowledge flows.[9] To explore the implications of return, we next examine the steady state of a simple extension of the model that allows for circulation.

At any point in time, the change in the Diaspora share mechanically depends on the emigration rate (e), the return rate (r), the

[8] From Eq. (5) we can see that the optimal Diaspora share converges to one half as δ approaches infinity. In other words, it will never be optimal for a country to have more than half its innovators abroad. Although in reality we expect the optimal Diaspora share to be well below one half, this finding is of interest because there are several countries for which the number of tertiary-educated nationals residing abroad is greater than the number residing at home (Docquier and Marfouk, 2005). These general emigrant shares are likely to underestimate the share of innovators, given the tendency for emigrant shares from poor countries to rise with education level. The model suggests that this is detrimental to knowledge production no matter how large the productivity gains are from emigrating and no matter how strong the diasporic connections. This result implies that countries must have a sufficient number of innovators at home to reap the benefits of emigrant-related productivity gains and diasporic connections.

[9] Agrawal et al. (2006) provide evidence of the impact of enduring social capital acquired during past co-location on subsequent knowledge flows.

growth rate of new Indian scientists (n), and the initial Diaspora share[10]:

$$d\left(\frac{D}{N}\right) = \frac{1}{N}dD - \frac{D}{N^2}dN = \frac{1}{N}(e(N-D) - rD) - \frac{D}{N}n$$
$$= e - (e + r + n)\frac{D}{N}. \quad (8)$$

Setting Eq. (8) equal to zero, we have an expression for the *steady-state* Diaspora share:

$$\left(\frac{D}{N}\right)^{ss} = \frac{e}{e + r + n}. \quad (9)$$

For a given steady-state Diaspora share and a given n, the steady state is consistent with an infinite number of (e, r) pairs. One possibility is that a given Diaspora share is observed with very low emigration and return rates, such that the Diaspora and the stock of scientists remaining in India have the character of "stagnant pools." However, we can observe the same Diaspora share with much higher emigration and return rates, such that the Diaspora and India-residing stocks have more the character of "circulating pools," innovators whom Saxenian (2006) calls the "New Argonauts" after the Greeks who sailed with Jason in search of the Golden Fleece. The nature of the India-residing stock is likely to have implications for the strength of their connections to domestic, Diaspora, and foreign scientists, with the relative strength of connections to innovators abroad increasing with the propensity to circulate.

Given perpetual circulation, the expected fraction of time that any Indian innovator will spend in the Diaspora will converge to the steady-state Diaspora share for any strictly positive return rate. Looked at from the viewpoint of innovators currently residing in India, the expected fraction of time spent abroad in the past is therefore increasing in the steady-state Diaspora share. An implication is that with a positive return rate, a higher Diaspora share is likely to be associated with stronger connections to foreign innovators.[11] This suggests a potential problem with inferences about optimal Diaspora size based on the static model.

We develop the static model on the premise of proportional co-location and Diaspora premiums that are independent of the size of the Diaspora itself. This independence allows us to estimate these premiums and then make inferences about the optimal size of the Diaspora. However, if a larger Diaspora share is associated with stronger connections to innovators abroad, then it is likely that the size of the Diaspora will affect the proportional co-location and Diaspora premiums. But when these premiums depend on the size of the Diaspora, we face the problem that we cannot use estimates of these premiums (based on a time period with a given Diaspora) to infer the size of the optimal Diaspora. We outline our method for identifying the importance of return in the empirical strategy section below.

2.3. Heterogeneous innovators and non-random selection

We have assumed that all innovators are equally productive. However, we can weaken this assumption without affecting the results if we assume that emigrants and returnees are random selections from the stocks of India-residing innovators and the Diaspora, respectively. The results are obviously affected, however, if emigrants and returnees are non-random selections from their respective pools.

Suppose, for example, that the most productive innovators have a higher probability of emigrating (possibly because they have a higher probability of qualifying for a visa such as the US H-1B). This positive selection will tend to augment the absence-related loss to India, suggesting an even lower optimal Diaspora.

Suppose further that returnees are a positive selection of the already positively selected Diaspora. It is possible that a few truly outstanding returnees, coming back with significantly enhanced productivity due to their time spent abroad, could have a major impact on Indian innovation. In this case, our model would give a misleading picture of the long-run effect of migration. We outline our tests for non-random selection in the empirical strategy section below.

2.4. Knowledge access and the value of an innovation

A core idea of the model is that knowledge access drives innovation. To keep the model as simple as possible, we have made the restrictive assumption that the way relationships facilitate knowledge access is the same for all innovators. One obvious concern is that the KFPF differs systematically based on the value of the innovation. For example, high-value innovations may draw relatively more on frontier knowledge through the Diaspora. As another example of how the KFPF may be context specific, Nanda and Khanna (2007) find that Diaspora connections are more important for Indian software entrepreneurs operating in weak institutional environments. We outline our method of testing for systematic differences in the KFPF in the empirical strategy section below.

3. Empirical strategy

To empirically implement the model, we follow the well-established approach of using patent citations as (noisy) indicators of knowledge flows between inventors.[12] The empirical challenge in estimating the impact of δ (Diaspora premium) and γ (co-location premium) on knowledge flows is to separate the effect of being in the same Diaspora and being co-located on knowledge flows from the underlying distribution of inventive activity across geographic and ethnic space. In other words, our objective is to estimate disproportionate levels of knowledge flow that are above and beyond a baseline level that would be expected given the distribution of overall inventive activity. For example, although we might observe high levels of knowledge flows between inventors residing in Bangalore, India and the Indian Diaspora residing in Silicon Valley, this may be due to the fact that the software industry is highly concentrated in those two locations. In other words, it might not be due to the ethnic connection between these two populations *per se* but rather simply reflect the underlying distribution of inventive activity in software across geographic space, which happens to be correlated with the distribution of the Indian Diaspora.

Building on the technique pioneered in JTH and modified for the purpose of ethnic matching in Agrawal et al. (2007), we choose a "control" patent to match every patent cited by a patenting Indian inventor. We select the control patents to match the technology class and vintage of each of the cited patents as closely as possible. Selecting a matched patent within the same technology class and from the same vintage (identical application year and proximate issue year) provides us with a benchmark that controls for the underlying distribution of inventive activity across time and technology space. So, in the example above, the baseline used for

[10] The emigration rate is the fraction of the stock of India-residing innovators ($N - D$) that emigrates each period, the return rate is the fraction of the innovator Diaspora (D) that returns each period, and the new innovator growth rate is the proportionate growth in the total stock of Indian innovators (N).

[11] When the return rate is zero, such that the current India-residing stock has spent no time abroad, the strength of the connection to foreign scientists is independent of the size of the Diaspora.

[12] See Jaffe and Trajtenberg (2002) for key developments in the use of patent citation data to track knowledge flows.

comparison would incorporate the high concentration of inventive activity in software in those two locations and thus only attribute knowledge flows above and beyond the expected level to the ethnic connection. We more fully discuss the process of selecting matched observation patents in Section 4.

To further clarify, assuming this matching procedure is successful, the cited and matched patents will have the same geographic distribution even where inventive activity is geographically concentrated within narrow technological specializations. Thus, if inventor co-location and co-membership in an ethnic Diaspora play no role in facilitating knowledge flows, knowing that the inventor on the focal patent and the inventor on the cited patent have a location or a Diaspora connection should be of no help in distinguishing an actual citation from a matched observation. On the other hand, if co-location and Diaspora membership are disproportionately associated with actual citations, we can use the estimated premiums as measures of the causal effects of location and Diaspora connections on knowledge flows.

The model points to the central empirical task: the identification of δ and γ parameters. If we find that δ is less than γ, then emigration is detrimental to knowledge flows. Even if δ is greater than γ, the gap will have to be large for a Diaspora to be beneficial.

We run the following regression to identity the key parameters:

$$\Pr(Citation = 1 | SameCountry, Diaspora) = a_0 + a_1\ SameCountry + a_2\ Diaspora + \varepsilon \quad (10)$$

The dependent variable throughout our analyses is *Citation*, which is an indicator variable assigned a value of 1 if the "citation" is an actual citation, thus reflecting a knowledge flow, or 0 if it is a matched observation. We use two main explanatory variables. *SameCountry* is an indicator variable assigned a value of 1 if at least one of the inventors on the cited patent is located in India (and thus is co-located in the same country as the inventors of the focal patent who are by construction all located in India) and 0 otherwise. (We also examine city-level co-location with the indicator variable *SameCity*.) *Diaspora* is an indicator variable assigned a value of 1 if at least one of the inventors has an Indian last name and none of the inventors are located in India.

If we randomly choose a cited/matched patent for which we know that both *SameCountry* and *Diaspora* equal 0, then we give an estimate of the probability that the observation is an actual citation by $\hat{a}_0$. However, if we know that the inventors are co-located, the estimate of the probability that the observation is an actual citation is given by $\hat{a}_0 + \hat{a}_1$. The proportionate increase in the probability that the observation is an actual cited patent is then $\frac{\hat{a}_1}{\hat{a}_0}$, which we take to identify the proportionate increase in the probability of a knowledge flow caused by co-location, that is, an estimate of γ. Similarly, $\frac{\hat{a}_2}{\hat{a}_0}$ provides an estimate of δ.

Co-location and Diaspora membership are unlikely to be equally important for all knowledge flows. Thus, we examine: (1) differences based on elapsed time between the focal patent and the cited patent (we expect that relationships are less important the longer the invention is in the public domain); (2) differences based on whether the knowledge is flowing across or within technological boundaries (we expect that relationships based on location and co-ethnicity are more important for inventors who do not share a technology specialization)[13]; (3) differences based on broad technology class (for example, owing to differences in the importance of non-codifiable knowledge, knowledge exchange in computing research might be less dependent on proximity than knowledge exchange in medical research)[14]; and (4) differences between "vintages" by comparing the co-location and Diaspora parameters for earlier versus later focal patents.[15]

We examine the importance of return (i.e., individuals who leave their home country and subsequently return) in two ways. First, we simply measure how many of the India-residing inventors are actually returnees. A finding that returns are rare will provide support for the constant parameters assumption. Second, we determine whether the co-location and Diaspora premiums are systematically different for returnees compared with inventors who never emigrated. Even if return is a significant phenomenon, a finding that the KFPFs are not significantly different for returnees and non-returnees will also provide support for the constant parameter model.

To determine the importance of non-random selection, we follow the prior literature and use forward citations (the number of citations a focal invention receives) as a proxy for the quality of the inventor. We discuss in more detail the use of forward citations in the next paragraph. To determine if emigrants are differentially selected, we look forward from the application dates of each focal patent to see if the inventors subsequently emigrated. We then compare the "quality" of the patents of non-emigrants to those of emigrants. Similarly, we compare the "quality" of the patents of returnees and non-returnees to make inferences about returnee selectivity.

Finally, we look for differences in the KFPF based on the value of an innovation. Our indicator of invention value is the number of citations received by a patent. It is well known that the distribution of patents in terms of their value is highly skewed (Scherer and Harhoff, 2000). In other words, a small fraction of patents accounts for the majority of value. Research has shown that the number of citations a patent receives correlates with several direct measures of patent value, including consumer-surplus (Trajtenberg, 1990), expert evaluation of patent value (Albert et al., 1991), patent renewal rates and infringement litigation (Harhoff et al., 1999; Lanjouw and Schankerman, 1999), and contribution to a firm's market value (Hall et al., 2005). We follow this interpretation and use the number of citations received by a patent as a proxy for its impact. Differentiating by the value of an innovation, we then test whether the co-location and Diaspora effects, i.e., the KFPFs, are systematically different for higher value innovations.

4. Data

We begin the construction of our sample by identifying a set of patents that are associated with inventions created by individuals based in India. Specifically, we identify all patents issued by the United States Patent and Trademark Office by 2004 (inclusive) that were applied for during the period 1981–2000 (inclusive) where all inventors listed on the patent are located in India.[16] Since we focus on knowledge flows proxied by citations, we also impose the restriction that focal patents make at least one citation. The majority of patents (84%) meet this criterion. We drop from the sample those that do not, either because they make no citations or because the citations they make are to patents issued before 1976 and thus not in our database. We are left with 831 patents.[17] These are our focal

[13] We measure technology co-specialization by the focal patent and the cited/citing patent sharing the same NBER two-digit technology classification.

[14] We divide focal patents in broad technological classes based on NBER one-digit technology classifications.

[15] The motivation for testing for such effects is that advances in communications technology may have changed the relative value of location-based and Diaspora-based relationships.

[16] We use information from the inventor country address field, not the assignee field.

[17] It is interesting to note that an additional 761 patents during the same time period list at least one inventor as located in India and at least one co-inventor (also on the patent) as residing outside of India. The co-inventor not residing in India is a member of the Indian Diaspora in approximately 14% of these cases.

patents. On average, they cite 6.7 patents, generating 5527 focal-cited patent pairs.

Since we are analyzing knowledge flows, we are not only interested in these focal inventions but also the prior ideas on which they build. Therefore, we use focal-cited patent pairs as our unit of analysis and also collect data associated with all of the patents that each focal patent cites. By taking this approach, we are using patent citations as a proxy for knowledge flows.

However, citations are not straightforward to interpret. Patents cite other patents as prior art, with citations serving to delineate the property rights conferred. Yet the applicant does not supply all citations; some are added by the patent examiner (Alcacer and Gittelman, 2006; Hegde and Sampat, 2007). Furthermore, 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. For example, Cockburn et al. (2002) report 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) survey cited and citing inventors to explore the meaning of patent citations and find that approximately one-quarter of survey responses corresponded to a "fairly clear spillover," approximately one-half indicated no spillover, and the remaining quarter indicated some possibility of a spillover. Those 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). Thus, we use citations here, but cautiously, recognizing that they are a noisy proxy for knowledge flows.

Next, for each cited patent, we identify a "control" patent that matches the cited patent on two dimensions: vintage and technology area. Specifically, we match on application year and the full six-digit primary US technology classification. If we cannot identify a suitable matched patent, then we drop the observation.

If we identify more than one suitable matched patent, then we select the patent that matches as many full secondary six-digit classifications as possible. If we identify more than one potential matched observation patent with "the best" match on technology classifications, then we select the one with the application date closest to that of the cited patent. Based on these criteria, we find matched patents for 86% of our cited patents. Thus, our sample consists of 9520 observations of which, by construction, half are focal-cited patent pairs and the other half are focal-matched patent pairs. In Table 1, we show that approximately 2% of the cited/matched patents are co-located with the focal patent (*SameCountry*) and approximately 4% of the cited/matched patents are invented by the Diaspora.

We identify inventors as being members of the Indian Diaspora based on their last names. 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. We then code these names based on their likelihood of being Indian. Of the 213,622 last names identified from the phone books, 38,386 names appear with a frequency of five or more. Of these, 13,418 match a proprietary database of US consumers (prepared by InfoUSA). One of the authors and an outside expert have 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 we use in this study includes only the 6885 last names coded as "extremely likely to be Indian."

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 ("extremely likely to be Indian") responded "yes" to the question: "Are you of Indian origin?" (Kapur, 2004).[18] Nor do we expect the frequency of false negatives to be large. Although we construct 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)). Furthermore, noise in our name data will bias our result downwards.

We have reason to believe that the Indian Diaspora is likely to stay connected to individuals in their home country. For example, 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).[19] Furthermore, more than one third of the Indian-born came after 1996 and more than half after 1990. 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 1022,552 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

Table 1
Descriptive statistics.

	Mean	Std. dev.	Min	Max
Cited patent is co-located with focal (SameCountry)	0.02	0.14	0	1
Cited patent is city co-located with focal (SameCity)	0.01	0.11	0	1
Cited patent is by Diaspora member	0.04	0.19	0	1
Application year of focal patent	1997.29	3.51	1981	2000
Lag[a]	7.93	5.92	0	27
Within-field knowledge flow[b]	0.62	0.48	0	1
Importance of focal patent[c]	2.88	7.36	0	112
Focal patent is by returnee	0.02	0.15	0	1
Focal patent is by future emigrant	0.03	0.18	0	1
Cited patent is cited by an inventor of focal patent	0.01	0.12	0	1
Cited patent is cited by an assignee of focal patent	0.03	0.16	0	1

$N = 9520$ observations.
[a] Years between the application date of the focal versus the cited patent.
[b] Probability the focal and cited patent are both assigned to the same two-digit NBER technology sub-category.
[c] Number of citations received by the focal patent.

[18] It is important to note that there may be differences in measurement error across our main regressors. Specifically, there is likely to be more error in the identification of members of the Diaspora than in the identification of co-location of inventors who are in the same country. However, these survey data provide us with some comfort that the number of false positives (foreign inventors coded as Diaspora who are actually not) in our Diaspora measure should be quite low. Given the large differences we report in the Results section between the estimated coefficient on *Diaspora* and *SameCountry*, this difference in errors is not likely to affect the results in a qualitative manner. We explore this issue in the results section by relaxing the Diaspora membership constraint, increasing the number of false positives, and decreasing the number of false negatives. The main result, that the knowledge access benefits of co-location outweigh those of the Diaspora, becomes stronger.

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

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. Moreover, survey evidence underlines the strong ethnic identification of the Diaspora in America: 53% visit India at least once every 2 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 (Kapur, 2004). When there are multiple inventors, we define a Diaspora patent as one where at least one inventor has an Indian last name and none of the inventors reside in India.

Although we construct our dataset from focal patents applied for during the period 1981–2000, the mean application year is 1997 (Table 1). These data are skewed towards the end of our study period due to the significant growth of patenting in India at that time. The average lag between the focal patent and the preceding cited patent is 8 years. Recall that the lag between focal and matched patents is precisely the same, by construction.

We compare various types of knowledge flows in terms of the degree to which they are mediated by co-location and Diaspora effects. These comparisons include: (1) flows within versus across fields, (2) flows associated with returnees (individuals who patented an invention outside of India and then returned to patent within India) versus those who show no evidence of ever having left India, (3) flows associated with future emigrants (individuals who patent in India and later patent abroad) versus others, and 4) flows associated with more versus less important inventions.

Table 1 shows that more than half (62%) of the focal-cited pairs represent within-field knowledge flows. Again, the fraction of focal-matched pairs that represent within-field knowledge flows is the same, by construction. In terms of the relative value or impact of the focal invention, the mean number of citations received by focal patents is approximately three. In terms of "circulation," returnees invent approximately 2.5% of the focal patents in our data. Finally, in terms of future emigrants, individuals who later leave India invent approximately 3% of the focal patents in our data.

Table 2 presents four panels of descriptive statistics concerning the urban characteristics of our sample. Panel A lists the 10 Indian city-regions with the most inventive activity. Although India is a large country with a dense population spread throughout the nation, inventive activity is very concentrated in just a few urban regions. In fact, the three most active city-regions (Mumbai, New Delhi, and Bangalore) account for approximately 54% of all Indian patenting activity during our sample period.

Panels B displays the 10 cities that are most cited by our sample of focal patents. New York, San Francisco, and Philadelphia are the three most heavily cited cities and account for approximately 16% of the foreign locations cited by our sample of focal patents. Panel C presents the 10 cities that are most cited by our sample of focal patents but where we only count citations to inventors who are members of the Diaspora. The two lists (Panel B and C) contain four cities in common: New York, San Francisco, Philadelphia, and Chicago, which are also the four top cities in Panel C. However, while they are similar, they are not identical. Although Tokyo Japan, Washington DC, Boston MA, Los Angeles CA, Norwich CT, and Austin TX are all on the most-cited list in Panel B, they do not appear on the list of most-cited Diaspora patents. Instead, Panel C contains: Raleigh NC, Houston TX, Indianapolis IN, Reading (UK), Charlotte NC, and Camberley (UK). In addition, it appears that citations to the Diaspora are more geographically concentrated than citations overall; the top 10 cities account for 52% of all citations to Diaspora members (Panel C), whereas the 10 most-cited cities overall (not conditioning on cites to the Diaspora) only account for 27% of citations.

Finally, Panel D lists the most common focal–Diaspora city pairs. Interestingly, almost 20% of all Indian-Diaspora linkages in our sample are between four cities: Bangalore–San Francisco, New Delhi–Chicago, New Delhi–New York, and Mumbai–New York. Collectively, the top 10 focal-Diaspora city pairs account for approximately one third of the overall number of focal-Diaspora links.

5. Results

Table 3 reports the OLS results for the full sample.[20] Focusing first on Specification (1), we find evidence of a large and statistically significant co-location effect (*SameCountry*) and a much smaller (though still statistically significant) Diaspora effect. The difference between the two effects is also positive and statistically significant at the 1% level. The implied estimate of the proportionate co-location premium is (0.395/0.490) = 0.806, whereas the implied estimate of the proportionate Diaspora premium is just (0.066/0.490) = 0.135. Interpreted through the lens of the model, the much larger co-location premium implies that the total access of India-residing inventors to knowledge is harmed by the absence of fellow Indian inventors. Furthermore, the very large co-location premium confirms the importance of localized knowledge flows.

We include an indicator variable (*SameCity*) representing instances of co-location at the city level in Specification (2). Conditioning on the level of country co-location, further co-location at the city level is not associated with a higher probability of citation.

We include controls for instances where the focal patent cites a patent that is assigned to the same assignee in Specifications (3) and (4). Interestingly, although the citation being an assignee self-citation does not increase the probability that it is a real rather than a matched citation, it does increase the probability if the citation is both an assignee self-citation and a citation to a member of the Indian Diaspora. In fact, in Specification (4), although the point estimate of the coefficient on Diaspora is positive and similar in magnitude to the previous specifications, it is no longer statistically significant once we include the interaction of assignee self-cite and Diaspora. Overall, this result indicates that inventors who are based in India and work for multinational firms disproportionately cite the Indian Diaspora who are employed by the same firm but based at facilities in other countries.

In terms of inventor-level self-citations, Specifications (5) and (6) include controls for instances where the focal patent and the cited patent are authored by the same inventor (and perhaps others). Not surprisingly, Indian inventors are more likely to cite their own work than a matched patent. However, even controlling for inventor self-cites, the statistical significance and magnitude of our two main parameters of interest remain largely unchanged. Similarly, when we include both inventor- and assignee-level self-cites in Specification (7), the magnitude and statistical significance of our two parameters of interest are qualitatively unchanged.

We allow for the co-location and Diaspora effects to vary by the citation lag and also by whether the citation occurs within or across NBER two-digit classifications in Specification (8).[21] There is no direct effect of lags and sub-category matches since we choose the control patents by matching on both timing and technology class. Surprisingly, the estimated coefficient on the interaction between *Lag* and *SameCountry* suggests that the co-location effect increases with the age of the patent. This is counterintuitive since we would expect co-location to matter more for newer inventions since ideas diffuse further over time (Jaffe et al., 1993).

[20] We find identical conditional probability estimates using a logit specification. We concentrate on the OLS results due to their ease of interpretability. Furthermore, all predicted values of our dependent variable fall within the unit interval.

[21] For example, within the general category of computers and communications, the two-digit classification system distinguishes between communications, computer hardware and software, computer peripherals, and information storage.

Table 2
City descriptive statistics.

Rank	City		Count	Share (%)	Cumulative share (%)
Panel A: top 10 focal patent cities					
1	MUMBAI Region, IN		184	20.86	20.86
2	NEW DELHI Region, IN		173	19.61	40.48
3	BANGALORE Region, IN		123	13.95	54.42
4	PUNE, IN		68	7.71	62.13
5	HYDERABAD, IN		67	7.60	69.73
6	LUCKNOW, IN		57	6.46	76.19
7	CHENNAI, IN		38	4.31	80.50
8	CALCUTTA Region, IN		26	2.95	83.45
9	AHMEDABAD, IN		24	2.72	86.17
10	KOCHI Region, IN		12	1.36	87.53
	Total Patent-Locations		883		
Panel B: top 10 outside india cited patent cities					
1	NEW YORK, NY MSA		515	7.37	7.37
2	SAN FRANCISCO, CA MSA		414	5.93	13.30
3	PHILADELPHIA, PA MSA		210	3.01	16.31
4	TOKYO, JP		145	2.08	18.38
5	WASHINGTON, DC MSA		133	1.90	20.29
6	CHICAGO, IL MSA		128	1.83	22.12
7	BOSTON, MA MSA		126	1.80	23.93
8	LOS ANGELES, CA MSA		126	1.80	25.73
9	NORWICH, CT MSA		62	0.89	26.62
10	AUSTIN, TX MSA		56	0.80	27.42
	Total Patent-Locations		6984		
Panel C: top 10 Diaspora cited patent cities					
1	SAN FRANCISCO, CA MSA		40	14.13	14.13
2	NEW YORK, NY MSA		37	13.07	27.21
3	CHICAGO, IL MSA		18	6.36	33.57
4	PHILADELPHIA, PA MSA		13	4.59	38.16
5	RALEIGH, NC MSA		12	4.24	42.40
6	HOUSTON, TX MSA		6	2.12	44.52
7	INDIANAPOLIS, IN MSA		6	2.12	46.64
8	READING, GB		6	2.12	48.76
9	CHARLOTTE, NC MSA		5	1.77	50.53
10	CAMBERLEY, GB		5	1.77	52.30
	Total Patent-Locations		283		
Panel D: top 10 focal patent city – Diaspora city pairs					
1	BANGALORE Region, IN	SAN FRANCISCO, CA MSA	28	7.98	7.98
2	NEW DELHI Region, IN	CHICAGO, IL MSA	16	4.56	12.54
3	NEW DELHI Region, IN	NEW YORK, NY MSA	14	3.99	16.52
4	MUMBAI Region, IN	NEW YORK, NY MSA	10	2.85	19.37
5	BANGALORE Region, IN	NEW YORK, NY MSA	9	2.56	21.94
6	HYDERABAD, IN	RALEIGH, NC MSA	9	2.56	24.50
7	KURUKSHETRA, IN	CHICAGO, IL MSA	7	1.99	26.50
8	MUMBAI Region, IN	READING, GB	6	1.71	28.21
9	NEW DELHI Region, IN	PHILADELPHIA, PA MSA	6	1.71	29.91
10	NEW DELHI Region, IN	SAN FRANCISCO, CA MSA	5	1.42	31.34
	Total Focal City – Diaspora City Dyads		351		

Note: Patent-Locations in Panel A exceed the 793 focal patents as each location on each patent is treated distinctly.
As such, a patent with two distinct locations will increase the count of Patent-Locations by two.
Put differently, each patent has on average 1.11 locations.

Finally, Specification (9) allows for a less strict definition of Diaspora membership, whereby we assign surnames that are more ambiguously Indian to the Diaspora (i.e., we include not only last names that are coded as "extremely likely to be Indian," as in our main sample, but also names coded as "could be either"). With this expanded sample, the point estimate of the coefficient on Diaspora decreases by almost half and is statistically insignificant as would be expected due to the attenuation bias that results from the increased noise in the measurement of Diaspora membership.

Table 4 shows the results for our base specification for five of the six NBER one-digit classifications (we leave out the sixth category, "Others," due to the very small number of observations). We find the previously identified pattern of large co-location effects and small Diaspora effects in most categories. However, we also find larger point estimates of the coefficient on *Diaspora* for both Electrical & Electronic and Mechanical, though only the former is statistically significant at the 10% level. Still, these estimates are much smaller than those of the coefficients on *SameCountry*. The single exception is Computers & Communications, which indicates no co-location effect. Perhaps India's international competitiveness in this sector, particularly information technology, involves drawing from a more global knowledge base, which is reflected in this finding.

In Table 5, we examine whether vintage mediates the co-location and Diaspora effects on knowledge flows. We take 1995 as the cutoff, but the results are not sensitive to this choice. The gap between the co-location and Diaspora parameters is somewhat greater for the more recent focal patents (both because the co-loca-

Table 3
OLS estimates of the KFPF.

Dependent variable = *Citation*	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SameCountry	0.395***	0.367***	0.371***	0.398***	0.322***	0.381***	0.317***	0.382***	0.383***
	(0.023)	(0.051)	(0.029)	(0.031)	(0.050)	(0.044)	(0.051)	(0.045)	(0.045)
Diaspora	0.066**	0.066**	0.064**	0.053	0.063**	0.054*	0.062**	0.088	
	(0.032)	(0.032)	(0.032)	(0.033)	(0.031)	(0.031)	(0.031)	(0.068)	
SameCity		0.038							
		(0.059)							
Assignee Self-Cite			0.061	0.055			0.051		
			(0.039)	(0.052)			(0.039)		
Assignee Self-Cite × SameCountry				−0.058					
				(0.069)					
Assignee Self-Cite × diaspora				0.270***					
				(0.098)					
Inventor Self-Cite					0.110*	0.260*	0.087		
					(0.063)	(0.153)	(0.067)		
Inventor Self-Cite × SameCountry						−0.239			
						(0.162)			
Inventor Self-Cite × Diaspora						0.196			
						(0.156)			
Lag × SameCountry								0.017***	0.017***
								(0.006)	(0.006)
Lag × Diaspora								−0.000	0.003
								(0.007)	(0.006)
Sub-Category Match × SameCountry								−0.026	−0.026
								(0.047)	(0.047)
Sub-Category Match × Diaspora								−0.017	0.018
								(0.067)	(0.047)
"Could be either" Diaspora[a]									0.036
									(0.027)
Constant	0.490***	0.490***	0.489***	0.489***	0.490***	0.490***	0.489***	0.489***	0.488***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
# Observations (actual & matched observation citations)	9520	9520	9520	9520	9520	9520	9520	8862	8862
R^2	0.012	0.012	0.013	0.013	0.013	0.013	0.013	0.014	0.014
# Clusters (focal patents)	793	793	793	793	793	793	793	751	751

Notes: Dependent variable is if the citation is an actual cited patent versus a matched observation patent.
The unit of analysis is the focal-cited patent pair with a matched observation for each cited patent.
The number of Focal Patents (and conversely total observations) decreases in Columns 8 and 9, due to missing sub-category data.
Focal patent cluster-adjusted standard errors in parentheses.
[a] "Could be either" includes names whose origins are more ambiguous and thus increases the risk of false positives while reducing the risk of false negatives.
* Significance at 10% level.
** Significance at 5% level.
*** Significance at 1% level.

Table 4
OLS estimates of the KFPF By NBER one-digit code.

Dependent variable = *Citation*	(1) Chemical	(2) Computers & Comm.	(3) Drugs & Medical	(4) Electrical & Electronics	(5) Mechanical
SameCountry	0.418***	0.061	0.440***	0.508***	0.511***
	(0.035)	(0.156)	(0.030)	(0.003)	(0.005)
Diaspora	0.027	0.078	0.054	0.175*	0.178
	(0.059)	(0.054)	(0.067)	(0.100)	(0.160)
Constant	0.490***	0.495***	0.485***	0.492***	0.489***
	(0.002)	(0.004)	(0.004)	(0.003)	(0.005)
# Observations (actual & matched observation citations)	2826	1682	2734	1170	282
R^2	0.016	0.002	0.022	0.011	0.019
# Clusters (focal patents)	389	105	298	107	77

Notes: Dependent variable is if the citation is an actual cited patent versus a matched observation patent.
The unit of analysis is the focal-cited patent pair with a matched observation for each cited patent.
Focal patent cluster-adjusted standard errors in parentheses.
* Significance at 10% level.
*** Significance at 1% level.

tion effect has risen and the Diaspora effect has fallen) but remains large even for older vintage focal patents.

As outlined in Sections 2.2–2.4, the interpretation of these results is made more complicated by return migration, non-random selection, and heterogeneous-valued innovations. To analyze the impact of returnees, we split the sample into returnees and non-returnees in the first two specifications of Table 6. The first thing to note is that returnees account for just 2.3% of our sample of focal patents. We are concerned that this significantly under-represents the true number of returnees since the identification of a returnee

Table 5
OLS estimates of the KFPF By "Vintage".

Dependent variable = *Citation*	Vintage	
	Recent (1) Application year for focal patent > 1995	Early (2) Application year for focal patent ⩽ 1995
SameCountry	0.402*** (0.025)	0.369*** (0.057)
Diaspora	0.063* (0.035)	0.097 (0.078)
Constant	0.490*** (0.002)	0.492*** (0.002)
# Observations (actual & matched observation citations)	7524	1996
R^2	0.013	0.010
# Clusters (focal patents)	588	205

Notes: Dependent variable is if the citation is an actual cited patent versus a matched observation patent.
The unit of analysis is the focal-cited patent pair with a matched observation for each cited patent.
Focal patent cluster-adjusted standard errors in parentheses.
* Significance at 10% level.
*** Significance at 1% level.

Table 6
OLS estimates of the KFPF by returnee/future emigrant status.

Dependent variable = *Citation*	(1) Non-returnees	(2) Returnees[a]	(3) Non-future emigrants	(4) Future emigrants[b]
SameCountry	0.397*** (0.024)	0.350*** (0.063)	0.396*** (0.023)	0.369** (0.171)
Diaspora	0.064* (0.033)	0.105 (0.124)	0.063* (0.033)	0.262 (0.225)
Constant	0.490*** (0.002)	0.483*** (0.015)	0.490*** (0.002)	0.488*** (0.009)
# Observations (actual & matched observation citations)	9286	234	9208	312
R^2	0.012	0.015	0.012	0.015
# Clusters (focal patents)	775	18	771	22
Mean forward cites to focal patent	2.863	3.598	2.366	18.083

Notes: Dependent variable is if the citation is an actual cited patent versus a matched observation patent.
The unit of analysis is the focal-cited patent pair with a matched observation for each cited patent.
Focal patent cluster-adjusted standard errors in parentheses.
* Significance at 10% level.
** Significance at 5% level.
*** Significance at 1% level.
[a] Returnees are identified as inventors who are observed to have previously patented outside of India.
[b] Future emigrants are identified as inventors who are subsequently observed to patent outside of India at a later date.

in the patent database requires that the individual previously patented abroad. This data limitation notwithstanding, we note that the co-location point estimate is slightly lower for returnees although the difference is not statistically significant. However, the estimated coefficient on *Diaspora* is similar. Most importantly, the gap between the co-location and Diaspora effects remains large. We find a similar result when we compare future emigrants with non-future emigrants in Specifications (4) and (3), respectively. Although the point estimate of the coefficient on *Diaspora* is measurably higher in the case of future emigrants compared to the other three specifications presented in that table, it not statistically significant and the co-location effect remains higher.

Table 6 also allows us to explore the nature of selection for returnees and emigrants. Our measure of inventor "quality" is the number of forward citations to the invention. As we describe in Section 3 above, we follow the prior literature in using forward citations as a proxy for invention quality. The last row in Table 6 gives the mean number of forward citations for the various subsamples. Comparing returnees and non-returnees by this measure, we find that returnees are of higher quality on average, although the difference is relatively small. In contrast, we find evidence that emigrants are highly positively selected. For our sample of focal patents, the mean number of forward citations is a little more than two for those who do not subsequently go on to emigrate and just over 18 for those who do. Taken together, these results suggest that emigrants are positively selected and returnees are negatively selected from the resulting (select) Diaspora pool. These findings on returnees and selection reinforce the inference based on the simple model: Inventor emigration harms knowledge access and domestic innovation.

The identification challenges created by positive selection have been a major focus of the literature on human capital externalities (Rosenthal and Strange, 2004). Although research has established that larger concentrations of human capital are associated with higher wages (Rauch, 1993), the finding could be due to more able workers being more likely to move and agglomerate. Recent work has focused on controlling for endogenous labor quality effects using longitudinal data. As with our data, there exists broad evidence that mobile workers are positively selected. In a recent study utilizing a rich panel of French workers, Combes et al. (2010) find an elasticity of mean wages to human capital density of 0.5. However, this elasticity falls by roughly one-third to 0.33 after controlling for endogenous labor quality, suggesting that positive selection explains some but not all of the density-wage association.

In Table 7, we examine whether invention impact mediates the co-location and Diaspora effects on knowledge flows. The results are striking. Focusing on the 88th percentile and above (Specification 2), we see a somewhat lower co-location effect and a substantially higher Diaspora effect as compared to the rest of the sample (Column 1) or the full sample (Table 3).[22,23] Further narrowing the sample to only the 93rd percentile and above (Specification (3)), we see an even greater Diaspora effect (almost 10 times the magnitude as that for the overall sample), and the co-location effect is no longer statistically significant. This Diaspora-oriented result continues to hold when we cap the sample even further along the tail of the distribution to include only the 95th percentile and above.

These results are particularly salient since prior research has shown that the value of innovations increases nonlinearly with the number of citations (Trajtenberg, 1990).[24] When we focus on the 95th percentile and above, the Diaspora effect slightly exceeds unity (see Section 2.1). Thus, our finding, that the estimated Diaspora effect rises dramatically and that the estimated co-location effect simultaneously falls substantially as we move out to the extreme right-hand side of the impact distribution, gives us pause in concluding that a Diaspora is never beneficial.

The small number of patents with even larger numbers of forward citations limits us from restricting attention to even higher

[22] We also look at the number of forward citations occurring within specified time windows—3 years, 5 years, and 10 years—and find similar results.
[23] The percentile cutoffs are not round numbers since they are dictated by the distribution of patents with certain numbers of citations received, which are discrete count values.
[24] An important caveat is that the evidence cited was taken from a single industry: computed tomography scanners.

Table 7
OLS estimates of the KFPF by "Quality" of focal patents.

Dependent variable = *Citation*	Slightly below average (1) Total Cites to Focal Patent < 6 (Below 88th percentile)	High (2) Total Cites to Focal Patent ⩾ 6 (88th percentile and above)	Very high (3) Total Cites to Focal Patent ⩾ 9 (93rd percentile and above)	Extremely high (4) Total Cites to Focal Patent ⩾ 12 (95th percentile and above)
SameCountry	0.402***	0.321***	0.258	0.172
	(0.023)	(0.104)	(0.156)	(0.198)
Diaspora	0.057*	0.231*	0.508***	0.505***
	(0.033)	(0.126)	(0.003)	(0.004)
Constant	0.490***	0.491***	0.492***	0.495***
	(0.002)	(0.003)	(0.003)	(0.004)
# Observations (actual & matched observation citations)	8448	1072	650	372
R^2	0.013	0.009	0.013	0.007
# Clusters (focal patents)	699	94	59	38

Notes: Dependent variable is if the citation is an actual cited patent versus a matched observation patent.
The unit of analysis is the focal-cited patent pair with a matched observation for each cited patent.
Focal patent cluster-adjusted standard errors in parentheses.
* Significance at 10% level.
*** Significance at 1% level.

quality patents. But the rising size of the Diaspora effect (both absolutely and relative to the co-location effect) as we restrict the sample to higher-quality focal patents raises the possibility that a Diaspora is beneficial where the welfare effect of high-quality inventions is large relative to the average invention. In the next section, we discuss further how this finding tempers our interpretation of the main findings reported in Table 3.

6. Conclusion

We find evidence of a large co-location premium for knowledge flows between Indian inventors associated with the "average" invention. We also find evidence of a Diaspora premium, but its size is much smaller (14% compared to 81%). Interpreted through the lens of a simple relationships-based model of knowledge access and innovation, the difference between the effects is a sufficient condition for emigration to be harmful to the domestic economy.

While our model abstracts from the possibility of return and also of the non-random selection of emigrants and returnees, we find that returnees are quite rare in our sample of Indian innovators and that their knowledge-flow characteristics are similar to innovators who never left. Our data also indicate that emigrant innovators are a highly positively selected sub-sample of the Indian innovator population and that returnees are negatively selected from the emigrant stock. Thus, our basic conclusion is robust to returnee and selection effects.

However, we temper this conclusion drawn from our main results with our additional finding that domestic access to knowledge facilitated by the Diaspora is relatively more important for high-value inventions. Given that the distribution of patents is highly skewed with respect to market value (and social value), the small fraction of patents for which the Diaspora effect is particularly important might actually represent a large fraction of the productivity gains that result from innovation. Thus, to fully understand the effect of emigration on domestic innovation in a poor country, we need to better understand the relative value of very important innovations compared to others.

The central assumption of our model is that innovation output depends on access to knowledge. This focus on knowledge access allows us to incorporate a range of widely discussed but difficult to quantify emigration-related impacts on the domestic economy, including the loss of local knowledge spillovers, the gains via Diaspora connections, and the implications of knowledge-worker circulation. A limitation of our approach, however, is that we investigate innovation indirectly through our measures of knowledge access. The most important next step is to more directly measure how migration flows affect national innovation. We are currently exploring this question using detailed information on the career paths and productivity of mobile scientists.

Two additional issues in our paper need further investigation. One is whether skilled migration indeed entails a tradeoff between a smaller domestic stock of innovators and larger international networks. We have not addressed the possibility that the domestic stock of innovation-producing talent might actually increase as a result of migration. This may occur because of two possible effects.

The first effect arises because the possibility of migration induces higher investments in education due to greater returns abroad (Beine et al., 2001). If these additional investments in human capital are sufficiently large but only a fraction of innovators can actually leave, then it is possible that the country will end up with a greater stock of human capital. Although the basic "brain gain" story has plausibility given the clearly forward-looking nature of the demand for skills, doubts remain as to it its quantitative importance. Commander et al. (2004) as well as Schiff (2005) have argued, for example, that the highest-ability individuals will invest in skills regardless of the prospect of emigrating, but these individuals will be particularly prone to being recruited away when the prospect of emigration is enhanced. Thus, increased investments may provide the largest boost to the supply of more moderate-ability individuals.

A second effect could arise whereby increases in financial remittances may increase investments in education investment by easing binding liquidity constraints (Yang, 2006). Here, too, are contrary effects. For instance, if parents are absent from the household as a result of migration, there could be less parental inputs into education acquisition and greater work pressure on remaining household members. While Yang (2006) presents evidence from the Philippines where the former effect dominates, in Mexico's case the second factor appears to dominate (McKenzie and Rapoport, 2006). This issue needs further investigation.

Another issue arises from the time period of the data (which ends in 2000 because of our use of forward citations) and the implications of right-censoring our data. The experiences of countries as varied as Ireland and South Korea and more recently

of China point to the importance of changing domestic economic conditions in catalyzing the "brain bank" effect; Diasporas have little impact on their home countries as long as their economies remain closed. India's economic liberalization and recent rapid growth rates are attracting some of its Diaspora back in tandem with new multinational R&D investments. This effect may become apparent in the patent data over time.

We began this paper by noting the controversy between those who think the emigration of knowledge workers is good for the national economy as it expands the global technology pool and those who are concerned about the harm to national innovation. Overall, we find it unlikely that a poor country with a reasonably functioning economy and working hard to absorb the massive stock of available technology is actually better off if a large fraction of its scarce talent resides abroad. To this end, our main empirical results suggest that, in terms of access to knowledge, the localization effect outweighs the Diaspora effect: Poor countries are better off if their highly skilled workers stay home.

However, we do not doubt that reallocation to higher productivity environments does increase global innovation and some of the fruits of that innovation surely do flow back to the poor-sending countries. Examples abound of emigrants from poor countries making great contributions to science. A few also do return to have transformative effects on their home countries, often as institution builders.[25] Furthermore, our findings suggest that access to knowledge provided by the Diaspora is particularly important for the highest value inventions, those in the extreme tail of the distribution. How important is this minority share of inventions to the overall economy of poor countries? This question sets the stage for future research and contributions to this lively and important debate on the role of migration for economic growth in poor countries.

Acknowledgments

We thank seminar participants at the University of Lille, Harvard Business School, National University of Ireland at Galway, and the University of Toronto for valuable comments. Iain Cockburn, Lee Fleming, Richard Freeman, Avi Goldfarb, Gordon Hanson, Ramana Nanda, Caglar Ozden, Tim Simcoe, and Will Strange provided especially detailed feedback. This research was funded by the Martin Prosperity Institute's Program on Innovation and Creative Industries, by the Social Sciences and Humanities Research Council of Canada (Grant No. 410-2004-1770 and Grant No. 537-2004-1006), and by Harvard University's Weatherhead Initiative grant. Their support is gratefully acknowledged. Errors and omissions are our own.

References

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. NBER Working Paper 12823.

Agrawal, A., Kapur, D., McHale, J., 2008. How do spatial and social proximity influence knowledge flows? Evidence from patent data. Journal of Urban Economics 64 (2), 258–269.

[25] For example, Dr. F.C. Kohli, who left India to study electrical engineering at Queen's University in Canada and then at MIT in the US and who became an active member of the Institute of Electrical and Electronics Engineers headquartered in New York, returned to India to lead Tata Consultancy Services (India's largest IT firm during most of the last quarter of the 20th century). More importantly, from a social welfare perspective, many cite him more generally as the "father of the Indian software industry." He credits much of his ability to accomplish what he did to his education and social network in North America. (The IT Revolution in India, F.C. Kohli: Selected Speeches and Writings, F.C. Kohli, 2005).

Albert, M., Avery, D., Narin, F., McAllister, P., 1991. Direct validation of citation counts as indicators of industrially important patents. Research Policy 20 (3), 251–259.

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.

Anderrson, R., Quigley, J., Wilhelmsson, M., 2004. University decentralisation as regional policy: the Swedish experiment. Journal of Economic Geography 4, 371.

Anderrson, R., Quigley, J., Wilhelmsson, M., 2009. Urbanization, productivity, and innovation: evidence from investment in higher education. Journal of Urban Economics 66, 2–15.

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

Basu, S., Weil, D., 1998. Appropriate technology and growth. The Quarterly Journal of Economics 113 (4), 1025–1054.

Beine, M., Docquier, F., Rapoport, H., 2001. Brain drain and economic growth: theory and evidence. Journal of Development Economics 64 (1), 275–289.

Caselli, F., Coleman II, W.J., 2001. Cross-country technology diffusion: the case of computers. The American economic review. Papers and Proceedings 91 (2), 328–335.

Cockburn, I., Kortum, S., Stern, S., 2002. Are All Patent Examiners Equal? The Impact of Examiner Characteristics on Patent Statistics and Litigation Outcomes. National Bureau of Economic Research, Working Paper 8980.

Cohen, W., Levinthal, D., 1989. Innovation and learning: the two faces of R&D. The Economic Journal 99 (397), 569–596.

Combes, P., Duranton, G., 2006. Labour pooling, labour poaching and spatial clustering. Regional Science and Urban Economics 36, 1–28.

Combes, P., Duranton, G., Gobillon, L., Roux, S., 2010. Estimating agglomeration economies with history, geology, and worker effects. In: Glaeser, Edward (Ed.), Agglomeration Economics. University of Chicago Press, pp. 15–65.

Commander, S., Kangasniemi, M., Winters, L.A., 2004. The brain drain: curse or boon? A survey of the literature. In: Baldwin, R.E., Winters, L.A. (Eds.), Challenges to Globalization: Analyzing the Economics. University of Chicago Press, pp. 235–278.

Docquier, F., Marfouk, A., 2005. International migration by educational attainment. In: Özden, C., Schiff, M. (Eds.), International Migration, Remittances, and the Brain Drain. The World Bank and Palgrave Macmillan, Washington DC and New York, NY.

Dumont, J.C., Lemaitre, G., 2005. Counting Immigrants and Expatriates in OECD Countries: A New Perspective. OECD Social, Employment, and Migration Working Papers No. 25.

Fallick, B., Fleischman, C., Rebitzer, J., 2006. Job-hopping in Silicon Valley: some evidence concerning the microfoundations of a high-technology cluster. Review of Economics and Statistics 88 (3), 472–481.

Freedman, M., 2008. Job hopping, earnings dynamics, and industrial agglomeration in the software and publishing industry. Journal of Urban Economics 64, 590–600.

Gerlach, H., Ronde, T., Stahl, K., 2009. Labour pooling in R&D intensive industries. Journal of Urban Economics 65, 99–111.

Hall, B.H., Jaffe, A., Trajtenberg, M., 2005. Market value and patent citations. RAND Journal of Economics 36 (1), 16–38.

Harhoff, D., Narin, F., Scherer, F.M., Vopel, K., 1999. Citation frequency and the value of patented inventions. The Review of Economics and Statistics 81 (3), 511–515.

Hegde, D., Sampat, B., 2007. Examiner citations, applicant citations, and the private value of patents. Economics Letters 105 (3), 287–289.

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

Jaffe, A., 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.

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

Kapur, D., McHale, J., 2005. Give us your best and brightest: the global hunt for talent and its impact on the developing world. Center for Global Development/Brookings Institution Press, Cambridge.

Keller, W., 2002. Geographic localization and international technology diffusion. The American Economic Review 92 (1), 120–142.

Kerr, W., 2008. Ethnic scientific communities and international technology diffusion. The Review of Economics and Statistics 90 (3), 518–537.

Klenow, P., Rodriguez-Clare, A., 2004. Externalities and Growth. NBER Working Paper 11009.

Kuhn, P., McAusland, C., 2006. The International Migration of Knowledge Workers: When is Brain Drain Beneficial. NBER Working Paper 12761.

Lanjouw, J.O., Schankerman, M.A., 1999. The Quality of Ideas: Measuring Innovation with Multiple Indicators. NBER Working Paper 7345.

Le, C.N., 2004. Socioeconomic Statistics & Demographics. Asian-Nation: The Landscape of Asian America. <http://www.asian-nation.org/demographics.shtml> (accessed 22.07.04).

McKenzie, D., Rapoport, H., 2006. Can Migration Reduce Educational Attainment? Evidence from Mexico. World Bank Policy Research Working Paper 3952.

Nanda, R., Khanna, T., 2007. Diasporas and Domestic Entrepreneurs: Evidence from the Indian Software Industry. Working Paper, Harvard Business School.
Rauch, J., 1993. Productivity gains from geographic concentration of human capital: evidence from cities. Journal of Urban Economics 34, 380–400.
Romer, P., 1990. Endogenous technical change. The Journal of Political Economy 98 (5 Part 2), S71–S102.
Rosenthal, S., Strange, W., 2001. The determinants of agglomeration. Journal of Urban Economics 50, 191–229.
Rosenthal, S., Strange, W., 2003. Geography, industrial organization, and agglomeration. Review of Economics and Statistics 85 (2), 377–393.
Rosenthal, S., Strange, W., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Urban and Regional Economics, vol. 4. Elsevier, Amsterdam, pp. 2119–2172.
Rosenthal, S., Strange, W., 2008. The attenuation of human capital spillovers. Journal of Urban Economics 64, 373–389.
Saxenian, A., 2005. From brain drain to brain circulation: transnational communities and regional upgrading in India and China. Studies in Comparative International Development 20 (2), 35–61.
Saxenian, A., 2006. The New Argonauts: Regional Advantage in a Global Economy. Harvard University Press, Cambridge, MA.
Scherer, F.M., Harhoff, D., 2000. Technology policy for a world of skew-distributed outcomes. Research Policy 29 (4–5), 559–566.
Schiff, M., 2005. Brain gain: claims about its size and impact on welfare are greatly exaggerated. In: Özden, C., Schiff, M. (Eds.), International Migration, Remittances, and the Brain Drain. World Bank and Palgrave Macmillan, Washington, DC, pp. 201–226.
Sukhatme, S., Mahadevan, I., 1987. Pilot Study on Magnitude and Nature of the Brain Drain of Graduates of the Indian Institute of Technology, Bombay. Indian Institute of Technology.
Trajtenberg, M., 1990. A penny for your quotes: patent citations and the value of innovations. RAND Journal of Economics 21 (1), 172–187.
World Bank, 2008. Global Economic Prospects 2008: Technology Diffusion in the Developing World. The World Bank, Washington, DC.
Yang, D., 2006. International Migration, Remittances, and Household Investment: Evidence from Philippine Migrants' Exchange Rate Shocks. NBER Working Paper 12325.