Return migrants' self-selection: Evidence for Indian inventors

Stefano Breschi [1/§], Francesco Lissoni [1/2], Ernest Miguelez [2/3]

[1] *ICRIOS – Università "L.Bocconi", Milan*

[2] *GREThA UMR CNRS 5113 – Université de Bordeaux*

[3] *AQR-IREA – Universitat de Barcelona*

[§] contact author: stefano.breschi@unibocconi.it

This version: *28 June 2018*

Abstract

Based on an original dataset linking patent data and biographical information for a large sample of US immigrant inventors with Indian names and surnames, specialized in ICT technologies, we investigate the rate and determinants of return migration. For each individual in the dataset, we both estimate the year of entry in the United States, the likely entry channel (work or education), and the permanence spell up to either the return to India or right truncation. By means of survival analysis, we then provide exploratory estimates of the probability of return migration as a function of the conditions at migration (age, education, patenting record, migration motives, and migration cohort) as well as of some activities undertaken while abroad (education and patenting). We find both evidence of negative self-selection with respect to educational achievements in the US and of positive self-selection with respect to patenting propensity. Based on the analysis of time-dependence of the return hazard ratios, return work migrants appear to be negatively self-selected with respect to unobservable skills acquired abroad, while evidence for education migrants is less conclusive.

Keywords: immigration, innovation, inventor data, patent data

JEL codes: F22, O15, O31

ACKNOWLEDGEMENTS: We gratefully acknowledge financial support from NBER and the French National Research Agency (TKC project - reference: ANR-17-CE26-0016).

1. Introduction

Return migration represents an important share of present-day total cross-border population flows. In 2008, the OECD International Migration Outlook, based on indirect estimation methods, suggested that 20% to 50% of adult immigrants to advanced countries might leave within five years after their arrival, albeit with much variation due to heterogeneity of sending-receiving country pairs, years of entry, and the definition itself of "return migrant" (OECD, 2008).[1]

Such high rates also affect high-skilled (highly educated) migrants. Based on a large sample of foreign recipients of a US doctorate in science and engineering, Finn (2014) calculates an average return rate – five years after graduation - of about 30%, with country-specific figures ranging from less than 10% for India and China, to over 40% for Western European countries. In addition, evidence from questionnaires on return intentions suggests, for migrants to United States and Germany, a U-shaped relationship between years of schooling and return rates (Dustmann and Görlach, 2016), that is a self-selection of return migrants with respect to very low and very high educational levels. OECD (2008) estimates on actual returns conform to this pattern, especially for the United States.

High-skilled return migration is especially relevant for innovation studies. From the viewpoint of migrants' home countries, returnee scientists, engineers, and other professionals can play a role in knowledge diffusion and new business creation (on entrepreneurs: Nanda and Khanna, 2010; Filatotchev et al., 2011; Luo et al., 2013; on scientists: Kahn and MacGarvie, 2012; Jonkers and Cruz-Castro, 2013; Trippl, 2013; Gibson and McKenzie, 2014; on managers: Nanda and Khanna, 2010; Choudhury, 2016). In this respect, high-skill return migration can act as potential compensating mechanism of the "brain drain" suffered by sending countries (Dustmann et al., 2011; Gibson and McKenzie, 2011).

As for host countries, their policy-makers, higher education institutions, and knowledge-intensive firms fret not only about attracting, but also about retaining the "best and brightest" among the foreign workers and students (Hawthorne, 2018; Teitelbaum, 2014; Wadhwa et al., 2009). This begs the question of whether returnees self-select positively not only with respect to their immediately observable skills, such as their educational level, but also with respect to harder-to-observe skills, such as their inventiveness, creativity, or entrepreneurial propensity, conditional on education.

More generally, the issue of skill-based self-selection of return migrants plays a crucial role in economic theories of migration as a lifetime investment, with important implications for the expected economic and social assimilation of both permanent and temporary migrants (Borjas and Bratsberg, 1996; Dustmann and Görlach, 2016).

Despite its relevance, return migration is an understudied topic, due to lack of data. National authorities commonly register the inflows of foreign-born and foreign nationals, but not their outflows, which makes it nearly impossible to know precisely how many immigrants later leave the country, and when, let alone their individual characteristics. Quantitative research then relies on longitudinal surveys or on complex manipulation of administrative panel data (Dustmann and Görlach, 2016).

[1] In what follows, unless otherwise stated, we will adopt Dustmann's and Weiss' (2007) definition of return migrants as those who settle back in their home country by their own choice, after having spent several years abroad. This echoes the definition provided for statistical purposes by the United States Statistical Division of "persons returning to their country of citizenship after having been international migrants (whether short-term or long-term) in another country and who are intending to stay in their own country for at least a year" (UN, 1998; as quoted by OECD, 2008), but hides more complex migration patterns, such as circular and repeat migration (Constant and Zimmermann, 2016).

Most surveys, however, concern specific, often low-skilled migrant groups (such as the *gastarbaiters* of the 1960s and 1970s, in the much used German Socio-Economic Panel) and/or focus on labour market determinants of return migration, such as unemployment (Bijwaard et al., 2014). Notable but rare exceptions concern academic scientists, whose return rates and individual characteristics can be obtained by combining archival and bibliometric data sources, as in Gaulé (2014) and Kahn and McGarvie (2012).

In a recent assessment of the emerging literature on migration and innovation, Kerr (2017) states that we know very little about return migration of workers engaged in innovation and entrepreneurship, except that it is rapidly growing in importance, and that "clever data work to […] quantify [it] would be most welcome" (Kerr, 2017; p.212). This paper answers the call. Based on an ambitious data-linkage project joining patent data and inventors' biographical information from a web-based, professionally oriented social network, we build a large sample of United States immigrant inventors of Indian origin, specialized in ICT technologies. This is a social group that both figures prominently in the recent debate on temporary work migration to the United States (most notably, on the use of H-1B visas; Kerr and Lincoln, 2010) and contributes significantly to international student mobility (OECD, 2017).

Our data-mining strategy allows us to identify only migrants entering the United States via work and education channels, most likely associated to temporary visas. Yet, we do not consider it a weak point, due to two well-established stylized facts:

1) The overwhelming importance of temporary channels as a source for high-skilled immigration into the United States, via the transformation of both temporary work and student visas into permanent ones (in contrast with countries such as Australia and Canada, where permanent visas for the highly skilled are more easily obtained upon entry; Koslowski, 2018).

2) The remarkable innovation impact of migrant scientists and engineers entering the United States with work and student visas, as opposed to those entering through the channel of family reunions, as documented by Hunt (2013).

While subject to a number of limitations, our dataset allows us to trace return migration from the United States with a degree of precision comparable to survey data, but on a much larger scale and with original information on its possible determinants. For each individual in the dataset, we both estimate the year of entry, the likely entry channel (work or education), and the permanence spell up either to the return to India or to 2016 (right-censoring year). By means of survival analysis, we provide estimates of the probability of return migration as a function of the conditions at migration (age, education, patenting record, migration motives, and migration cohort) as well as of some activities undertaken while abroad (education and patenting).

Our results, albeit exploratory, find rather different patterns for work and education migrants. Considering the former, we find that the Indian inventors' return risk is positively associated to their age and education at migration as well as to their propensity to patent while in the US. As for education migrants, the return risk correlates negatively with the education level they attain. We also find some evidence of negative (positive) time-dependence for work (education) return migrants, which we interpret as indicative of negative (positive) self-selection with respect to unobservable skills acquired in the host country.

We proceed as follows. In section 2, we present in a rather succinct way our database building strategy (more details in the Appendix), introduce our own definitions of migrant and return migrant, and propose some descriptive evidence. When necessary, we discuss some conceptual and

methodological issues concerning the definition of return migrant. In section 3 we present our model specification and discuss how it serves the purpose of investigating skill-based self-selection in return migration. In section 4, we perform the related econometric exercise, and comment the results. Section 5 concludes, with special focus on further research plans and some tentative policy implications.

2. Data: methodology and descriptive statistics

2.1 Methodology

Our dataset originates from an ambitious data-linkage project between patent and inventor data gathered from Patentsview (http://www.patentsview.org/web/) and biographical information extracted from a large number of LinkedIn profiles. Patentsview is a data repository recently made available by the United States Patent & Trademark Office (USPTO), which provides, among other things, disambiguated data on all the inventors of USPTO granted patents from 1975 onward, irrespective of their country of residence. LinkedIn, a well-known professional-oriented social network, represents an unparalleled source of information on the international mobility of individuals, as the members' public profiles include information on names and (possibly) locations of their education institutions and employers, along with graduation and recruitment years (Ge et al., 2016; Zagheni and Weber, 2015).

As a pilot project, we focused on a subset of high-skilled migrants in the United States, namely Indian inventors with ICT patents. This is a distinctive social group, due both to its inventive contribution (Kerr and Lincoln, 2010; Breschi et al., 2017) and to its implication in two important temporary migration channels, namely highly qualified temporary work (most notably, through the H1-B visa system; Kerr et al., 2015; Kapur and McHale, 2005) and education (Finn, 2014; Kapur and McHale, 2005). It is also a highly represented group on LinkedIn, which in 2016 registered well over 100 million members in the US and over 30 millions in India, with the two countries standing at the top of LinkedIn world rankings for both membership and traffic.[2]

We extracted from PatentViews all the patents granted to the 179 largest US public firms in the ICT industry, from 1975 to 2016, and the relative inventors, for a total of 262,847 distinct individuals.[3] We then proceeded to the ethnic analysis of such inventors' names and surnames, based on Global Name Recognition, a name search technology produced by IBM (from now on, IBM-GNR) and adapted to our purposes by Breschi et al. (2017). This allowed us to identify inventors of presumed Indian origin (from now on, Indian inventors), for a total of 24,017 individuals, representing 9.1% of all inventors employed by the companies in our sample. Each Indian-named inventor was then matched to one LinkedIn profile, based on rule-based name- and company-matching, with extensive manual checking. This exercise yielded 10,839 inventors with a valid LinkedIn account (around 45% of the original sample). For details, see sections C. and D. of the Appendix.

We then proceeded to codify three major sets of variables, respectively concerning education, employment and patent records. On that basis, we also estimated the inventors' year of birth as well as their migrant, non-migrant and return migrant status.

[2] Unofficial statistics from: https://www.statista.com/statistics/272783/linkedins-membership-worldwide-by-country/ (last visited: April, 2018)

[3] The definition of ICT industry follows the one provided by the OECD (https://www.oecd.org/sti/ieconomy/1835738.pdf). More details in section B. of the Appendix

We coded information on education according to the 2011 version of Unesco's International Standard Classification of Education (ISCED), for the educational levels from 3 (Upper Secondary) to 8 (Doctoral or equivalent).[4] After treating jointly ISCED levels 5 and 6 (respectively: Short-cycle Tertiary and Bachelor), and distinguishing between Masters of Arts and/or Science and MBAs, we ended up with the following classification: Upper Secondary Education, Bachelor, Master, MBA and PhD, plus a residual Unclassified category. We then geo-localized as many education institutions as possible at the country level by means of Google Maps, and obtained at least one geo-localization per inventor. (For full details, see section E of the Appendix).

As for employment, we recorded the start and end years of each related employment spell, as well as the employer's name. We geo-localized the latter, at the country level, only on the basis of the information provided by the LinkedIn profile, with no further attempt to use GoogleMaps, which would prove useless for multinationals with several branches and affiliates in multiple countries. Thus, our estimates on migration and return migration for work reasons have to be considered extremely conservative. In section F of the Appendix we discuss some possible ways to improve them, by capturing more return moves, based on a more sophisticated treatment of LinkedIn information.

As for the inventive activities of each inventor, we geo-localized them at the country level on the basis of the inventor's address as reported on his/her various patents, and dated them on the basis of the patent's priority year (De Rassenfosse et al., 2013).[5] Based on the unique inventor ID provided by PatentViews, we could then calculate the number of patents signed by each inventor on each year, either in India or abroad.

Coming to the inventor's year of birth, our preferred option was to estimate it on the basis of education information, with reference to the lowest-level education achievement among those reported in the LinkedIn profile, its year of completion, and the presumed age at start (see section G in the Appendix; see also Gaulé, 2014). For the inventors whose profile did not report any information on the timing or level of education, we estimated the year of birth based on the average age of the other inventors in the same patent cohort (that is, the inventors who filed their first patent in the same year). In most cases, the age so calculated is around 32, which is close to general estimates by Jones (2009).

After dropping the inventors whose LinkedIn profiles did not provide sufficient information for estimating neither the educational level nor year of birth, we remained with 8,982 observations (Table A5 in the Appendix). For these, we estimated the accuracy of our Patentsview-LinkedIn match based on around 1,000 LinkedIn profiles of Indian ICT professionals that report patent information. Based on around 800 "true positives" (successful matches of a LinkedIn profile to an inventor in Patentsview, with coherent patent information) and 30 "false positives" (successful matches, but with discordant patent information), we calculated a 96.4% precision rate and a 77% recall rate. The high precision suggests that the education, employment, and age information in our dataset are rather accurate (that is, it is unlikely that they refer to the wrong inventor). However, the low recall rate suggests that our sample possibly suffers of truncation problems, to the extent that the

[4] See: http://ec.europa.eu/eurostat/statistics-explained/index.php/International_Standard_Classification_of_Education_(ISCED) (last visited: March, 2018)

[5] We obtain the priority year of the patent from its priority date, namely the date of filing of the first USPTO application or, in case of patents extended to the United States but first filed abroad, the first application worldwide.

excluded inventors may share some characteristics associated to the phenomenon of our interest (return migration).[6]

We finally proceeded to identify migrant and return migrant inventors to/from the United States, as described in detail by Figures A6 and A7 in the Appendix. We proceeded by elimination, first dropping from our sample all the inventors without any education, employment or patenting record within India, who may include second generation migrants or members of the historical Indian diaspora in the former British Commonwealth. Second, we dropped all those we consider non-migrants, namely the inventors without any education, employment or patenting record outside India. Among the remaining inventors, we considered as "education migrants" all the inventors whose LinkedIn profile reports the enrolment in a foreign higher education institution as the first event in their life taking place outside India, and this occurs earlier than any patenting activity abroad. Similarly, we considered as "work migrants" all the inventors whose LinkedIn profile reports the recruitment by a foreign-based company as the first event taking place outside India and/or who have at least one patent abroad and this dates back to before any enrolment in a foreign higher education institution.

Finally, we restricted our attention to migrants whose first move outside India occurred in the United States. This left us with 3,943 "education migrants" and 1,589 "work migrants" from India to the United States, for a total of 5,532 individuals. For the former, we considered as migration year the starting year of the first education programme undertaken in the United States. For the latter, we similarly defined the migration year as the beginning year of the first working spell in the United States or, alternatively, the priority year of the first patent. When distinguishing between "work" and "education" migrant inventors it is important to keep in mind that the distinction refers only to the individuals' condition at migration time. Nothing impedes that a work migrant will at some point enter a Master or PhD programme in the United States, or that an education migrant will start working there. Indeed, the first case is rather frequent, and the second is very frequent.

Coming to return migration, we record as a returnee every migrant reporting an employment or a patent in India after having moved to the United States. We do not record return events related to further education in India, but we suspect these to be very few. However, we record employment in higher education. As for the return year, this coincides either with the start of the relevant employment spell or the priority year of the relevant patent. All migrants for whom we do not observe any return event are considered as still living in the United States in 2016, our final year of observation. For sake of simplicity, at this stage of our research, we do not code any event following the first return to India. Similarly, we ignore any move from the United States to another country, different from India. For example, we will treat an Indian student in the United States who leaves for the United Kingdom after graduation as if he/she was staying in the United States. This implies that

[6] In section I of the Appendix, we further investigate the properties of our sample of 8,982 inventors. We first compare their patent records to those of other Indian-named inventors in our initial dataset, and find no significant differences for what concerns the average number of patents granted, conditional on the year of the first patent. However, based on the year of the first patent, inventors with a LinkedIn profile appear to be younger than those without it. Second, we compare the inventors for whom we found a LinkedIn profile (whether complete or not) to all others, and find that inventors who patent exclusively in India have a significantly higher probability of being matched with a LinkedIn profile than inventors who patent exclusively in the US or both in India and in the US. These diagnostics suggest that, based on our data, we may possibly underestimate migration from India, especially for more recent calendar years, due to the relative over-representation of India-based inventors versus US-only- and US-plus-India-based ones, the former group being more likely to contain non-migrants and the latter more likely to contain migrants. Reasoning along similar lines, we may risk over-estimating return migration, since the propensity to have a LinkedIn profile is higher for US-plus-India-based inventors than for US-only-based inventors. More generally, the younger the inventors, the more representative our sample.

we ignore circular migration. A cursory look at our data, however, reveal only very few instances of this type.

Albeit imperfect, our coding of return events (and, in consequence, permanence abroad) does not compare unfavourably with similar coding one can find in the literature. Borjas' (1989) classic study based on the 1972-1978 Survey of Scientists and Engineers simply recorded as returnees all foreign respondents to the 1972 questionnaire who had left the sample by 1978. Gaulé (2014), who relied on the several editions of Directory of Graduate Research of the American Chemical Society from 1993 to 2007, first identified as potential returnees all foreign faculty and postdocs who appear at least once in the Directory and then disappear. He then looked manually in bibliographic and web resources for information on the likely motives for the disappearance (so to distinguish between return to the home country, while not ceasing the academic career, and moves to industry, third countries and deaths). To our knowledge, the only accurate survey of return moves is provided by Gibson and McKenzie (2014), but for a very small sample.[7]

Even much used resources for studying low-skill return emigration, such as the German Socio-Economic Panel (GSOEP), are far from faultless or resort to measuring return intentions, rather than actual moves.[8]

2.2 Descriptive statistics

In what follows we produce a number of descriptive statistics that serve the double purpose of checking the information contents and quality of our data and of providing some basic evidence on the phenomenon under study.

Figures 1a,b report the distribution of the age at migration respectively for education and work education migrants. We notice that the overwhelming majority of the former move to the US at 23 or 24, which is compatible with the age for starting a Master course or possibly a PhD. The very sparse observations for ages less than 19 are either due to errors in our calculation of migrants' year of birth or to the very few Indian migrants who move to the US for the Bachelor studies. As for those very few apparently moving at older ages, especially over 30, they may be mature postgraduate students or professionals taking MBAs courses.

As for work migrants, the figure 1b shows a high peak at 32, which is a statistical artefact that results from the inclusion in this category of migrants of many inventors with two characteristics. First, for want of better information, we estimate their age based on the priority year of their first patent. Second, they appear on such patent with a US address and this is the earliest evidence we have of their migration move. Yet, we notice that the age distribution is rather symmetric around 32. This is compatible with migrants in this group moving abroad after completing their education in India and

[7] Gibson and McKenzie (2014) survey around 800 high-achieving secondary school graduates from New Zealand, Tonga and Papua New Guinea, 200 of whom undertook academic careers. In this subgroup, 78% moved abroad, with a 25-30% return rate.

[8] As explained by Bönisch et al. (2013), the basic information on return migration provided by GSOEP consists in nonresponse items accompanied by the "moved abroad" motivation. This amounts to under-reporting, as observed by Constant and Massey (2002), who find that a much larger number of individuals in the panel leave it for one or more years, without providing a motivation explicitly related to a move back home; and hence resort to code as returnees all absentees for three or more years. Kirdar (2009) reports similar problems for more recent issues of the survey. As many surveys of low-skilled migrants, the GSOEP collects information on return intentions. Similar information for the highly-skilled in collected by Baruffaldi and Landoni (2012). While useful for testing theoretical models of temporary migration, return intentions may be different from *de facto* choices. For example, the 2000-2013 trends for return migration and return intentions calculated by Finn (2014), for a longitudinal cross-section of foreign doctoral graduates in the US, are markedly different.

starting their careers there, as it happens with many H1-B visa holders, as well as with being an employee of an Indian firm, temporarily detached to the United States. We also notice that, when excluding from the work migrants all inventors whose ages was determined by the year of the first patent, the shape of the distribution does not change much, since the modal value remains at 32 and with the symmetry is preserved (figure A13 in the Appendix).

Figure 1a. Estimated age at migration, education channel (percentage distribution of all education migrants to the United States)

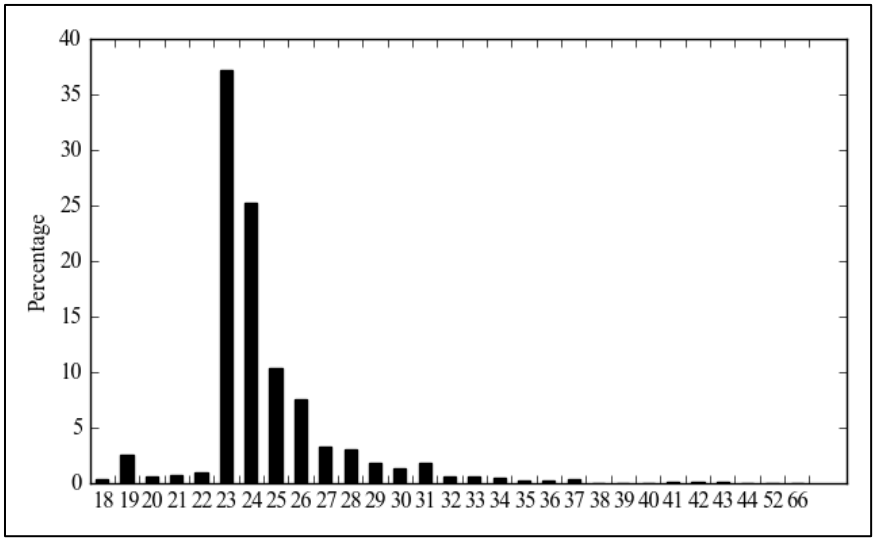


Figure 1b. Estimated age at migration, work migrants (percentage distribution of work migrants to the United States)

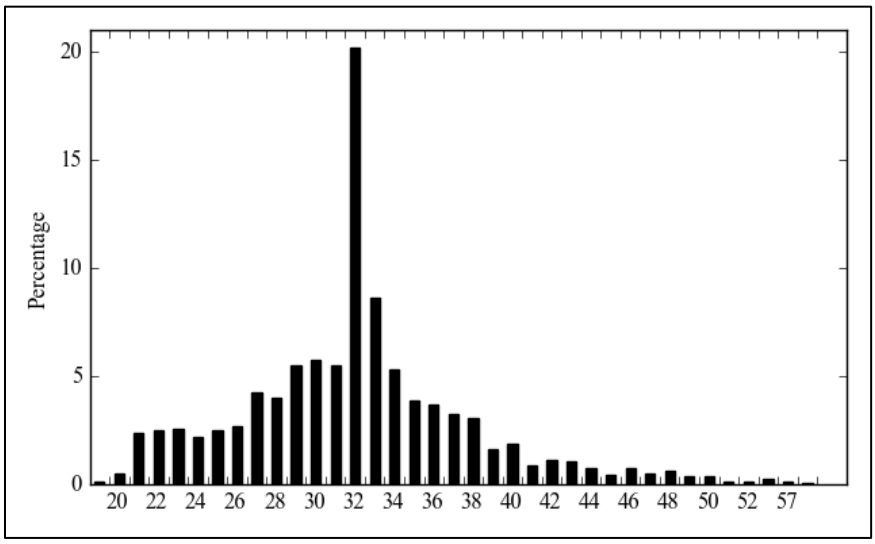


Table 1 provides a breakdown of our dataset by migration motives and cohorts (decades during which migration occurred). Two features emerge. First, most migrants in our sample belong to the 1990s and 2000s cohorts. This is broadly compatible with historical records of high-skilled Indian migration to the United States (Desai et al., 2005), but also possibly emphasized by the

characteristics of our LinkedIn records, namely right truncation at 2016 and under-reporting for the earlier cohorts (the older an individual, the less likely he/she is to maintain a LinkedIn profile).

Second, the importance of the education channel relative to the work one is both evident for early cohorts and declining over time. This trend again is broadly compatible with the history of graduate and post-graduate education in India since the 1960s, whose offer and quality was extremely limited until the 1990s (so that the early Indian migrants seeking a job in science or engineering usually got their graduate education in the host country; Kapur, 2010). But it may be accentuated, once again, by under-reporting for early cohorts and its correlation with educational levels (the more likely an individual is to have migrated through the work channel, which is associated to lower education levels, the less likely he/she is to maintain a LinkedIn profile, especially for older individuals). These observations suggest that our data are more reliable for the 1990s and 2000s cohorts, which concern 4,362 individuals, namely 79% of migrants in our database.

Table 1: Migrants to the United States by cohort and channel

Channel	1960	1970	1980	1990	2000	2010	Total
Education	19	102	697	1739	1315	71	3943
% column	*100*	*95.3*	*95.2*	*85.9*	*56.3*	*22.8*	*71.3*
Work	0	5	35	286	1022	241	1589
% column	*0.0*	*4.7*	*4.8*	*14.1*	*43.7*	*77.2*	*28.7*
All channels	19	107	732	2025	2337	312	5532
% column	*100*	*100*	*100*	*100*	*100*	*100*	*100*
% row	*0.3*	*1.9*	*13.2*	*36.6*	*42.2*	*5.6*	*100*

Figures 2a,b provide further details on the education levels of both education and work migrants. We first remark how the overwhelming majority of the former and the relative majority of the latter hold a Master degree. This suggests that PhD holders and academic scientists, for which Finn (2014), Gaulé (2014) as well as Kahn and McGarvie (2012) have provided some evidence, are not a representative sample of migrant inventors in the ICT industry. We also notice that the share of Doctorate holders is higher for education-based migrants, while the share of Bachelor holders is higher for work-based ones, which is in line with our selection criteria for the two categories.

Figure 3 reports the total return rates for all migrants in our sample (irrespective of length of stay), by migration channel. For comparative purposes, the return rates are calculated both according to the definition of returnee we adopted above (first job or patent back in India, as per LinkedIn profile) and to a purely patent-based definition (first patent back in India, irrespective of other information). The latter corresponds to what found in most of the available literature on the international mobility of inventors, which relies exclusively on patent data and can observe a cross-border move only for inventors with at least two patents, in as many different countries (for example: Oettl and Agrawal, 2008). We notice immediately that this definition severely underestimates return rates (black bars), compared to the one based also on job information (grey bars), whatever migration channel we consider. In fact, the latter include among the returnees also the inventors with no more than one patent in their career (either in the United States or in India) but education or employment in a different country than the one where that only patent was signed. More generally, it also counts as returnees the inventors whose entire patent production occurred in one country, but whose education or career took place also elsewhere.

When comparing migration channels, figure 3 reports a 7-point difference in the return rate of work-migrant inventors compared to education ones. This may be due to the different types of visas used to enter the United States, both in terms of initial validity length and renewal ease, but also to different efforts that work and education migrant may make to convert their temporary visas into permanent ones. Different type of migrants may also be differently exposed to the opportunity of establishing social ties in the United States, which may influence their propensity to return at each point in time.

Figure 2a. Highest educational attainment, percentage distribution - Education migrants

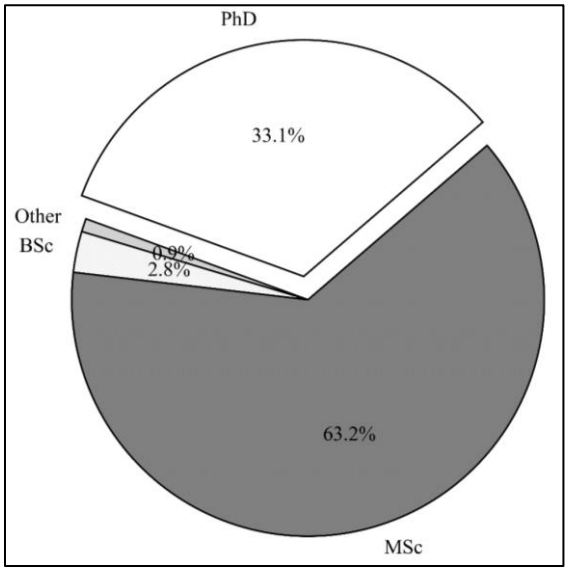


Figure 2b: Highest educational attainment, percentage distribution of work migrants

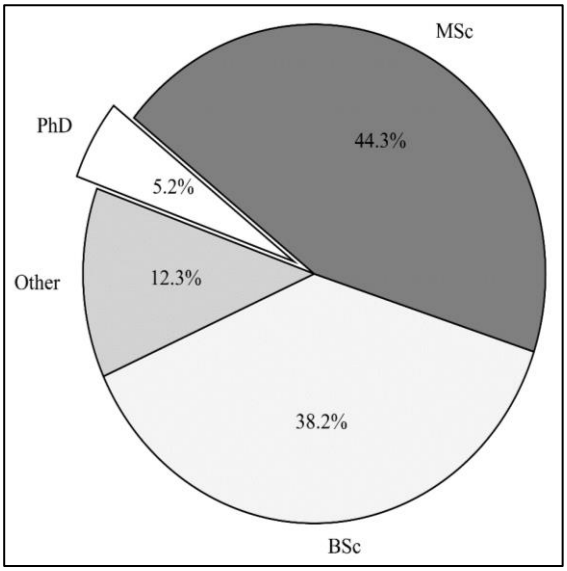
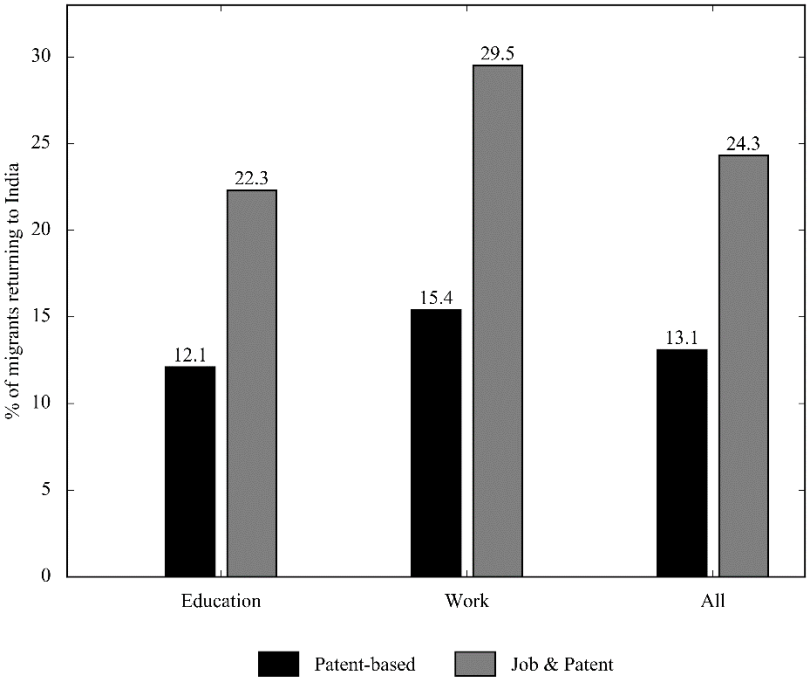


Figure 3: Total return rates by migration channel (irrespective of the length of stay in the United States)



Figures 4a,b report the total return rates (based on both patent and job information) for different cohorts of migration to the United States. The return rates for education migrants appear to be increasing, and this despite the longer observation interval for older cohorts (which intuitively should lead to more accumulation of returns). However, for cohorts before 1990, the number of observations is rather limited and, as discussed in the previous subsection, the probability of under-reporting by return migrants rather high. As for the 2010 cohort, once again we are faced with very few observations, which makes the very high return rate figure extremely unreliable. Once again, we can trust only the data for the 1990 and 2000 cohorts, which still exhibit different return rates.

Contrary to education migrants, the return rates of work migrants appear rather stable, especially for recent cohorts.

Figure 4a: Percentage of education migrants returning to India, by cohort (irrespective of the length of stay)

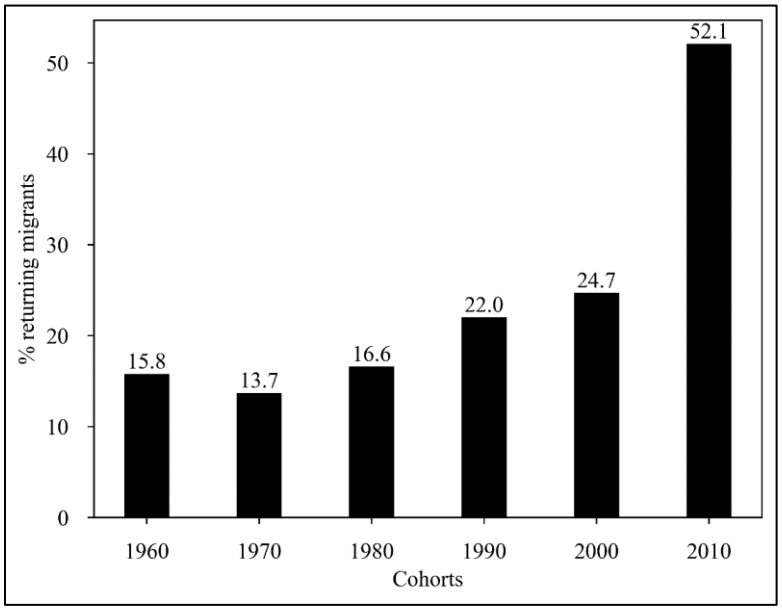


Figure 4b: Percentage of work migrants s returning to India, by cohort (irrespective of length of stay)

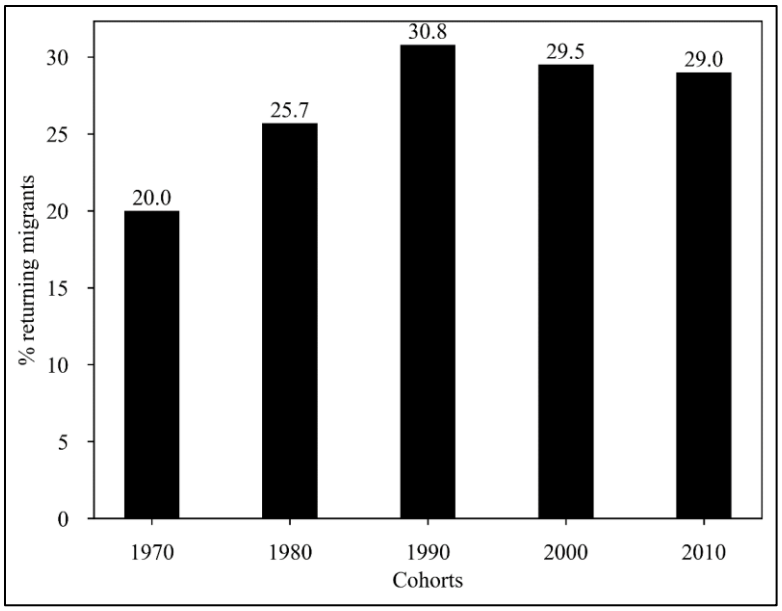


Figure 5 reports the Kaplan-Meier estimators for work and education migrants from the 1990 and 2000 cohorts, with time measured yearly. We notice that the survival (stay) rate for work migrants is both lower and more rapidly decreasing over time than for education ones. We also notice that the stay rate after 10 years since migration for education migrants (slightly less than 90%) is very close to what reported by Finn's (2014) for Indian PhD graduates in the United States. We take it as a sign of the reliability of our data.

Figure 5. Stay rates over time (years since migration), by migration channels (1990 and 2000 cohorts)

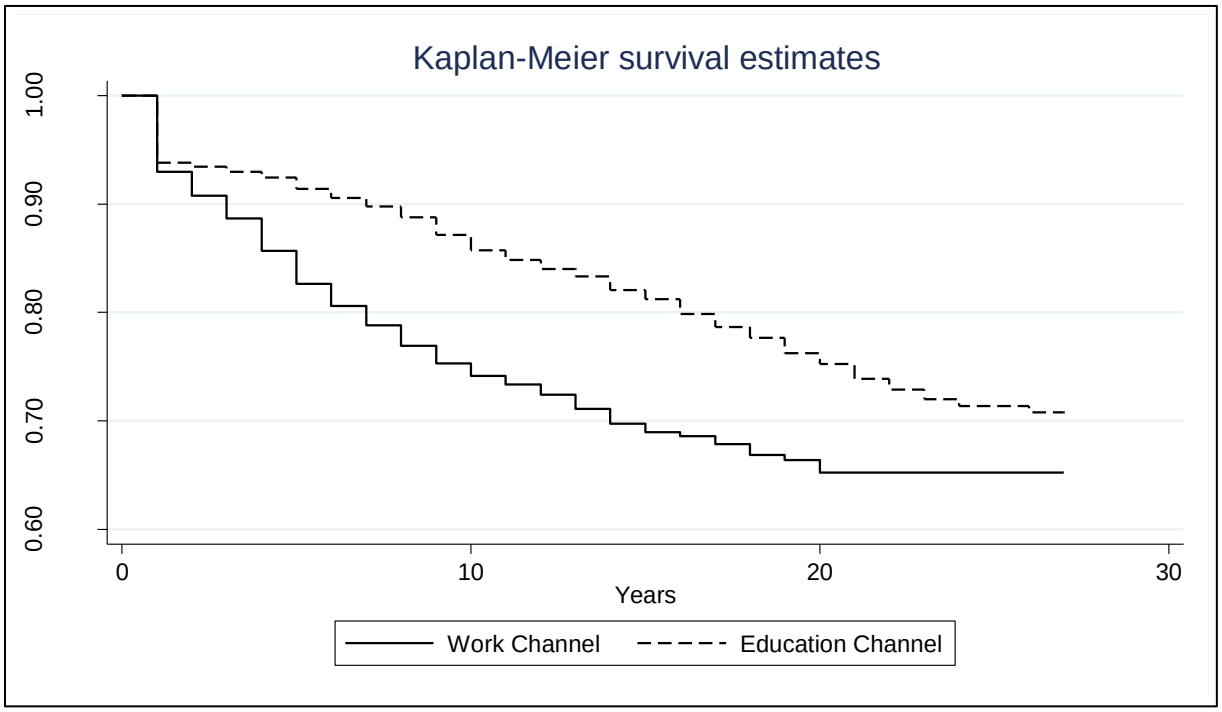


Table 2 provides detailed information on the return time for migrant inventors in the 1990 and 2000 cohorts. Returnees in the first cohort leave the United States, on average, 11 years after their arrival. The minimal return time is zero (which implies a return to India less than a year after entry in the United States) and the value of the first quartile is 5.5. This indicates that 25% of the returnees in the 1990 cohort go back to India either on the same year of their arrival or not later than five and a half years here after it. An additional 25% leave in between 5.5 and 11 years after their arrival, followed by 25% more who leave in between 11 and 16 years. The maximum stay, for returnees, is 25 years. When splitting the 1990 cohort between work and education returnee migrants, the former exhibit shorter stay period, both on average and according to the quartile distribution. As for the 2000 cohort, this exhibits on average shorter stays than the 1990 one (which may be due to shorter exposure to the return risk), but also less striking differences between work and education migrants.

Table 2: Average time to return by cohort

	All channels	Education	Work
		Cohort 1990	
# of inventors in cohort	2025	1739	286
# of returnees	471	383	88
mean	10.58811	11.22193	7.829545
std	6.643841	6.71642	5.560992
min	0	0	0
25%	5.5	7	3
50%	11	12	8
75%	16	16	13
max	25	25	19
		t-test 4.952 (p-value 0.000)	
		Cohort 2000	
# of inventors in cohort	2337	1315	1022
# of returnees	626	325	301
mean	3.889776	4.132308	3.627907
std	3.897362	4.316674	3.374574
min	0	0	0
25%	0	0	0
50%	3	3	3
75%	7	8	6
max	16	16	15
		t-test 1.635 (p-value 0.103)	

Note: cohort 1990 includes inventors who migrated to the US between 1990 and 1999; cohort includes inventors who migrated to the United States between 2000 and 2009.

3. Specification

We exploit our data to explore the extent of skill-based self-selection in return migration of the highly-skilled. Skill-based self-selection was first investigated by Borjas (1989) in order to provide an explanation for two common stylized facts concerning the education and income levels of migrants. First, stock data on foreign-born vs native populations recurrently show that the former are, on average, better educated than the latter, for most of traditional destination countries. Second, when observing a cohort of foreign-born over time through successive censuses, it is often found that, starting from a lower average wage or income level, migrants catch up relatively quickly. Regardless of whether migrants are positively self-selected at entry, with respect to their education and/or unobservable skills, negative self-selection may contribute to explain this evidence, to the extent that return migrants escape successive censuses, therefore leaving behind them, in the host country, only the best and brightest of their respective immigration cohorts.

Borjas' and Bratsberg's (1996) provide a classic treatment of the topic, in which they show that different remuneration levels of skills in the host and home countries jointly determine whether migrants will be positively (negatively) self-selected upon arrival and, conversely, negatively (positively) selected upon return. In other words, return migration is expected to reinforce the sign of skill-based self-selection at entry. Dustmann and Görlach (2016) provide the last in a series of refinements of this basic idea, which describes the migrant's behaviour at destination (including his/her investment in the acquisition of education and skills) as resulting from the same lifetime optimization plan that determines the return decision and timing.

Other, less dominant theories of return migration stress the fact that many migrants neither move permanently to the host country nor return home once and for all after a prolonged spell abroad. Instead, they move back and forth the home and the host country (or several host countries), possibly in response of economic shocks (Constant et al., 2013). In this case, we should not expect any positive or negative self-selection, the economic shocks being orthogonal to skill levels.

Empirical studies on return migration can be categorized according to two criteria: i) whether they observe and explain the actual duration of migration spells, from entry to return, or simply compare the characteristics of stayers and returnees; ii) whether they focus on observed return moves or on return intentions.

With respect to i), empirical studies fall in one or the other category depending on data availability and, to a lesser extent, on their theoretical focus. On the data side, most studies simply do not have longitudinal information on individual migrants, that is that have no records on the entry and return dates. Based on this limited information, they can only apply linear probability or logit/probit models, and investigate the determinants of the probability to return, irrespective of when this occurs. When longitudinal data is available, instead, one can apply duration analysis (also known as survival or event history analysis; Allison, 2014). This has two advantages over linear probability or logit/probit models. First, it is not inherently static and therefore it allows considering time-varying covariates, so to study how intervening changes in the migrant's characteristics may affect the return decision. Second, and more importantly, duration analysis allows estimating the probability to return conditional on not having yet done so at a specific point in time (return hazard). By derivation, one can explain or predict the timing of the return decision, and not just the probability of its occurrence. This also implies that, by means of duration analysis, we can test whether the probability to return is time-dependent, either positively or negatively. According to Constant and Massey (2002), negative time dependence may be indicative of negative skill-based self-selection (where skills are unobservable). The longer a migrant stays in the host country, the more country-specific skills he/she accumulates, which are hard to transfer and/or are less remunerated at home, *ceteris paribus*. This makes return increasingly less likely. At the same time, to the extent that migrants vary in the speed at which they accumulate the local skills, early returnees would necessarily be those who, at a given point in time, have accumulated fewer local skills.

Coming to the distinction between studies based on observed return moves or declared return intentions, this often boils down, once again, to data availability, with survey data being much better at recording the latter than the former (see above our discussion on how we record return moves). However, some recent literature suggests that data on return intentions serve better the purpose of testing lifetime income maximization models. This is because, according to such theories, the actual timing of return is decided contextually to the education and skill investments, both of them depending on return intentions.

The data structure for our regression exercises is a panel one, with each inventor *i* being observed repeatedly since his/her immigration year until the minimum between his/her return year (when he/she exits the panel) and 2016, our last observation year. In this way, we have a large number of right-censored observations, but no left-censored ones. In what follows we exploit this feature of our data and estimate the determinants of actual return decisions by means of discrete time duration analysis. Given the exploratory nature of our exercise, we do not put forward any claim of having established causal links. We care instead for producing much needed evidence on return frequency and timing, and its association to observable and unobservable skills (that is, self-selection based on education, patenting activity, and time spent in the United States).

Following Jenkins (2005), we assume a proportional hazard function, which, in a discrete time setting as ours, results in a complementary log-log (cloglog) model, as follows:

$$h(t,x)_i = 1 - \exp[-\exp(c(t) + \beta_i X_i)]$$

where $c(t)$ represents a generic inventor's baseline probability to return home after a migration spell t (duration), conditional on not having yet returned, and $\beta_i X_i$ is a scaling factor depending on specific inventor *i*'s characteristics X_{it} (some of which are time-variant). As for *t*, we measure it either as the number of years (plus 1) spent in the US since immigration or, for conducting robustness checks on education migrants only, since the end of their first education spell in the United States.

Concerning the baseline hazard ratio $c(t)$, we adopt two alternative specifications. First we follow Constant and Massey (2002) enter t with a quadratic term, as follows:

$$c(t) = \alpha_1 t + \alpha_2 t^2 \qquad (1)$$

This parametric specification may allow us to test for any time dependence of the hazard ratio, and its sign, in a rather immediate and intuitive way, on the basis of estimates for α_1 and α_2. But it comes at the cost of imposing a specific functional form to $c(t)$.

Second, we experiment with a non-parametric specification (as in Gaulé, 2014) and make use of fixed effects, as follows:

$$c(t) = \eta_1 t_1 + \cdots + \eta_N t_N \qquad (2)$$

where $(t_1 \ldots t_N)$ is a set of duration dummies, corresponding to migration spells lasting from 1 to N years (and N is the longest spell observed in our data). This model has the advantage of not imposing any functional form to the hazard ratio, but it produces so many estimated coefficients that, in order to appreciate any time dependence of the hazard ration, one needs a graphical representation.

Based on the evidence from figures 4 to 6, plus table 2, in the previous section, we expect time to affect differently the hazard ratio of work- and education-based migrant inventors. Hence, we run separate regressions for the two type of migrant inventors. We also restrict our regressions to the two most populated migration cohorts in our sample, namely the 1990s and the 2000s ones, for which data are more reliable. We also right-censor our data at 2016 for a matter of convenience. This makes the longest possible duration equal to 27 years.

Coming to our choice of regressors X_i, they include both a set of time-invariant variables that describe the migrant's conditions at entry in the United States, and a set of time variant ones that describe his/her activities during his/her permanence there (see table 3 for descriptive statistics).

As for conditions at entry we consider the inventor's age, educational level, migration cohort and patenting experience at migration, both of which we expect to be positively associated to the return hazard, as they may proxy for the inventor's stronger attachment or professional insertion in India and may affect negatively his/her chance to renew the initial temporary visa. We measure age in years (*Age at migration*) and education with the dummy variable *Master or more at migration* (the reference case being given by migrants with no more than a Bachelor at migration; as for Doctorate holders, they are too few to create a meaningful separate category, so we treat them as Master holders). Due to our restriction of the analysis to just two migration cohorts, we control for them with just a dummy for the 2000s one (1990s as reference). As for patenting experience, we measure it with the cumulative number of patents signed at the time of migration (*Patent stock at migration*).

As for activities in the United States we consider:

- the migrant's student status (*Student*), which is a dummy taking value one for all the years comprised between the start and end years of an education spell in the United States, whatever its level, and zero otherwise;
- the migrant's educational attainment while in the United States, as measured by the dummy variables *Master* and *PhD*, which takes value zero before the year of completion of, respectively, the migrant's master or doctoral studies, and one thereafter;
- the migrant's productivity as an inventor while abroad, which we measure with the *cumulative number of patents* from entry into the United States up to observation time *t*.

We expect the student status to lower return hazard, as it guarantees the migrant the renewal of his/her temporary visas. As for the educational attainment, based on the existing evidence of Indian graduates' low return rates, we also expect a negative impact on the return hazard. In other words, we expect negative self-selection based on education. As for the number of patents filed in the United States, we would expect negative self-selection, but the interpretation of this variable is complicated by the fact that not all migrants in our sample, once in the United States, pursue an inventor career, but may move on to management, entrepreneurship or academia. (We come back on this issue when commenting the results).

4. Results

Table 3 reports separate descriptive statistics for the education and work migration channels. We notice some important differences between education and work migrants, besides the age at migration.

Table 3. Descriptive statistics, by migration channel

	Education channel					Work channel				
	Obs	*Mean*	*Std. Dev.*	*Min*	*Max*	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
Migration cohort	50211	1993.1	4.630	1990	2000	15333	1996.8	4.648	1990	2000
Age at migration	50211	24.32	2.652	18	52	15333	31.87	5.956	18	62
Master or more at migration	50211	0.09	0.283	0	1	15333	0.34	0.473	0	1
Current student status	50211	0.20	0.403	0	1	15333	0.04	0.192	0	1
Master in the US	50211	0.66	0.474	0	1	15333	0.04	0.202	0	1
PhD in the US	50211	0.20	0.400	0	1	15333	0.01	0.097	0	1
MBA in the US	50211	0.08	0.267	0	1	15333	0.04	0.201	0	1
Patents at migration	50211	0.01	0.114	0	5	15333	0.03	0.354	0	12
Cumulative # patents US	50211	3.83	10.64	0	261	15333	4.71	9.07	0	162

First, work migrants are considerably more likely to leave India after graduating at the master level, which most education migrants move to the United States precisely for getting that same degree. As for getting the PhD, this happens almost exclusively to education migrants. In this respect, it is

important to remark that this may happen on top of getting a Master but also an alternative to it, with the latter case being the most frequent.[9]

Both education and work migrants exhibit a rather low average number of patents before moving to the United States, but the figures are higher for the latter. At a closer inspection, our data reveal that most migrants in our sample leave India without having filed any patent there. In fact, only about 1% of education migrants and 4% of work migrants have a non-null patent record before migrating. As for the cumulative number of patents filed while in the United States, its average value is higher for work migrants than for education ones (around 5 against 4). When looking at the underlying distribution (unreported in the table), we notice that only 2% of work migrants never files any patent while in the United States, while the same figure for education migrants amounts to 14% (the overwhelming majority of these individuals patent only when they go back to India, while a tiny minority may have patent before migrating). As for those who filed at least one patent in United States, the differences between work and education migrants are much less striking, albeit education migrants exhibit more variability (witness the standard error reported in table 3). In both subsamples over a third of migrants file just one patent while in the United States and as many file from two to five (followed by a very long tail for values higher than ten), but education migrants are slightly more likely to file just one patent, or two-to-five, as well as more than one hundred.

We notice an important difference between education and work migrants with respect to the number of patents filed while in the United States, which on average is higher for the latter. As for the very high maxima that we observe for this variable, they correspond to very senior principal scientists in large ICT companies.[10]

Table 4 reports the results of our regressions, which we run separately for education and work migrants. The first two columns refer to parametric specification (1) of the baseline hazard ratio *c(t)*, while the other two refer to the non-parametric specification (2). In both cases, we calculate the estimated odds ratios, which we read as the marginal effects of the covariates on the return hazard ratio (Jenkins, 2005).

We first ask to what extent return migrants appear to be self-selected with respect to either their observable skills, such as education and patenting activity. We then move on to analyse the sign of time dependence of the hazard ratio.

Concerning education, we first notice that the odds ratio for *Master or more at migration* is greater than one in all columns of table 4, but it is significant in only one case (for education migrants in column 1). Hence, there is evidence of return migrants being positive selected with respect to education they got in India, but it is rather weak. On the contrary, all return migrants appear to be negatively selected with respect to the education obtained in the United States. For education migrants, both *Master in the US* and *PhD in the US* have estimated odds ratios largely inferior to one (the reference case being migrants obtaining only a Bachelor degree, or not completing their graduate studies).

However, the difference between the underlying coefficients is non-significant, which suggests that, for individuals holding either a Master or a PhD, graduate education is all that matters, and more advanced or research-oriented degrees do not convey any particular advantage to migrants

[9] It is very likely, however, that we largely over-estimate the number of PhD holders without a Master. This due by many LinkedIn members to report only their highest educational achievements (such as a Doctorate), and not the previous ones (such as a Master).

[10] These are the cases, respectively, of education migrant Durga Malladi of Qualcomm (261 patents) and work migrant Alok Srivastava, an independent consultant with activities both in India and the United States (162 patents).

intending to stay in the United States, nor to those with return intentions. As for those holding both a Master and a PhD, however, the two effects may sum up, which reinforces the negative selection effect of education on return migrants

As for work migrants, neither *Master in the US* nor *PhD in the US* are significant, and what really seems to count to increase their chances to stay in the United States is getting a MBA, whose coefficientis way less than 1, although significant only at 95%. Notice that *MBA in the US* also appears significant in one of the regressions for education migrants, but with an odds ratio closer to one.

Coming to patenting activity, inventors that leave India with substantial patenting experience are definitely those with the higher return hazard, witness the size of the odds ratio of Patents at migration for both education and work migrants (respectively, well over 2 and close to 1.5). Whether this result can be interpreted as evidence of positive self-selection (in contradiction with the education-based negative self-selection) is doubtful. The number of individuals in our sample with at least one patent at migration is very limited and for several of them we may over-estimate the occurrence of return.[11]

Table 4. Event history analysis of return risk, discrete time analysis; by migration channel

	(1)	(2)	(3)	(4)
	Education channel	Work channel	Education channel	Work channel
Time from migration	0.881***	0.883***		
	(0.0201)	(0.0307)		
Time from migration^2	1.005***	1.002		
	(0.000830)	(0.00195)		
Migration cohort = 2000	1.779***	1.423***	1.867***	1.424***
	(0.138)	(0.168)	(0.150)	(0.170)
Age at migration	0.872***	0.899***	0.977	0.904***
	(0.00565)	(0.00467)	(0.0159)	(0.0115)
Master or more at migration	1.623***	1.154	1.180	1.138
	(0.227)	(0.136)	(0.176)	(0.139)
Current student status	0.595***	0.160***	0.459***	0.173***
	(0.0681)	(0.0809)	(0.0908)	(0.0884)
Master in the US	0.432***	0.724	0.568***	0.719
	(0.0444)	(0.215)	(0.0709)	(0.216)
PhD in the US	0.552***	1.259	0.585***	1.430
	(0.0744)	(0.763)	(0.0805)	(0.835)
MBA in the US	0.866	0.401**	0.711**	0.403**
	(0.148)	(0.169)	(0.124)	(0.171)
Patents at migration	2.525***	1.429***	2.320***	1.431***
	(0.358)	(0.0842)	(0.301)	(0.0822)
Cumulative # patents US	1.001	1.011**	0.999	1.012**
	(0.00429)	(0.00528)	(0.00524)	(0.00528)
Observations	50,211	15,333	50,211	15,094
Times dummies	NO	NO	YES	YES
# unique inventors	3054	1308	3054	1308
Chi2	11757	4625	11347	4604

[11] Many individuals with patents at migration are considered returnees on the basis of their patenting activity, with the patent apparently marking their return ("return patent") to India following closely the event (job, education or patent) marking their original migration to the United States. For education migrants, it may well be the "return patent" was actually invented before the migration event, but filed afterward, so we are facing a false positive case of return migration. For work migrants, besides false positives, we may face the cases of inventors temporarily detached in the United States, for very short periods.

LogL	-3623	-1684	-3442	-1664

Notes: Inventor-level clustered standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

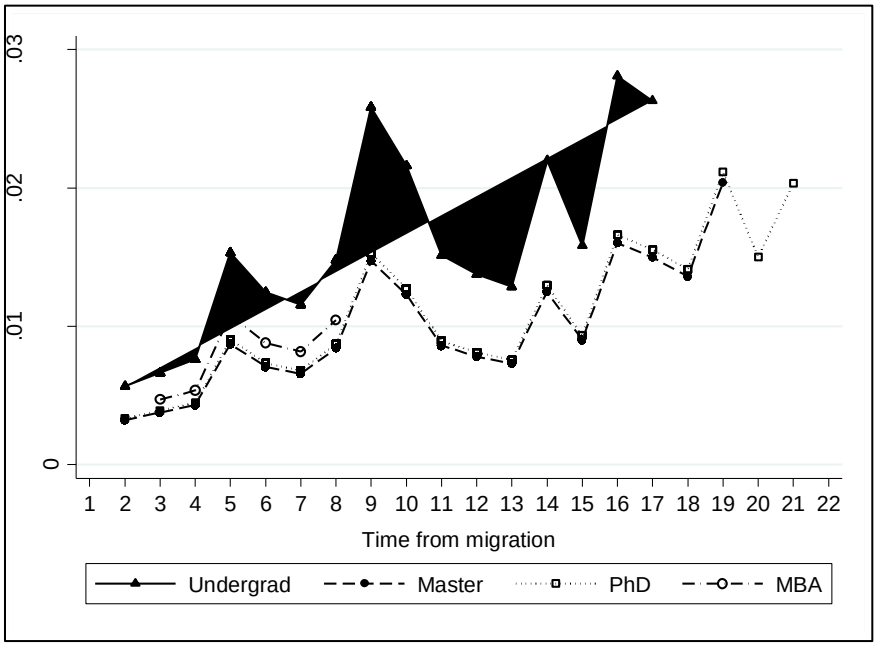
As for the patenting activity in the United States (*Cumulative # patents US*), we also find it to be positively related to the return hazard, with odds ratios barely larger than one, and not significant for education migrants. However, rather than to positive self-selection, this result may be related to specialization. In fact, inventors in our database range from the occasional to the professional ones, the former having signed one or very few patents before or after migration, the latter displaying instead a significant patent record, one that possibly span over several years. In the absence of information on the migration strategies adopted by individuals in our sample, nor on the opportunity and constraints that may shape them, we can speculate what follows. Professional inventors are more likely to move to the United States on a strictly temporary basis and for the specific task of undertaking inventive activities there, possibly on request to their employer in India, which organise their two-way trip. Occasional inventors may be instead a more heterogeneous group, which includes a large number of individuals moving to the United States on their own initiative, rather than their employer's, and more determined to turn an originally temporary visa into a permanent one. They will be at once more open towards different career options and less bound by the original visa arrangements. For example, they may move out of the R&D laboratory and stop producing patents, possibly to undertake managerial functions or an entrepreneurial career, thus getting more chances to stay in the United States. This interpretation fits with the size and significance ratio of the MBA in the US variables, upon which we commented above. Notice that this explanation applies better to work migrants than education ones, all of them entering the United States via a higher education programme and therefore more likely to be occasional, rather than professional inventors. This is coherent with the odds ratios for *Cumulative # patents US* being *de facto* equal to one in the regressions for education migrants.

Moving to time dependence of the hazard ratio, the estimated odds ratios in columns (1) and (2) suggest it to be negative and monotonic for work migrants (the coefficient for the time-squared is not significant), but possibly non-monotonic for education ones (the coefficient for the time-squared is significant and the odds ratio is greater than one).

Following Constant and Massey (2002), we interpret the negative time dependence of the return hazard ratio as indicative of some negative self-selection with respect to unobservable skills the migrant acquires through experience in the host country, and are not as well rewarded back at home. Admittedly, Constant's and Massey's interpretation of time-dependence of the hazard ratio is rather speculative, since other factors besides skill accumulation may intervene, such as increasing investments in real estate or social capital, both of which increase the opportunity cost of return. Still, the negative time dependence we find for work migrants is coherent with the possibility that those among them who stay longer in the United States also engage in managerial functions or undertake entrepreneurial careers. This implies developing skills for which the US-India remuneration gap may be higher than for skills purely related to R&D-performing tasks, and that may facilitate replacing the temporary visa with a permanent one.

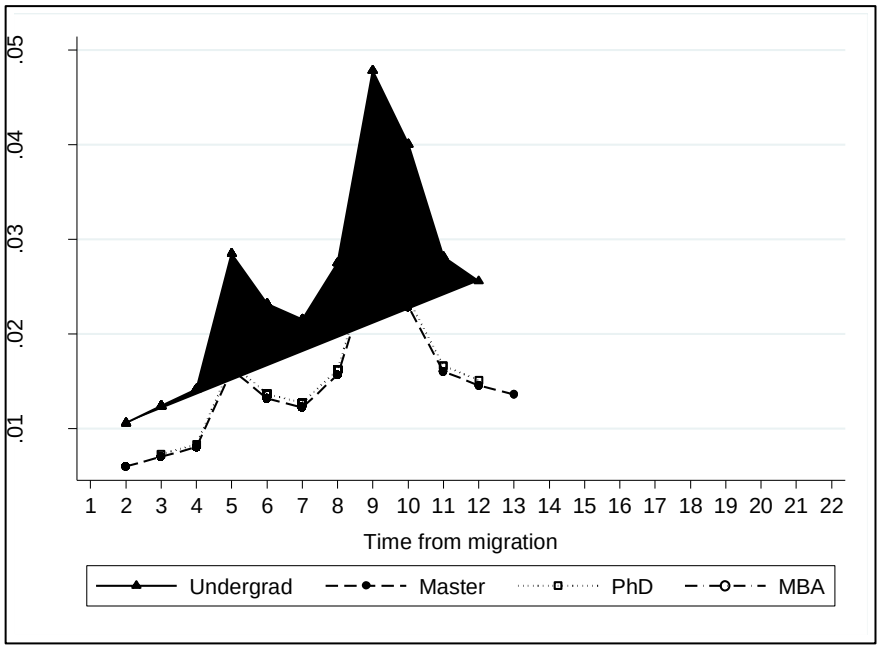
As for the time pattern of education migrants' return hazard ratio, regression in column (1) is not very enlightening. First, it results from imposing a parametric form to *c(t)*; second, it requires to understand whether opposite signs of the estimated coefficients for α_1 and α_2 implies some non-monotonicity, which is not immediately clear in the case of non-linear estimation methods such as cloglog. For this reason, we prefer relying on the results of the non-parametric estimation of column (3). Based on such results, Figures 6a,b reports the within-sample estimates of the total hazard ratio *h(t)* as a function of time and for different educational levels, by migration cohort.

Figure 6a. Estimated hazard ratios since entry in the United States, by education level - Education migrants, 1990 cohort



Within sample estimations from regression (3) in Table 4, for Age at migration =23 and Student status=0 (all remaining regressors at mean values)

Figure 6b. Estimated hazard ratios since entry in the United States, by education level - Education migrants, 2000 cohort



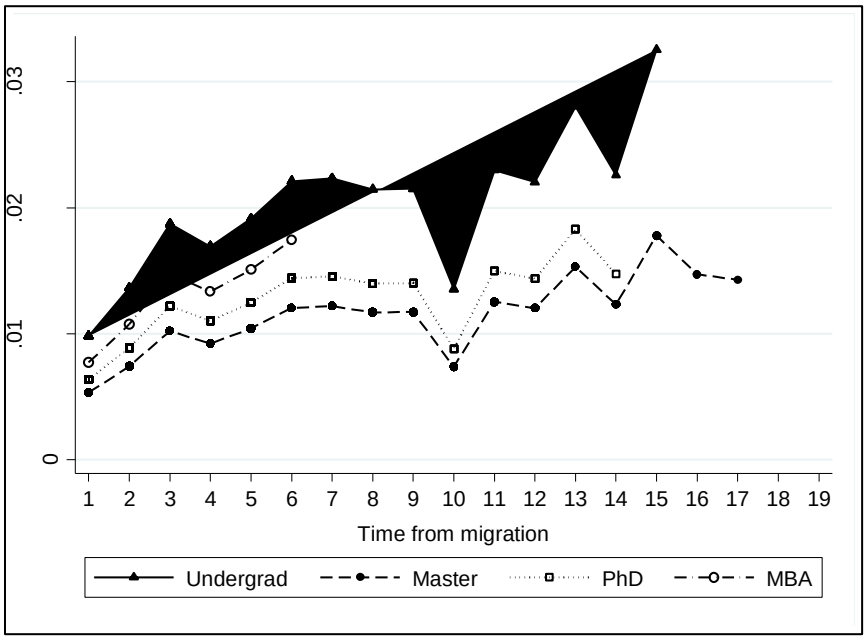
Within sample estimations from regression (3) in Table 4, for Age at migration =23 and Student status=0 (all remaining regressors at mean values)

Both figures suggest the return hazard to follow an inverted U-shaped function of time over the first 13 years of permanence in the United States. After then, we cease to observe migrants in the 2000

cohort, due to right truncation, while the return rate for the 1990 cohort start increasing again, albeit erratically. The hazard ratios for the early years after entry, however, may be underestimated. This is because we produced the graph by setting *Current student status* equal to zero, while in reality it should be equal to one from entry in the United States until graduation (notice that the odd ratios for *Current student status* in table 4 is always greater than one). As a partial remedy, we have replicated regression (3) in Table 4, but with duration *t* counted from the end of migrant's first student spell in the United States. Results for the estimated return hazard ratios are reported in Figures 7a,b, which we can compare to Figures 6a,b. We notice how the estimation of return hazard ratios with respect to time now changes: the inverted U-shape profile we initially observed is significantly smoothed and the return hazard ratio appears first to increase, then to flatten down.

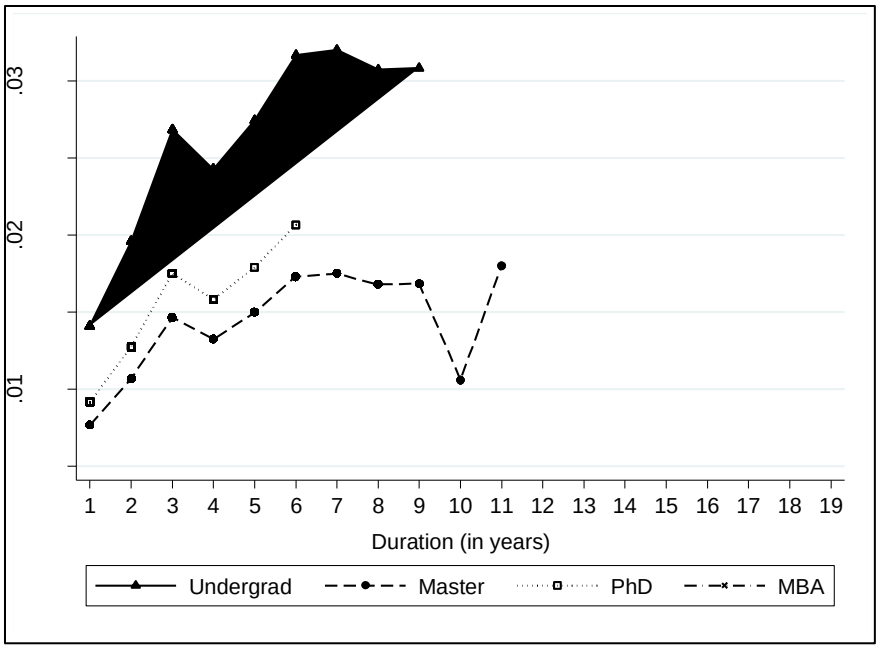
Overall, however, we find some signs of a positive time dependence of the return hazard on time, for education migrants, which may imply that positive self-selection with respect to unobservable skills. We further discuss these results in the Conclusions.

Figure 7a. Estimated hazard ratios since completion of studies in the United States, by education level - Education migrants, 1990 cohort



Within sample estimations (unreported regression), for Age at migration =23 and Student status=0 (all remaining regressors at mean values)

Figure 7b. Estimated hazard ratios since completion of studies in the United States, by education level - Education migrants, 2000 cohort



Within sample estimations (unreported regression), for Age at migration =23 and Student status=0 (all remaining regressors at mean values)

5. Conclusions

Return migration is a much under-studied topic, especially when it comes to its implications for innovation in both the host and home countries of country. Lack of data is a major cause for this situation, due to the virtual absence of official statistics and the technical difficulties that stand in the way of large-scale data mining.

In this paper, we have presented the outcome of an ambitious attempt to overcome such difficulties, based on linking inventor information from patent data to biographical information from an important web-based social network. We focused on Indian inventors with professional experiences of various lengths at one or more US ICT company, and obtained rather reliable data for those among them who moved to the United States in the 1990s and 2000s. Based on biographical information, we could draw a clear distinction between work and education migrants and analyse separately the related return events. In particular, we applied event history analysis and explored the issue of returnees' self-selection with respect to observable and unobservable skills.

Both the distinction between work and education migrants and the study of self-selection may contribute to evaluate the effectiveness of the United States' migration policies, with special reference to scientists, engineers and other innovation-relevant professional categories.

As stressed by Koslowski (2018), the United States' immigration policies are often compared unfavourably to those of countries such as Canada and Australia, whose selective, point-based visa system is held responsible for their success in attracting a high proportion of high-skilled migrants. But the comparison is biased by its exclusive focus on migrants first entering their host countries with permanent visas, which account for a very limited share of entries in the United States. When considering migrants entering with temporary visas, whether work- or education-based, the United States appear the most attractive country, also in view of the large share of temporary migrants turning into permanent ones over the years. In this respect, it becomes crucial to estimate the stay rates of highly skilled permanent immigrants, which our study on Indian migrants finds rather high

and in accordance with the limited evidence available in the literature, especially for education migrants.

Besides assessing the highly skilled migrants' length of stay, it is crucial to assess whether the host countries manage to retain the best and brightest among them, namely those who can contribute most to innovation. In this respect, Wadhwa et al. (2009) give voice to widespread concerns on the difficulties supposedly met by the United States in this respect. Our results, albeit exploratory, go against such concerns for work migrants and leave room for debating on education migrants.

Concerning work migrants, Indian returnees in our sample appear to be negatively selected with respect to education as well as, most likely, to the working experience they accumulate in the United States (as inferred by the negative time dependence of their hazard ratios). Admittedly, we also find a positive relationship between the return hazard and the number of patents they produce while in the United States, but we have suggested how this may have to do more with specialization in managerial functions or entrepreneurship, than positive self-selection.

As for education migrants, Indian returnees in our sample are also negatively selected with respect to education, but appear also increasingly at risk to return, the longer their permanence in the United States, especially over the first 10 years after migration. This can be interpreted as positive self-selection with respect of unobservable skills, at least over the first few years after graduation. But we should bear in mind that our return migration measure does not distinguish between individuals who settle permanently back in their home country, or get engaged in circular migration patterns and/or parallel professional activities in the home and host country.

Further research is clearly needed to both assess the strength of these initial results as well as to extend them. Further codification of the information contained in our dataset will let us assess the quality and location of the educational institutions attended by migrants, so to test whether the return hazard is positively or negatively associated to the prestige of the institution and/or its links with a vibrant labour market for the highly skilled. We also plan to disambiguate fully the name of companies reported by work migrants in their LinkedIn profiles, so to distinguish between intra-company and inter-company mobility. We expect the former to generate short-term temporary migrants, not much exposed to the risk of turning permanent, while the latter should be at the origin of longer stays and more interesting phenomena of negative vs positive self-selection.

More generally, our methodology may be extended to other countries of origin of migrants besides India, and to other professional categories besides those related to ICT.

While a large amount of the knowledge we may gather on highly skilled return migration will pass through the refinement and sharing our data, we think that some ad hoc theorizing is also necessary, so to adopt the emerging theoretical literature on temporary and circular migration we discussed in section 2 to the specificities of STEM workers and students.

References

Allison, P.D., 2014, Event history and survival analysis: Regression for longitudinal event data, vol. 46. SAGE publications.

Baruffaldi, S.H., Landoni, P., 2012. Return mobility and scientific productivity of researchers working abroad: The role of home country linkages. Research Policy, 41(9), 1655-1665.

Bijwaard, G.E., Schluter, C., Wahba, J., 2014. The impact of labor market dynamics on the return migration of immigrants. Review of Economics and Statistics, 96(3), 483-494.

Borjas, G.J., 1989. Immigrant and emigrant earnings: A longitudinal study. Economic inquiry, 27(1), 21-37.

Borjas, G.J., Bratsberg, B., 1996. Who Leaves? The Outmigration of the Foreign-Born. The Review of Economics and Statistics, 165-176.

Breschi, S., Lissoni, F., Miguelez, E., 2017. Foreign-origin inventors in the US: Testing for Diaspora and Brain Gain Effects. Journal of Economic Geography, 17(5), 1009–1038.

Bönisch, P., Gaffert, P., Wilde, J., 2013. The impact of skills on remigration flows. Applied Economics, 45(4), 511-524.

Choudhury, P., 2016. Return migration and geography of innovation in MNEs: a natural experiment of knowledge production by local workers reporting to return migrants. Journal of Economic Geography, 16(3), 585-610

Constant, A., Massey, D.S., 2002. Return migration by German guestworkers: Neoclassical versus new economic theories. International migration, 40(4), 5-38.

Constant, A.F., Nottmeyer, O., Zimmermann, K.F., 2013, The economics of circular migration. In: A.F. Constant, K.F. Zimmermann (Eds.). International handbook on the economics of migration. Edward Elgar, Cheltenham.

Constant, A.F., Zimmermann, K.F., 2016. Diaspora economics: New perspectives. International Journal of Manpower, 37(7), 1110-1135.

De Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., de la Potterie, B.v.P., 2013. The worldwide count of priority patents: A new indicator of inventive activity. Research Policy, 42(3), 720-737.

Desai, M., Kapur, D., McHale, J., 2005. The fiscal impact of high skilled emigration: flows of Indians to the US, mimeo, Harvard University.

Dustmann, C., Görlach, J.-S., 2016. The economics of temporary migrations. Journal of Economic Literature, 54(1), 98-136.

Dustmann, C., Weiss, Y., 2007. Return migration: theory and empirical evidence from the UK. British Journal of Industrial Relations, 45(2), 236-256.

Finn, M.G., 2014. Stay rates of foreign doctorate recipients from US universities, 2011, Oak Ridge Institute for Science and Education (ORISE), Oak Ridge TN.

Gaulé, P., 2014. Who comes back and when? Return migration decisions of academic scientists. Economics Letters, 124(3), 461-464.

Ge, C., Huang, K.W., Png, I.P.L., 2016. Engineer/scientist careers: Patents, online profiles, and misclassification bias. Strategic Management Journal, 37(1), 232-253.

Hawthorne, L., 2018, International Student Mobility: Sending Country Determinants and Policies. In: M. Czaika (Ed.). High-Skilled Migration: Drivers and Policies. Oxford University Press, Oxford.

Hunt, J., 2011. Which Immigrants Are Most Innovative and Entrepreneurial? Distinctions by Entry Visa. Journal of Labbaor Economics, 29, 417-457.

Jenkins, S.P., 2005, Survival analysis. In: U.o.E. Institute for Social and Economic Research (Ed.), Colchester, UK.

Jones, B.F., 2009. The burden of knowledge and the "death of the Renaissance man": is innovation getting harder? The Review of Economic Studies, 76(1), 283-317.

Kapur, D., 2010, Indian Higher Education. In: C.T. Clotfelter (Ed.). American Universities in a Global Market. University of Chicago Press Chicago IL.

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 Washington, DC.

Kerr, W.R., 2017, US high-skilled immigration, innovation, and entrepreneurship: empirical approaches and evidence. In: C. Fink, E. Miguelez (Eds.). The International Mobility of Talent and Innovation: New Evidence and Policy Implications. Cambride University Press, Cambridge UK.

Kerr, W.R., Lincoln, W.F., 2010. The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention. Journal of Labor Economics, 28(3), 473-508.

Kırdar, M.G., 2009. Labor market outcomes, savings accumulation, and return migration. Labour Economics, 16(4), 418-428.

Koslowski, R., 2018, Shifts in Selective Migration Policy Models. In: M. Czaika (Ed.). High-Skilled Migration: Drivers and Policies. Oxford University Press, Oxford.

Nanda, R., Khanna, T., 2010. Diasporas and domestic entrepreneurs: Evidence from the Indian software industry. Journal of Economics & Management Strategy, 19(4), 991-1012.

OECD, 2008, International Migration Outlook. Organisation for Economic Co-operation and Development, Paris.

OECD, 2017, International Migration Outlook. Organisation for Economic Co-operation and Development, Paris.

Teitelbaum, M.S., 2014, Falling behind?: Boom, bust, and the global race for scientific talent. Princeton University Press.

UN, 1998, Recommendations on Statistics of International Migration, Revision 1. United Nations Statistics Division, New York.

Wadhwa, V., Saxenian, A., Freeman, R.B., Salkever, A., 2009. Losing the world's best and brightest: America's new immigrant entrepreneurs, Part 5. mimeo (http://ssrn.com/abstract=1362012).

Zagheni, E., Weber, I., 2015. Demographic research with non-representative internet data. International Journal of Manpower, 36(1), 13-25.

Appendix – Data Methodology

A. Data Sources

The data set used in this paper is the result of a linkage between USPTO patent and inventor data gathered from Patentsview[12] and biographical information extracted from a large number of LinkedIn profiles. Patentsview is a data repository and data visualization tool recently made available by the USPTO, which provides disambiguated data on inventors of USPTO granted patents from 1975 onward. LinkedIn, a well-known social networking system, reports a very large number of users' public profiles that include information on the users' educational curricula and careers (name and possibly locations of education institutions and employer), thereby allowing to trace (return) migration with a scale and degree of precision unmatched by other sources of data[13].

LinkedIn data are subject to a number of limitations. First, resume information is self-reported by individuals and therefore subject to misreporting or even cheating. Second, the choice of creating an account in a professional social network might be correlated with factors affecting the propensity to move (and migrate), thus leading to biased results. Third, we used LinkedIn "public" profiles, namely those who are publicly visible on the internet without being logged into LinkedIn. Hence, our data exclude those profiles for which the account holder chose to keep the profile as "private" and thus visible only from within the system and/or for paying subscribers. In spite of these limitations, we argue that LinkedIn data represent an unparalleled source of information on the international mobility of individuals, both as students and as workers (Ge et al., 2016; Zagheni and Weber, 2015). In what follows, we describe in detail the methodology used to build our sample and we report some tests on the accuracy of information coming from LinkedIn.

B. Sample selection

For the purposes of the present paper, we extracted all the patents granted to the 179 largest US public firms in the ICT industry, from 1975 to 2016. The definition of ICT industry follows the one provided by the OECD[14]. To select our sample of firms, we proceeded as follows. For each SIC code contained in the OECD definition, we extracted from Compustat the list of public US firms active in that SIC and we matched them to the USPTO patent data. As company names reported in patents (i.e. patent assignees) may be written in different ways, we used two sources of information in order to disambiguate them: (a) the concordance tables between Compustat GVKEY codes and patent assignees provided by the NBER patent data project website[15]; (b) the PTMT Custom Bibliographic Patent Data Extract DVD produced by the USPTO, which provides first-named assigned owner at grant as harmonized for spelling variations[16]. From the resulting sample, we dropped all firms with less than 200 patents granted and that either disappeared (because of exit or acquisition) or were delisted before 2005. It is important to stress that for this paper, we kept patents of parent companies by simply disambiguating their names, but we did not collect patents of their subsidiaries with different names from the parent company. For example, ADC Telecommunications Oy and ADC Telecommunications Inc. were considered as the same company. However, patents of Codenoll Techology Corporation, which was acquired by ADC Telecommunications in 1996, have not been

[12] http://www.patentsview.org/web/
[13] LinkedIn data used in this paper were obtained in June 2016.
[14] https://www.oecd.org/sti/ieconomy/1835738.pdf
[15] https://sites.google.com/site/patentdataproject/
[16] https://www.uspto.gov/web/offices/ac/ido/oeip/taf/data/misc/data_cd.doc/custom_extract_dvd/

collected and consolidated with those of the parent company. Moreover, each company included in our sample was considered as active from the date of foundation to the date of exit (most often because of acquisition). Thus, for example, we considered ADC Telecommunications as an active independent company from 1974 to 2010, given that it was acquired by TE Connectivity in December 2010. Finally, for each of the 179 firms thus identified we selected the inventors of their patents using Patentsview, for a total of 262,849 distinct individuals. The complete list of the 179 firms considered in the paper is reported at the end of this appendix.

C. Ethnic analysis of inventor names: identification of Indian-origin inventors

We then proceeded to the ethnic analysis of such inventors' names and surnames, based on Global Name Recognition, a name search technology produced by IBM (from now on, IBM-GNR) and adapted to our purposes by Breschi et al. (2017). This allowed us to identify inventors of presumed Indian origin (from now on, Indian inventors), for a total of 24,017 individuals, representing 9.1% of all inventors employed by the companies in our sample. It is worth noting that this share is higher than the one reported in Kerr (2008). He estimates that the share of Indian inventors residing in the US with a patent application in the period 1975-2004 in the Computers technology field (i.e. the field closer to our sample) is equal to 6.9%. The difference with our estimates might be due both to the different time span covered (our sample includes patents granted up to December 31 2016) and to the different methodologies and data sources used to assign ethnicity (i.e. IBM GNR vs. Melissa)[17]. Moreover, our sample includes also inventors that, even though patented for US companies, do not reside in the US. Yet, the difference still persists even if we restrict the attention to US residing inventors. In this case, Indian origin inventors are 19,222 out of a total 211,480 inventors (i.e. about 9% of all US residing inventors).

D. Matching Indian inventors and LinkedIn profiles

Indian inventors were matched with the employees of the 179 ICT firms in our sample having a LinkedIn profile. The linkage was accomplished by matching first and last name of inventors and employees, on the one hand, and employer and patent assignee names, on the other hand. In other words, for each inventor having made patents with a given company we searched for an individual with the same (or a very similar) first and last name reporting the same company as an employer in the LinkedIn resume. Given that patent assignees and employer names were unlikely to match exactly, due to spelling variations, abbreviations and so on, we implemented a Python script using fuzzy matching techniques and regular expressions. Similar techniques were used to match names of individuals appearing in patent documents and in LinkedIn profiles. To this purpose, we preliminarily standardized the names in the two sets (e.g. removing special characters, such as dots, commas, hyphenations, semicolon etc., converting UTF8 characters into latin characters, removing suffixes such as Jr, PhD, and so on). Using these standardized names, we first performed an exact match between the names of inventors and the names of employees from the LinkedIn profiles. When an exact match was not found, we computed the Jaro-Winkler similarity[18] between the full name of inventors and LinkedIn profiles and we kept only those matches with a name similarity higher than 0.85. For those cases where inventors were matched to multiple LinkedIn profiles[19] because of homonyms, we used the city, state or country reported in the LinkedIn profile and in patents (when

[17] For further details on the methods used to identify the ethnicity of inventors, see Breschi et al. (2017).

[18] https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

[19] In some cases, this problem was due to the fact that the same person opened up multiple profiles. In those cases, we picked up among the different profiles opened by the same individual the one containing more information, under the assumption that this is the profile currently maintained and updated by the person.

available) to improve our matching algorithm. We dropped all cases in which we were unable to unambiguously link an inventor to a unique LinkedIn resume. This exercise yielded 10,839 inventors matched with a LinkedIn account (around 45% of the original Indian inventor sample). This preliminary matched sample was further processed to drop false positives and improve accuracy. In what follows, we describe the methodology used to extract and code information from LinkedIn resumes, as well as the steps undertaken to minimize measurement errors.

E. Classification of educational attainments and country of education

For each matched inventor, we extracted from the LinkedIn resume the information on their educational attainments and we coded the level of education according to the ISCED standard (2011 version)[20]. In particular, we coded the following education levels:

1) ISCED level 3: upper secondary education
2) ISCED level 5 – level 6: short-cycle tertiary education, Bachelor's or equivalent
3) ISCED level 7: Master's or equivalent level
4) ISCED level 8: Doctoral or equivalent level

Given our focus on inventors (i.e. scientists and engineers), we also distinguished between Master of Sciences (generaly in electrical and electronic engineering, computer science or related fields) and Master in Business Admnistration (MBA). It must be pointed out that information on education (like most other types of information) contained in LinkedIn resumes consists of free, unstructured text fields. As a consequence, the assignment of a given educational attainment to the corresponding ISCED level must be done by implementing some type of text classification algorithm. To this purpose, we implemented a Python script, which uses regular expressions and a list of keywords, capturing possible variations in which a certain degree title is written (e.g. a Bachelor of Engineering can be found written as such, but also as BEng, B. Eng, B.E. or other similar variations). We denoted as *unclassified* all those titles which we were unable to classify in any of the ISCED levels. They include a miscellanea of diplomas (e.g. Diploma of Information Technology) and professional certifications (e.g. Certificate IV Web Design, Project Management Professional PMP, and so on) that do not easily fit into any of the ISCED categories. Moreover, some of the matched LinkedIn resumes did not report any information on the educational attainment.

For each education level, we also coded the starting and end year and the country of the school where the education title was achieved. In few cases, resumes reported only the starting year of education. In those cases, we estimated the end year by using the average duration of the corresponding education level (e.g. four years in the case of Bachelor)[21].

Regarding the country of education, this was found by geocoding school names. To this purpose, we implemented a simple Python script which fed the name of the school into Google Maps, using the Google Maps Geocoding API. In few cases, Google returned more than one country match. We manually cleaned and checked those cases. Still in other cases, Google was unable to return a valid address, as information contained in the school name was not sufficiently detailed to allow accurate geocoding. We did not make any further check in those cases and we considered as missing the information on the school country. Overall, it is important to stress that this exercise might be prone to some (possibly limited) measurement errors. Given that the full address and city of schools is unknown and the only information we can provide is the school name, the accuracy of Google

[20] http://uis.unesco.org/en/topic/international-standard-classification-education-isced

[21] More specifically, the average duration in years of the different educational levels for the inventors in our sample is: 5 (High school), 2.5 (short-cycle tertiary education), 4 (Bachelor), 3 (Master), 2 (MBA), 5 (PhD).

geocoding is high whenever the school name is sufficiently unique and distinctive (e.g. Bocconi University, Insead and so on), but it is likely to be lower for school names, such as St. James or St. Joseph School. Since geocoding errors are less likely to occur for university names, we manually checked all geocoding results related to ISCED level 3 educational institutions (i.e. secondary schools).

F. Employment history and employment country

Similarly to what done in the case of education, we extracted the start and end year of each employment spell as well as the name of the employer as reported in the resume. Given that our interest is on the mobility of inventors across countries, particularly from India to the US and return, and not across firms, we did not disambiguate employers' names appearing in the LinkedIn resumes. Rather, we focused on the job location of each employment spell. In this respect, it is important to note that reporting the job location is not compulsory when filling the employment history of a LinkedIn resume. To illustrate this issue, we report below the employment history of two different inventors in our sample as they are reported in their respective LinkedIn resumes (see Tables A1 and A2). Both inventors report to have worked for Broadcom Corp. at some time during their working career. However, whereas one resume reports the job location at Broadcom in Bangalore, the other does not report any information on the job location.

In order to track the mobility of inventors from India to the US and return, we took the job location (if not missing) «self-reported» by the inventor in her resume and we coded whether the location was in the US (e.g. San Jose, Bay Area) or in India (e.g. Bangalore). Out of a total 35,456 employment spells recorded by the inventors in our sample, 7,743 reported the job location, namely around 22% of all job spells.

Table A1: Inventor A, resume reporting job location in employment history

Job title	Employer	Job location	Period
Sr. System Engineer	Motorola Solutions	Dallas/Fort Worth Area	1997-2000
Member of Technical Staff	Iospan Wireless	San Jose, Bay Area	2000-2002
Student	University of Texas at Austin	Austin/Texas Area	2003-2005
Member of Technical Staff	Texas Instruments	Dallas/Fort Worth Area	2005-2006
Director (Technical ATD)	Broadcom	Bangalore	2006-Present

Table A2: Inventor B, resume not reporting job location in employment history

Job title	Employer	Job location	Period
Mixed Signal Design Engineer	Crystal Semiconductor	.	1997-1999
Staff Design Engineer	Level One Communications (Intel Corp)	.	1999-2000
Director of Engineering, Broadcom Distinguished Engineer	Broadcom Corporation	.	2000-2014
Director, Touch and Sensing Hardware	Apple	.	2014-Present

In this paper, *we only considered «self-reported» job locations in assessing mobility and migration events*. When job location was missing, we did not consider the corresponding employment spell in assessing inventors' mobility from India to the US and return.

As illustrated above, only slightly less than a quarter of all employment spells reports the job location in the resume. Hence, our approach is fairly accurate (under the assumption that the location reported in the resume corresponds to the actual job location), but it is likely to under-estimate the extent of mobility and migration.

In principle, one might improve upon this method at the cost of somewhat lower accuracy. In particular, when the information on the job location is missing, one can estimate the likelihood of the job location to be in India (or more generally in a certain country) by exploiting information on other LinkedIn profiles reporting the same employer *and* the job location[22]. Given a certain employment spell whose job location is unknown, one can compute the fraction of all its employees with a LinkedIn profile (not just inventors, but any LinkedIn profile holders) who associated such employer to an Indian address (or to the address of a focal country).

To illustrate the idea, consider the employment spells of the inventor reported in Table A3. This is still another case in which the inventor did not report information on the location of jobs. For each employer reported in her resume, one can extract all LinkedIn resumes (i.e. not just inventors, but any LinkedIn profile), who meet two conditions:

i. The resume reports the same employer name (i.e. the invidual reported to have worked for the focal employer);
ii. The resume reports information on the job location.

For example, given that inventor C reported *Art of Living* as one of her employers (see Table A3) but it did not report the job location, in order to estimate the probability that the location was in India, one can extract from LinkedIn all resumes that also reported *Art of Living* as an employer *and* reported the job location in the resume.

Table A3: Inventor C, resume not reporting job location in employment history

Job title	Employer	Job location	Period
Founding Engineer	Ipcell	.	1998-2000
Sr Mgr, Software Development	Cisco	.	2000-2008
Sr Product Manager, Marketing, Business Development Manager	Cisco	.	2008-2012
State Coordinator, Texas and Director, YES for Schools	*Art of Living*	.	2002-Present
Senior Consultant	Context BI	.	2012-Present

Table A4 illustrates the cross-country distribution of all LinkedIn resumes that reported an employment spell at *Art of Living* and also specified the location of the job. Out of 164 individuals who mentioned *Art of Living* as an employer *and* also reported job location, 90 of them (55% of total)

[22] Differently from schools (i.e. universities and other educational institutions) who have generally a unique location in a single country, large firms have operations, plants and subsidiaries in multiple countries. As a consequence, whereas geocoding schools through their names is likely to yield a reliable unique address, this is not the case for large firms. Put it differently, one cannot estimate the missing job locations by geocoding company names, such as Broadcom, Texas Instruments and so on.

were located in India. Hence, one can take this number as a rough estimate of the likelihood that the job location of inventor C in Table A3 was actually in India. Following this approach, one can also establish progressively looser thresholds of this probability, e.g. at 100%, 90%, 70%, 50%, and so on. For example, in the case illustrated in Table A3, the probability that the job location of inventor C when employed at *Art of Living* was in India is higher than 50%, but lower than 70%[23].

Table A4: Cross-country distribution of all *Art of Living* employees reporting job location in their LinkedIn resumes

Country	Number of employees located in country	% of total
India	90	54.9
Canada	16	9.8
Germany	13	7.9
Australia	12	7.3
Other countries	33	20.1
Total	164	100.0

As already mentioned above, in this paper (and in this appendix), we *rely exclusively* upon «self reported» job locations, namely on geographical information regarding job location that was *explicitly* reported in the resume (i.e. as in the case of Table A1). In this respect, our results need to be considered as conservative estimates of the true return migration.

G. Estimating age and year of birth

Using information on the starting year of education, we were also able to estimate the year of birth of the matched inventors. To this purpose, we assumed that inventors started a given educational programme at the most typical age for the corresponding educational level. More precisely, we assumed that the starting age was:

a) 14 for ISCED level 3 (upper secondary education)
b) 19 for ISCED level 5-6 (Bachelor or equivalent)
c) 23 for ISCED level 7 (Master or equivalent[24])
d) 25 for ISCED level 8 (Doctoral or equivalent)

To estimate the age of an individual, we followed the above list in a hierarchical order. Thus, for example, for an individual who reported to have started a high school cycle in 2000, we assumed that birth year was 1986 (i.e. =2000-14), irrespective of the other attainments achieved later in the life. Similarly, for an individual who *did not* report information on high school, but reported to have started a Bachelor in 2000, we assumed that she was 19 years old in that year and therefore was born in 1981 (i.e. =2000-19).

This approach has some obvious limitations. First, although probably correct on average, the estimated year of birth is greater than the actual one for those inventors who started a formal

[23] The Art of Living Foundation is a volunteer-based, humanitarian and educational non-governmental organizations (NGO). It was founded in 1981 by Ravi Shankar. The Art of Living Foundation is spread over 156 countries. Its headquarter is in Bangalore. Not surprisingly, thus, the majority of individuals with a LinkedIn account reporting job location at Art of Living in their resumes declared a job located in India. Source: https://en.wikipedia.org/wiki/Art_of_Living_Foundation.

[24] In case the only information on educational attainment was related to MBA, we assumed a starting age at 27, as this looks the most typical age of MBA applicants.

education programme later on in their life cycle and for those inventors who did not follow the typical sequence of studies, BSc→MSc→PhD. For example, the estimated year of birth of an inventor reporting to have started a PhD in 2000 (without reporting any other information on secondary education, BSc and MSc) is 1975, given that we assume that the average age of a first year PhD student is 25. To the extent that the inventor actually started her PhD at 30, her true year of birth is 1970 and as a consequence we are under-estimating his actual age[25]. Similarly, for an individual who started a MSc in 2000 *after* obtaining a PhD in 1995, we assume that birth year is 1977 (i.e. =2000-23), whereas her actual birth year is 1970 (i.e. =1995-25).

Second, estimating the year of birth is not possible for those inventors who either did not report any education information in the resume or whose *only* education attainment is *unclassified*, given that in this case there is not an age benchmark. In our sample of 10,839 inventors with a matched LinkedIn profile, there were 1,391 inventors whose resume did not report any information on education, and 1,585 inventors whose only educational attainment was *unclassified*.

Given our focus on educational level and education country to assess the extent of self-selection in return migration, we simply dropped from our sample the 1,391 inventors whose LinkedIn resume did not report any information on education. Regarding the 1,585 inventors whose only educational attainment was unclassified, we estimated the year of birth in the following way. For each of them, we extracted the application year of their first patent at the USPTO. From the sample of inventors for which we were able to estimate the year of birth based on education, we identified all inventors whose first patent application was made in the same year and we computed the average age of those inventors. Finally, we used this average age to estimate the year of birth. For example, given an individual whose first patent was made in 2000 and whose year of birth was unknown, we extracted all inventors whose first patent was in 2000 and for which the year of birth was estimated using educational attainments. As the average age of inventors whose first patent was made in 2000 is 32, we assumed that the year of birth of the focal inventor is 1978 (i.e. 2000 - 32). Once again, although probably correct on average, this approach is likely to be prone to some measurement errors. To this purpose, section K below reports descriptive statistics on the distribution of the inventors in our final sample by year of birth and age at the first patent.

H. Dropping incomplete and inconsistent profiles

An extensive follow-up checking was performed in a semi-automated way to improve the accuracy of our matching between inventors and LinkedIn profiles. In the first place, given our focus on educational attainments to assess the extent of self-selection in return migration, we dropped from the list of 10,839 matched inventors, 1,391 inventors whose LinkedIn resume did not report any information on education. Second, we dropped 279 inventors whose estimated age at the first patent was either less than 21 (i.e. the age at the completion of a short-cycle of tertiary education) or greater than 66 (i.e. age at retirement) or whose first patent was granted before the first reported education title. These cases were dropped because they are likely to be false positives, namely matched to the wrong LinkedIn profile. In addition to this, other 187 inventors were dropped from our sample as their LinkedIn resume did not report any employment history.

Table A5 summarizes the outcome of our matching exercise between USPTO inventors and LinkedIn resumes. Out of 24,017 inventors of Indian origin, we could match 10,839 unique LinkedIn profiles. For 1,857 of them (i.e. =1,391+279+187), however, the information contained in the LinkedIn

[25] The opposite case of individuals starting formal education at an age lower than the typical age for a certain education level is arguably less common.

resumes was either incomplete or inconsistent and the corresponding inventors were dropped from our sample. Overall, our final sample consists of 8,982 inventors, which represent 37.4% of all Indian inventors.

Table A5: Indian Inventors in ICT: LinkedIn matching outcome

	Number	% of all Indian inventors
Indian inventors matched with a LinkedIn profile	10,839	45.13
of which:		
Profile has no info on education attainments [a)]	1,391	5.79
Profile has no info on employment history [b)]	187	0.78
Profile has inconsistent info on age [c)]	279	1.16
Profile has complete info on education and employment and consistent info on age	**8,982**	**37.40**
Indian inventors not matched with a LinkedIn profile	13,178	54.87
All Indian inventors	24,017	100.00

[a)] the matched profile does not contain any information on the educational attainments of the inventor;
[b)] the matched profile does not contain any information on the employment history of the inventor
[c)] the age of the inventor, estimated on the basis of the educational attainment, at the time of the first patent is either lower than 21 or greater than 66, or the first patent was applied before the first reported education title.

I. Accuracy of match: precision and recall

In order to assess the accuracy of our matching, we exploited the fact that some inventors report in their LinkedIn resumes information on the patents made. In particular, we could identify 1,049 cases of Indian inventors for whom the match with LinkedIn was "certain", as the inventor herself reported information on the patents made in the LinkedIn profile. Using this subset, we were able to assess the rate of errors generated by our matching algorithm. For this test, we restricted attention to the 8,982 matched inventors for whom we have complete education and employment history and consistent information on age. In particular, we computed two types of statistics.

First, we evaluated the rate of "false positives" (Type 1 error). They correspond to those cases in which an inventor is matched by our algorithm to a "false" LinkedIn profile, i.e. the algorithm assigns the inventor to a profile, which is not the correct one. More specifically, we computed the so-called "precision rate", defined as:

$$\frac{\#\text{ of true positives}}{\#\text{ of true positives} + \#\text{ of false positives}} \qquad (1)$$

Of the 1,049 "certain" matches, our matching algorithm was able to assign a LinkedIn profile to 838 cases. Of them, 808 were "true positives" (i.e. the matched profile was the correct one) and 30 were "false positives" (i.e. the matched profile is a false one). Overall, the precision rate is equal to 808/838=0.964. This means that, when our algorithm assigns a LinkedIn profile to an inventor, it does so correctly in about 96.4% of cases.

Secondly, we evaluated the rate of "false negatives" (Type 2 error). They correspond to those cases in which our algorithm fails to find a match even when there is a valid one, i.e. the inventor has a

LinkedIn profile, but our algorithm is unable to match it. More specifically, we computed the so-called "recall rate", defined as:

$$\frac{\text{\# of true positives}}{\text{\# of true positives + \# of false negatives}} \qquad (2)$$

Of the 1,049 "certain" matches, our matching algorithm was able to assign a correct LinkedIn profile to 808 cases (true positives), but it failed to find a valid match in 241 cases (false negatives). Overall the recall rate is equal to 808/1,049=0.77. This means that our algorithm is able to find a valid match for about 77% of all inventors who have a LinkedIn profile.

J. Comparison between matched and unmatched inventors

A further control needed to ensure the representativeness of our sample is comparing the inventors matched with a LinkedIn profile and the inventors not matched. To this purpose, we restricted again attention to the 8,982 matched inventors for whom we have complete education and employment history and consistent information on age[26].

In particular, we carried out three types of tests. In the first place, we tested to what extent matched and unmatched inventors differ in terms of patent productivity. To this purpose, we carried out a simple t-test on the average number of patents produced by the inventors in the two groups. Results reported in Table A6 show that patent productivity of matched and unmatched inventors does not differ in a statistically significant way.

Table A6: Average number of patents of matched and unmatched inmventors

Matched			Unmatched			
Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.	t-test (p-value)
8982	7.33	15.54	15035	7.29	18.11	0.185 (0.853)

Second, we assessed to what extent the sample of matched inventors includes more recent cohorts, under the assumption that younger people have more incentives or simply a higher propensity to register a LinkedIn profile than relatively older people. Ideally, we would like to compare the age profile across the two subsets. However, while this can be somehow estimated from education data for the matched inventors, there is no way to retrieve this information for the unmatched ones. As a second best solution, therefore, we computed for each subset of inventors the distribution by application year of the first patent at the USPTO. As shown in Figure A1, although the two distributions appear quite similar, the sample of matched inventors seems to include individuals with a relatively more recent patenting history than the sample of unmatched inventors. A Kolmogorov-Smirnov two-sample test (0.256, p-value 0.10) allows to reject the hypothesis that the distributions of the two samples are the same. As a consequence, keeping in mind that the date of the first patent is not perfectly correlated with the age of the inventor, we can reasonably conclude that our sample of matched inventors includes relatively younger individuals.

[26] Note that the unmatched cases include inventors who may actually have a LinkedIn profile, which for various reasons we have been unable to match. As noted above, we estimate a recall rate of 77%, meaning that we are unable to match the LinkedIn profile for about 23% of all those who actually have one. For this reason, what we are assessing here, strictly speaking, is not the probability that an inventor has or has not a Linkedin profile, but the probability that the inventor has been included in our final sample. At the same time, it is also correct to say that the majority of the inventors in the unmatched subset is composed of individuals that are truly absent from LinkedIn.

Third, we tested to what extent there might be a different propensity to have a LinkedIn account across groups of inventors. For instance, inventors might be more likely to sign up to keep in touch if they are away from US, or conversely more likely to do it if in the US because they need to do it for work. In order to test this type of conjectures, we split the population of Indian inventors into four mutually exclusive groups:

1. Inventors who patented only in India
2. Inventors who patented only in the US
3. Inventors who patented both in India and in the US (and possibly other countries)
4. Others

Table A7 reports the number of inventors in each group as well as the fraction of inventors with a matched LinkedIn profile.

Figure A1: Distribution of LinkedIn matched and unmatched inventors by application year of the first patent at the USPTO

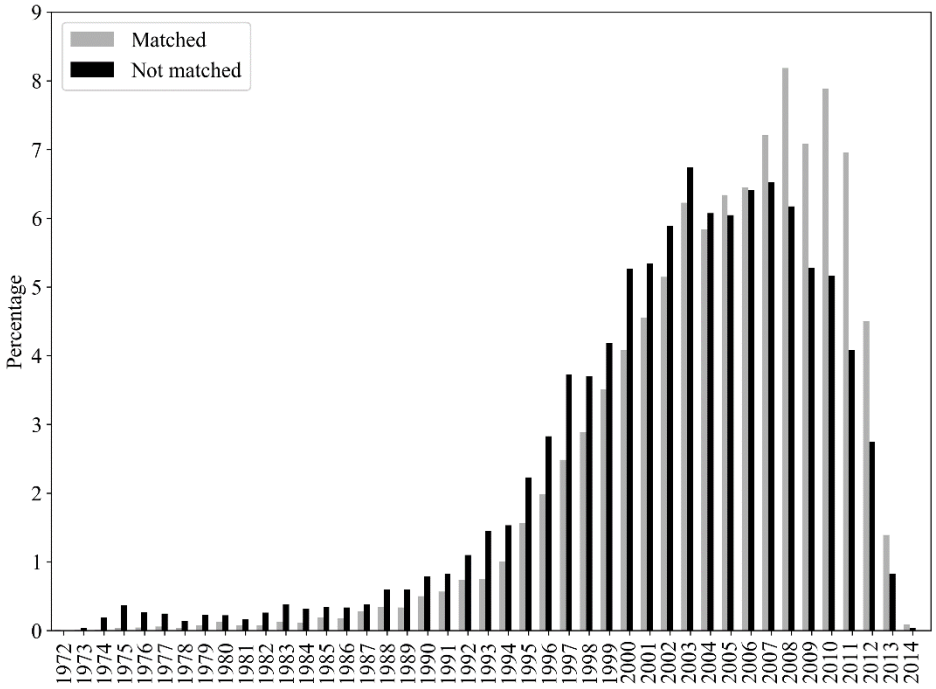


Table A7: Fraction of inventors with a matched LinkedIn profile

Group	Number	Number with a matched LinkedIn profile	% with a matched LinkedIn profile
1. Inventors who patented only in India	4,324	2,003	46.3
2. Inventors who patented only in the US	17,392	6,088	35.0
3. Inventors who patented both in India and in the US	1,457	593	40.7
4. Others	844	298	35.3

All Indian inventors	24,017	8,982	37.4

A simple z-test of proportions indicates that inventors that patent exclusively in India have a significantly higher probability of being matched with a LinkedIn profile than both inventors patenting exclusively in the US (z-score=13.776, p-value=0.000) and inventors that patent both in India and in the US (z-score=3.732, p-value=0.000). This evidence is consistent with the use of LinkedIn by inventors resident in India to signal their skills and "promote" themselves in the job market. Moreover, it is also consistent with the broader pattern of LinkedIn usage by country. As a matter of fact, with 35 million accounts India is second only to the US (with 128 million profiles) in terms of registered members of Linkedin (as of the first quarter of 2016)[27].

K. Age distribution of inventors

In this section, we provide some descriptive statistics on the age distribution of inventors in our sample. This is once again relevant to assess the reliability of our sample, given that age was estimated based on the education attainmnents of inventors. Also in this case, we focus attention on the 8,982 inventors included in our sample.

As described above (section F), age of inventors was estimated on the basis of ISCED education levels. When the educational attainment could not be classified in any of the ISCED levels, we estimated age on the basis of the average age at the first patent. Figure A2 reports the percentage distribution of inventors according to the way in which age was estimated. For the vast majority of inventors in our sample (64%), age was estimated on the basis of the start year of the BSc, as this was the first educational attainment reported in their resume. For an additional 12% of inventors age was estimated on the basis of the start year of MSc (as this was the first educational attainment reported in their resume). For just 4% of all inventors the source of information to estimate age was the start year of the secondary school, as this is a type of information that relatively few individuals mention in their resumes. Moreover, for about 16% of all inventors whose educational attainment could not be classified in any ISCED level, we were forced to estimate age by taking the average age of inventors at the time of their first patent.

Figure A2: Percentage distribution of inventors in the final sample by source of information used to estimate year of birth (8,982 obs.)

[27] https://www.statista.com/statistics/272783/linkedins-membership-worldwide-by-country/

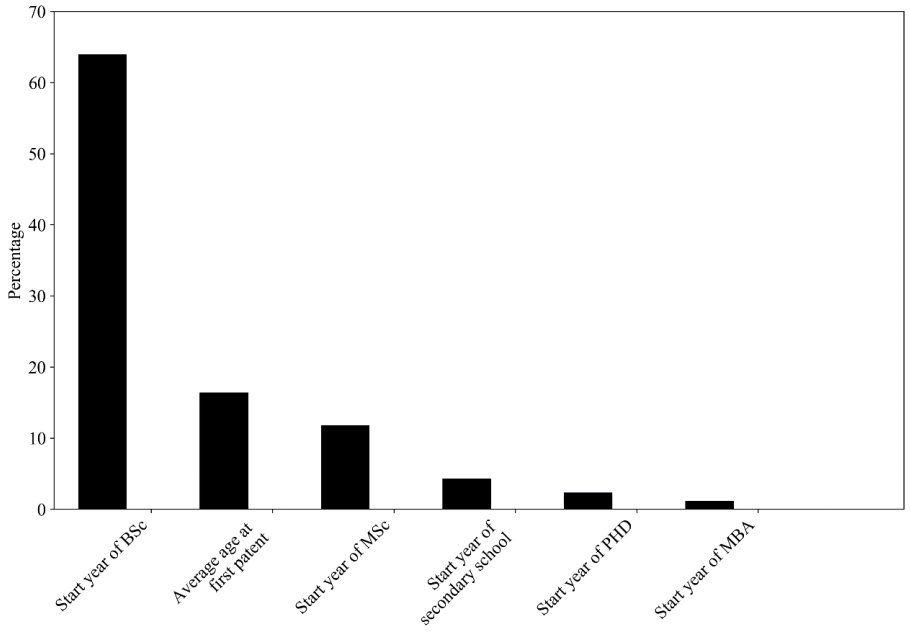


Figure A3 shows the percentage distribution of the inventors included in our final sample by estimated year of birth. The bulk of inventors are born between early 1970s and mid-1980s, with a modal value in 1978. Around 67% of all inventors in our sample are born between 1970 and 1985, whereas an additional 25% are born between 1960 and 1969. Overall, this evidence suggests, as already noted above, that our sample of inventors consists of relatively young individuals (i.e. the modal inventor is 40 years old in 2018).

Figure A3: Frequency distribution of inventors in the final sample by estimated year of birth (8,982 obs.)

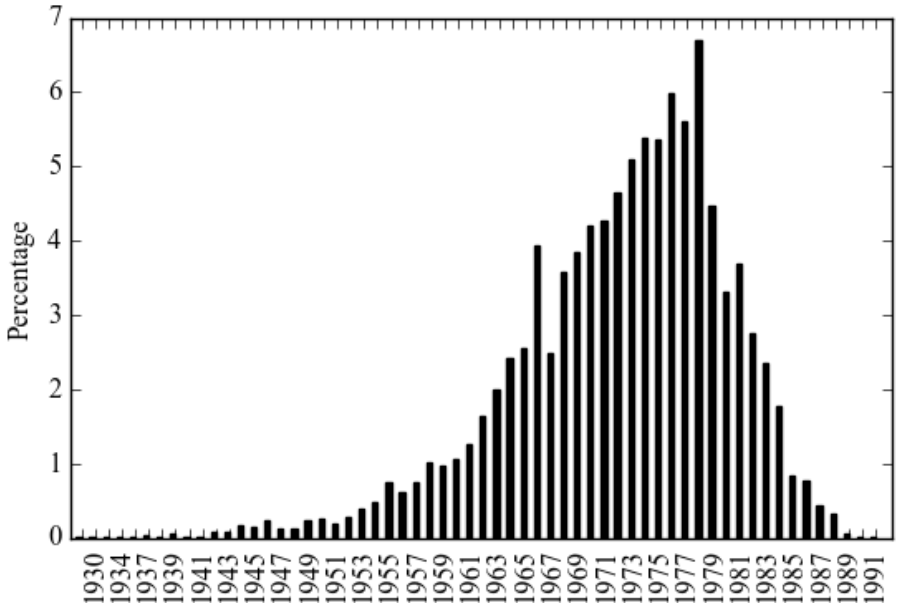


Figure A4 illustrates the percentage distribution of inventors in our sample by age at the first UPSTO patent application. The modal age is 32: 18% of all inventors made their first patent application at

this age. More generally, the distribution is remarkably concentrated between 25 and 35: inventors in this age range account for 77% of all inventors in our sample[28]. These results are broadly consistent with those reported by Jones (2009).

Finally, Figure A5 shows the average age at the time of the first patent by year of the first patent. Some variation is observed for older cohorts (i.e. inventors who made their first patent in the '70s and in the '80s), yet these cohorts include very few individuals (see above Figure A1). Apart from that, no clear pattern is detectable in the data. The average age at the first patent of individuals who made their first patent in the '90s and '00s, which represent the bulk of our sample, was around 32 with no significant variation.

Figure A4: Percentage distribution of inventors in the final sample by age at the first patent (8,982 obs.)

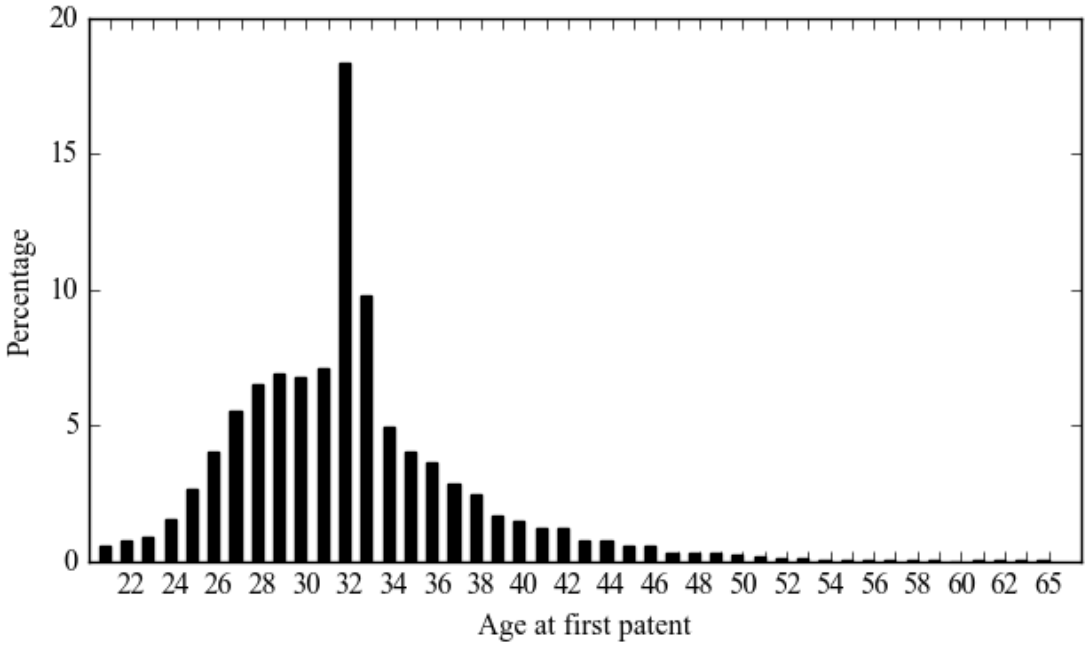
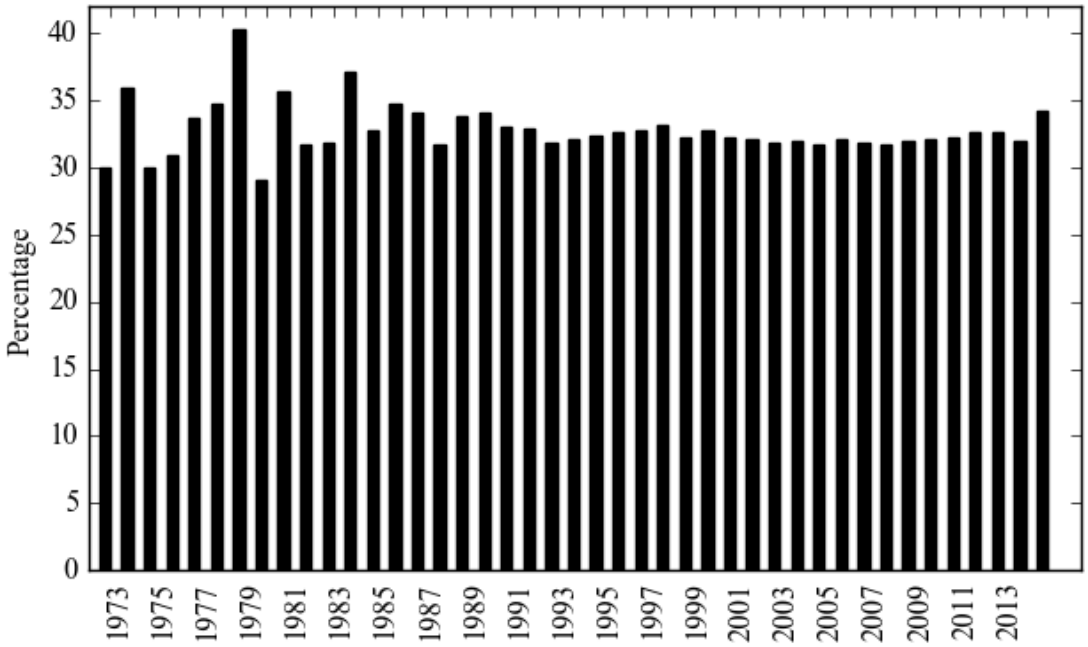


Figure A5: Average age at the first patent by year of first patent application (8,982 obs.)

[28] In evaluating the peak at 32, note however that the distribution includes also 1,473 inventors for whom age was estimated on the basis of the average age at the time of the first patent, which peaks around that age for most cohorts.



L. Coding migration events and migrant inventors

In this section, we describe the methodology used to code migration events and to identify migrant inventors. To this purpose, we exploited three types of information on location of inventors:

1. Country of address reported in patents (from USPTO)
2. Country of job locations reported in the resume (from LinkedIn)
3. Country of educational institutions where education was attained (from LinkedIn)

As far as 2. and 3. are concerned, we already illustrated above the way in which location was extracted from LinkedIn records. Information on 1. was obtained from Patentsview.

In order to identify migrant inventors, we proceeded as follows. In the first place, we split the sample of 8,982 inventors in two mutually exclusive subsets. The first subset includes inventors who, *at any time* in their career, either made a patent, were educated or «self reported» a job location in India. The second subset includes inventors who never made patents, were educated or reported a «self reported» job location in India. Note that, as our sample consists of inventors, the second subset includes Indian-origin inventors that for sure made patents in other countries, but India.

The first subset comprises potential migrants, whereas we label inventors in the second subset as «false positives» (with respect to migration). The reason for this labelling is the following: these inventors have an Indian origin, have made patents outside India, but did not leave any trace of activity in India, particularly with respect to education. Even though they might include true migrants, they might also consist of second-generation Indians born and educated outside India. Out of 8,982 inventors, we labelled 1,445 of them as «false positives» and we dropped them from our sample. As argued above, some of these inventors might be true migrants and not second-generation Indian inventors. Yet, we cannot discriminate the former from the latter on the basis of available information. For example, an inventor who in her resume reported only a PhD attained in the US, did not report any job location in India, and never made patents in India is considered as a «false

positive», even though she might have achieved a BSc in India without reporting such educational attainment in her resume.

With this caveat in mind, our sample of potential migrants, after dropping 1,445 «false positives», is reduced to 7,537 inventors. This sample was further split into two mutually exclusive subsets. The first subset includes inventors who never reported in their career an educational attainment, an employment (i.e. job location) or a patent made in a country different from India. The second subset is defined in a complementary way and it includes inventors who, *at any time* in their career, either made a patent, were educated or reported a job location outside India. We label the first subset as «non migrants» to indicate that on the basis of available information these inventors were active only in India and did not migrate during their career. Out of 7,537 potential migrants, 1,672 were labeled as «non migrants» and, given our focus on return migration, were dropped from the sample. Our sample of migrants thus consists of 5,865 inventors.

Our sample of «migrant» inventors was further split into two mutually exclusive groups reflecting the motives for which individuals migrated. In particular, we distinguished two major reasons for migration: education and work. Accordingly, the first subset includes inventors whose *first event* outside India was the attainment of an educational title. Similarly, the second subset includes inventors whose *first event* outside India was either a patent or a job in a country different from India. Overall, out of 5,865 migrant inventors, we identified 4,161 «education migrants» and 1,704 «work migrants».

Figure A6 summarizes the process followed to identify the sample of migrant inventors.

Figure A6: Identification of migrant inventors

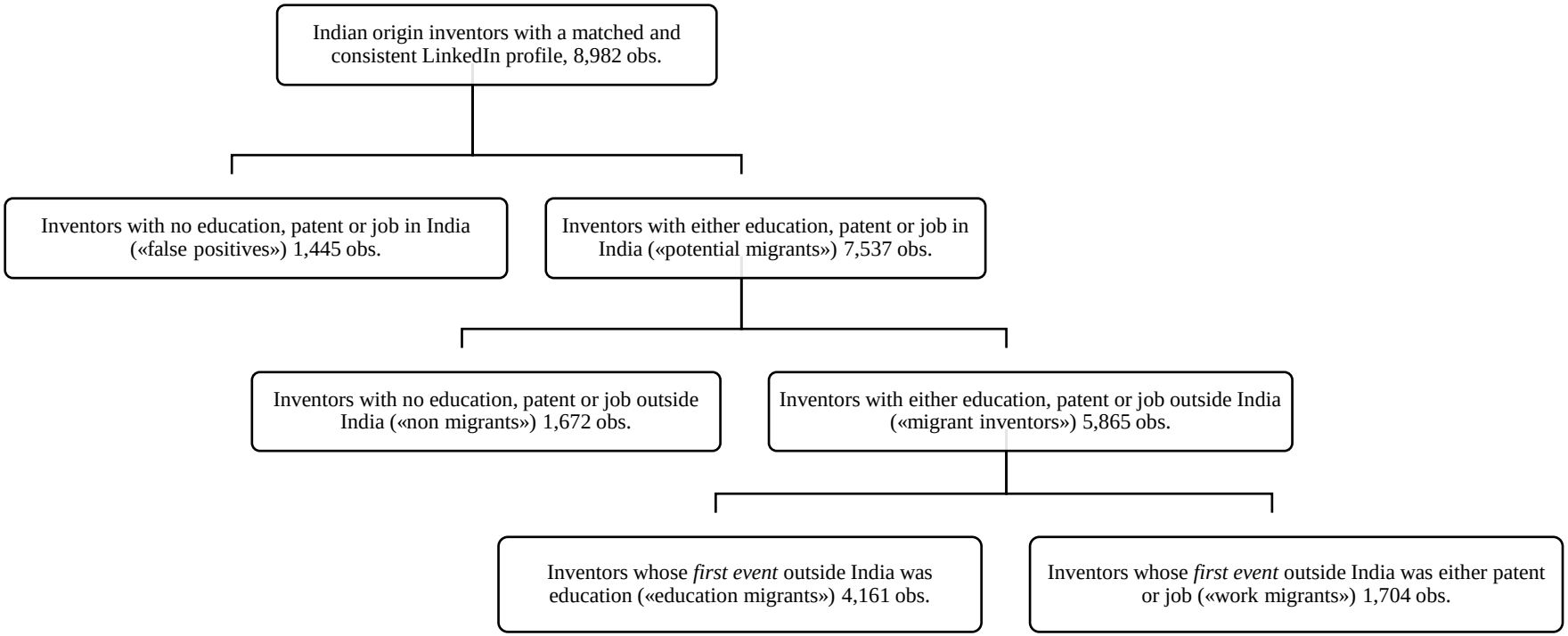
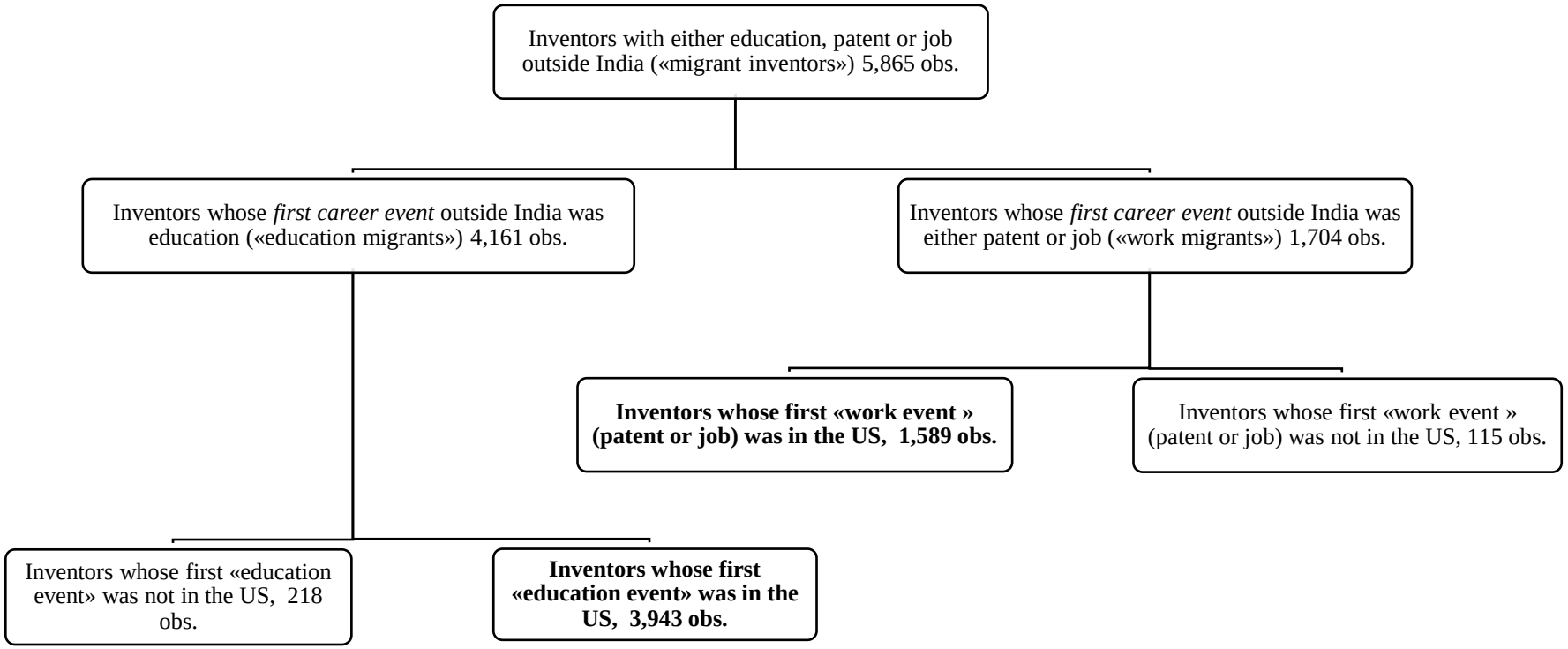


Figure A7: Identification of migrants to the US



«Education migrants» can be further split into two distinct categories: i) migrant inventors that never made any patent outside India (317); ii) migrant inventors that patented outside India *after* being educated abroad (3,844). Similarly, «work migrants» can be further split into two distinct categories: i) migrant inventors that did not take any education outside India (1,253); ii) migrant inventors that took education outside India *after* either patenting or taking a job outside India (451).

As a final step, we further split the sample of 5,865 migrant inventors into two mutually exclusive groups. The first group comprises migrant inventors whose first event outside India was in the US. The second group consists of migrant inventors whose first event outside India was in a country different from the US. Figure A7 illustrates this further selection step. Out of 5,865 migrant inventors, 5,532 (i.e. 94% of all migrant inventors) are defined as «migrants to the US», whereas 333 are defined as «migrants to other countries». Of the 5,532 migrants to the US, 3,943 migrated for education motives, whereas 1,589 migrated for work reasons. In what follows, we focus on the subset of «migrants to the US».

M. Coding migration year

Once coded migration events and identified migrant inventors (to the US), we defined the year in which migration took place. The identification of the year of migration differs according to the migration motive. For inventors whose migration motive was education, we assumed that migration occurred at the beginning of the first education programme undertaken by the inventor in the US. For example, for an inventor whose first event outside India was a MSc in the US started in 1981, migration year was set equal to 1981.

For inventors whose migration motive was work, the migration year was similarly defined as the date of the first event occurring outside India. As for «work migrants» two possible events, i.e. patent or employment, can mark the starting of migration, the year of migration was defined accordingly. Thus, for inventors whose first event outside India was a patent made in the US, migration year was set equal to the application year of the first patent in the US. Instead, for inventors whose first event outside India was an employment in the US, migration year was set equal to the starting year of the corresponding employment spell. Out of 1,589 migrant inventors to the US for work reasons, the first event in the US was a patent for 1,280 (i.e. 81%) of them.

It is important to emphasize the asymmetry in estimating the migration date for inventors whose migration motive was education as compared to inventors who migrated for work reasons. Whereas this estimate is likely to be fairly accurate for inventors who moved for education reasons, this is less likely to the case for inventors who moved for work reasons. As noted above, for the majority of the latter the first event in the US that we could detect on the basis of available information was a patent application. Yet, it might be that these inventors moved to the US before this application and we are simply unable to spot the time of the move because inventors' resume does not report sufficiently detailed and accurate information to date the migration event more precisely.

For descriptive purposes, Figure A8 reports the percentage distribution of migrant inventors to the US, who migrated for education motives, by migration year. Most of the migration for these reasons was concentrated in the two decades 1990-1999 and 2000-2009. Of all Indian inventors that migrated to the US for education reasons, 44% of them did it in the period 1990-1999, and 33% did in the period 2000-2009.

Figure A9 illustrates the percentage distribution of migrant inventors to the US whose first event in the US was a patent. The distribution appears quite different from the one observed for education

migrants. Only 16% of inventors of the inventor who migrated in this way did so in the decade 1990-1999, whereas 68% of them did it in the decade 2000-2009. Finally, Figure A10 shows the percentage distribution of migrant inventors to the US whose first observable event was an employment. Keeping in mind that the number of inventors in this subset is lower than in the other two cases, one can notice a sort of cyclical pattern. A first peak is observed in the years from 1997 to 2001 (corresponding to the development of the dot com economy), whereas a second peak is observed in the years 2011 and 2012.

Figure A11 reports the percentage distribution of the 3,943 migrants for education motives by age at migration. Around 84% of all migrant inventors for education motives had an age at migration comprised between 23 and 27, suggesting that the vast majority of those who moved to the US for this reason went there to attain either a MSc or a PhD.

Figure A8: Percentage distribution of «education migrants to the US» by migration year (3,943 obs.)

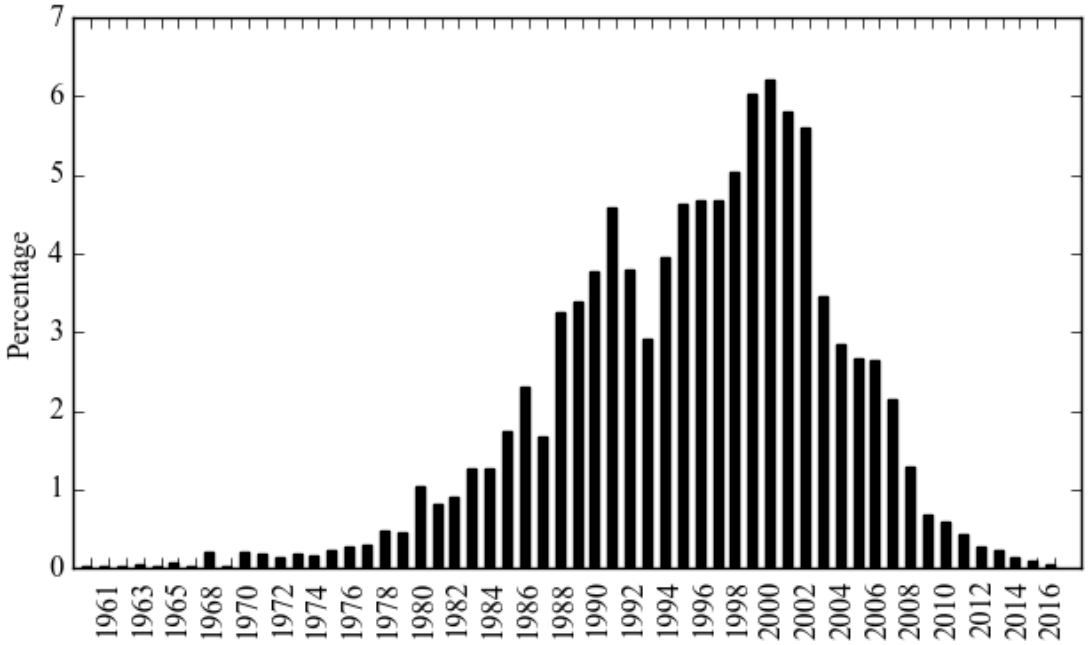


Figure A9: Percentage distribution of «work migrants to the US» whose first event in the US was a patent by migration year (1,280 obs.)

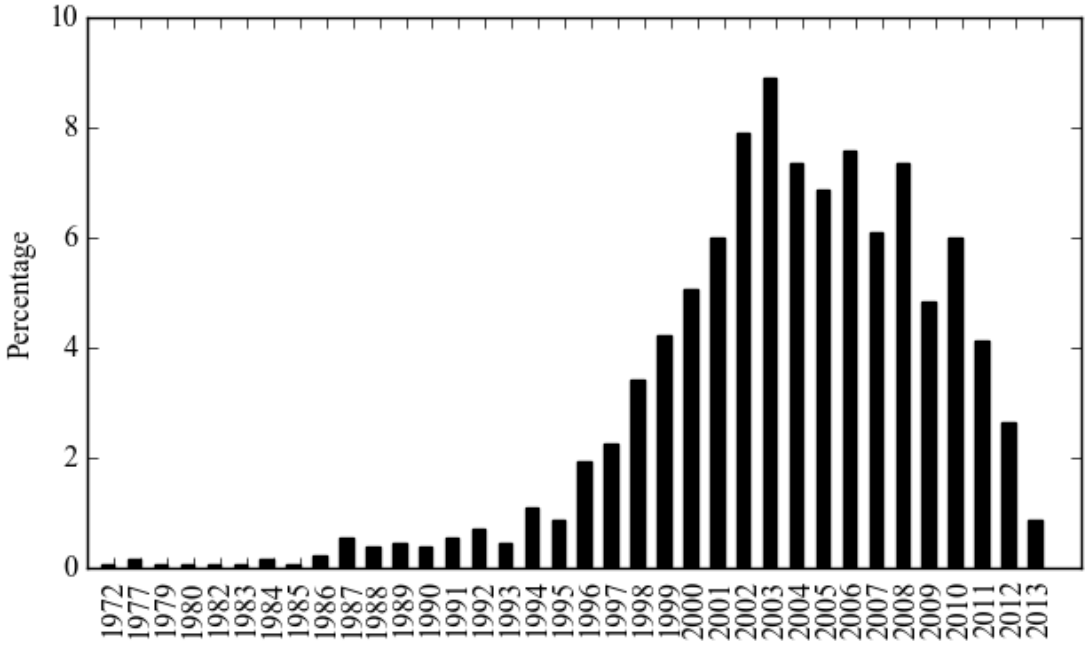


Figure A10: Percentage distribution of «work migrants to the US» whose first event in the US was an employment by migration year (309 obs.)

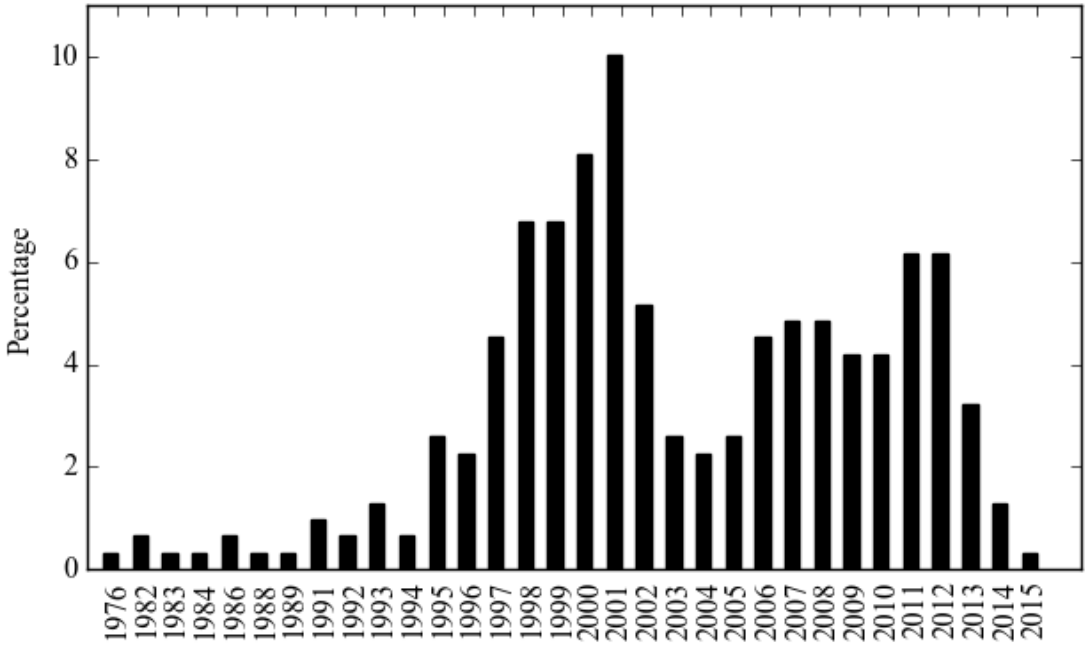
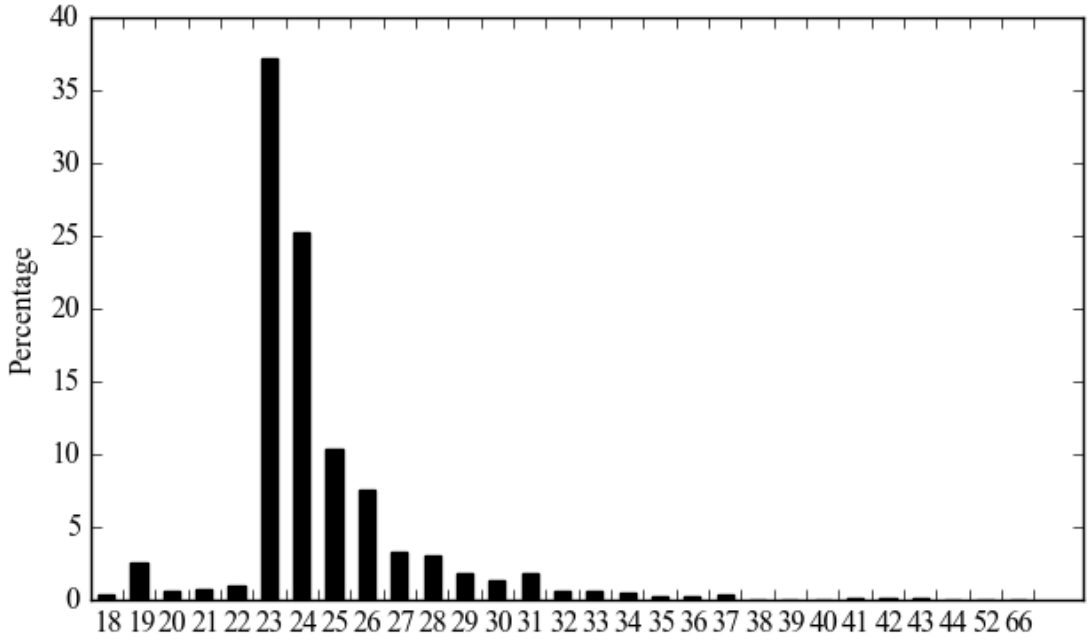


Figure A11: Percentage distribution of «education migrants to the US» by age at migration (3,943 obs.)



Note that, given the way in which we coded the year of birth, the age at migration should in principle display only three values, i.e. 19, 23, 25 and 27, depending on the type of education level used for the estimation. Yet, due to misreporting errors in the data, the age at migration is not necessarily the same for all inventors who moved to the US to attain a certain education level. To illustrate this issue, consider the case of the inventor shown in Table A8. The resume reported three educational attainments. For each of them, the resume reported the start but not the end year. As the starting date of the BSc in India was in 1988, we accordingly estimated that the inventor was born in 1969 (i.e. =1988-19). Yet, the resume also reports that the inventor started a MSc degree programme in the US in 1990. Hence, the inventor's migration year was set equal to 1990 and her age at migration was equal to 21 (i.e =1990-1969), whereas the age at the start of the MSc is in general equal to 23.

Table A8: Inventor D, migrant to the US for education motives

University	Start year	End year	Degree
Indian Institute of Technology, Kharagpur	1988	.	B.Tech. (Honors) in Computer Science and Engineering
The University of Texas at Austin	1990	.	M.S. in Computer Science
The University of Texas at Austin	1995	.	Ph.D. in Computer Science

Despite all our efforts to carefully clean and check raw data, a few errors, inconsistencies and more generally noise are still present in the data. Notwithstanding this, we believe that the general pattern reported in Figure A11 is reassuring about the quality of the data used in our analysis. Moreover, some deviations from the general pattern might be due to genuine devitations from the typical educational pattern. This is particularly the case of inventors whose age at migration is greater than 27. Consider for example the inventor reported in Table A9. The inventor started a BSc in India in

1949 and her birth year was accordingly estimated as 1930 (i.e. =1949-19). In 1963, when she was 33 years old, she started (i.e. migrated to) a PhD in the US.

Table A9: Inventor E, migrant to the US for education motives

University	Start year	End year	Degree
University of Lucknow	1949	1952	B.Sc. Physics
Brooklyn Polytechnic	1963	1968	Ph.D. Electrical Engineering

Figure A12 reports the age at migration for the work migrants to the US. Not surprisingly, we observe a substantial difference with the distribution of age at migration of inventors who migrated for education reasons. Work migrants tend to be significantly older than education migrants at the time of migration. In comparing the two distributions, however, one should keep in mind that our ability to estimate the year of birth based on educational attainment was lower for work migrants than for education migrants. For a substantial fraction of the latter, we had to estimate the year of birth as the average age at the time of the first patent (see discussion above and Figure A2), which is necessarily a rather crude estimate. Out of 3,943 education migrants to the US we estimated age on the basis of the average age at the time of the first patent for 35 inventors (i.e. less than 1%). On the other hand, out of 1,589 work migrants to the US we had to estimate age on the basis of the average age at the time of the first patent for 194 inventors (i.e. about 12% of them).

As a robustness check, Figure A13 plots the percentage distribution of work migrants by age at migration, excluding the 194 inventors for whoem age was estimated as the average age at the time of the first patent. Once again, we observe that the modal value is at an age of 32 and that the distribution appears concentrated on older ages than the distribution of education migrants.

Figure A12: Percentage distribution of «work migrants to the US» by age at migration (1,589 obs.)

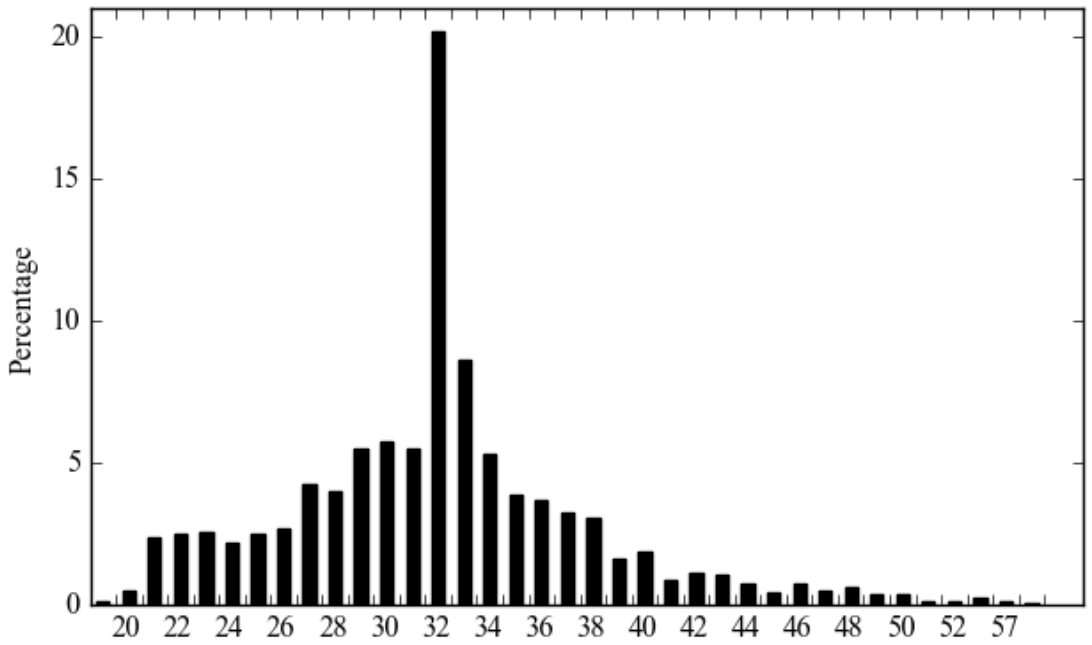
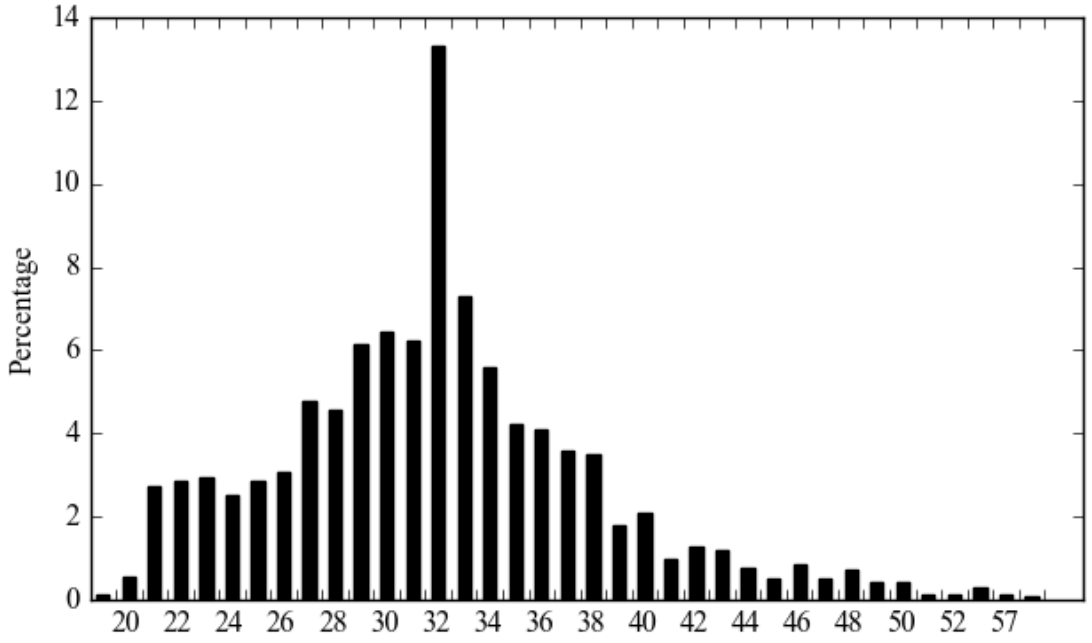


Figure A13: Percentage distribution of «work migrants to the US» by age at migration excluding inventors for whom age was estimated as average age at the first patent (1,395 obs.)



N. Coding return migration and return year

The final methdological step consisted of coding the events of return migration. To this purpose, we exploited again information on the location of the three types of activities included in our data, i.e. education, patenting, and employment. For each of the 5,532 «migrants to the US», we identified «returnees to India» by looking at their career path. An inventor was defined as a «returnee to India» when she either made a patent, attained education or reported a job located in India in a year *following the one of migration to the US*. Return year was set equal to the date of the first event (if any) taking place in India after migration. For example, for an inventor who migrated to the US in 1990 and subsequently made a patent in India in 1995, the return year was set equal to 1995.

O. Indian-born and second-generation Indians

As explained above, we defined Indian-born inventors or «potential migrants» as those inventors who met the following conditions:

1) given and family names were classified as having an Indian-origin, and
2) *at any time* in their career, the inventors either made a patent, were educated or «self reported» a job location in India.

The condition that the inventors had to show some experience in India *at any time* during their career might introduce some false positives in the sample of Indian-born «potential migrants». For instance, consider the case of a second-generation Indian inventor born and educated in the US, who at some point starts working or patenting in India. This individual will be considered as a potential migrant, whereas in fact she is not. Although this is arguably a relatively uncommon case, we cannot completely rule out it. A potential solution to this issue would be to include in our sample only inventors who attained a BSc in India. The problem with this solution is that many Indian-born

inventors *do not* report information in the resume on the BSc (and thus where this has been attained), including only information on the MSc or higher degree often attained in the US. According to the logic explained above, we should treat them as second-generation migrants and exclude them from the pool of potential migrants, whereas in fact they are. In other words, we would generate many false negatives reducing our sample size.

Rather than trying to solve the issue, we keep our definition and in this section we show that the concern illustrated above is likely to be rather limited. To this purpose, we focus attention on the 5,532 migrants to the US. For each of them, we attempted to reconstruct their career path *before* migrating to the US. In particular, we assessed to what extent the sample of migrants to the US includes inventors who were *active* in some way in India *before* the year of migration to the US. Specifically, for each inventor we recorded the year of the *first event* (i.e. education, patent or a self-reported employment) in India and we compared it with the year of migration to the US. To the extent that the first event in India was preceding the year of migration, we can exclude that the inventor is a second-generation Indian who at some point returned to India. Out of 5,532 migrants to the US 5,230 (i.e. 94.5% of all migrants) were active in India in the sense specified above (i.e. they either made a patent, attained education or had a job) *before* the migration event.

Note that one should not consider the other 302 inventors, for which we have no trace of educational or professional activity in India *before* the migration, as false positives, namely second-generation Indian inventors who went to India for professional reasons. Rather, the majority of them are likely to be genuine migrants, who simply did not record in their resume any experience in India made *before* the choice of migration. The case discussed above of the inventor attaining a BSc in India, without reporting it in the resume, and recording only the MSc or the PhD attained in the US fits this picture.

To dig more into this problem, we further split the 302 potential migrants without any trace of experience in India *before* migration: of them, 122 are returnees to India, while 180 are inventors who attained in India education for which we could not define the start and end dates, which provide the crucial information to define *before* migration events. Of the 180 inventors who attained education in India at some unknown date, 4 attained a BSc, 17 a MSc, 4 a PhD, and 158 other unclassified education titles[29]. Moreover, a casual inspection reveals that most of the unclassified education titles relate to secondary school or college level education. Of the 122 returning inventors, 7 attained a BSc, 13 a MSc, 2 a PhD, and 33 other unclassified education titles in India at some unknown date. Overall, of the 120 returning inventors, 49 got some education in India at some unknown date. As it is quite reasonable to assume that second-generation US born Indian inventors are unlikely to go and get any education in India, it is likely that 229 (=180+49) out of 302 potential false positives are actually genuine migrants. Following this logic, the potential problem of having false positives in the sample of potential migrants is restricted to only 73 individuals, i.e. around 1.3% of all migrants to the US.

[29] Please note that the sum is greater than 180 since some inventors reported multiple educational attainments.

List of US public ICT companies used in the paper

IDX	Company name	IDX	Company name
0	3COM CORP	89	JUNIPER NETWORKS INC
1	ACTEL CORP	90	L-3 COMMUNICATIONS HLDGS INC
2	ADC TELECOMMUNICATIONS INC	91	LATTICE SEMICONDUCTOR CORP
3	ADOBE SYSTEMS INC	92	LEVEL 3 COMMUNICATIONS INC
4	ADTRAN INC	93	LEXMARK INTL INC -CL A
5	ADVANCED MICRO DEVICES	94	LINEAR TECHNOLOGY CORP
6	AFFYMETRIX INC	95	LORAL SPACE & COMMUNICATIONS
7	AGERE SYSTEMS INC	96	LSI CORP
8	AGILENT TECHNOLOGIES INC	97	LUCENT TECHNOLOGIES INC
9	AKAMAI TECHNOLOGIES INC	98	MAXIM INTEGRATED PRODUCTS
10	ALTERA CORP	99	MAXTOR CORP
11	AMETEK INC	100	MCI INC
12	AMKOR TECHNOLOGY INC	101	MENTOR GRAPHICS CORP
13	AMPHENOL CORP	102	METHODE ELECTRONICS -CL A
14	ANALOG DEVICES	103	METROLOGIC INSTRUMENTS INC
15	APPLE INC	104	MICREL INC
16	APPLIED MICRO CIRCUITS CORP	105	MICROCHIP TECHNOLOGY INC
17	ARRIS GROUP INC	106	MICRON TECHNOLOGY INC
18	AT&T CORP	107	MICROSEMI CORP
19	AT&T INC	108	MICROSOFT CORP
20	ATHEROS COMMUNICATIONS INC	109	MICROVISION INC
21	ATI TECHNOLOGIES INC	110	MINDSPEED TECHNOLOGIES INC
22	ATMEL CORP	111	MITEL NETWORKS CORP
23	AUTODESK INC	112	MKS INSTRUMENTS INC
24	AVANEX CORP	113	MOLEX INC
25	AVAYA INC	114	MONOLITHIC POWER SYSTEMS INC
26	BEA SYSTEMS INC	115	MOTOROLA INC
27	BECKMAN COULTER INC	116	NATIONAL INSTRUMENTS CORP
28	BELL & HOWELL OPERATING CO	117	NATIONAL SEMICONDUCTOR CORP
29	BELLSOUTH CORP	118	NCR CORP
30	BIO-RAD LABORATORIES INC	119	NETLOGIC MICROSYSTEMS INC
31	BMC SOFTWARE INC	120	NETWORK APPLIANCE INC
32	BROADCOM CORP -CL A	121	NETWORKS ASSOCIATES
33	BROCADE COMMUNICATIONS SYS	122	NOVELL INC
34	CA INC	123	NUANCE COMMUNICATIONS INC
35	CADENCE DESIGN SYSTEMS INC	124	NVIDIA CORP
36	CASCADE MICROTECH INC	125	OMNIVISION TECHNOLOGIES INC
37	CERTICOM CORP	126	ORACLE CORP
38	CIENA CORP	127	PITNEY BOWES INC
39	CIRRUS LOGIC INC	128	PLANTRONICS INC
40	CISCO SYSTEMS INC	129	PMC-SIERRA INC
41	CITRIX SYSTEMS INC	130	POLYCOM INC
42	COGNEX CORP	131	POWER INTEGRATIONS INC
43	COHERENT INC	132	QLOGIC CORP
44	COMMVAULT SYSTEMS INC	133	QUALCOMM INC
45	CONEXANT SYSTEMS INC	134	QUANTUM CORP
46	CORNING INC	135	QWEST COMMUNICATION INTL INC
47	CREDENCE SYSTEMS CORP	136	READ-RITE CORP
48	CREE INC	137	RED HAT INC
49	CYPRESS SEMICONDUCTOR CORP	138	RESEARCH IN MOTION LTD
50	DALLAS SEMICONDUCTOR CORP	139	ROGERS CORP
51	DELL INC	140	SANDISK CORP
52	DIEBOLD INC	141	SCIENTIFIC-ATLANTA INC
53	DIGIMARC CORP	142	SEAGATE TECHNOLOGY
54	DIRECTV GROUP INC	143	SENSORMATIC ELECTRONICS

55	EBAY INC
56	ECHOSTAR CORP
57	ELECTRONIC DATA SYSTEMS CORP
58	ELECTRONICS FOR IMAGING INC
59	EMC CORP/MA
60	EMULEX CORP
61	EXTREME NETWORKS INC
62	F5 NETWORKS INC
63	FAIRCHILD SEMICONDUCTOR INTL
64	FEI CO
65	FINISAR CORP
66	FIRST DATA CORP
67	FORMFACTOR INC
68	FOUNDRY NETWORKS INC
69	FREESCALE SEMICONDUCTOR INC
70	GATEWAY INC
71	GENESYS TELECOMM LABS INC
72	GOOGLE INC
73	HARMAN INTL INDUSTRIES INC
74	HARRIS CORP
75	HEWLETT-PACKARD CO
76	HUTCHINSON TECHNOLOGY INC
77	I2 TECHNOLOGIES INC
78	IMMERSION CORP
79	INFINERA CORP
80	INTEGRATED DEVICE TECH INC
81	INTEL CORP
82	INTERMEC INC
83	INTERSIL CORP
84	INTL BUSINESS MACHINES CORP
85	INTL RECTIFIER CORP
86	INTUIT INC
87	IOMEGA CORP
88	IXYS CORP
144	SIGMATEL INC
145	SILICON GRAPHICS INC
146	SILICON IMAGE INC
147	SILICON LABORATORIES INC
148	SILICON STORAGE TECHNOLOGY
149	SILICONIX INC
150	SKYWORKS SOLUTIONS INC
151	SPANSION INC
152	STANDARD MICROSYSTEMS CORP
153	STORAGE TECHNOLOGY CP
154	SUN MICROSYSTEMS INC
155	SYBASE INC
156	SYMANTEC CORP
157	SYMBOL TECHNOLOGIES
158	SYMYX TECHNOLOGIES INC
159	SYNAPTICS INC
160	SYNOPSYS INC
161	TEKTRONIX INC
162	TELECOMMUNICATION SYS INC
163	TELLABS INC
164	TERADYNE INC
165	TEXAS INSTRUMENTS INC
166	TRIQUINT SEMICONDUCTOR INC
167	UNISYS CORP
168	UNIVERSAL DISPLAY CORP
169	UNIVERSAL ELECTRONICS INC
170	VARIAN INC
171	VIASAT INC
172	WESTERN DIGITAL CORP
173	WORLDCOM INC-CONSOLIDATED
174	XEROX CORP
175	XILINX INC
176	YAHOO INC
177	ZILOG INC
178	ZORAN CORP