Foreign-origin inventors in the USA: Testing for Diaspora and Brain Gain Effects

Stefano Breschi
CRIOS – Università Bocconi, Milan

Francesco Lissoni
GREThA CNRS UMR 5113, Université de Bordeaux
CRIOS – Università Bocconi, Milan

Ernest Miguelez
GREThA CNRS UMR 5113, Université de Bordeaux
AQR-IREA, University of Barcelona

contact author: francesco.lissoni@u-bordeaux.fr

This is a pre-print of an article published in *Journal of Economic Geography*. The definitive publisher-authenticated version: Breschi, S., Lissoni, F., Miguelez, E. (2017) "Foreign-origin inventors in the USA: testing for diaspora and brain gain effects", *Journal of Economic Geography*, 17(5): 1009–1038. DOI: https://doi.org/10.1093/jeg/lbw044

Abstract : We assess the role of ethnic ties in the diffusion of technical knowledge using a database of patents filed by US-resident inventors of foreign origin, identified by name analysis. We consider 10 leading source countries, both Asian and European, of highly skilled migration to the USA and test whether foreign inventors' patents are disproportionately cited by (i) co-ethnic migrants ('diaspora' effect), and (ii) inventors residing in their country of origin ('brain gain' effect). We find evidence of the diaspora effect for the Asian but not the European countries, with the exception of Russia. A diaspora effect does not necessarily translate into a brain gain effect, most notably for India where no such effect is detected. Neither does a brain gain effect occur solely in conjunction with a diaspora effect. Overall, diaspora and brain gain effects carry less weight than other channels of knowledge transmission, most notably co-invention networks and multinational companies.

Acknowledgements: Unique identifiers for inventors in the EP-INV database come from the APE-INV project (Academic Patenting in Europe), funded by the European Science Foundation. The pilot project for assigning inventors to specific countries of origin was funded by the World Intellectual Property Organization (WIPO), which also made available to us the WIPO-PCT dataset. We received useful suggestions by participants to the following conferences : MEIDE (Santiago de Chile, November 2013), PATSTAT (Rio de Janeiro, November 2013), EUROLIO (Utrecht, January 2014), EPIP (Brussels, September 2014), AAG (Chicago, April 2015) and "Migration & Development" (Washington, May 2015); as well by participants to seminars at University College Dublin, London School of Economics, CRIOS-Bocconi, Kassel University, Collegio Carlo Alberto (Turin), IMT (Lucca), LUISS (Rome), GREThA-Bordeaux, UC Davis and UC Berkeley. Gianluca Tarasconi contributed decisively to the creation of the Ethnic-Inv dataset. Diego Useche provided valuable research assistance. We owe the tip on the IBM-GNR system to Lars Bo Jeppesen, while Curt Baginski assisted us in its implementation. Lissoni and Miguelez acknowledge financial support from the Regional Council of Aquitaine (Chaire d'Accueil programme and PROXIMO project).

1. Introduction

Recent research on the international mobility of scientists and engineers has seen the convergence of two streams in the literature. First, studies of the geography of innovation have explored the role of social ties in facilitating knowledge diffusion and in determining their spatial reach, including, in the case of migrant and foreign-origin inventors, the ties established with other members of their ethnic community (Agrawal et al. 2008). Second, migration and development studies have explored the extent to which highly skilled migrants contribute to innovation in their home countries through international knowledge flows (Kapur 2001, Kerr 2008, Agrawal et al. 2011), foreign direct investment (Foley and Kerr 2013), and entrepreneurial returnee migration (Nanda and Khanna 2010, Saxenian 2006).

While this convergence in the literature has enabled major advances, there is a clear need to consolidate the research field. We do not know, for example, the extent to which the social ties between migrants in the countries of destination extend to their countries of origin and thus contribute to international knowledge transfer. We also have little understanding of the differences across migration corridors. Here, existing studies focus primarily on what, in recent years, have been the fastest growing corridors (from China and India to the US) and tend to overlook the importance of Europe, not just as a region of destination, but also of origin. As of 2010/11, the top 10 contributors to the stock of highly educated migrants to OECD countries included the UK, Germany, and Poland. With over 3.6 million people, the combined stock of these three countries was 60% higher than that of India (top of the ranking) and more than twice that of China (third in the ranking, after the Philippines) (OECD 2015). According to the same source, Germany, Italy, and France have greater highly skilled emigration rates than China and India (between 6 and 9% compared to less than 5%), while the rates for the UK, Poland, and Romania stand at 11, 17, and 21%, respectively (Mayr and Peri 2008, Schiff and Wang 2013).

Finally, official statistics do not provide details on the specific skills or jobs of the highly educated. Hence, the gathering of micro-evidence at the cross-country level is essential.

We contribute to this emerging literature by producing and analysing an extensive dataset of foreign-origin, US-based inventors from five Asian (China, India, Iran, Japan, and South Korea) and five European countries

(France, Germany, Italy, Poland, and Russia). All data are novel and come from EP-INV, a database of uniquely identified inventors listed on patent applications at the European Patent Office (EPO), combined with name analysis based upon IBM-GNR©, a commercial database. Foreign-origin inventors include both foreign nationals and first or subsequent generation migrants who have acquired the US nationality, but still may contribute to knowledge diffusion, based on ethnic affiliation.

We analyse the knowledge flows generated by these inventors, as measured by forward citations to their patent applications. We state a "diaspora effect" to exist when US-resident inventors of the same foreign origin have a higher propensity to cite one another's patents, compared to patents by other inventors, other things being equal. We state a "brain gain effect" to exist when US-resident foreign-origin inventors are disproportionately cited by inventors in their home countries, so that the latter stand to gain from high skilled migration. We find evidence of a diaspora effect for Asian inventors, but not for their European counterparts, with the exception of Russian and (to a much lesser extent) German inventors. In general, ethnic ties appear to act as substitutes of co-location at the city level and of proximity in the social network of inventors. Their marginal effect does not appear to be as large as those of city-level co-location and short social distance.

In the cases of China, India and Russia, the diaspora effect presents an international dimension, as migrant inventors in countries other than the US also enjoy privileged access to knowledge produced by co-ethnic, US-based inventors. This, in turn, translates into a brain gain for China and Russia, though not for India. South Korea also presents a brain gain effect from its US-resident inventors. In the case of the advanced countries of France, Italy, and Japan, the brain gain effect is channelled through multinational enterprises. We detect no brain gain effect for Germany.

In what follows, we first survey the literature on migration and knowledge flows, with special emphasis on patent-based studies (section 2). We then present our research questions and data (section 3) and our results for the diaspora (section 4) and brain gain effects (section 5). Section 6 concludes. Appendixes discussing methodological issues and presenting robustness checks are available as additional online material.

2. Background literature

2.1 Localized knowledge flows and the role of social ties

Localized knowledge flows are a central concept in the geography of innovation (Breschi 2011). In the form of pure externalities, they are present in both Marshallian and Jacobian location theories (Henderson 1997, Ellison et al. 2010). Yet, their importance has been questioned both by New Economic Geography models (Krugman, 1991 and 2011) and by evolutionary location theories (Boschma and Frenken 2011). A key point of contention in this debate has been that of their measurement, fraught with technical and conceptual difficulties.

These difficulties were first addressed by Jaffe et al. (1993), who introduced the use of patent citations along with a simple, yet influential, methodology (from now on, the JTH method). The method makes use of two sets of patent pairs. The first includes a sample of cited patents and all their corresponding citing patents, excluding self-citations at the company level (cited-citing or "case" pairs); the second includes the same sample of cited patents, but here the citing ones are replaced by controls that have the same technological classification and priority year (cited-control or "control" pairs). After geo-localising patents at the city, state, or country level, a simple test of proportions is conducted to demonstrate that the proportion of co-localized cases is significantly higher than that of co-localized controls. The method can be generalized by means of regression analysis, with the probability of a citation occurring as the dependent variable, and the stacked sets of cited-citing and cited-control patent pairs as observations (Singh and Marx 2013). Subsequent technical refinements of the JTH method have involved the level of detail chosen for the technological classification of patents (Thompson and Fox-Kean, 2005; Henderson et al., 2005) and the origin of patent citations (applicant vs. examiner; Thompson, 2006; for a critical discussion of this approach, see Alcacer and Gittelman, 2006; Breschi and Lissoni, 2005b). (*19-23*)

Later studies have modified the JTH method as they seek to identify the actual mechanisms underpinning localized knowledge flows and their economic characteristics. Breschi and Lissoni (2005a, 2009) show that a large proportion of localized patent citations are self-citations at the individual level, associated with inventors that move between or consult across firms in the same location or region. Other localized citations occur between individuals located at short geodesic distances in co-inventorship networks. Agrawal et al. (2006) show that social ties of this kind, once established locally, may be resistant to physical distance.

Subsequent studies have sought to uncover other forms of social ties besides those established through collaboration. Agrawal et al. (2008) assess the importance of ethnic ties in the US-resident Indian community, described as a close-knit diaspora (Kapur 2001). Drawing on a database of Indian surnames, the authors identify a large number of ethnic Indian, US-resident inventors of patents issued by the US Patent and Trademark Office (USPTO). They then extend the JTH method by including inventors' co-ethnicity among their explanatory variables. Indian inventors are found to be more likely to cite one another's patents than patents by non-Indians. Moreover, co-ethnicity and co-location seem to act as substitutes, with Indian inventors activating their ethnic connections to reach beyond their metropolitan area. Likewise, Almeida et al. (2015) find evidence for Indian inventors that excessive reliance on ~~an~~ intra-ethnic citations is associated to lower inventive productivity.

Agrawal et al. (2011) extend Agrawal et al.'s (2008) data and methodology to the case of international knowledge flows, based on a sample of USPTO patents issued to inventors based in India, and their backward citations. Physical proximity (co-location in India) appears to exert a much stronger effect than co-ethnicity, suggesting that the Indian diaspora is not a major knowledge source for the home country. This conclusion, however, does not hold for patents in Computers and Communications, patents owned by multinational firms, and for very important (highly cited) patents. It is at this juncture that studies in the geography of innovation converge with research on migration and development.

2.2 *Migrant contribution to innovation in countries of origin*

Migration studies have traditionally sought to identify potentially positive returns on emigration for countries of origin. Early research focused specifically on remittances; recently, the consequences of highly skilled migration on knowledge transfer have taken centre stage (Bhagwati and Hanson 2009). This takes three, non-mutually exclusive, forms:

(i) *"Ethnic-driven" knowledge flows.* Migrant scientists and engineers may retain social contacts with professional associations and educational institutions in their home countries, and transmit scientific and technical skills either on a friendly or contractual basis (Meyer and Brown 1999, Meyer 2001);

(ii) *Transfers by multinational companies*, due either to internal mobility or collaboration (Blomström and Kokko 1998, Veugelers and Cassiman 2004, Branstetter et al. 2015);

(iii) *Returnees' direct contributions.* Highly skilled migrants may decide to move back to their home countries or set up entrepreneurial activities there, while maintaining contact with knowledge sources in their countries of destination (Wadhwa, 2007, and references therein; for a critique: Kenney et al. 2013).

While case studies on these phenomena abound, large-scale quantitative evidence is scant and almost entirely focused on the US as destination and on China and India as origins. This ignores the fact that highly skilled migration to the US also originates from Western Europe, South Korea, and Japan (Docquier and Marfouk 2006, Widmaier and Dumont 2011; see also Freeman, 2010).

Some progress has been made by William Kerr and co-authors, who combine information from the USPTO patent database with that from the Melissa database, a commercial repository of names and surnames of US residents, classified according to nine broad ethnic groups. In the case of *"ethnic-driven" knowledge flows*, Kerr (2008) examines foreign-based citations to patents filed by US-resident inventors (excluding company self-citations). The citations are distributed over ~100k cells, each cell consisting of two-plus-two dimensions: the ethnicity of the citing foreign inventor and that of the cited US inventor; and, the technology class of the citing and cited patents. Co-ethnic cells exhibit a higher citation count than mixed cells, controlling for technology. This outcome points to a brain gain effect.

As for transfers by multinational companies, Foley and Kerr (2013) investigate the specific role of ethnic inventors in the US in relation to the activities of US multinational companies in their home countries. Migrant inventors are found to act as substitutes of local intermediaries, thus reducing their companies' costs of entering foreign markets. Using information about the nationality of inventors listed on Patent Cooperation Treaty (PCT) patents, Miguelez (2016) estimates the impact of migrant inventors on the extent of international technological collaborations between countries of origin and destination, as measured by co-patenting activity. His findings suggest a positive and significant impact for all countries of origin, not just China and India. Similarly, Branstetter et al. (2015) find that foreign multinationals are responsible for the majority of USPTOpatents filed from India and China, signed or co-signed by local inventors. Although these

patents are of lower quality than those produced at home by the same multinationals, the quality gap is narrowing in the case of China, which is indicative of effective knowledge transfer. The same does not apply to India.

As far as *returnees' direct contributions* are concerned, Agrawal et al. (2011) and Alnuaimi et al. (2012), based on studies of Indian inventors, manage to identify only a handful of returnees, suggesting that in the case of India, at least, these are not a massive source of knowledge transfer. Choudhury (2016), drawing on data from the Indian R&D facilities of one US multinational, finds that the most inventive employees are those working under returnee managers. This may be indicative of the latter's role as knowledge brokers between headquarters and subsidiaries.

2.3 Data issues

The increasing availability of inventor data has led several scholars to improve the quality and transparency of their data mining efforts. A key issue here is that of name disambiguation, that is, assigning a unique ID to inventors whose name or address might be reported differently on several patent documents (Marx et al. 2009, Raffo and Lhuillery 2009, Martínez et al. 2013, Li et al. 2014, Ventura et al. 2015, Pezzoni et al. 2014, Ge et al. 2016). This has important implications for migration studies (more details in Appendix 1).

Ideally, a good disambiguation algorithm should minimize both false negatives (maximize "recall") and false positives (maximize "precision"). In practice, a trade-off exists, with high recall being much harder to achieve than high precision. High precision/low recall algorithms underestimate the number of personal self-citations and overestimate co-ethnic citations, as one self-citing ethnic inventor might be mistaken for two co-ethnic inventors citing one other. This latter bias can vary according to the inventors' country of origin, as disambiguation algorithms are language-sensitive.

To date, patent-based studies of migration and innovation have ignored these issues. Kerr (2008) and extensions employ non-disambiguated data; Agrawal et al. (2008, 2011) and Almeida et al. (2015) provide no details on disambiguation; and Alnuaimi et al. (2012) resort to "perfect matching", which functions as an extreme high precision/low recall algorithm.

Issues of precision and recall also emerge when assigning inventors to a country of origin or ethnic group based on names/surnames. Agrawal et al. (2008), for example, identify Indian inventors using a very narrow list of Indian surnames, considered as being both highly frequent in India and indicative of recent migration status. This, however, tends to limit attention to first-generation migrants and to assume that the strength of ethnic ties weakens with time. While this might be true, the assumption is not precise about the generational timing of this decay and it ignores the possibility of "ethnic revival" and "reverse brain drain" effects (Kuznetsov 2010, Kuznetsov 2006, Zweig 2006). Information on inventors' nationality, as used by Miguelez (2016), is an extremely practical substitute of name analysis, but also constitutes a low recall algorithm (established migrants that acquire the host country's nationality turn out as false negatives).

Technical concerns also arise with patent applicants. All studies claim to control for company self-citations; yet they remain silent on the methodologies adopted to identify companies and business groups. This contrasts with recent harmonization efforts (Peeters et al. 2010, Du Plessis et al. 2009, Thoma et al. 2010). The use of raw or poorly treated applicant data is equivalent to applying a high precision/low recall disambiguation technique and leads to underestimation of company self-citations and overestimation of knowledge externalities. Internationally, it undervalues the role of multinationals as carriers of knowledge, and overvalues that of inventors' social ties.

3. Research questions and data

Below, we formulate our research questions and describe our dataset, keeping complexity to a minimum (details in Appendixes 1 and 2).

3.1 Research questions: diaspora and brain gain effects

We are interested in exploring how membership in the same foreign-origin community affects the diffusion of technical knowledge, both within the country of destination (CoD) and towards the country of origin (CoO). Emerging naming conventions, as reviewed in section 2, refer to within-community ties as "ethnic" or "co-ethnic" – imperfect terms that we nevertheless also adopt (for want of a better alternative). However, when referring to individual inventors, we opt for "foreign-origin inventors", "inventors from the

same country of origin" (both expressions including second- and subsequent-generation migrants) or, where more appropriate, "migrant inventors".

Ethnic ties exist independently of professional experiences and/or physical proximity. They may have been forged in the CoD (reflecting homophilic tendencies; Currarini et al., 2009) or inherited from the home country (as in chain migration). In both cases, they represent an instance of vitality and relevance of a foreign-origin community, to which we will refer as a diaspora.[1]

We state a diaspora effect to exist when inventors from the same CoO and active in the same CoD have a higher propensity to cite one another's patents than those of other inventors, ceteris paribus, and excluding self-citations at the company level. We test for the effect by adapting the JTH method (see section 2). We consider all cited patents signed by at least one foreign-origin inventor in the US, with citing and control patents having been filed by inventors (both local and of foreign origin) also located in the US. We then estimate the simple model:

$$Probability\ of\ citation = f(co-ethnicity;\ spatial\ distance; social\ distance;\ controls) \qquad (1)$$

where the observations are patent pairs and the binary dependent variable takes value one if the two patents in the pair are linked by a citation. The main variable of interest, *co-ethnicity*, is a dummy variable equal to one when both patents in the pair have been invented by at least one inventor from the same CoO. Spatial distance is determined on the basis of the inventors' addresses and measured both in terms of co-location and as a continuous variable. Social distance refers to geodesic distances in the network of inventors (Breschi and Lissoni, 2009). When one or both patents in a pair have multiple inventors, we consider minimum social and spatial distances. The other regressors refer primarily to the characteristics of the patents in the pair (in particular, the citing/control patents), based on the considerable body of literature examining the determinants of patent citations (Harhoff et al. 2003, Hall et al. 2005). We provide full details of our sampling scheme and specification in the next two subsections. We also conduct various robustness

[1] "Diaspora" is also a somewhat imperfect term, used here to conform to current conventions. Dufoix (2008) shows how the term has progressively lost its original meaning in reference to Jewish history (the emphasis being on the absence of a home country) and is now used when speaking of any widely dispersed ethnic community (often in reference to its ties with the home country). In the economics of migration, the term is used even more casually, simply to indicate any stock of migrants (Beine et al., 2011).

checks. This includes replacing the JTH citing-control patent methodology with one derived by Thompson (2006), which consists in making use of inventor-added citations as cases and of examiner-added ones as controls, based on the assumption that the former, albeit noisy, may be more revealing of direct knowledge exchanges between inventors.

We state a brain gain effect to exist when patents by foreign-resident inventors from a given CoO are disproportionally cited by home-resident inventors (inventors residing in the same CoO). We consider citations mediated by ethnic ties separately from other citation sources, including self-citations of returnee inventors and multinational companies. To do so, we adapt once again the JTH method.

We sample all cited patents signed by foreign-origin inventors in the US, and as citing and control patents we select only those signed by inventors residing outside the US. We retain all patent pairs by the same inventor (most likely a returnee inventor) as well as pairs from the same company or business group, but control for them. We then estimate the following regression, by modifying equation (1) as follows:

$$Probability\ of\ citation = \\ = f(home\ country;\ returnee;\ same\ company; spatial\ \&\ social\ distance;\ controls) \quad (2)$$

The dependent variable is the same as in (1), but the main regressor of interest is now *home country*, a dummy variable that takes value one if at least one inventor of the citing (control) patent resides in the cited inventor's CoO. *Returnee* and *Same company* are also dummy variables, which take value one if both patents in the pair have been signed by the same inventor, back in his CoO, or filed by the same company or business group, respectively. Other controls are as in (1), with some adaptations.[2]

Notice that countries with strong education systems, but limited inventive activity of international standing (such as India, Russia, and China), may have fewer inventors of local origin at home than abroad. This suggests the possibility of some intra-ethnic global knowledge flows to exist, similar to trade flows between countries hosting the same ethnic minorities (Rauch and Trindade 2002, Felbermayr et al. 2010). We will refer

[2] Most notably, spatial distance cannot be measured with co-location dummies, since, by construction, the inventors of cited and citing patents do not reside in the same country. Notice that networks of inventors may span across countries, which justifies including social distance in (2). Personal self-citations may occur (social distance =0) as when an Indian returnee inventor cites his own prior art, which he filed when abroad. Still, these are rare cases. Even more rare is the case of a migrant inventor who does not return, but move to a new CoD, and cite his own prior art from there.

to this as an "international diaspora" effect and test for it by re-inserting the co-ethnicity dummy in the regression.

3.2 Data

3.2.1 Patent and inventor data

Our data result from matching the names and surnames of inventors in the EP-INV inventor database (Coffano and Tarasconi, 2014) with information obtained by Global Name Recognition, a name search technology produced by IBM (from now on, IBM-GNR©). We track all inventors residing in the US, from one of the following CoO: China, India, Iran, Japan, and South Korea (for Asia); and France, Germany, Italy, Poland, and Russia (for Europe). These countries figure among the top 20 sources of highly skilled migrants to the US according to OECD/DIOC data, release 2005/6 (Widmaier and Dumont 2011). Moreover, none of these countries has English or Spanish as official languages, which are the most widely spoken languages in the US and complicate the name analysis exercise.

The EP-INV database contains information on uniquely identified inventors listed on patent applications filed at the EPO from 1978 to around 2014. Disambiguation is performed by means of the *Massacrator* algorithm, introduced by Breschi and Lissoni (2009) and, subsequently, refined by Pezzoni et al. (2014) and EP-INV users' feedbacks (including: Cowan and Zinovyeva, 2013; Sterzi, 2013; Nathan, 2015; Akcigit et al., 2016). Appendix 1 succinctly describes both the algorithm and how it was adapted to the needs of the present study.

Using USPTO data, as opposed to EPO, would appear a more natural choice for a study on US-resident inventors. In fact, Li et al. (2014) provide disambiguated data. For our purposes, however, the EP-INV dataset has richer information contents, especially on inventors' addresses, which come complete of harmonized street and zip code, from the OECD REGPAT database (Maraut et al. 2008). This provides crucial information for our disambiguation algorithm. In addition, mastering our own disambiguation algorithm allows us to calibrate it according to our needs (see Appendix 1).[3]

[3] For example, the original NBER database (Hall et al., 2001) on USPTO patents provides the street address for just 11.5% of observations in the 'Inventors' file.

Harmonization of applicant names is performed using the OECD HAN Database, as employed in recent PatStat releases. However, as this is far from perfect, we also carried out an *ad hoc* reconstruction of business groups, using Bureau van Dijk's Zephyr database on Mergers & Acquisitions.[4]

The *IBM-GNR system* is a commercial product using information collected by the US immigration authorities in the first half of the 1990s. When fed with either a first name or a surname, IBM-GNR returns a list of Countries of Association (CoA) plus statistical information on the strength of the association. Consider for instance the inventor Rajiv Laroia. His first name, Rajiv, is associated with seven countries, including India, the UK and the Netherlands. As far as the cross-country distribution of the name (labelled "significance") is concerned, IBM-GNR suggests that around 80% of individuals named Rajiv originate from India, 10% from the UK, and around 1% only from each of the other countries. In the case of the within-country distribution (labelled "frequency"), Rajiv is deemed very common in India (in the top decile), but not elsewhere (5th decile in the UK, bottom decile in the Netherlands). The surname Laroia is associated with just two countries, India (99% significance) and France (1% significance).[5]

We treat this information using an additional, original algorithm (Ethnic-Inv), as described in Appendix 2. Briefly, we select one and only one CoO by selecting the CoA most closely associated with the inventor's name and surname. We use three indicators: (a) the frequency of the first name in English- and Spanish-speaking countries (the two most spoken languages in the US); (b) the product of the significance of the first name and the surname, for each CoA; and (c) the stand-alone significance of the surname, for each CoA. The higher (b) and (c), the more likely it is that a CoA actually corresponds to the inventor's CoO. The opposite holds for (a), since an inventor with, say, a typical Indian surname, but with John or Luis as a first name, is unlikely to be a first-generation immigrant to the US. He may be second-generation, but with no close ties to the diaspora, since his parents did not choose an ethnic name, opting for a distinctly local one. In the case of Rajiv Laroia, his surname present high values of (b) and (c) when associated to India, while his name is a high-frequency one in India, a zero-frequency one in Spanish-speaking countries, and a low-

[4] On OECD-HAN, see Thoma et al. (2010). Manual checks are necessary both for companies in different countries and for multinational groups, whose boundaries change over time.

[5] As the original dataset included only non-US citizens, the US itself is never listed among possible CoA.

frequency name in English-speaking countries (and only in those that host Indian minorities). We conclude he is either a first-generation Indian migrant or an insider member of the Indian community in the US.

Our algorithm assigns a specific weight to each of our three indicators (a)-(c), which we obtain by calibration against a benchmark dataset on the nationality of inventors resident in the US, based on PCT patent applications (Miguelez and Fink 2013). We retain the weights for a "high recall" calibration, that is, one that minimizes false negatives (foreign-origin inventors from the selected CoO mistaken for locals), albeit at the price of low precision. We do so in order to avoid a bias in favour of positive co-ethnicity effects in equations (1) and (2). When no CoO can be selected (no association is sufficiently strong), inventors are treated indifferently as locals or as foreigners from an unknown CoO.

Nationality, however, is not the ideal benchmark, as it tends to be overly restrictive. Migrants can, for example, acquire nationality if they reside long enough in the US, and any child born in the US to foreign parents is a US citizen in accordance with the *jus soli* principle (the former is the case of inventor Raijv Laroia, who we know to be an Indian-born US national). Indeed, some ethnic minorities may be composed largely of destination-country nationals, and yet remain cohesive over several generations.

Table1 HERE

Table 1 shows the percentage of inventors in our database from each of the ten selected CoO (column 1) alongside the analogous percentage of inventors listed as nationals from the same countries in PCT data (column 2). Figures in column (1) are always higher than those in column (2), as expected. Columns (3) and (4) report the results of z-test on proportions, which indicates these differences always to be significant. They can also be very large, as for Germany, Italy, Poland, and, above all, Iran. This latter case is highly instructive since, as Iranians have very distinctive names and surnames, so no large error can be attributed to our algorithm. More likely, we explain the difference with the Iranian inventors from the generation that fled the Islamic revolution in the 1980s and their descendants. Both are now US nationals, and yet they form

quite a distinct community, one that could play a key role in their home country, should there be a change of regime (Modarres 1998, Modarresi 2001, Mostofi 2003).

In sum, nationality as an indicator of foreign origin is imperfect. Yet, in the absence of a better alternative, we use it to calibrate our algorithm as well as to conduct robustness checks.

3.2.2 Sampling

We select all patent applications from the EP-INV database, with priority years between 1990 and 2010, and for which at least one inventor, resident in the US, reported a CoO among the ten selected. Our initial sample includes 88,522 inventors and 174,160 patents. Of these we retain only those applications receiving at least one forward citation from another EPO patent application (either directly, or indirectly, via an equivalent patent in its family).[6] In this way, we build a "national" and an "international" sample, which we use to investigate the diaspora and brain gain effects, respectively.

For the national sample, we retain all cited-citing pairs in which the citing patent comprises at least one US-resident among its inventors. We then exclude all self-citations at the applicant level, as well as all self-citations at the inventor level, where the self-citing inventor belongs to one of the 10 CoO selected. For each citing patent, we randomly select a control patent that satisfies the following conditions:

1. it does not cite the cited patent,
2. it has the same priority year and is classified in the same IPC groups as those of the citing patent[7],
3. it comprises at least one US-resident among its inventors.

This gives us 1,043,320 observations, half of which are cited-citing pairs, the other half cited-control pairs. They combine 89,986 cited patents, 195,595 citing patents, and 279,623 controls. Table 2 (part 1) reports details by CoO. As expected, the majority of observations are for the two largest CoO, China and India. The only European country presenting a similar order of magnitude is Germany.

[6] On the use of patent families for citation analysis, see Harhoff et al. (2003). For definitions of patent families, see Martinez (2011).

[7] As the same patent may be assigned to several IPC groups, our matching criteria require the citing patent and its control to be classified in the same number of IPC groups, and to share them all.

Table 2 HERE

For the international sample, we retain all cited-citing pairs in which the citing patent has no US-resident inventors. For each citing patent, we randomly select a control patent that satisfies conditions 1. and 2. above, and does *not* include any US-residents among its inventors.[8]

This gives us 1,048,258 observations (105,059 cited patents; 266,629 citing patents; and 390,519 controls). Table 2 (part 2) shows that the CoO distribution of the cited inventors is very much the same as that of the national sample.

In the regression setting, observations are "stacked" and flagged by means of the binary variable *Citation* (equal to one for cited-citing pairs, zero for cited-control pairs). Our dependent variable is then the probability of *Citation*=1, which we estimate by means of a Linear Probability Model (LPM; Logit estimates, which provide similar results, are available on request).[9]

As for regressors, for all patent pairs in the two samples, we produce the following dummy variables:

1. *Co-ethnicity*: =1 if at least one inventor in the cited patent and one inventor in the citing (control) patent are from the same CoO.

2. *Social distance S* (with S=0,1,2,>3,+∞): =1 if the geodesic distance between the cited patent and the citing (control) patent is equal to *S*. Formally: $S = min\ (S_{ij})$ where S_{ij}=geodesic distance between inventor *i* (*i*=1…I) on the cited patent and inventor *j* (*i*=1…*J*) on the citing (control) patent, as calculated over the entire network of inventors, for all inventors on the cited and the citing (control) patents. Notice that for *i*=*j* →S=0. If *i* and *j* belong to disconnected network components then: S=+∞. For each *t* we calculate a

[8] Notice that the cited patent may include, alongside the US-resident inventor(s), one or more foreign residents. This means that in our regressions we have to control for the distance between the inventors of the citing/control patents and both the US- and the foreign-resident inventors.

[9] LPM is easy to interpret, as its estimated coefficients can be read directly as marginal effects. This is particularly valuable in specifications like ours, which are loaded with interactions. Following Long (1997) and Wooldridge (2003), we consider LPM to be a good approximation of logit and probit models for probabilities between 0.2 and 0.8,. In our case, the baseline probability of the citation event is 0.5, by construction. The predicted probabilities are as follows: only 1% lower than 0.2 (with no negative predictions) and less than 4% higher than 0.8 (with less than 1% nonsensical, higher-than-1 predictions). Several of the papers we cite adopt the same strategy.

different network of inventors, based on co-inventorship patterns of all patents with priority years from *t*-5 to *t*-1.[10]

3. *Miles*: shortest distance (in miles) between the two patents, based on their inventors' addresses. We take the log of this value with the addition, in some specifications, of a quadratic term.[11]

4. Characteristics of the citing (control) patent, as suggested by Singh and Marx (2013): technological field dummies (OST-7 classification, as in Coffano and Tarasconi, 2014), number of *claims*, number of *backward citations* to prior art and to non-patent literature (*NPL citations*), as well as technological proximity to the cited patent (number of overlapping IPC-7 codes – *overlap IPCs 7* – and number of overlapping full IPC codes, out of all codes assigned to the patents - *overlap IPCs*).

For patent pairs in the national sample we also calculate:

5. *Same MSA* and *Same State*: =1 if at least one inventor in the cited patent and one inventor in the citing (control) patent are located in the same metropolitan statistical area (MSA) or US State, respectively.

For patent pairs in the international sample:

6. *Home country*: =1 if at least one inventor in citing (control) patent is located in the CoO of one of the inventors of the cited patent.

7. *Same country*: =1 if at least one inventor in the cited patent and one inventor in the citing (control) are located in the same country, outside the US.[12]

8. Other measures of country proximity, such as, border-sharing (*Contiguous countries*), *Former colonial relationship*, *English* as a common official language, and *Similarity to English*, a language similarity index ranging from 0 to 1, adapted from Miguelez (2016).

[10] This amounts to assuming that social ties generated by co-inventorship decay after 5 years, unless renewed by further co-patenting. For more details, see Breschi and Lissoni (2009).

[11] For each combination of inventors *i* and *j*, we calculate the great-circle distance between the centroid of the respective ZIP codes; we then retain the minimum distance. In case of missing ZIP codes, the centroid of the city was used (or the county, if the city's was missing, too).

[12] Co-inventors of a given patent may be located in different countries. In the international sample no inventor of the citing (control) patent can be located in the US, but nothing impedes two inventors in the cited and citing (control) patents from both being located outside the US and in the same country, which is not necessarily the CoO of the inventor(s) of the cited patent.

9. *Same company*: =1 if applicants of the cited and the citing (control) patents are the same.

10. *Returnee*: =1 if the inventor of the cited and the citing (control) patents are the same (notice that this implies *Social distance 0* = 1)

Table 3 reports the descriptive statistics for all variables in both samples (details by country available on request). For the brain gain regressions we did not retain the observations relating to Iran and Poland, given the small numbers involved. This reduces the international sample from 1,048,258 to 1,004,950 observations.

Table 3. HERE

Notice that a cited patent enters our sample as many times as the number of citations it receives. The same applies to each patent citing more than one cited patent, though this tends to be less frequent. This necessitates correcting for non-independence of errors, which we do by clustering errors by cited patent.

4. Results: within-US knowledge flows and the diaspora effect

Table 4 reports the results of three specifications of equation (1), without distinguishing by CoO. The first specification reproduces Agrawal et al.'s (2008) basic exercise; the second and third introduce social distance between inventors. Two further specifications (unreported) include further controls, first for patent characteristics, including technology fixed effects, then for spatial distance.

Table 4 HERE

The estimated coefficients in column (1) present the same sign and are of the same order of magnitude as those in Agrawal et al. (2008): co-ethnicity positively affects the probability of observing a citation link between two patents, but its marginal effect is smaller than that of MSA co-location. The interaction term

between co-ethnicity and co-location is negative. This suggests that a diaspora effect exists, and it is a substitute for co-location.

When controlling for social distance on the network of inventors (column 2), the estimated coefficients for co-location fall sharply, as the former affects negatively the probability of citation but is positively correlated with spatial distance, in line with previous findings of Breschi and Lissoni (2009). The marginal effect of co-ethnicity also falls, but not so noticeably (the interaction terms remain unaltered). At first sight, this suggests that co-ethnicity is not correlated with social distance as strongly as co-location.

Estimates in column (3), where we interact social distance on the network of inventors and co-ethnicity, qualify this result. Here, the interaction terms are positive and significant for social distances higher than three degrees. This indicates a substitution effect. Social ties based on ethnicity only matter when those based on professional experience (co-inventorship) are too loose. However, social distance on the network of inventors is generally associated with larger marginal effects than co-location or co-ethnicity.

Controlling for the patent's characteristics (*claims*, *backward citations*, *NPL citations*, *overlap IPCs 7*, and *overlap IPCs*) does not alter the coefficients of interest greatly, which is indicative of the robustness of the refined JTH sampling scheme we adopted. Adding controls for spatial distance (*Same State* and *ln_miles*, also in quadratic form) further alters the estimated co-efficient of *co-location*, but does not change the *social distance* and *co-ethnicity* estimates. (For both specifications, results are available upon request)

In Table 5 we allow the estimated coefficient of co-ethnicity to vary across CoO, first without any interaction with MSA co-location (column 1), then with interaction (column 2). The importance of co-ethnicity for the probability of citation varies across CoO, with its estimated coefficient being clearly positive and significant for Asian countries (albeit unstable across specifications for Japan and Iran), Russia, and Germany (again unstable). Marginal effects appear to be largest for Russia, followed, in descending order, by China, Iran, India, South Korea, Japan, and, at some distance, by Germany. As for the interaction term, this is negative and significant only for China and India, and either positive or negative, but never significant for all the other CoO. This suggests that, overall, the substitution effects between physical and ethnical proximity are driven

mostly by Chinese and Indian inventors. The coefficients of social distance and other controls (unreported) do not differ much from those in Table 4.

Table 5 – HERE

As a robustness check, we re-examine our evidence by replicating the Thompson's (2006) case-control methodology, as adapted by Singh and Marx (2013). We first assign to all the cited-citing patent pairs (and relative cited-control pairs) two new dummy variables (*applicant* and *examiner*), which take value one or zero according to the origin of the citation.[13] We then interact *co-ethnicity* with the new dummies in the cited-citing patent pair. Finally, we test whether the estimated coefficients for the two interaction terms are the same (F-test). If the hypothesis is rejected, and the coefficient for *co-ethnicity*applicant* is larger than that for *co-ethnicity*examiner*, we can conclude that ethnicity matters. Columns 1 and 2 in table 6 report the results of two regressions very similar to those in tables 4 and 5, respectively, but with the interactions we just described and the F-test results just below each pair of coefficients. For ease of exposition, results for the second regression (column 2) are arranged over five columns and two lines. The sample reduces from 1,043,320 to 1,005,592 observations, due to lack of information on the origin of several citations. Our findings are mostly in line with those we obtained with the JTH methodology. For the general co-ethnicity dummy (column 1) as well as for China, Germany, India, and Japan (column 2), the coefficient for *co-ethnicity*applicant* is larger than that for *co-ethnicity*examiner* and we reject the null hypothesis. For France, Italy, and Poland, on the contrary, the hypothesis cannot be rejected, again in line with our previous results. Contrary to what we expected, we cannot reject the hypothesis for Iran and Russia, but even in these cases the coefficient for *co-ethnicity*applicant* is larger than that for *co-ethnicity*examiner* (for Russia, we are close to rejecting the hypothesis at 90%). The only odd case is that for Korea, for which the co-ethnicity effect is significantly stronger in the examiner citation case.

[13] Differently than Thompson (2006) and Singh and Marx (2013), however, we do not deal with citations reported on documents by one national patent office only (in their case, the USPTO; in ours, the EPO). This would make the applicant vs examiner distinction highly dependent on the specific procedures of that particular office. We consider instead all documents in a patent family. We then define a citation as coming from the applicant if and only if it appears as such on all documents in the family (all examiners throughout patent offices worldwide ignored the cited prior art). We consider a citation to be coming from the examiner if it appears as such on at least one document in the family (at least one examiner took notice of the cited prior art). More details in Appendix 3.

Table 6 –HERE

Cross-country differences in the estimated diaspora effect may depend either on the demographic composition of ethnic groups (shares of first vs second- and subsequent-generation migrants) or on their social structure (social cohesiveness). Some of these characteristics depend, in turn, on how well we calibrate our algorithm for each specific CoO. The lower the precision, the more likely we are to mix first or second generation migrants with locals with the same ancestry, but no connections (e.g. young Italian PhD students at Yale with Italian-Americans in New Jersey), or with migrants from different CoO, but a common language (e.g. French vs Quebecois; or Germans vs Austrians and Swiss). In Appendix 2, we compare, among other things, our data with US census data on ancestry. We find measurement errors to be most likely for German inventors, followed at a considerable distance by Italians and, at an even further distance, by French and Polish.

One way to assess the relative weight of substantive factors vs measurement errors is to employ a different definition of foreign-origin inventor. In Table 7 we exploit information on inventors' nationality, which is a more stringent definition (although not necessarily more appropriate, as discussed in section 3). This reduces the sample to around a fifth of its initial size. We then run two sets of regressions: in the first, we maintain co-ethnicity as our explanatory variable of interest; in the second, we replace it with co-nationality. When comparing the estimated coefficients for co-ethnicity and co-nationality across the same specifications (columns 1 and 3, and columns 2 and 4, respectively), we note that, in general, the latter is larger. This suggests that our definition of foreign-origin inventors may present the errors described above. However, our results do not change substantially. Coefficients for Poland remain negative, while those for France and Italy do not become significant (although we observe a change of sign for France). For Russia, both co-ethnicity and co-nationality are positive and significant, and do not differ much.

The last column in Table 7 reports the results of a Wald test on the null hypothesis of the coefficients for co-ethnicity and co-nationality to be equal. The results are counter-intuitive, because small (large) differences in the coefficients often correspond to very small (large) standard errors. The hypothesis is only rejected at

the 95% confidence interval for India and (almost) at 90% for China, whose coefficients are in any case large and significant in both regressions, thus confirming that a diaspora effect exists. Overall, this suggests that, with the exception of Russia and (to less extent) Germany, no European country exhibits a diaspora effect, and this is not just a statistical artefact due to measurement error problems.

Table 7 –HERE

To probe further the robustness of our results, we estimate separate regressions for different technological classes of cited patents. We expect the ratio between first- and subsequent-generation migrant inventors to be higher in science-based technologies, where inventors are likely to be PhD holders and possibly academics, two social categories in which foreign-born US residents are over-represented (Auriol 2010, Scellato et al. 2015). Science-based technologies include primarily *Pharmaceuticals & Biotech*, followed by *Instruments*, *Chemicals & Materials*, *Electrical engineering & Electronics* and *Industrial processes*; with *Mechanical engineering & Transport* and *Consumer goods & Civil engineering* being considered traditional.[14]

Figure 1 reports the *Co-ethnicity* coefficients from each regression (full results in Appendix 4, Table A4.6). It shows that Pharma & Biotech is the technological class with the most instances of a positive and strongly significant coefficient (six CoO out of ten), followed by Chemicals & Materials (four CoO), Electrical engineering & Electronics, Industrial Processes (three CoO each) and Instruments (two). Mechanical engineering & Transport has just one case and Consumer goods none. This is in line with our expectations. Notice also that, for each of the first five technological classes we have many more observations than in the last two, a disproportion that would have not been observed had we sampled local as opposed to foreign-origin inventors (patents in the EP-INV dataset are quite evenly distributed across the seven classes). This is in line with the over-representation of foreign-born inventors in US high technologies.

Figure 1 – HERE

[14] On this ranking, see Callaert et al. (2006) and Lissoni (2012).

Appendix 4 reports the results of additional robustness checks. First, in Tables A4.2 and A4.3, we test whether our results depend exclusively on the most important US high-tech clusters, which attract a disproportionate number of highly skilled migrants (Kerr 2009). Second, in Table A4.4, we consider the possibility of cohort effects, with different generations of migrant inventors having different propensities to share knowledge with members of their communities. In both case our main results remain unchanged. Third, in table A4.5 we consider the possibility that the high significance of several coefficients in Tables 4 and 5 depends exclusively on our very large sample size. We apply the bootstrap techniques described by Greene (2008, p.596) and Wooldridge (2003, p.378) to specifications (2) in Table 4 and (1) in Table 5. While standard errors increase, estimated coefficients remain significant for India and China, as well as for Russia (with only one exception).

5. Results: international knowledge flows and the brain gain effect

Coming to the brain gain effect, column (1) in Table 8 reports the results of our baseline regression. Of the three countries with the strongest diaspora effect (China, Russia, and India), only the former two also exhibit and a positive and significant coefficient for *Home country*. As for the other countries, the coefficient is positive and significant only for South Korea and France. This suggests that the diaspora effect does not translate necessarily into brain gain, and vice versa.

Table 8 and Figure 2 – HERE

We are also interested in assessing how much of the brain gain effect may pass through multinationals, rather than intra-ethnic spillovers. In this respect, white bars in Figure 2 show the percentage of home-resident inventors (of either the citing or the control patents) who work for the company that owns the cited patent, for the ten CoO in our sample. Figures are high for the most advanced, innovative countries (around 60% for Germany and Japan, 40% for France), while the opposite is true for the other countries (1 to

3% for China, India, and Russia; South Korea and Italy present intermediate values). This suggests that multinationals may carry different weight across countries.

We test for this in column (2) of Table 8. We observe that, when interacting *Home country* with *Same company*, the positive effect of *Home country* for France disappears, while the coefficient of the interaction term is positive and significant. A similar pattern can be detected for other advanced countries, including Italy and Japan, but not for South Korea (where the interaction term is negative) nor, more interestingly, for any BRIC country. This suggests that US-resident foreign-origin inventors from advanced countries transfer knowledge back home mainly through the multinationals they may work for, rather than through personal contacts.

Interestingly, Germany behaves neither like France and the other advanced countries, nor like the BRICs and South Korea. That is, neither its inventors nor companies seem to have privileged access to knowledge produced by migrant inventors in the US. We explored the possibility that this result might be due to measurement errors, caused by the presence of many German inventors in Swiss companies and/or confusion between German, Swiss, and Austrian inventors when using our algorithm. But this appears not to be the case.[15]

We finally explore the role of returnee inventors as a brain gain channel. In all columns of Table 8 we observe a positive and significant coefficient for *Returnee*. However, returnees are very few in number (0.1% of all observations vs. 3% for *Same company*), so they are an unlikely channel for massive knowledge flows.

Another important cross-country difference may refer to absorptive capacities. Grey bars in Figure 2 report the percentage of home-resident native inventors (of both citing and control patents), by country of origin. For the most advanced countries we observe values over 70%, indicating that native inventors are disproportionately more active at home than abroad. The opposite hold for the less advanced ones. This implies that while the former host within their borders the vast majority of potential beneficiaries of co-ethnic knowledge flows from the US, the same does not hold for the latter. Indeed, migrant inventors from

[15] We re-ran regressions in Table 8 by extending value =1 for Home country to all cases of inventors located in Austria and Switzerland. We also restricted the regressions to the case of Pharma & Biotech technologies, in which Swiss companies are over-represented. Always to no avail.

less advanced countries are so many, and so dispersed around the world that an 'international diaspora' effect may exist, to the benefit of several countries of destination, instead or besides the home country. We test for this in column (3) of Table 8. There we replace *Home country* with *Co-ethnicity*, which indicates the existence of an ethnic tie irrespective of the inventor's country of residence. Results remain the same for China and Russia, but not for India, whose coefficient is now positive and significant. This suggests that, for this country, an "international diaspora" may exist, along with no benefits for the home country. [16]

In order to explore this finding further, Table 9 reports the results of a regression exercise limited to just the BRIC countries in our sample. We allow for the simultaneous presence, among the regressors, of *Home country* and *Co-ethnicity*, plus their interaction. For China and Russia, the coefficients of both variables remain positive and significant, while for India it is so only for *Co-ethnicity* (the interaction terms are never significant). This is further evidence that, in the case of China and Russia, the brain gain and the international diaspora effect co-exist, while for India no brain gain is detected. Notice that this result is compatible with findings by Agrawal et al. (2011) and Branstetter et al. (2015).[17]

Table 9– HERE

6. Discussion and conclusions

Drawing on patent and inventor data, we have investigated whether ethnic ties help in the diffusion of technical knowledge among foreign-origin inventors active in the same country of destination (diaspora effect) and back to their home country (brain gain effect). Our study has focused on the US as destination

[16] Table A4.7 of appendix 4 reports the results of a robustness check run by replacing, where available, Co-nationality with either Home country or Co-ethnicity. Although the much smaller sample size means several coefficients lose significance, the main results do not change. In general, the same applies to regressions by technology (results are available on request).

[17] For the international sample, we did not perform the robustness check based on Thompson's (2006) applicant vs examiner approach. In the international sample, in fact, the share of applicant citations drops dramatically, since patent offices outside the US do not impose any duty of candor rule (see appendix 3 for a detailed discussion). In addition, when the USPTO examines patent applications according to the PCT procedure – which is quite likely in the international sample - it treats all citations inserted by foreign patent offices as they were from applicants, while in reality most come instead from examiners of such offices (Alcacer et al., 2009). Therefore, the distinction between applicant-added and examiner-added citations becomes too blurred to be useful.

country and on five Asian and five European countries of origin, selected from among the main sources of highly skilled migration to the US.

Our empirical exercise has exploited a large, original dataset, based on disambiguated inventor data and the linguistic analysis of names and surnames. We also conducted robustness checks based upon inventors' nationality for a sizeable subsample.

We find evidence of a diaspora effect for all Asian countries in our sample (China, India, South Korea, and, to a lesser extent, Japan and Iran) and for two European countries (Russia and, to a much lesser extent, Germany). However, the marginal effect of co-ethnicity is secondary to that of proximity in physical space (co-location at the city level) and in the social network of inventors. In addition, co-ethnicity ties appear to be relevant for social-network-distant inventors. Substitutability holds too, for spatial proximity, especially for Chinese and Indian inventors, as already found (for India) by Agrawal et al. (2008).

In the case of the brain gain effect, ethnic ties do not necessarily imply a knowledge transfer to the home country. Specifically, we find no evidence for one of the main diasporas in the US, namely the Indian one. This may be attributable more to the absorptive capacities of the country of origin than to the international dimension of the diffusion process under consideration. In fact, for both India and the other BRIC countries in our sample, we find evidence of an international diaspora effect, which presents certain analogies with findings in the trade literature (Felbermayr et al. 2010). In contrast, any brain gain effect for France, Italy, and Japan, seems mediated by multinationals.

Despite imperfections in our name-based method for identifying migrants, our results appear robust enough to rule out major problems of measurement error. Still, while we highlight differences between the migrants' countries of origin, we can only speculate as to their causes. We suspect them to lie in the cohort composition of foreign-origin communities or to their composition by migration channel. In the first case, we refer to the different ratio between first- and subsequent-generation migrants, which is higher for Asian countries of origin (plus Russia), as opposed to European ones. As ethnic ties may be stronger for first-generation migrants, this could explain some of the observed differences in the diaspora effect. As for

channels, migration from the BRICs may be occurring more frequently via the higher education system, and that from advanced countries via multinationals (Kerr et al. 2016) .

We intend dedicating our future research efforts to assessing the validity of these intuitions and investigating their policy implications.

Future research plans also include examining the role of ethnic ties in the formation of inventor networks, so as to reconsider their role in determining collaboration-based social proximity. Additionally, we wish to extend the analysis conducted herein and consider Europe as a region of destination. This, among other things, should help shedding light on the policy-sensitive topic of the comparative attractiveness of Europe and the US as destinations for migrant scientists and engineers (Cerna and Chou 2014, Guild 2007).

References

Agrawal, A., Cockburn, I. and McHale, J. (2006) 'Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships', *Journal of Economic Geography,* 6(5), 571-591.

Agrawal, A., Kapur, D. and McHale, J. (2008) 'How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data', *Journal of Urban Economics,* 64(2), 258-69.

Agrawal, A., Kapur, D., McHale, J. and Oettl, A. (2011) 'Brain Drain or Brain Bank? The Impact of Skilled Emigration on Poor-Country Innovation', *Journal of Urban Economics,* 69(1), 43-55.

Akcigit, U., Baslandze, S. and Stantcheva, S. (2016) 'Taxation and the International Mobility of Inventors', *American Economic Review,* (forthcoming).

Alcacer, J. and Gittelman, M. (2006) 'Patent citations as a measure of knowledge flows: The influence of examiner citations', *The Review of Economics and Statistics,* 88(4), 774-779.

Alcácer, J., Gittelman, M. and Sampat, B. (2009) 'Applicant and examiner citations in US patents: An overview and analysis', *Research Policy,* 38(2), 415-427.

Almeida, P., Phene, A. and Li, S. (2015) 'The Influence of Ethnic Community Knowledge on Indian Inventor Innovativeness', *Organization Science,* 26(1), 198-217.

Alnuaimi, T., Opsahl, T. and George, G. (2012) 'Innovating in the periphery: The impact of local and foreign inventor mobility on the value of Indian patents', *Research Policy,* 41(9), 1534-1543.

Auriol, L. (2010) *Careers of doctorate holders: employment and mobility patterns*, 2010/04, Paris: OECD Publishing.

Beine, M., Docquier, F. and Özden, Ç. (2011) 'Diasporas', *Journal of Development Economics,* 95(1), 30-41.

Bhagwati, J. and Hanson, G. (2009) *Skilled immigration today: prospects, problems, and policies,* Oxford: Oxford University Press.

Blomström, M. and Kokko, A. (1998) 'Multinational corporations and spillovers', *Journal of Economic surveys,* 12(3), 247-277.

Boschma, R. and Frenken, K. (2011) 'The emerging empirics of evolutionary economic geography', *Journal of Economic Geography,* 11(2), 295-307.

Branstetter, L., Li, G. and Veloso, F. (2015) 'The Rise of International Co-invention' in Jaffe, A. B. and Jones, B. F., eds., *The Changing Frontier: Rethinking Science and Innovation Policy*, Chicago: University of Chicago Press.

Breschi, S. (2011) 'The geography of knowledge flows' in Cooke, P., Asheim, B. T., Boschma, R., Martin, R., Schwartz, D. and Tödtling, F., eds., *Handbook of Regional Innovation and Growth*, Cheltenham: Edward Elgar Publishing.

Breschi, S. and Lissoni, F. (2005) '"Cross-Firm" Inventors and Social Networks: Localized Knowledge Spillovers Revisited', *Annals of Economics and Statistics / Annales d'Économie et de Statistique,* (79/80), 189-209.

Breschi, S. and Lissoni, F. (2005) 'Knowledge networks from patent data' in Moed, H. F., Glänzel, W. and Schmoch, U., eds., *Handbook of quantitative science and technology research*, Berlin: Springer Science+Business Media, 613-643.

Breschi, S. and Lissoni, F. (2009) 'Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows', *Journal of Economic Geography,* 9(4), 439-468.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K. and Thijs, B. (2006) 'Traces of prior art: An analysis of non-patent references found in patent documents', *Scientometrics,* 69(1), 3-20.

Cerna, L. and Chou, M.-H. (2014) 'The regional dimension in the global competition for talent: Lessons from framing the European Scientific Visa and Blue Card', *Journal of European Public Policy,* 21(1), 76-95.

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

Coffano, M. and Tarasconi, G. (2014) 'Crios-Patstat Database: Sources, Contents and Access Rules', *Center for Research on Innovation, Organization and Strategy, CRIOS*, available: [accessed

Cowan, R. and Zinovyeva, N. (2013) 'University effects on regional innovation', *Research Policy,* 42(3), 788-800.

Currarini, S., Jackson, M. O. and Pin, P. (2009) 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica,* 77(4), 1003-1045.

Docquier, F. and Marfouk, A. (2006) 'International migration by educational attainment (1990-2000)' in Özden, Ç. and Schiff, M., eds., *International migration, remittances and the brain drain*, New York: The World Bank - Palgrave Macmillan, 151-199.

Du Plessis, M., Van Looy, B., Song, X. and Magerman, T. (2009) *Data production methods for harmonized patent indicators: Assignee sector allocation,* Luxembourg: EUROSTAT Working Paper and Studies.

Dufoix, S. (2008) *Diasporas,* Berkeley CA: University of California Press.

Ellison, G., Glaeser, E. L. and Kerr, W. R. (2010) 'What causes industry agglomeration? Evidence from coagglomeration patterns', *The American Economic Review,* 100(3), 1195-1213.

Felbermayr, G. J., Jung, B. and Toubal, F. (2010) 'Ethnic networks, information, and international trade: Revisiting the evidence', *Annals of Economics and Statistics/Annales d'Économie et de Statistique,* 97/98, 41-70.

Foley, C. F. and Kerr, W. R. (2013) 'Ethnic innovation and US multinational firm activity', *Management Science,* 59(7), 1529-1544.

Freeman, R. B. (2010) 'Globalization of scientific and engineering talent: international mobility of students, workers, and ideas and the world economy', *Economics of Innovation and New Technology,* 19(5), 393-406.

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

Greene, W. H. (2008) *Econometric analysis (6th edition),* Pearson Education.

Guild, E. (2007) 'EU Policy on Labour Migration: A First Look at the Commission's Blue Card Initiative', *CEPS Policy brief,* (145).

Hall, B. H., Jaffe, A. B. and Trajtenberg, M. (2001) 'The NBER patent citation data file: Lessons, insights and methodological tools', *National Bureau of Economic Research Working Paper Series,* No. 8948.

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

Harhoff, D., Scherer, F. M. and Vopel, K. (2003) 'Citations, family size, opposition and the value of patent rights', *Research Policy,* 32(8), 1343-1363.

Henderson, R., Jaffe, A. and Trajtenberg, M. (2005) 'Patent citations and the geography of knowledge spillovers: A reassessment: Comment', *American Economic Review,* 95(1), 461-464.

Henderson, V. (1997) 'Externalities and industrial development', *Journal of urban economics,* 42(3), 449-470.

Jaffe, A. B., Trajtenberg, M. and Henderson, R. (1993) 'Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations', *The Quarterly Journal of Economics,* 108(3), 577-598.

Kapur, D. (2001) 'Diasporas and technology transfer', *Journal of Human Development,* 2(2), 265-286.

Kenney, M., Breznitz, D. and Murphree, M. (2013) 'Coming back home after the sun rises: Returnee entrepreneurs and growth of high tech industries', *Research Policy,* 42(2), 391-407.

Kerr, S. P., Kerr, W., Özden, Ç. and Parsons, C. (2016) 'Global talent flows', *National Bureau of Economic Research Working Paper Series,* No. 22715.

Kerr, W. (2009) 'The Agglomeration of US Ethnic Inventors' in Glaeser, E. L., ed. *Agglomeration Economics* Chicago: The University of Chicago Press.

Kerr, W. R. (2008) 'Ethnic Scientific Communities and International Technology Diffusion', *Review of Economics and Statistics,* 90(3), 518-537.

Krugman, P. (2011) 'The new economic geography, now middle-aged', *Regional Studies,* 45(1), 1-7.

Krugman, P. R. (1991) *Geography and trade,* Cambridge MA: MIT press.

Kuznetsov, Y., ed. (2006) *Diaspora networks and the international migration of skills: how countries can draw on their talent abroad,* World Bank Publications, Washington, DC.

Kuznetsov, Y., ed. (2010) *Talent Abroad Promoting Growth and Institutional Development at Home: Skilled Diaspora as Part of the Country,* World Bank, Washington, DC (http://hdl.handle.net/10986/10117).

Li, G.-C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., Yu, A. Z. and Fleming, L. (2014) 'Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)', *Research Policy,* 43(6), 941-955.

Lissoni, F. (2012) 'Academic patenting in Europe: An overview of recent research and new perspectives', *World Patent Information,* 34(3), 197-205.

Long, J. S. (1997) *Regression models for limited and categorical dependent variables,* Thousand Oaks, CA: Sage.

Maraut, S., Dernis, H., Webb, C., Spiezia, V. and Guellec, D. (2008) *The OECD REGPAT database: a presentation*, OECD Publishing.

Martínez, C. (2011) 'Patent families: When do different definitions really matter?', *Scientometrics,* 86(1), 39-63.

Martínez, C., Azagra-Caro, J. M. and Maraut, S. (2013) 'Academic Inventors, Scientific Impact and the Institutionalisation of Pasteur's Quadrant in Spain', *Industry and Innovation,* 20(5), 438-455.

Marx, M., Strumsky, D. and Fleming, L. (2009) 'Mobility, skills, and the Michigan non-compete experiment', *Management Science,* 55(6), 875-889.

Mayr, K. and Peri, G. (2008) 'Return Migration as a Channel of Brain Gain', *National Bureau of Economic Research Working Paper Series,* No. 14039.

Meyer, J.-B. (2001) 'Network Approach versus Brain Drain: Lessons from the Diaspora', *International Migration,* 39(5), 91-110.

Meyer, J.-B. and Brown, M. (1999) 'Scientific diasporas: A new approach to the brain drain', *MOST discussion Paper No. 41, UNESCO - Paris*.

Miguelez, E. (2016) 'Inventor Diasporas and the Internationalization of Technology', *The World Bank Economic Review,* (forthcoming - 10.1093/wber/lhw013).

Miguelez, E. and Fink, C. (2013) *Measuring the International Mobility of Inventors: A New Database*, World Intellectual Property Organization-Economics and Statistics Division.

Modarres, A. (1998) 'Settlement patterns of Iranians in the United States', *Iranian Studies,* 31(1), 31-49.

Modarresi, Y. (2001) 'The Iranian community in the United States and the maintenance of Persian', *International journal of the sociology of language,* (148), 93-116.

Mostofi, N. (2003) 'Who we are: The perplexity of Iranian-American identity', *Sociological Quarterly,* 44(4), 681-703.

Nanda, R. and Khanna, T. (2010) 'Diasporas and domestic entrepreneurs: Evidence from the Indian software industry', *Journal of Economics & Management Strategy,* 19(4), 991-1012.

Nathan, M. (2015) 'Same difference? Minority ethnic inventors, diversity and innovation in the UK', *Journal of Economic Geography,* 15(1), 129-168.

OECD (2015) *Connecting with Emigrants. A Global Profile of Diasporas*, Paris.

Peeters, B., Song, X., Callaert, J., Grouwels, J. and Van Looy, B. (2010) *Harmonizing harmonized patentee names: an exploratory assessment of top patentees, EUROSTAT Working Paper and Studies,* Luxembourg.

Pezzoni, M., Lissoni, F. and Tarasconi, G. (2014) 'How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation', *Scientometrics,* 101(1), 1-28.

Raffo, J. and Lhuillery, S. (2009) 'How to play the "Names Game": Patent retrieval comparing different heuristics', *Research Policy,* 38(10), 1617-1627.

Rauch, J. E. and Trindade, V. (2002) 'Ethnic Chinese networks in international trade', *Review of Economics and Statistics,* 84(1), 116-130.

Saxenian, A. (2006) *The new argonauts: Regional advantage in a global economy,* Cambridge MA: Harvard University Press.

Scellato, G., Franzoni, C. and Stephan, P. (2015) 'Migrant scientists and international networks', *Research Policy,* 44(1), 108-120.

Schiff, M. and Wang, Y. (2013) 'North–South Trade-related Technology Diffusion and Productivity Growth: Are Small States Different?', *International Economic Journal,* 27(3), 399-414.

Singh, J. and Marx, M. (2013) 'Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity', *Management Science,* 59(9), 2056-2078.

Sterzi, V. (2013) 'Patent quality and ownership: An analysis of UK faculty patenting', *Research Policy,* 42(8), 564-576.

Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B. H. and Harhoff, D. (2010) 'Harmonizing and combining large datasets–An application to firm-level patent and accounting data', *National Bureau of Economic Research Working Paper Series,* 15851.

Thompson, P. (2006) 'Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations', *The Review of Economics and Statistics,* 88(2), 383-388.

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

Ventura, S. L., Nugent, R. and Fuchs, E. R. H. (2015) 'Seeing the Non-Stars:(Some) Sources of Bias in Past Disambiguation Approaches and a New Public Tool Leveraging Labeled Records', *Research Policy,* 44(9), 1672-1701.

Veugelers, R. and Cassiman, B. (2004) 'Foreign subsidiaries as a channel of international technology diffusion: Some direct firm level evidence from Belgium', *European Economic Review,* 48(2), 455-476.

Wadhwa, V., Rissing, B., Saxenian, A. and Gereffi, G. (2007) *Education, Entrepreneurship and Immigration: America's New Immigrant Entrepreneurs, Part II.*

Widmaier, S. and Dumont, J.-C. (2011) *Are recent immigrants different? A new profile of immigrants in the OECD based on DIOC 2005/06, OECD Social, Employment and Migration WP,* Paris: OECD Publishing.

Wooldridge, J. (2003) *Introductory Econometrics: A Modern Approach* Nashville TN: South-Western College Publishing.

Zweig, D. (2006) 'Competing for talent: China's strategies to reverse the brain drain', *International Labour Review*, 145(1-2), 65-90.

This version: *03 March 2021*

TABLES

Table 1 – Comparison of EP-INV and WIPO-PCT data, by country of origin/nationality

	% of US-resident inventors of:				
	foreign origin, active in 2000 [i]	foreign nationality, 1995-2005 [ii]		z [iii]	p-value [iv]
	(1)	(2)	(1)/(2)	(3)	(4)
China	3.938	3.763	1.05	2.57	0.01
Germany	2.181	1.038	2.10	29.62	0.00
France	0.782	0.589	1.33	6.92	0.00
India	3.872	2.984	1.30	14.38	0.00
Iran	0.366	0.110	3.33	18.94	0.00
Italy	0.470	0.228	2.06	13.30	0.00
Japan	0.599	0.483	1.24	4.62	0.00
Korea	0.564	0.482	1.17	3.30	0.00
Poland	0.204	0.111	1.83	7.41	0.00
Russia	0.587	0.469	1.25	4.80	0.00

[i] source: EP-INV database

[ii] source: WIPO-PCT dataset (see Miguelez and Fink, 2013).

[iii] Normalized difference between (1) and (2)

[iv] p-values for z-test on $H_0 : (1) = (2)$

Table 2. National and international samples: nr of patents, pairs, and observations; by country of origin of cited patents' inventors

	cited patents		citing patents		cited-citing pairs		obs [iii]
	Nr	%	nr	%	nr	%	nr
1. National sample (citations from within the US)							
China	27,496	25.35%	73,747	20.81%	124,674	23.90%	249,348
Germany	17,542	16.18%	62,991	17.77%	87,785	16.83%	175,570
France	6,913	6.37%	26,637	7.52%	33,085	6.34%	66,170
India	33,172	30.59%	97,439	27.49%	162,017	31.06%	324,034
Iran	2,984	2.75%	12,421	3.50%	14,522	2.78%	29,044
Italy	4,255	3.92%	18,847	5.32%	23,332	4.47%	46,664
Japan	4,929	4.54%	19,944	5.63%	24,086	4.62%	48,172
Korea	5,217	4.81%	20,431	5.77%	25,887	4.96%	51,774
Poland	1,757	1.62%	6,993	1.97%	8,032	1.54%	16,064
Russia	4,184	3.86%	14,939	4.22%	18,240	3.50%	36,480
Total [i]	108,449	100.00%	354,389	100.00%	521,660	100.00%	1,043,320
Total [ii]	89,986		195,595		437,737		875,474
2. International sample (citations from outside the US)							
China	31,321	24.75%	88,675	21.87%	128,122	24.44%	256,244
Germany	21,512	17.00%	72,694	17.93%	88,782	16.94%	177,564
France	8,246	6.52%	29,305	7.23%	34,050	6.50%	68,100
India	37,984	30.02%	114,872	28.33%	158,233	30.19%	316,466
Iran	3,309	2.62%	12,007	2.96%	13,317	2.54%	26,634
Italy	5,019	3.97%	19,834	4.89%	23,114	4.41%	46,228
Japan	6,189	4.89%	23,281	5.74%	26,575	5.07%	53,150
Korea	5,957	4.71%	21,072	5.20%	24,512	4.68%	49,024
Poland	2,089	1.65%	7,382	1.82%	8,337	1.59%	16,674
Russia	4,911	3.88%	16,330	4.03%	19,087	3.64%	38,174
Total [i]	126,537	100.00%	405,452	100.00%	524,129	100.00%	1,048,258
Total [ii]	104,991		266,130		444,916		889,832

[i] *Total = sum of observations by country of origin (same patent may be recorded under >1 country)*

[ii] *Total = sum of distinct observations*

[iii] *Nr observations per country = Nr cited-citing pairs * 2*

Table 3. National and international samples: descriptive statistics

	Obs	Mean	Std. Dev.	Min	Max
1. National sample (citations from within the US)					
Citation	1,043,320	0.50	0.50	0	1
Co-ethnicity	1,043,320	0.13	0.33	0	1
Same MSA	1,043,320	0.14	0.34	0	1
Same State	1,043,320	0.22	0.41	0	1
Miles	1,043,320	933.71	877.68	0	5085
Soc. Dist. 0	1,043,320	0.01	0.09	0	1
Soc. Dist. 1	1,043,320	0.01	0.09	0	1
Soc. Dist. 2	1,043,320	0.01	0.08	0	1
Soc. Dist. 3	1,043,320	0.01	0.09	0	1
Soc. Dist. >3	1,043,320	0.24	0.43	0	1
Soc. Dist. ∞	1,043,320	0.73	0.44	0	1
claims	1,043,320	8.50	12.80	0	259
backward citations	1,043,320	4.58	3.15	0	87
NPL citations	1,043,320	1.33	2.45	0	57
overlap IPCs 7 digits	1,043,320	1.13	1.47	0	27
overlap IPCs 7 digits / all IPCs	1,043,320	0.28	0.28	0	1
overlap IPCs	1,043,320	0.83	1.57	0	53
2. International sample (citations from outside the US) [(i)]					
Citation	1,004,950	0.50	0.50	0	1
Co-ethnicity	1,004,950	0.10	0.30	0	1
Home country	1,004,950	0.09	0.29	0	1
Same company	1,004,950	0.03	0.17	0	1
Returnee	1,004,950	0.001	0.02	0	1
Contiguous countries	1,004,950	0.03	0.18	0	1
Former colonial relationship	1,004,950	0.20	0.40	0	1
Same country	1,004,950	0.04	0.19	0	1
English	1,004,950	0.17	0.38	0	1
Similarity to English	1,004,950	0.25	0.26	0	1
Miles	1,004,950	4,452.46	1,936.59	0	11,498.1
Soc. Dist. 0	1,004,950	0.00	0.06	0	1
Soc. Dist. 1	1,004,950	0.01	0.07	0	1
Soc. Dist. 2	1,004,950	0.00	0.07	0	1
Soc. Dist. 3	1,004,950	0.00	0.07	0	1
Soc. Dist. >3	1,004,950	0.20	0.40	0	1
Soc. Dist. ∞	1,004,950	0.78	0.42	0	1
claims	1,004,950	9.90	11.82	0	442
backward citations	1,004,950	4.00	3.20	0	98
backward NPL citations	1,004,950	1.00	2.07	0	76
overlap IPCs 7 digits	1,004,950	1.09	1.28	0	32
overlap IPCs 7 digits / all IPCs	1,004,950	0.31	0.30	0	1
overlap IPCs	1,004,950	0.79	1.38	0	54

[(i)] *For regression analysis we did not retain all observations related to Iranian and Polish inventors (43,308 obs). This reduces the sample size from 1,048,258 to 1,004,950 observations*

Table 4 – Probability of citation from within the US, as a function of co-ethnicity, spatial & social distance, and controls -– LPM regression

	(1)	(2)	(3)
Same MSA	0.135***	0.0894***	0.0887***
	(0.00232)	(0.00236)	(0.00238)
Co-ethnic	0.0510***	0.0421***	-0.000590
	(0.00195)	(0.00193)	(0.00721)
Co-ethnic * MSA	-0.0161***	-0.0165***	-0.0125***
	(0.00449)	(0.00436)	(0.00461)
Soc. Dist. 1		-0.0442***	-0.0431***
		(0.00412)	(0.00446)
Soc. Dist. 2		-0.137***	-0.134***
		(0.00655)	(0.00716)
Soc. Dist. 3		-0.224***	-0.219***
		(0.00688)	(0.00766)
Soc. Dist. >3		-0.392***	-0.402***
		(0.00287)	(0.00302)
Soc. Dist. ∞		-0.428***	-0.433***
		(0.00274)	(0.00284)
Co-ethnic *Soc. Dist. 1			0.00515
			(0.0101)
Co-ethnic *Soc. Dist. 2			0.00394
			(0.0140)
Co-ethnic *Soc. Dist. 3			0.00275
			(0.0140)
Co-ethnic *Soc. Dist. >3			0.0629***
			(0.00737)
Co-ethnic *Soc. Dist. ∞			0.0335***
			(0.00716)
Technology F.E.	no	no	yes
Constant	0.475***	0.892***	0.898***
	(0.000440)	(0.00289)	(0.00298)
Observations	1,043,320	1,043,320	1,043,320
R2	0.010	0.023	0.023
F	3457	7879	4991

Clustered standard errors (at the cited patent level) in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 5 – Probability of citation from within the US, as a function of co-ethnicity by Country of Origin, spatial & social distance, and controls -– LPM regression

	(1)	(2)	(2-cont.)	(2-cont.)
Same MSA	0.0314***	0.0342***		
	(0.00347)	(0.00353)		
Same State	0.0216***	0.0215***		
	(0.00280)	(0.00280)		
ln(Miles)	-0.00603***	-0.00606***		
	(0.000667)	(0.000667)	*Co-ethnicity* Same MSA*	
China co-ethnic	0.0562***	0.0615***	China * Same MSA	-0.0302***
	(0.00242)	(0.00272)		(0.00622)
Germany co-ethnic	0.0105**	0.00875	Germany*Same MSA	0.0102
	(0.00504)	(0.00561)		(0.0130)
France co-ethnic	-0.0105	-0.00402	France * Same MSA	-0.0345
	(0.0110)	(0.0125)		(0.0273)
India co-ethnic	0.0344***	0.0364***	India * Same MSA	-0.0121*
	(0.00249)	(0.00279)		(0.00639)
Iran co-ethnic	0.0508**	0.0389	Iran * Same MSA	0.0425
	(0.0241)	(0.0294)		(0.0514)
Italy co-ethnic	0.0125	0.0189	Italy * Same MSA	-0.0375
	(0.0347)	(0.0403)		(0.0535)
Japan co-ethnic	0.0278*	0.0336**	Japan * Same MSA	-0.0334
	(0.0142)	(0.0162)		(0.0366)
Korea co-ethnic	0.0345***	0.0424***	Korea * Same MSA	-0.0373
	(0.0133)	(0.0148)		(0.0331)
Poland co-ethnic	-0.0434	-0.0519	Poland * Same MSA	0.0476
	(0.0416)	(0.0484)		(0.0899)
Russia co-ethnic	0.0710***	0.0594***	Russia * Same MSA	0.0635
	(0.0157)	(0.0178)		(0.0409)
Co-ethnicity Same MSA*	*No*	*Yes (see right)*		
Constant	0.669***	0.668***		
	(0.00538)	(0.00538)		
Social distance dummies	yes	yes		
Citing patent characteristics	yes	yes		
Technology F.E.	yes	yes		
Observations	1,043,320	1,043,320		
R2	0.080	0.080		
F	2139	1601		

Clustered standard errors (at the cited patent level) in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 6 – Probability of citation from within the US, applicant vs examiner -– LPM regression

	(1)	(2)				
	All	CN	DE	FR	IN	IR
Co-ethnic * Applicant	0.059***	0.078***	0.042***	0.0031	0.046***	0.055
	(0.0026)	(0.0035)	(0.0079)	(0.018)	(0.0042)	(0.042)
Co-ethnic * Examiner	0.024***	0.035***	-0.013**	-0.018	0.025***	0.044
	(0.0019)	(0.0029)	(0.0060)	(0.013)	(0.0027)	(0.028)
F-test[ii]	*146.8*	*106.7*	*33.5*	*0.92*	*19.8*	*0.05*
Prob.	*0.00*	*0.00*	*0.00*	*0.34*	*0.00*	*0.82*
		IT	JP	KR	PL	RU
Co-ethnic * Applicant		0.059	0.053**	0.0081	-0.057	0.094***
		(0.071)	(0.023)	(0.022)	(0.056)	(0.024)
Co-ethnic * Examiner		-0.022	0.0042	0.045***	-0.032	0.050**
		(0.022)	(0.017)	(0.017)	(0.054)	(0.020)
F-test		*1.23*	*3.06*	*1.91*	*0.12*	*2.10*
Prob.		*0.27*	*0.08*	*0.17*	*0.73*	*0.15*
Controls [i]	yes			yes		
Observations	1,005,592			1,005,592		
R2	0.080			0.080		
F	2731			1512		

[i] Both regressions (1) and (2) include controls for : Physical and social distance between inventors ; Citing patent's characteristics ; Technology F.E.

[ii] F-tests (p-values in parentheses) on *Ho:* $\beta_{Co\text{-}ethnic}$**applicant citation =* $\beta_{Co\text{-}ethnic}$**examiner citation*

Clustered standard errors (at the cited patent level) in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 7 – Probability of citation from within the US, as a function of co-ethnicity or co-nationality [(i)] -– LPM regression

	CO-ETHNICITY		CO-NATIONALITY		(2) vs (4)
	(1)	(2)	(3)	(4)	Wald test [(ii)]
Same MSA	0.0313***	0.0311***	0.0311***	0.0310***	
	(0.00672)	(0.00671)	(0.00673)	(0.00673)	
Same State	0.0362***	0.0364***	0.0361***	0.0364***	
	(0.00487)	(0.00486)	(0.00486)	(0.00486)	
ln(Miles)	-0.00230*	-0.00228*	-0.00241*	-0.00236*	
	(0.00123)	(0.00123)	(0.00123)	(0.00123)	
Co-ethnic / co-national	0.0564***		0.0637***		
	(0.00298)		(0.00347)		
China		0.0704***		0.0750***	2.68
		(0.00356)		(0.00412)	(0.102)
Germany		0.0196*		0.0305**	1.27
		(0.0111)		(0.0132)	(0.259)
France		-0.0111		0.00673	1.00
		(0.0197)		(0.0231)	(0.317)
India		0.0416***		0.0514***	5.31
		(0.00490)		(0.00634)	(0.021)
Iran		0.159**		0.265**	0.88
		(0.0754)		(0.131)	(0.347)
Italy		0.0485		0.0596	0.07
		(0.0477)		(0.0368)	(0.784)
Japan		0.0549**		0.0338	1.15
		(0.0260)		(0.0327)	(0.284)
Korea		0.0308		0.0529*	1.68
		(0.0243)		(0.0292)	(0.195)
Poland		-0.251**		-0.243	0.00
		(0.126)		(0.160)	(0.959)
Russia		0.101***		0.104***	0.01
		(0.0290)		(0.0333)	(0.940)
Social distance dummies	yes	yes	yes	yes	
Citing patent characteristics	yes	yes	yes	yes	
Technology F.E.	yes	yes	yes	yes	
Constant	0.710***	0.712***	0.711***	0.712***	
	(0.00918)	(0.00917)	(0.00919)	(0.00919)	
Observations	237,696	237,696	237,696	237,696	
R2	0.078	0.079	0.078	0.078	
F	1246	872.4	1243	866.4	

[(i)] Co-ethnicity in columns 1 and 2 ; co-nationality in columns 3 and 4 - Clustered robust standard errors (at the cited patent level) in parentheses, *** p<0.01, ** p<0.05, * p<0.1

[(ii)] Last column provides Chi-sq statistics (p-values in parentheses) for Wald tests on *H0:* $\beta_{Co\text{-}ethnicity} = \beta_{Co\text{-}nationality}$ across regressions (2) and (4)

Table 8 – Probability of citation from outside the US, as a function of inventors' country of residence (Home country) and Country of Origin (Co-ethnicity) [(i)] -– LPM regression

	HOME COUNTRY		CO-ETHNICITY
	(1)	(2)	(3)
Same company	0.214***	0.210***	0.212***
	(0.00508)	(0.00543)	(0.00536)
Home country / Co-ethnicity [(ii)]:			
China	0.0407***	0.0398***	0.0396***
	(0.00642)	(0.00651)	(0.00514)
Germany	-0.00165	-0.00108	0.000900
	(0.00282)	(0.00293)	(0.00281)
France	0.0150**	0.00761	0.0276***
	(0.00662)	(0.00703)	(0.00669)
India	0.00989	0.00798	0.0284***
	(0.0102)	(0.0106)	(0.00663)
Italy	-0.00721	-0.0162	-0.00741
	(0.0122)	(0.0125)	(0.0112)
Japan	0.00416	0.00118	0.00470
	(0.00539)	(0.00575)	(0.00586)
Korea	0.0923***	0.0992***	0.0976***
	(0.0114)	(0.0117)	(0.0115)
Russia	0.128***	0.135***	0.119***
	(0.0347)	(0.0353)	(0.0207)
Home country / Co-ethnicity # Same company [(ii)]:			
China # Same company		0.0382	0.0109
		(0.0389)	(0.0264)
Germany # Same company		-0.00435	-0.00812
		(0.0104)	(0.0105)
France # Same company		0.0635***	0.0255
		(0.0193)	(0.0187)
India # Same company		0.0440	0.00941
		(0.0347)	(0.0246)
Italy # Same company		0.111**	0.0743*
		(0.0461)	(0.0386)
Japan # Same company		0.0360**	0.0304*
		(0.0159)	(0.0159)
Korea # Same company		-0.112***	-0.130***
		(0.0378)	(0.0381)
Russia # Same company		-0.156	-0.0489
		(0.137)	(0.0800)
Returnee	0.122***	0.117***	0.113***
	(0.0179)	(0.0182)	(0.0183)
Country proximity controls	yes	yes	yes
Social distance dummies	yes	yes	yes
Citing patent characteristics	yes	yes	yes
Technology F.E.	yes	yes	yes
Constant	0.358***	0.359***	0.359***
	(0.0106)	(0.0106)	(0.0106)
Observations	1,004,950	1,004,950	1,004,950
R2	0.123	0.124	0.124
F	3020	2432	2443

[(i)] The table includes neither Iran nor Poland, as the number of observations is negligible

[(ii)] « Home country » effect in columns 1 and 2; Co-ethnicity in column 3

Clustered standard errors (at the cited patent level) in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 9 – Probability of citation from outside the US, as a function of inventors' country of residence (Home country) and Country of Origin (Co-ethnicity): BRICs only -- LPM regression

	(1)		
	Home country	Co-ethnicity	Home country # Co-ethnicity
China	0.0316*	0.0359***	-0.0314
	(0.0178)	(0.00740)	(0.0203)
India	-0.0771	0.0337***	0.0603
	(0.0471)	(0.00810)	(0.0491)
Russia	0.160*	0.110***	-0.148
	(0.0942)	(0.0255)	(0.101)
Controls [(i)]		yes	
Observations		621,283	
R2		0.120	
F		2071	

[(i)] Regression (1) includes controls and fixed effects as in regression (1) of Table 8.
Clustered standard errors (at the cited patent level) in parentheses; *** p<0.01, ** p<0.05, * p<0.1

This version: *03 March 2021*

APPENDIXES

Appendix 1 – Inventor names' disambiguation

We discuss here a few technical issues concerning name disambiguation, both general and specific of studies based on name and surname analysis, like ours. We also present succinctly one key feature of Massacrator 2.0, the name disambiguation algorithm at the basis of the EP-INV database, which we use for our research (for a detailed description, see Pezzoni et al., 2014).

Name disambiguation algorithms can be roughly classified into two groups: rule-based and Bayesian. Here we deal only with the former (for the latter, see: Li et al., 2014, and Ventura et al., 2015).[18]

A key element of rule-based name disambiguation algorithms consists in measuring the edit or phonetic distance between similar names/surnames, and setting some thresholds under which different names/surnames are considered the same ("matching"). Further information contained in the patent documents, as well as benchmarking is then used to validate the matches ("filtering"). Ideally, a good algorithm would minimize both "false negatives" (maximise "recall") and "false positive" (maximise "precision").

Precision and recall rates are measured as follows:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

where: $tp\ (fp) = \text{number of true (false)positives} \quad ; \ tn\ (fn) = \text{number of true (false)negatives.}$

False negatives occur whenever two inventors, whose names or surnames have been spelled or abbreviated differently on different patents, are treated as different persons. False positives occur when homonyms and quasi-homonyms are treated as the same person. Unfortunately, a trade-off exists between the two objectives, which requires making choices based on the consequences of each type of error for the subsequent analysis.

The three most important consequences for the analysis of ethnic citations are:

1. High precision/Low recall algorithms lead to underestimating the number of personal self-citations and overestimating that of co-ethnic citations. This is because all variants of the same inventor's name and surname will be, most likely, classified as belonging to the same ethnic group (for example, "Vafaie Mehrnaz" and "Vafaie Mehranz" will be both classified as Iranian, but a low recall algorithms may end up treating them as different persons, when instead they are one). When considering the two most important countries of origin of migrant inventors in the US, China and India, and before disambiguating inventors, we calculate a co-ethnic citation rate of respectively 20.5 and 15.2, which drop at 18.8 and 13.3 if we recalculate it after disambiguation. When applying the JTH methodology, this problem can be magnified by the presence of very prolific inventors, who are responsible for a large number of both cited and citing patents, and thus have the potential to generate a large number of false co-ethnic citations.
2. High precision/Low recall algorithms may also lead to underestimating the number of returnee inventors. If one Russian inventor patent as "Yavid Dimitriy" and as "Yavid Dimitriy" in Russia, he will

[18] The wave of interest for disambiguated inventor data has produced several open access inventor datasets. Two of them are: (i) the EP-INV dataset, originally developed for the identification of academic inventors, but comprising all inventors of patent applications filed at the European Patent Office from 1978 to around 2014 (http://www.esf-ape-inv.eu/index.php?page=3#EP-INV); and (ii) the US Patent Inventor Database, developed by Lee Fleming and associates, which contains USPTO data (http://dvn.iq.harvard.edu/dvn/dv/patent)

not be counted as a returnee (but his self-citations will be counted as a knowledge flow mediated by ethnicity). However, we suspect this to be a relatively minor problem, as figures of returnee inventors appear too low for their order of magnitude to change with a change in algorithms.

3. When applied to inventor sets from different countries of origin, the same matching rules return different results in terms of pre-filtering precision and recall, due to cross-country differences in the average length of text strings containing names and surnames, and in the relative frequency of common names and surnames. Chinese and Korean names and surnames, for example, are both short (which makes it arduous to tell them apart on the sole basis of edit distances) and heavily concentrated on a few, very common ones (such as Wang or Kim). The opposite holds for Russian surnames.

Three complementary strategies may help tackling these problems. The first one consists in making the best possible use of the contextual information contained in patents (that is, to correct for matching errors at the filtering stage). The second consists in using different algorithms to produce more than one datasets, each of which with different combinations of precision and recall, and using them to test the robustness of results. The third one consists in calibrating the disambiguation algorithm by collecting information on linguistic specificities of each country of origin, and exploit them at the matching stage. The information retrieval and computational costs increase when moving from the first to the third strategy. For this reason, Massacrator 2.0 does not follow the third one.

Massacrator 2.0's matches inventors on the basis of edit distances between all tokens comprised in the inventors' name-and-surname text strings, and then filters the matches by exploiting information on both the inventors and their patents.[19]

Massacrator 2.0 does not produce a unique dataset, but several ones, each of which is calibrated against a benchmark dataset in order to return a different combination of precision and recall. For this paper we started from the "balanced" calibration (which returns a precision rate of 88%, and a recall of 68%, when tested against a benchmark of French inventors) and slightly modified it. The modification consists in considering as positive cases (that is, the same person) all matched inventors whose patents are linked by at least one citation, irrespective of other filter criteria. This presumably allows for higher recall, and directly address the problem of over-estimation of ethnic citations.

To the extent that this modification induces higher recall at the price of lowering precision, it may lead to over-estimating the phenomenon of returnee inventorship (when the same inventor is first found to be active away from her country of origin, and then back to it). As shown in the paper's descriptive statistics, we find very few cases. Whether true or false positives, they are unlikely to affect our findings.

[19] As an example, consider "Dmitriy Yavid", a Russian inventor with a 2-token name-and-surname text string, and his fellow countryman "Sergei Vladimirovich Ivanov", with a 3-token name-and-surname string. As all of their tokens are pretty different, the two inventors will not be matched. Instead, "Dmitriy Yavid" and "Dimitriy Victorovich Yavid" will be matched, as, of the former's two tokens, one is identical to a token in the latter's, and another differs for just one character. The "Dmitriy Yavid" - "Dimitriy Victorovich Yavid" match will be then retained as valid if the two inventors' patents are either similar in contents, citation patterns, priority year, location in space, or property regime (same applicant); or if the two inventors have common co-inventors, or co-inventors who worked together. Otherwise they will be discarded as false matches.

Appendix 2 – Ethnic classification of inventors

When fed with a name and/or a surname, the *IBM-GNR system* returns a list of Countries of Association (CoAs) and two main scores:[20]

- "frequency", which indicates to which percentile of the frequency distribution of names or surnames the name or surname belongs to, for each CoA;
- "significance", which approximates the frequency distribution of the name or surname across all CoA.[21]

The IBM-GNR list of CoAs associated to each inventor is too long for being immediately reduced to a unique country of origin for each inventor in our database. This operation requires filtering a large amount of information through an *ad hoc* algorithm, one that compares the frequency and significance of the two lists of CoAs associated, respectively, to the inventor's name and surname to the inventor's "country of residence" at the moment of the patent filing (which we obtain from the inventor's address in the EP-INV dataset). Figure A2.1 illustrates the type of information provided by IBM-GNR, the position of our algorithm in the information processing flow, and the final outcome. Notice that we refer to "country of association" (CoA) when considering the raw information from IBM-GNR, and to "country of origin" when considering the final association between the inventor and one of the many CoAs proposed by IBM-GNR (or one of our "meta-countries" based on linguistic association). The full description of the algorithm is as follows:

I. We consider only inventors in the EP-INV database with at least one patent filed as US residents, or who cite at least one patent filed by US residents, and we assign them to either one of the 10 CoO of our interest, or leave her "unassigned" (which means she may be either a US "native" – whatever this might mean - or a migrant from other countries)

II. The 10 CoO of our interest are China, India, Iran, Japan, and South Korea (for Asia) and France, Germany, Italy, Poland, and Russia (for Europe). They share two characteristics: they belong to the top 20 CoO of highly skilled migrants in the US, according to OECD/DIOC stock figures for 2005/06 (Widmaier and Dumont, 2011); and their official language is neither English nor Spanish, which is a prerequisite for our algorithm to make sense when applied to migration into the US.[22]

III. For each inventor, we consider three indicators:
   a. The frequency of her first name(s) in English- and Spanish-speaking CoA [23]
   b. The product of the significances attached to her name and to the surname, for each CoA coinciding with one of the 10 CoO of our interest. Notice that, in principle, we could find that an inventor is associated to more than one of the 10 CoO of our interest, either via her name or her surname (for example, a French inventor of Italian descent may have a French name and an Italian surname). However, these cases are very few.

[20] Information on IBM-GNR reported here comes from IBM online documentation (http://www-01.ibm.com/support/knowledgecenter/SSEV5M/SSEV5M_welcome.html?lang=en; last visit: 19/1/2015) as well as: Patman (2010) and Nerenberg and Williams (2012). E-mail and phone exchanges with IBM staff were also decisive to facilitate our understanding. Still, being IBM-GNR a commercial product partly covered by trade secrets, we did not have entire access to its algorithms and we had to reconstruct them by deduction. For an application to a research topic close to ours, see Jeppesen and Lakhani (2010).

[21] For example, an extremely common Vietnamese surname such as Nguyen will be associated both to Vietnam and to France, which hosts a significant Vietnamese minority; but in Vietnam it will get a frequency value of 90, while it France it will get only, say, 50, the Vietnamese being just a small percentage of the population. When it comes to significance, the highest percentage of inventors names Nguyen will be found in Vietnam (say 80), followed by France and several Asian countries, with much smaller values.

[22] Language is an issue to the extent that our tools cannot distinguish English-speaking migrant inventors from US ones, nor Spanish-speaking migrants from one country of origin or another. This is why we cannot include in our analysis important origin countries such as the UK, Canada, Mexico and Cuba. We also have not yet included Ukraine and Taiwan, as this will require merging them with Russia and China, respectively. Two other countries in the top 20 list we have not included are Vietnam (too few observations among inventors) and Egypt (whose migrants into the US we cannot tell apart from those from other Arab-speaking countries).

[23] The intuition is as follows. An inventor with a typical Indian surname, such as Laroia, but named John or Luis is unlikely to be a recent Indian migrant into the US; this is because John and Luis are high-frequency names, respectively, in English-speaking and Spanish-speaking countries (among which we count US). More likely, he will be born in the US, possibly from mixed parents. On the contrary, Rajiv Laroia is more likely to be a first -generation Indian immigrant, as Rajiv is high-frequency name in India, a zero-frequency name in Spanish-speaking countries, and a low-frequency name in English-speaking countries that host Indian minorities.

c. The significance attached to the surname in the CoA associated to indicator n.2. [24]

As a result, we will have, for each inventor, one (or very few) candidates CoO and three indicators of potential success of this "candidacy".

IV. We set six possible threshold values for indicator n.1 (from 10 to 100, with steps of 20), eleven threshold values for indicator n.2 (from 0 to 10000, with steps of 1000), and six threshold values for indicator n.3 (from 50 to 100, with steps of 10). We consider 102 combinations of such threshold values ("calibrations"), and for each combination we assign each inventor to one or another CoO (or to no CoO at all). Each inventor is therefore associated to one vector of 102 dummies (one for each calibration) and a specific CoO, with dummy=1 indicating that the inventor comes for that CoO, and dummy=0 that she does not (no CoO assigned).[25]

V. We apply steps I. to IV. also to inventors in the WIPO-PCT database by Miguelez and Fink (2013), which report the inventors' nationality, which we use as benchmark to evaluate the precision and recall rates obtained by each calibration, for each CoO. We then identify Pareto-optimal calibration, namely the calibrations whose precision rate cannot be improved upon without losing out on the recall rate, and viceversa (blue dots in figures A2.2, which report the calibration results for China and Italy). Notice that the Pareto-optimal calibrations are not necessarily the same for all CoO; again from figure A2.2, one can see that the distribution of Pareto-Optimal calibrations for China is more convex than the one for Italy. In other words, the sharpness of trade-off between precision and recall differ across CoO: while for Italy we can attain a 70% precision rate only at the cost of reducing the recall rate to 10%, for China we reduce the latter only to 60%. The precision-recall trade-off can be considered a measure of the quality of our algorithm, per country. In general, quality is higher for Asian countries (with the exception of Iran) than for the European ones.

VI. Finally, we retain for our analysis two calibrations per CoO: a "high recall" calibration (one that ensures the highest recall value, conditional on precision being at least 30%); and a "high precision" calibration, one that requires precision to be no less than 70% . High recall values may include a large number of false positives (inventors wrongly assigned to one or another of the 10 CoO of interest), but also accommodate for a looser definition of migrant inventors, one that includes late-generation migrants. The latter's validity depends on the strength of ties binding such migrants to other US residents of the same descent and/or to their countries of origin (on which we have no *a priori* information).

In the present version of the paper, we make use only of "high recall" calibration results. To further compare data quality across CoO, we inspect the frequency distribution of values taken by indicator n.2 (figure A2.3). The more right (left) skewed the distribution, the better (worse) the quality: the most striking comparison here is between India and Italy, with the former clearly exhibiting higher quality. According to this measure, too, quality is generally higher for Asian countries (with the exception of Iran) than for European ones.

[24] The intuition is as follows: the indicator n.2 may have a high value due exclusively to a very high value of the significance for the name, with a moderate value for the significance of the surname. We wish the latter not to be too low.

[25] Keeping with the example from the previous footnotes, Rajiv Laroia will be associated to CoO=India, with a vector containing n<102 zeroes and 102-n ones. The ones are all associated with "high recall" combinations of high threshold values for indicator n.1 and low threshold values for nr.2 and nr.3 (such as, respectively, 70-5000-60; see figure 1), while the zeroes will be associated with "high precision" combinations (low threshold values for indicator n.1 and high threshold values for nr.2 and nr.3; such as, respectively, 30-8000-80). Rajiv Laroia will be confirmed having CoO=India only in the high recall case, but not in the high precision case (for which indicator nr.1 is too high). In practice, the high precision combination leaves the door open to Rajiv Laroia's CoO being the UK, and to Rajiv Laroia being possibly of Indian descent, but with no ties to India or to Indian migrants in the US.

Figure A2.1 From inventor data to the Ethnic-INV database

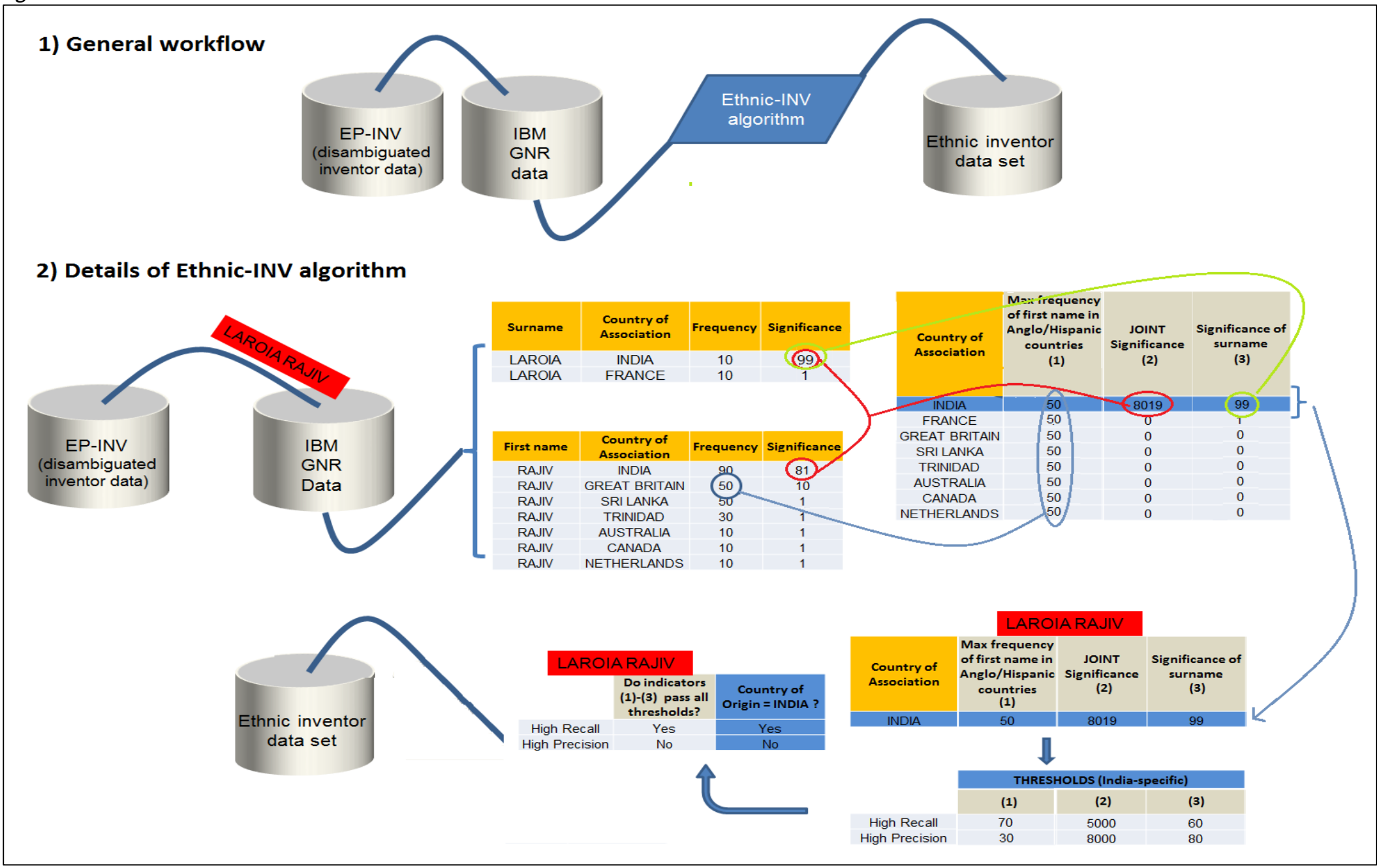


Figure A2.2 - Ethnic-INV algorithm calibration results: China and Italy

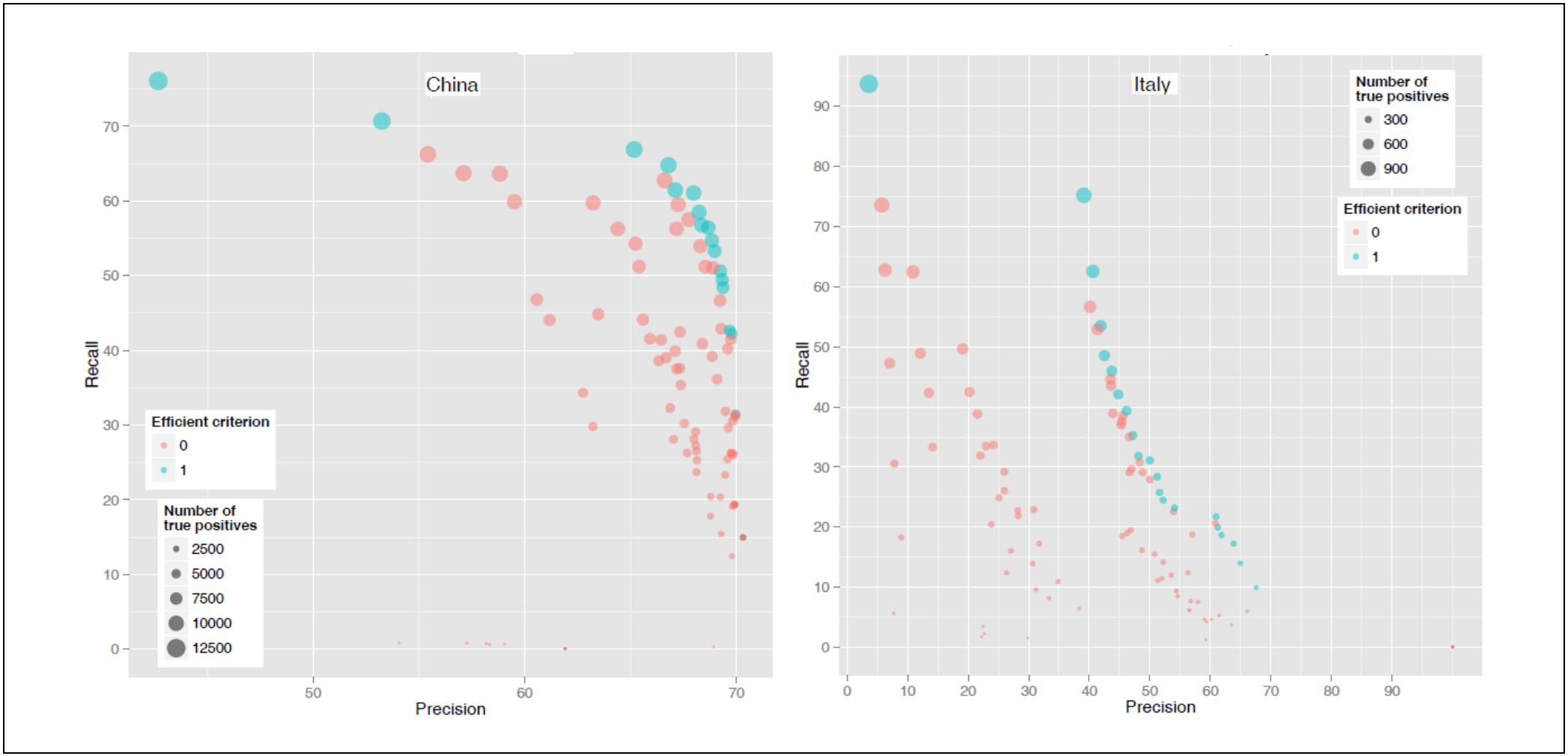
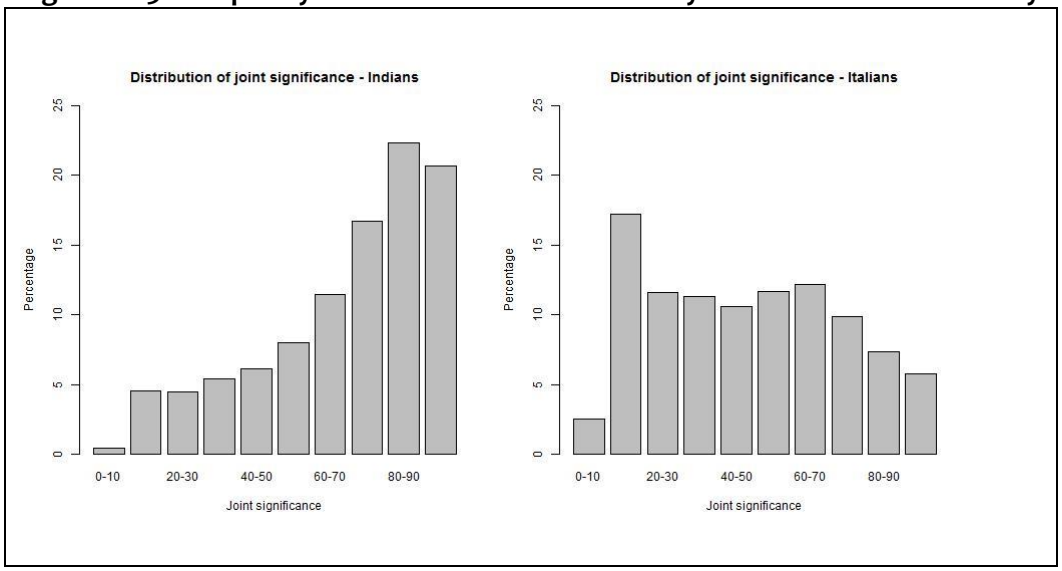


Figure A2.3 - Frequency distribution of values taken by indicator n.2: India vs. Italy



This is confirmed by a comparison between the distribution by CoO of our inventors and comparable distribution obtained from censual data. Table A2.1 reports information drawn from IPUMS-USA data for year 2000 (https://usa.ipums.org/usa/), namely:

- The percentage share of US residents with 4+ years of college education, born outside the US, by country of birth (aged 15 and above)
- The percentage share of US residents (all education levels, aged 15 and above), born in the US but of foreign ancestry, by ancestors' country.[26]

The two shares are compared to the shares of inventors of foreign origin in our database, for inventors with at least one patent in year 2000. The same information is displayed in figure A2.4, with ancestry information on the right axis.

Table A2.1 – Comparison of EP-INV and censual data for year 2000; by Country of Origin

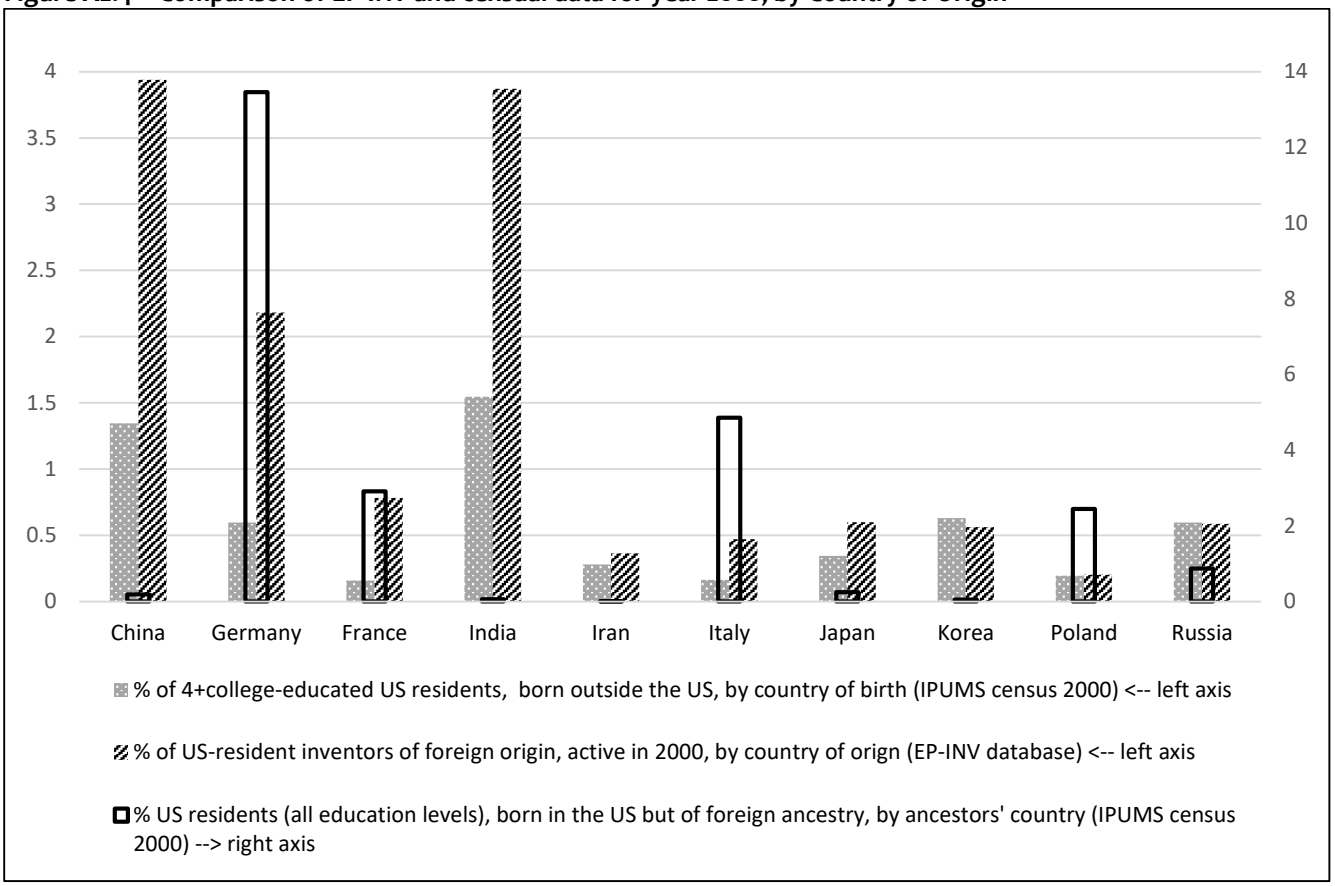
	% 4+college-educated US residents, born outside the US, by country of birth §	% US residents (all education levels), born in the US, by ancestors' country §	% US-resident inventors of foreign origin, active in 2000, by country of origin §§
China	1.346	0.189	3.938
Germany	0.598	13.457	2.181
France	0.159	2.912	0.782
India	1.547	0.067	3.872
Iran	0.28	0.016	0.366
Italy	0.164	4.861	0.470
Japan	0.345	0.252	0.599
Korea	0.631	0.059	0.564
Poland	0.196	2.452	0.204
Russia	0.597	0.874	0.587

§ source: IPUMS-USA census data

§§ source: EP-INV database

[26] Ancestry is an information provided by census respondents, which is subsequently recoded but not verified by census officials; respondents with mixed ancestry typically pick one, or rarely two, according to their own identity feelings; and census official recode, but not check the information.

Figure A2.4 – Comparison of EP-INV and censual data for year 2000; by Country of Origin



College-educated US residents are the best proxy for inventors we can get from censual data, based on the reasonable assumption that most inventors hold a college degree (especially in science-based fields, which we know to be the most affected by immigration). As for the share of US-born residents of foreign ancestry, this is indicative of the presence of many non-English surnames, and possibly names, which may induce the Ethnic-Inv algorithm to classify an inventor as of foreign origin, when in fact he or she maybe the descendant of 19th-20th century migrants.

We observe the share of college-educated foreign born to be very similar to that of inventors of foreign origin for Iran, Korea, Poland, Russia, and, to less extent, Japan. We take it as a suggestion that the Ethnic-INV algorithm does a relatively good job in these cases.

For China and India, the percentage of foreign-origin inventors is much higher than that of college-educated US residents; but we can explain that with the recent migration boom of scientists and engineers, as confirmed by many sources in the literature. At the same time, we observe that the percentage of US-residents with Chinese or Indian foreign ancestry is relatively small, which rules out a misclassification of the latter in the Ethnic-Inv database. The opposite holds for Germany, France and Italy, where again the percentage of foreign-origin inventors is much higher than that of foreign-born college-educated residents, but:

(1) the literature does not suggest, as for China and India, a recent migration wave of scientists and engineers;
(2) the percentage of US residents of foreign ancestry is very high, which suggests misclassification in the Ethnic-Inv database.

The problem appears to be particularly severe for Germany, where the difference between college-educated and inventors is very large, and the percentage of US residents of German ancestry is very high.

We further check the reliability of our data by comparing them to both WIPO-PCT data (which, as said above, provide information on nationality of inventors) and to estimates by Kerr (2008), who also uses a name-based ethnicity assignment algorithm, based on a different source than IBM-GNR (and for a more limited spectrum of countries of origin).

Table A2.2 reports the shares of inventors of foreign origin active in the US in 2000 (same as in table A2.1) with the shares of foreign inventors active in the US between 1995 and 2005, from the WIPO-PCT database. It is identical to table 1 in the main text. For all countries of interest, the share of inventors of foreign origin according to EP-INV is larger than the equivalent share of foreign inventors. Columns (3) and (4) report the results of z-test on proportions, which indicates these differences always to be significant. This is expected, as long-term migrants have the possibility to acquire US nationality over the years (and a cursory look at WIPO-PCT data suggests this to be the case, with some prolific inventors who declare different nationalities in their early vs late patents).

Table A2.2 – Comparison of EP-INV and WIPO-PCT data, by country

	% of US-resident inventors of:				
	foreign origin, active in 2000 [(i)]	foreign nationality, 1995-2005 [(ii)]		z [(iii)]	p-value [(iv)]
	(1)	(2)	(1)/(2)	(3)	(4)
China	3.938	3.763	1.05	2.57	0.01
Germany	2.181	1.038	2.10	29.62	0.00
France	0.782	0.589	1.33	6.92	0.00
India	3.872	2.984	1.30	14.38	0.00
Iran	0.366	0.110	3.33	18.94	0.00
Italy	0.470	0.228	2.06	13.30	0.00
Japan	0.599	0.483	1.24	4.62	0.00
Korea	0.564	0.482	1.17	3.30	0.00
Poland	0.204	0.111	1.83	7.41	0.00
Russia	0.587	0.469	1.25	4.80	0.00

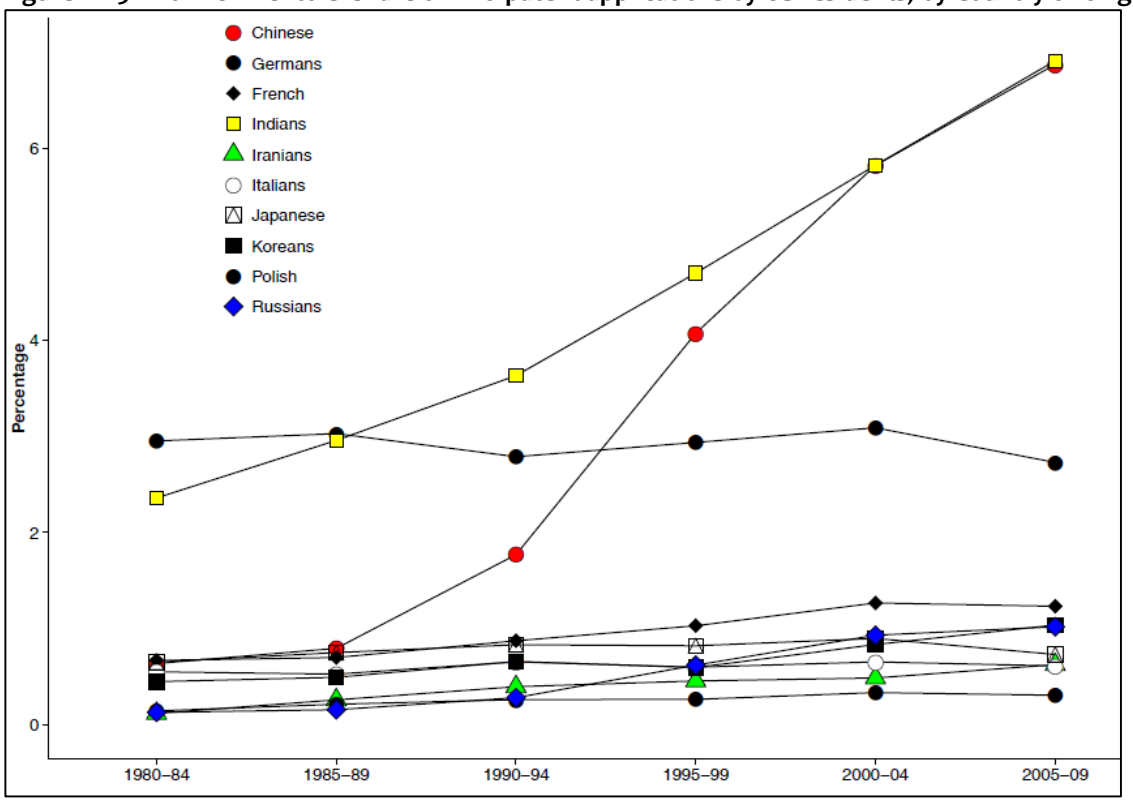
[(i)] source: EP-INV database
[(ii)] source: WIPO-PCT dataset (see Miguelez and Fink, 2013).
[(iii)] Normalized difference between (1) and (2)
[(iv)] p-values for z-test on $H_0 : (1) = (2)$

Still, we observe cross-country variations that may be due either to lack of precision in the Ethnic-INV algorithm or to differences in the propensity to acquire nationality for each migrant community (this in turn may be due to the average time spent by the migrants in the US, the number of US-born second-generation migrants, and the frequency of mixed marriages). In particular, we notice larger differences, in relative terms, for Germany, Italy, and Poland, where the share of foreign nationals is about double the share of foreign-origin inventors. Still, the differences for both Italy and Germany are much more limited than the ones observed in table A2.1 (comparison with college-educated foreign residents).

With a 3:1 ratio, Iran is a special case, as we know that neither Iran is an historical country of origin of US immigrants; nor Iranian surnames lack of distinctiveness. Hence, we conclude that many Iranian inventors may be part, or the immediate descendants, of the migration wave following the 1979 revolution, later to acquire (or obtain at birth by *ius soli*) the US citizenship.

We finally compare our data with those published by Kerr (2008) for a more limited set of countries of origin (China, India, Japan, Korea and Russia) and patents granted by the USPTO.[27] Figure A2.5 reports the share EPO patent applications by US residents of foreign-origin inventors, over the total of US residents' applications, from 1980 to 2010, for the 10 CoO of our interest. The observed trends are very similar, with the only exception of Indian inventors' patents in the 2000s, for which Kerr observes a decline and we do not. As for values, they are in the same order of magnitude but with our data exhibiting generally lower shares especially for Russia (from little more than 0% to around 1%, as opposed to 3% to 4.5% for Kerr), and with the exception of India (our share being overall 1% point higher then Kerr).

Figure A2.5 – Ethnic inventors' share of EPO patent applications by US residents; by Country of Origin



[27] Kerr considers "ethnic groups", as defined by the Melissa database for ethnic marketing, rather than specific CoO, namely: Chinese, Indian, Japanese, Korean and Russia, which correspond more or less to our CoO; Vietnam, which we do not consider; and European and Hispanic, which are too large aggregations of CoO for being of our interest.

Appendix 3 – Applicant vs examiner citations

Appendix 3 details the way in which we divide citations as coming from the applicants or from the examiners. This is used in section 4 of the main paper as a robustness check, and it is based on the assumption that the applicant-added citations, albeit noisy, may be more revealing of direct knowledge exchanges between inventors, while the examiner-added citations result from a less informed, less precise prior art search on bibliographic databases (Thompson, 2006).

The Patstat database from which EP-INV originates provides information on the origin of each citation to prior-art patents, as follows:

0 - SEA- citation introduced during search (that is, search for prior art, which occurs between the filing data and the publication date)
1 - APP- citation introduced by the applicant (after search)
2 -EXA- citation introduced during examination (which occurs after the search report is published)
3 - OPP - citation introduced during opposition
4 - 115- citation introduced according to Art 115 EPC
5 - ISR -citation from the International Search Report
6 - SUP - citation from the Supplementary Search Report
7 - CH2 - citation introduced during the Chapter 2 phase of the PCT
8- PRS – "Pre-Search" citations (available before official publication)

Citations with SEA can be further categorized with letters, which provide various information. The most important category for the present discussion being "D", which indicate a "document cited in the application". Specifically, the EPO Guidelines for Examination mention that: "When the search report cites documents already mentioned in the description of the patent application for which the search is carried out, these should be denoted by the letter "D"".

The sample considered in our database consists of citations from EPO patents to other EPO patents. However, the citation may be either a direct one or an indirect one, through one or more members of the patent families of citing and/or cited patents.

As a mode of example, Figure A3.1 illustrates the concept of indirect citation through patent family. It refers to the case of a patent with a USPTO priority and an equivalent EPO application, which cites another USPTO patent also having an EPO patent equivalent. The citation comes from two sources. On the one hand, the citation is reported in the International Search Report (ISR) contained in the patent document published after that the original USPTO patent was extended through an application to the WIPO-PCT system. On the other hand, the citation is also reported in the USPTO patent publication of the priority patent filing. The latter reports as origin of the citation both the applicant (APP) and a pre-search report. Notice that, of the two documents belonging to the same citing family, the one published by WIPO (WO2004004603) is the oldest one, which means that either the examiner spotted the citation before the applicant or that the citation from the International Search Report has been treated as an applicant one by the USPTO examiner. This may happen when the USPTO examines patent applications via the PCT (Patent Cooperation Treaty) procedure, as it treats all citations inserted by foreign patent offices as they were coming from the applicant, despite most of them coming instead from examiners of other patent offices (Alcacer et al., 2009).

Figure A3.1. Example of EPO application with USPTO priority

CITING patent (EP20030763301) and its family			
Application		Publication	
Number	Filing Date	Number	Date
EP20030763301	02JUL2003	EP1517653	30MAR2005
EP20030763301	02JUL2003	EP1517653	02MAY2007
AU20030248837	02JUL2003	AU2003248837	23JAN2004
JP20040519977	02JUL2003	JP2005532120	27OCT2005
JP20040519977	02JUL2003	JP4953571	13JUN2012
WO2003US21229	02JUL2003	WO2004004603	15JAN2004
WO2003US21229	02JUL2003	WO2004004603	03JUN2004
US20030612833 (*)	01JUL2003	US2004073190	15APR2004
US20030612833 (*)	01JUL2003	US7314484	01JAN2008

(*) Priority patent (provisional filing thereof, dated 2002)

Origin: ISR

Origin: APP, PRS

CITED patent (EP19980912985) and its family			
Application		Publication	
Number	Filing Date	Number	Date
AT19980912985T	18MAR1998	AT343981	15NOV2006
CA19982285473	18MAR1998	CA2285473	03DEC1998
CA19982285473	18MAR1998	CA2285473	12JUN2007
DE1998636320	18MAR1998	DE69836320	14DEC2006
DE1998636320T	18MAR1998	DE69836320	18OCT2007
EP19980912985	18MAR1998	EP0971645	19JAN2000
EP19980912985	18MAR1998	EP0971645	02NOV2006
ES19980912985T	18MAR1998	ES2276457	16JUN2007
JP19990500645	18MAR1998	JP2002514117	14MAY2002
JP19990500645	18MAR1998	JP4094684	04JUN2008
WO1998US05470	18MAR1998	WO9853765	03DEC1998
US19970820213(*)	18MAR1997	US5824054	20OCT1998

(*) Priority patent

Adopting the logic illustrated above, we collected the origin and, where available, the category of citations using information from PatStat. Given that we reconstruct citations through patent families and in most cases each patent family contains more than one member, the relationship between citing and cited patents is a M:N relation. In the example of Figure A3.1, two patents (M=2) in a given family cited the same patent (N=1) in another family. In other, more general cases the two patent families (i.e. the citing and the cited one) may be linked by more complex relations (i.e. it may be M>1 and N>1). The key problem thus becomes to adopt some criterion to decide whether the citation has been inserted by the applicant or whether it has been added by the examiner. We proceeded as follows.

1. We adopted a rule according to which we consider a citation as having been inserted by the applicant if the category of the citation is "D" or the origin of the citation is "APP". All the other origins and categories were considered as pointing to an examiner citation.
2. For each citing-cited patent pair, we counted how many times the citation was added by the examiner and how many times it came from the applicant. In the example above, the citation was added twice by the examiner (i.e. ISR plus PRS) and once by the applicant (i.e. APP).
3. We created two dichotomous variables, "applicant" and "examiner", which take value 1 if the count of citations from, respectively, the applicant and the examiner is greater than zero. Note that the two variables are not mutually exclusive, i.e. they can both take value one for the same citation pair. This is the case of the example reported above, in which the citation came both from the examiner and from the applicant.

4. Finally, we repeated step 3. above by taking into account only the first citing patent in the family, i.e. the patent (or patents) with the oldest filing date (possibly corresponding to the priority date). In the example above, the first patent was US20030612833 with (priority) date July 1 2003. Considering only this patent, the citation came once from the applicant (i.e. APP) and once from the examiner (i.e. IRS).

Tables A3.1 reports summary statistics for the national sample (which we used for testing the diaspora effect) and the international one (brain gain effect). We notice immediately that the percentage of applicant citations is much larger in the national sample than in the international one (60.1% vs 41.5%). If we consider citations only from applicants, the difference is even more striking (40.5% against 22.9%). This is due both to differences in the origin of the citations, and to differences in the citation behaviour of US vs non-US applicants.

Table A3.1. Summary statistics of citation origin

National sample	
Number of unique citations pairs	503,103
Pairs with info on origin/category	486,054
% of citations from applicant [(a)]	60.1
% of citations from examiner [(b)]	59.5
% of citations *only* from applicant [(c)]	40.5
%of citations *only* from examiner [(d)]	39.9
% of citations *both* from examiner & applicant	19.6
International sample	
Number of unique citations pairs	444,916
Pairs with info on origin/category	430,160
% of citations from applicant [(a)]	41.5
% of citations from examiner [(b)]	77.1
% of citations *only* from applicant [(c)]	22.9
%of citations *only* from examiner [(d)]	58.5
% of citations *both* from examiner & applicant	18.6

(a) It counts a citation as coming from the applicant irrespective of whether the same citation was also added by the examiner in the same or in some other member of the citing patent family.
(b) It counts a citation as coming from the examiner irrespective of whether the same citation was also added by the applicant in the same or in some other member of the citing patent family.
(c) It counts a given citation as coming only from the applicant if no citation was added by the examiner in the same or in some other member of the citing patent family..
(d) It counts a given citation as coming from the examiner if no citation was added by applicant in the same or in some other member of the citing patent family..

Table A3.2 reports the distribution of citations across origin (applicant vs examiner), by source (patent authority). In the national sample, the citations from patent authorities other than the USPTO are mostly examiner citations, while the opposite is true for the USPTO. This is expected, due to the peculiarity of the US system (duty of candour rule). But when we move to the international sample, the proportion of applicant vs examiner citations from the USPTO is reversed. Thus, non-US applicants (from whom most citations in the international sample come from) provide, in relative terms, many fewer citations than US ones, as they do not conform to the duty of candour rule.

Table A3.2. Origin (applicant vs examiner) of patent citations, by source (patent authority)

Patent Authority	Applicant only	Examiner only	Both

National Sample				
EPO	35.9%	59.6%	4.5%	*100*
USPTO	54.6%	35.6%	9.8%	*100*
WIPO	25.7%	69.2%	5.2%	*100*
Others	1.6%	98.1%	0.2%	*100*
International Sample				
EPO	19.8%	73.0%	7.1%	*100*
USPTO	33.7%	61.6%	4.7%	*100*
WIPO	22.8%	69.4%	7.9%	*100*
Others	9.9%	86.9%	3.3%	*100*

Table A3.3 reproduces the results of an OLS regression that replicates on our data the exercise conducted by Lampe (2012). Lampe (2012) finds that "applicants withhold between 21% and 33% of relevant citations known to them" (p.320). He obtains this result by comparing the citations introduced by examiners on a focal patent to the citations introduced by the focal patent's applicant in its own previous applications. He finds that, very often, applicants cite some prior art in a patent filing at time *t* (which proves they were aware it), but do not do it at time *t'>t*, when it is the examiner who does it, thus providing evidence that the prior art was relevant.

Thus, in Table A3.3, the observation set pools all citing-cited pairs from both the national and the international samples, but retains only those in which the cited patent was already "known" to the applicant (it appears among the backward citations of the same applicant's prior patents). The dependent variable is binary one, =1 if the citation comes from the applicant (at least one applicant citation in the patent family). All regressors are dummies, indicating whether:

- the citation comes from the international sample
- the citing applicant is located outside the US (applicant address is a non-US one)
- the priority patent in the citing family was first applied for at the USPTO, before 2001 (that is, before the USPTO started releasing information on the origin of the citations)

The regression also includes an interaction term between the first two regressors.

We notice that, in line with the descriptive statistics, the citations from the International sample are less likely to originate from the applicant, and the same applies to the citation from non-US firms, with the two effects being complements.

Table A3.3. OLS - Probability citation comes from applicant

Constant	0.78	(0.0)*
International Sample	-0.13	(0.0)*
Non US firm	-0.03	(0.0)*
Non US firm x Intl Sample	-0.03	(0.01)*
Pre-2001 US application	-0.32	(0.0)*
Nr obs	247957	
AIC	311613.67	
BIC	311666	

Standard Errors in parentheses

* indicates significance at p = 0.01

Summing up, when dealing with the national sample, we can interpret the applicant vs examiner citations as in Thompson (2006) and the related, US-centric literature. This is because most citations come from the USPTO, which ensures both a sizeable proportion of applicant citations and no measurement error in the origin attribution (that is, most applicant citations come indeed from the applicant, albeit it is unclear whether it was the inventor or the attorney to produce them). But when we move to the international sample, we have little hope of reproducing Thompson (2006) results, since the share of applicant citations

drops dramatically, and a very high proportion of such citations is most likely to come from examiners of patent offices outside the US and not at all by applicants.

Appendix 4 – Regression analysis (Diaspora effect): Further robustness checks

We deal with the disparities in the precision of our Ethnic-Inv algorithm by running some robustness checks. First, we exploit information on the nationality of inventors, for the subset of inventors who also have patents in the WIPO-PCT database. Based on information on patent families provided by PatStat, we first identified all patents in the WIPO-PCT database that are equivalents of EP-INV patents in our sample. Within each pair of equivalent patents we name-matched inventors on the EPO patent to inventors on the WIPO-PCT one: around 90% of positive matches result from perfect name string matching, the remaining from a combination of Soundex matching of surname and first given name (around 9%), 2-gram string matching or manual checking (less than 125). This allowed us to assign a nationality to all inventors in the EP-INV database with at least one patent in the WIPO-PCT database. We then retain only the cited patents (and the related citing and control ones) in which the inventors' countries of origin and of nationality coincide. This reduces the sample to around one fifth of the initial one (see table A4.1). Notice that the distribution by CoO/Nationality is very similar in the two samples. For results and related comments, see table 6 in the paper.

Table A4.1. National and international samples based on Country of Origin (full sample) vs Nationality-based samples (for robustness checks); by CoO/Nationality of cited inventors

	National sample (citations from within the US)				Int'l sample (citations from outside US)			
	Full sample		Nationality sample		Full sample		Nationality sample	
	obs.	%	obs.	%	obs.	%	obs.	%
China	249,348	23.9	84,644	35.61	256,244	25.5	55,192	33.79
Germany	175,570	16.83	26,400	11.11	177,564	17.67	21,788	13.34
France	66,170	6.34	14,912	6.27	68,100	6.78	11,218	6.87
India	324,034	31.06	67,310	28.32	316,466	31.49	46,810	28.66
Iran	29,044	2.78	2,608	1.1	-			
Italy	46,664	4.47	8,018	3.37	46,228	4.6	5,890	3.61
Japan	48,172	4.62	13,328	5.61	53,150	5.29	10,190	6.24
Korea	51,774	4.96	10,402	4.38	49,024	4.88	7,406	4.53
Poland	16,064	1.54	2,604	1.1	-			
Russia	36,480	3.5	7,470	3.14	38,174	3.8	4,826	2.95
Total	**1,043,320**	**100**	**237,696**	**100**	**1,004,950**	**100**	**163,320**	**100**

Second, we test whether our results depend exclusively from the most important high-tech clusters within the US, which are likely to attract a disproportionate number of highly skilled migrants. We focus on the top six MSAs by number of patent applications in our sample (S.Francisco, S.José, NY, Dallas, Boston, and S.Diego) and on the top ten MSA pairs with the highest number of citations running in one or another direction (that is, the ten most important city corridors for citation flows; see table A4.2). We then control for the fixed effects of either the top MSAs or the top corridors (table A4.3). Our main results remain unaltered.

Table A4.2. Top-10 cross-MSA citation corridors

MSA name	MSA name	Citations (both directions)
San Jose-Sunnyvale-Santa Clara, CA	San Francisco-Oakland-Fremont, CA	8931.80
New York-Northern New Jersey-Long Island, NY-NJ-PA	San Francisco-Oakland-Fremont, CA	7194.53
New York-Northern New Jersey-Long Island, NY-NJ-PA	Boston-Cambridge-Quincy, MA-NH	6846.82
New York-Northern New Jersey-Long Island, NY-NJ-PA	San Diego-Carlsbad-San Marcos, CA	6834.77
San Francisco-Oakland-Fremont, CA	Boston-Cambridge-Quincy, MA-NH	6702.78
New York-Northern New Jersey-Long Island, NY-NJ-PA	San Jose-Sunnyvale-Santa Clara, CA	5909.32
San Francisco-Oakland-Fremont, CA	San Diego-Carlsbad-San Marcos, CA	5059.78
New York-Northern New Jersey-Long Island, NY-NJ-PA	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	4866.75
San Jose-Sunnyvale-Santa Clara, CA	Boston-Cambridge-Quincy, MA-NH	4496.28
New York-Northern New Jersey-Long Island, NY-NJ-PA	Chicago-Joliet-Naperville, IL-IN-WI	3638.95

Table A4.3. Probability of citation from within the US, as a function of co-ethnicity, controlling for inventor's location (top MSA or top corridor fixed effects) -- LPM regression

	(1)	(2)	(3)	(4)
Same MSA	0.0364***	0.0363***	0.0296***	0.0296***
	(0.00349)	(0.00349)	(0.00347)	(0.00347)
Same State	0.0264***	0.0264***	0.0167***	0.0166***
	(0.00294)	(0.00294)	(0.00283)	(0.00283)
ln(Miles)	-0.00548***	-0.00550***	-0.00775***	-0.00778***
	(0.000680)	(0.000679)	(0.000714)	(0.000713)
Co-ethnic	0.0402***		0.0400***	
	(0.00173)		(0.00173)	
China		0.0563***		0.0560***
		(0.00242)		(0.00242)
Germany		0.0101**		0.00977*
		(0.00502)		(0.00505)
France		-0.0105		-0.0109
		(0.0110)		(0.0110)
India		0.0342***		0.0343***
		(0.00248)		(0.00249)
Iran		0.0521**		0.0506**
		(0.0240)		(0.0241)
Italy		0.0108		0.0117
		(0.0343)		(0.0347)
Japan		0.0275*		0.0279*
		(0.0142)		(0.0142)
Korea		0.0347***		0.0345***
		(0.0133)		(0.0133)
Poland		-0.0467		-0.0435
		(0.0416)		(0.0416)
Russia		0.0705***		0.0713***
		(0.0157)		(0.0157)
Top MSA FE	yes	yes	no	no
Top corridors FE	no	no	yes	yes
Soc.dist dummies	yes	yes	yes	yes
Patent characteristics	yes	yes	yes	yes
Technlogy FE	yes	yes	yes	yes
Constant	0.668***	0.668***	0.677***	0.677***
	(0.00547)	(0.00547)	(0.00548)	(0.00548)
Observations	1,043,320	1,043,320	1,043,320	1,043,320
F	2,438	1,813	2,074	1,599
R2	0.081	0.081	0.081	0.081

Third, we consider the possibility of cohort effects, with different generations of migrant inventors (from the same CoO) having different propensities to share knowledge with members of their communities. In order to control for that, we run two regressions, with year fixed effects (where the year corresponds to the priority date of the cited patents; table A4.4). Our main results remain unchanged.

Table A4.4. - Probability of citation from within the US, as a function of co-ethnicity, controlling for inventor's cohort (using year of citing patent, adding dummies representing six five-year periods) -– LPM regression

	(1)	(2)	(3)
Same MSA	0.0318***	0.0318***	0.0345***
	(0.00347)	(0.00347)	(0.00353)
Same State	0.0201***	0.0201***	0.0200***
	(0.00281)	(0.00281)	(0.00281)
ln(Miles)	-0.00613***	-0.00613***	-0.00615***
	(0.000667)	(0.000667)	(0.000667)
Co-ethnic	0.0383***		0.0412***
	(0.00174)		(0.00194)
Co-ethnic * MSA			-0.0174***
			(0.00432)
China		0.0531***	
		(0.00243)	
Germany		0.0102**	
		(0.00506)	
France		-0.0117	
		(0.0110)	
India		0.0332***	
		(0.00249)	
Iran		0.0508**	
		(0.0240)	
Italy		0.0125	
		(0.0344)	
Japan		0.0275*	
		(0.0142)	
Korea		0.0292**	
		(0.0133)	
Poland		-0.0450	
		(0.0413)	
Russia		0.0711***	
		(0.0157)	
Social distance dummies	yes	yes	yes
Citing patent characteristics	yes	yes	yes
OST FE	yes	yes	yes
Year FE	yes	yes	yes
Constant	0.714***	0.713***	0.713***
	(0.00541)	(0.00541)	(0.00540)
Observations	1,043,320	1,043,320	1,043,320
F	2552	1885	2460
R2	0.081	0.082	0.081

Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Fourth, we consider the possibility that the high significance of several coefficients in tables 3 to 5 may depend on the very large number of observations in our sample – which may decrease the variance of the estimators. We run again the regressions in table A4.5 with samples of reduced size, by applying the bootstrap technique described by Greene (2008, p.596) and Wooldridge (2002, p.378). As reported in table A4.5, the coefficients are maintained, but the standard errors increase as the size of the subsamples diminishes. Despite this, significance is always maintained for India and China, as well as for Russia with the exception of the last case (smallest sample). In regressions 4 and 8, with many dummies, not all subsamples lead to convergence, so results are based on a smaller set of replications. Estimates based of 1% subsample do not include the last column, since any of the subsample was able to converge.

Table A4.5. Probability of citation from within the US, as a function of co-ethnicity, bootstrap regressions -- LPM

	(1)[a]	(2)[b]	(3)[a]	(4)[b]	(5)[a]	(6)[a]
	sample of 10% size, 50 reps.		sample of 5% size, 50 reps.		sample of 1% size, 50 reps.	
Same MSA	0.0341***	0.0314***	0.0341***	0.0314***	0.0341***	0.0314***
	(0.00369)	(0.00366)	(0.00443)	(0.00435)	(0.00924)	(0.00909)
Same State	0.0215***	0.0216***	0.0215***	0.0216***	0.0215***	0.0216***
	(0.00237)	(0.00237)	(0.00343)	(0.00341)	(0.00750)	(0.00752)
ln(Miles)	-0.00605***	-0.00603***	-0.00605***	-0.00603***	-0.00605***	-0.00603***
	(0.000645)	(0.000647)	(0.000804)	(0.000803)	(0.00193)	(0.00192)
Co-ethnic	0.0433***		0.0433***		0.0433***	
	(0.00151)		(0.00236)		(0.00554)	
Co-ethnic * MSA	-0.0177***		-0.0177***		-0.0177*	
	(0.00375)		(0.00517)		(0.0106)	
China		0.0562***		0.0562***		0.0562***
		(0.00198)		(0.00295)		(0.00761)
Germany		0.0105**		0.0105		0.0105
		(0.00469)		(0.00666)		(0.0154)
France		-0.0105		-0.0105		-0.0105
		(0.0111)		(0.0125)		(0.0288)
India		0.0344***		0.0344***		0.0344***
		(0.00252)		(0.00316)		(0.00638)
Iran		0.0508**		0.0508		0.0508
		(0.0244)		(0.0342)		(0.0675)
Italy		0.0125		0.0125		0.0125
		(0.0330)		(0.0446)		(0.0883)
Japan		0.0278**		0.0278		0.0278
		(0.0132)		(0.0192)		(0.0427)
Korea		0.0345***		0.0345*		0.0345
		(0.0127)		(0.0183)		(0.0391)
Poland		-0.0434		-0.0434		-0.0434
		(0.0471)		(0.0521)		(0.132)
Russia		0.0710***		0.0710***		0.0710
		(0.0151)		(0.0218)		(0.0464)
Soc.dist. dummies	yes	yes	yes	yes	yes	yes
Patent characteristics	yes	yes	yes	yes	yes	yes
OST FE	yes	yes	yes	yes	yes	yes
Constant	0.668***	0.669***	0.668***	0.669***	0.668***	0.669***
	(0.00573)	(0.00575)	(0.00730)	(0.00728)	(0.0170)	(0.0170)
Observations	1,043,320	1,043,320	1,043,320	1,043,320	1,043,320	1,043,320
Wald chi2	82830	95009	85540	121191	0.080	0.080
R2	0.080	0.080	0.080	0.080	-713590	-713519

[a] Specification as in column 5 of table 4

[b] Specification as in column 1 of table 5

Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table A4.6 – Probability of citation from within the US, as a function of co-ethnicity, by technological class of cited patents -– LPM regression

	Electrical engineering; Electronics	Instruments	Chemicals; Materials	Pharma & Biotech.	Industrial processes	Mechanical engineering; Transport	Consumer goods; Civil engineering
Same MSA	0.0399***	0.0420***	0.0135**	0.0152**	0.0312***	0.0295**	-0.00487
	(0.00529)	(0.00597)	(0.00633)	(0.00620)	(0.00895)	(0.0133)	(0.0176)
Same State	0.00181	0.0131***	0.0369***	0.0555***	0.00413	-0.0128	-0.0150
	(0.00393)	(0.00450)	(0.00516)	(0.00505)	(0.00679)	(0.0105)	(0.0139)
ln(Miles)	-0.00648***	-0.00733***	-0.00419***	-0.000453	-0.0119***	-0.0183***	-0.0182***
	(0.00110)	(0.00120)	(0.00112)	(0.00110)	(0.00167)	(0.00245)	(0.00340)
China	0.0454***	0.0220***	0.0658***	0.0594***	0.0243***	0.0159	-0.0269
	(0.00470)	(0.00562)	(0.00334)	(0.00312)	(0.00832)	(0.0157)	(0.0304)
Germany	-0.00489	0.00643	0.0318***	0.0222***	0.000138	-0.00130	-0.0577**
	(0.0106)	(0.00805)	(0.00933)	(0.00822)	(0.0130)	(0.0177)	(0.0251)
France	0.0344	-0.0445**	-0.0254	-0.0267*	-0.0832**	-0.0537	-0.0447
	(0.0222)	(0.0212)	(0.0171)	(0.0156)	(0.0361)	(0.0582)	(0.0577)
India	0.0332***	0.00472	0.0495***	0.0393***	0.00433	0.0230*	-0.0435**
	(0.00338)	(0.00563)	(0.00487)	(0.00469)	(0.00820)	(0.0134)	(0.0210)
Iran	0.0338	0.0661	0.0463	0.142**	0.128*	-0.0592	-0.567***
	(0.0303)	(0.0405)	(0.0735)	(0.0704)	(0.0691)	(0.0701)	(0.122)
Italy	-0.0172	-0.0252	0.0130	0.0520	-0.0625	0.0888	-0.155
	(0.0332)	(0.0476)	(0.0322)	(0.0552)	(0.0517)	(0.0937)	(0.0979)
Japan	-0.0306	0.0244	0.0424*	0.0522***	0.0324	-0.0575	0.117
	(0.0261)	(0.0287)	(0.0217)	(0.0197)	(0.0431)	(0.0804)	(0.145)
Korea	0.0217	0.0681**	0.0205	0.0129	0.0736**	0.0255	-0.0107
	(0.0234)	(0.0281)	(0.0197)	(0.0199)	(0.0364)	(0.0692)	(0.119)
Poland	-0.0181	-0.0159	0.0241	-0.130**	0.275**	-0.347**	0.217
	(0.0955)	(0.0827)	(0.0642)	(0.0581)	(0.115)	(0.151)	(0.248)
Russia	0.0463**	0.0583*	0.128***	0.103***	0.0360	0.0281	0.0601
	(0.0233)	(0.0308)	(0.0278)	(0.0268)	(0.0475)	(0.0923)	(0.0997)
Social distance dummies	yes	yes	yes	yes	yes	yes	yes
Citing patent characteristics	yes	yes	yes	yes	yes	yes	yes
OST FE	yes	yes	yes	yes	yes	yes	yes
Constant	0.645***	0.614***	0.609***	0.576***	0.574***	0.620***	0.597***
	(0.00989)	(0.00911)	(0.00870)	(0.00868)	(0.0148)	(0.0216)	(0.0401)
Observations	338,598	314,880	300,338	364,106	118,550	44,796	23,252
R2	0.079	0.091	0.105	0.099	0.128	0.112	0.096
F	1136	1071	1231	1236	655.0	319.4	137.7

Clustered standard errors in parentheses ; *** p<0.01, ** p<0.05, * p<0.1

Table A4.7 – Probability of citation from outside the US, as a function of "home-country" effect, co-ethnicity or co-nationality (also by Country of Origin) -– LPM regression

	HOME COUNTRY	CO-ETHNICITY	CO-NATIONALITY
	(1)	(2)	(3)
Same company	0.194***	0.195***	0.194***
	(0.0131)	(0.0127)	(0.0129)
Home country / Co-ethnicity / Nationality [§]:			
China	0.0421***	0.0377***	0.0427***
	(0.0110)	(0.00897)	(0.00996)
Germany	0.00668	0.00674	0.0110
	(0.0106)	(0.00899)	(0.00954)
France	0.0106	0.0182	0.0146
	(0.0143)	(0.0138)	(0.0136)
India	-0.0376*	0.00949	-0.00492
	(0.0227)	(0.0145)	(0.0174)
Italy	0.0296	0.0114	0.0382
	(0.0368)	(0.0297)	(0.0307)
Japan	0.0127	0.0181	0.0172
	(0.0132)	(0.0128)	(0.0128)
Korea	0.126***	0.117***	0.124***
	(0.0254)	(0.0250)	(0.0252)
Russia	0.0723	0.107**	0.0848
	(0.0819)	(0.0485)	(0.0627)
Home country / Co-ethnicity / Nationality # Same company [§]:			
China # Same company	-0.120*	0.00945	-0.0132
	(0.0711)	(0.0543)	(0.0648)
Germany # Same company	-0.0159	-0.0185	-0.0149
	(0.0187)	(0.0183)	(0.0181)
France # Same company	0.0570*	0.0464	0.0509*
	(0.0305)	(0.0295)	(0.0295)
India # Same company	0.0693	-0.00967	0.00206
	(0.0689)	(0.0574)	(0.0639)
Italy # Same company	0.100	0.0564	0.0770
	(0.0698)	(0.0706)	(0.0709)
Japan # Same company	0.0745**	0.0631*	0.0711**
	(0.0333)	(0.0332)	(0.0334)
Korea # Same company	-0.153	-0.176*	-0.179*
	(0.107)	(0.105)	(0.105)
Russia # Same company	-0.219**	-0.0541	-0.185***
	(0.0863)	(0.107)	(0.0633)
Returnee	0.140***	0.139***	0.138***
	(0.0272)	(0.0271)	(0.0271)
Country proximity controls	yes	yes	yes
Social distance dummies	yes	yes	yes
Citing patent characteristics	yes	yes	yes
Technology F.E.	yes	yes	yes
Constant	0.446***	0.446***	0.446***
	(0.0183)	(0.0182)	(0.0182)
Observations	163,320	163,320	163,320
R2	0.098	0.098	0.098
F[§§]	.	425.5	.

[§] « Home country » in columns 1 ; co-ethnicity in columns 2 ; co-nationality in columns 3

[§§] F-statistic not computed by our software package due to near-collinearity of some predictors (in particular, the Technology F.E.)

Clustered robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

References

Alcácer, J., Gittelman, M. and Sampat, B. (2009) 'Applicant and examiner citations in US patents: An overview and analysis', *Research Policy,* 38(2), 415-427.

Greene, W.H., 2008, Econometric analysis (6th edition). Pearson Education.

Jeppesen, L.B., Lakhani, K.R., 2010. Marginality and problem-solving effectiveness in broadcast search. Organization science, 21(5), 1016-1033.

Kerr, W.R., 2008. The Ethnic Composition of US Inventors. Harvard Business School Working Paper No. 08-006.

Li, G.-C., Lai, R., D'Amour, A., Doolin, D.M., Sun, Y., Torvik, V.I., Yu, A.Z., Fleming, L., 2014. Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). Research Policy, 43(6), 941-955.

Miguelez, E., Fink, C., 2013. Measuring the International Mobility of Inventors: A New Database, World Intellectual Property Organization-Economics and Statistics Division.

Nerenberg, S., Williams, K., 2013, The Case for Analytical Name Scoring over Name VariantExpansion - IBM® InfoSphere Global Name Management report. IBM Corporation, Armonk, NY

Patman, F., 2010, Advanced Global Name Recognition Technology - IBM® InfoSphere Global NameManagement report. IBM Corporation Armonk, NY

Thompson, P. (2006) 'Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations', *The Review of Economics and Statistics,* 88(2), 383-388.

Ventura, S.L., Nugent, R., Fuchs, E.R.H., 2015. Seeing the Non-Stars:(Some) Sources of Bias in Past Disambiguation Approaches and a New Public Tool Leveraging Labeled Records. Research Policy, (forthcoming).

Widmaier, S., Dumont, J.-C., 2011, Are recent immigrants different? A new profile of immigrants in the OECD based on DIOC 2005/06. OECD Publishing, Paris.

Wooldridge, J., 2003, Introductory Econometrics: A Modern Approach South-Western College Pub.