Breschi, Stefano; Lissoni, Francesco; Miguelez, Ernest

Working Paper

Foreign inventors in the US: Testing for Diaspora and Brain Gain Effects

CReAM Discussion Paper Series, No. 09/15

Provided in Cooperation with:
Rockwool Foundation Berlin (RF Berlin)

Suggested Citation: Breschi, Stefano; Lissoni, Francesco; Miguelez, Ernest (2015) : Foreign inventors in the US: Testing for Diaspora and Brain Gain Effects, CReAM Discussion Paper Series, No. 09/15, Centre for Research & Analysis of Migration (CReAM), Department of Economics, University College London, London

This Version is available at:
https://hdl.handle.net/10419/295517

CReAM

Centre for Research & Analysis of Migration

Discussion Paper Series

CPD 09/15

Foreign inventors in the US:
Testing for Diaspora and Brain Gain Effects

Stefano Breschi, Francesco Lissoni and Ernest Miguelez

Centre for Research and Analysis of Migration
Department of Economics, University College London
Drayton House, 30 Gordon Street, London WC1H 0AX

www.cream-migration.org

Foreign inventors in the US: Testing for Diaspora and Brain Gain Effects

Stefano Breschi [1], Francesco Lissoni [1/2/§], Ernest Miguelez [2/3/4]

[1] *CRIOS – Università Bocconi, Milan*

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

[3] *AQR-IREA (Barcelona, ES)*

[4] *CReAM (London, UK)*

[§] contact author: francesco.lissoni@u-bordeaux.fr

This version: *27 July 2015*

Abstract

We assess the role of ethnic ties in the diffusion of technical knowledge by means of a database of patent filed by US-resident inventors of foreign origin, which we identify through name analysis. We consider ten important countries of origin of highly skilled migration to the US, both Asian and European, 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 Asian countries, but not for European ones, with the exception of Russia. Diaspora effects do not translate necessarily into a brain gain effect, most notably for India; nor brain gain occurs only in presence of diaspora effects. Both the diaspora and the brain gain effects bear less weight than other knowledge transmission channels, such as co-invention networks and multinational companies.

Keywords: migration, brain gain, diaspora, diffusion, inventors, patents

JEL codes: F22, O15, O31

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 mobility of scientists and engineers has been marked by the convergence of two previously unconnected streams of literature. First, research in the geography of innovation has explored the role of social ties in facilitating knowledge diffusion, and in determining its spatial reach. Among such ties, a good deal of attention has been paid to those binding migrant scientists and engineers from India to the US (Agrawal et al., 2008; Almeida et al., 2014). Second, migration and development scholars have explored to what extent highly skilled migrants contribute to innovation in their home countries. This may occur through international knowledge flows (Kapur, 2001; Kuznetsov, 2006) as well as by facilitating foreign direct investments (Foley and Kerr, 2013) or by entrepreneurial returnee migration (Nanda and Khanna, 2010; Saxenian, 2006).

While this convergence has produced important advances, it remains to explore to what extent the two phenomena are intertwined, namely whether social ties between migrants are either a necessary or a sufficient condition for generating substantial knowledge feedbacks to home countries. In addition, both streams of literature have focussed almost exclusively on migration flows from China and India to the US. This overlooks the fact that Europe is an equally important source of highly skilled migration (Docquier and Marfouk, 2006; Widmaier and Dumont, 2011).[1]

We contribute to this emerging literature by analysing the forward citation patterns of patent applications filed by foreign inventors in the US from five Asian countries (China, India, Iran, Japan, and South Korea) and five European ones (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 for name disambiguation.

We test for the existence of "diaspora" and "brain gain" effects. We state a diaspora effect to exist when migrant inventors from the same country of 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

[1] Europe is also an important destination region, albeit far distanced from the US. This has also been disregarded by the literature, with just a few exceptions (such as Niebhur, 2010, and Nathan, 2015).

when the migrants' patents are disproportionately cited by inventors active in their countries of origin, so that the latter stand to gain from high skilled migration. We find evidence of the diaspora effect for Asian inventors, but not for the European ones, with the exception of Russians. Diaspora ties, however, appear to carry less weight than social ties established through co-inventorship (as in Breschi and Lissoni, 2009).

In the case of China, India and Russia, the diaspora effect takes an international dimension, as migrant inventors in countries other than the US also have privileged access to knowledge produced by co-ethnic, US-based inventors. However, this translates into brain gain only for China and Russia, as well as South Korea. As for advanced countries such as France, Italy, and Japan, brain gain passes through multinational enterprises. We detect no effect of sorts 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 the results of our empirical exercise (section 4). Section 5 concludes. A substantial set of appendixes on methodological issues and robustness checks is attached (for online publication).

2. Background literature

2.1 Localized knowledge flows and the role of social ties

Localized knowledge flows are a key topic in the geography of innovation (Breschi, 2011). Under the form of pure externalities, they play a key role in Marshallian and Jacobian location theories (Ellison et al., 2007; Henderson, 1997). 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 the debate has been that of measurement, which is fraught with technical as well as conceptual difficulties.

These were first tackled by Jaffe et al. (1993), who introduced the use of patent citations along with a simple, yet influential methodology (from now on, JTH test). The test makes use of two sets of patent pairs. The first one includes a sample of cited patents and all the related citing ones, with exclusion of self-citations at the company level (cited-citing or "case" pairs); the second includes the same sample of cited patents, with

citing ones replaced by controls with 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 carried out, one that proves the share of co-localized cases to be significantly higher than the share of co-localized controls. The test can be generalized by means of regression analysis, with the probability of a citation to occur as the dependent variable, and the stacked sets of cited-citing and cited-control patent pairs as observations (Singh and Marx, 2013).[2]

Further research has refined the JTH test in the direction of uncovering the actual mechanisms behind localized knowledge flows and their economic characteristics. Breschi and Lissoni (2005a, 2009) show that a large share of localized patent citations are self-citations at the individual level, associated to inventors who move or consult across firms, in the same location or region. Other localized citations occur between socially close inventors, namely inventors located at short geodesic distances on co-inventorship networks. Collaboration forces inventors to share knowledge or create strong enough obligations to share also with collaborators' neighbours. Agrawal et al. (2006) show that social ties of this kind may resist to physical distance, as when inventors who have relocated keep being cited by former co-inventors who have not.

This line of research has evolved in the direction of uncovering other forms of social ties besides the collaboration ones, and of exploring their relationship with spatial distance. To this end, Agrawal et al. (2008) asses the importance of ethnic ties in the US-resident population of Indian inventors, which the literature describes as a closely-knit diaspora (Kapur, 2001). Based on an Indian surname database, the authors identify a large number of ethnic Indian, US-resident inventors of USPTO patents. Second, they apply and extend the JTH methodology, by including inventors' co-ethnicity among the explanatory variables. Indian inventors are found to be more likely to cite one another's patents than patents by non-Indian ones. Besides, co-ethnicity and co-location seem to act as substitutes, with Indian inventors activating their ethnic connections to reach outside their metropolitan area. Almeida et al. (2014) also rely on an *ad hoc* collection of surnames to identify

[2] Technical refinements of the JTH test also concern 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 (Alcacer and Gittelman, 2006; Breschi and Lissoni, 2005b; Thompson, 2006).

Indian inventors in the US semiconductor industry. They find both evidence of intra-ethnic citations, as well as some indications that reliance on such citations is correlated to inventors' productivity.

Agrawal et al. (2011) extend the Agrawal et al.'s (2008) data and methodology to the case of international knowledge flows and find that patents by Indian inventors in the US do not seem to attract a higher-than-average rate of citations from the inventors' home country. The only (weak) exceptions are patents in Electronics, and patents owned by multinational firms. Overall, these results go in the direction of suggesting that the Indian diaspora is not a major source of knowledge feedbacks for the home country. It is at this point that studies in the geography of innovation tradition blend with research on migration and development.

2.2 Migrants' contribution to innovation in origin countries

Migration studies have traditionally looked for possible positive returns on emigration for origin countries. Early research placed special emphasis on financial remittances and their role in capital formation. More recently, due to the increasing importance of highly skilled migration, more attention has been paid to contributions to knowledge stock and innovation (Bhagwati and Hanson, 2009).[3]

These may come in 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, 2001; Meyer and Brown, 1999)

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

(iii) *Returnees' direct contribution.* Migrant scientists and engineers may decide to move back to or set up entrepreneurial activities in their home countries, while keeping in touch with knowledge sources in the destination countries (Wadhva, 2007a,b, and references therein).

[3] Still another stream of literature, which we do not address in this paper, focuses on the positive contribution of highly skilled migration to the accumulation of human capital in origin countries (Mountford, 1997; Stark and Wang, 2002).

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

A series of papers by William Kerr and co-authors has made progress by exploiting two sources of information:

- the NBER Patent Data File, by Hall et al. (2001), which includes information on name, surnames, and addresses of inventors
- the Melissa ethnic-name database, a commercial repository of names and surnames of US residents, classified according to nine broad ethnic groups, the most distinctive ones being the Asian ones

As for knowledge flows, Kerr (2008) focusses on citations running from patents filed at the USPTO in the last quarter of the 20th century by inventors from outside the US to patents filed by US residents (company self-citations excluded). By crossing the ethnicity of the inventors and the technological class of patents from both within and outside the US, Kerr produces over 100k cells and counts the citations falling in each cell. A negative binomial regression, with citations as the dependent variable and cells as observations, shows that co-ethnic cells exhibit a higher citation count than mixed ones, controlling for technology. The result is interpreted as evidence of brain gain by migrants' sending countries.

Foley and Kerr (2013) exploit the same database to investigate the specific role of ethnic inventors in relation to multinational companies' activities in origin countries. They find evidence that migrant inventors may act as substitutes of local intermediaries, thus diminishing their companies' costs of engaging into foreign direct investments.

As for returnee inventors, Agrawal et al. (2011) manage to identify just a handful. Similarly, Alnuaimi et al. (2012) examine around 3500 USPTO patents assigned to over 500 India-located patentees (local firms, subsidiaries of foreign companies, and universities) over 20 years, and find very few inventors once active in subsidiaries of foreign companies who then move to local firms. This suggests that, as far as India is concerned, returnees are not a massive source of knowledge transfer.

As for multinationals, Choudhury (2015) finds evidence that local employees in Indian R&D labs of large multinationals exhibit higher inventiveness when directed by returnee managers, which the author interpret as indicative of the latter's role as brokers of knowledge produced at the headquarters for use as subsidiaries. Branstetter et al. (2015) highlight the transfer role of co-invention between Indian and Chinese employees of multinational subsidiaries and their colleagues in the US.

A more recent contribution by Miguelez (2014) exploits the information on inventors' nationality contained in PCT patent applications up to 2011. The author estimates the impact of foreign inventors on the extent of international technological collaborations between origin and destination countries, as measured by co-patenting activity. Findings suggest a positive and significant impact for all countries of origin, and not just the largest ones, such as China and India.

2.3 Data issues

The importance assumed by inventor data in the innovation literature has pushed several scholars to improve the quality and transparency of their data mining efforts. A key topic is that of name disambiguation, which consists in assigning a unique ID to inventors whose name or address is reported differently on several patent documents (Ge et al., 2015; Li et al., 2014; Martínez et al., 2013; Marx et al., 2009; Pezzoni et al., 2014; Raffo and Lhuillery, 2009; Ventura et al., 2015). We discuss technical details in Appendix 1. Here we point out some substantive implications for migration studies.

Ideally, a good disambiguation algorithm would minimize both false negatives (maximise "recall") and false positives (maximise "precision"). In practice, a trade-off exists. High precision/low recall algorithms lead to underestimating the number of personal self-citations and overestimating co-ethnic ones. These biases may vary according to the inventors' countries of origin, as disambiguation algorithms are sensitive to country-specific details.

So far, patent-based studies on migration and innovation have ignored these issues. Kerr (2008) and extensions make use of non-disambiguated data. Agrawal et al. (2008, 2011) and Almeida et al. (2014) do not provide details on disambiguation, while Alnuaimi et al. (2012) resort to "perfect matching", which works as an extreme high precision / low recall algorithm.

Precision and recall issues also appear when assigning inventors to a country of origin, based on their names/surnames. Some studies discuss openly this issue, and usually decide to go for maximizing precision. For example, Agrawal et al. (2008) identify Indian inventors based on a very narrow list of Indian surnames, which are both highly frequent in India and validated by experts as indicative of recent migration status. This implies a tendency to limit the attention to first-generation migrants, which in turn hides the assumption that the strength of ethnic ties weakens with time. While making sense, the assumption is not very precise about the generational timing of the decay and does not consider the possibility of "ethnic revival" phenomena and "reverse brain drain" policies (Kuznetsov, 2006, 2010; Zweig, 2006). Information on inventors' nationality is a valid substitute of name analysis, but it can also be regarded as a high precision/low recall algorithm. In fact, long term migrants who acquire nationality of the country of residence can be considered as false negatives.

Technical concerns also arise when dealing with patent applicants. All studies claim to control for company self-citations. Yet, they are silent on the methodologies they follow in order to identify companies and business groups. This is in contrast with recent data harmonization efforts (Du Plessis et al., 2009; Peeters et al., 2010; Thoma et al., 2010).

Using raw or poorly treated applicant data can be equated to applying a high precision/low recall disambiguation technique. With localization studies, this leads to underestimating company self-citations and overestimating the co-location of knowledge externalities. At the international level, it underplays the role of multinationals as carriers of knowledge, and overemphasize that of social ties between inventors.

3. Propositions and data

In this section we first formulate our research questions by means of a set of empirical propositions. We then describe our dataset. Despite attaching great importance to data mining methodology, we relegate most details in Appendix 2.

3.1 Research questions: diaspora and brain gain effect

We are interested in exploring how social ties between migrant inventors from the same country of origin affect the diffusion of technical knowledge. Emerging naming conventions, as reviewed in section 2, label such ties as "ethnic" or "co-ethnic", a synthetic though imperfect adjective we will also adopt, for want of better alternatives. However, when referring to the individuals themselves, we will opt for expressions such as "inventors of foreign origin", "inventors from the same country of origin" (both expression involving second- and further-generation migrants) or, when more appropriate, "migrant" inventors.

Social ties between migrants are interesting insofar as they may exist independently of or precede shared professional experiences and/or physical proximity. They may have been formed in the destination country (as a result of homophily in the choice of acquaintances and friends; Currarini et al., 2009) or inherited from the home country (as with chain migration). In both cases, they represent an instance of vitality and relevance of a community of expatriates, to which we will refer as a diaspora. We state a diaspora effect to exist when inventors from the same country of origin and active in the same country of destination have a higher propensity to cite one another's patents, as opposed to patents by other inventors, other things being equal and excluding self-citations at the company level. We test for its existence by adapting the JTH methodology, as described in section 2, and building a sample of cited-citing & cited-control patent pairs. Cited patents are all signed by at least one foreign inventor in a given destination country (in our case, the US), while citing and control patents are signed by inventors (foreign and local) also located in the destination country.[4] 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 dependent variable is a binary one, which is =1 if the two patents in the pair are linked by a citation. The main variable of interest, *co-ethnicity*, is a dummy =1 when both patents in the pair have been invented by at least one inventor from the same country of origin. As for

[4] We follow the legal jargon in distinguishing between the *filing* of a patent, an action undertaken by the patent applicant (most often a company), and its *signing*, which is an action undertaken by the inventor (always a physical person). Filing grants ownership (a pure economic right), while signing grants authors' rights (a mix of moral and economics rights).

spatial distance, based on the addresses of inventors, we measure it both in terms of co-location and as a continuous variable (as in Marx and Singh, 2013). Social distance refers to geodesic distances on the network of inventors (Breschi and Lissoni, 2009). Whenever one or both patents in a pair have multiple inventors, we consider minimum social and spatial distances. As for further controls, they mostly refer to the characteristics of patents in the pair (especially the citing/control patents), based on the large literature on the determinants of patent citations (Hall et al., 2005; Harhoff et al., 2003). We provide full details of our sampling scheme and specification in the next two subsections.

Ethnic ties may also play a role at the international level. Most importantly, they may induce a brain gain effect, by which residents in the countries of origin of migrant inventors cite disproportionately the latter's patents. We are interested in considering them separately from other brain gain sources, such as returnee inventors' self-citations, and in weighing their importance against multinational companies' self-citations.

We adapt once again the JTH methodology. We still consider all cited patents signed by at least one foreign inventor in the US, but now citing and control patents are those signed by inventors from outside the US. We keep in the sample all patent pairs by the same inventors (returnee inventors) as well as pairs from the same company or business group, but control for them. We then proceed to regression analysis, by modifying equation (1) as follows:

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

where the dependent variable is the same as in (1), and 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 country of origin of the foreign inventor of the cited one. *Returnee* and *Same company* are also dummies, which is =1 if both patents in the pair have been signed by the same inventor or filed by the same company or business group, respectively. Other controls are as in (1), with adaptations.[5]

[5] Most notably, spatial distance cannot be measured with co-location dummies, since by construction inventors of cited and citing patents do not reside in the same city. Notice that networks of inventors may span across countries, which justifies including social distance in (2), of which returnee is a special case (with social distance =0).

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 then abroad.[6] We control for this by re-inserting the co-ethnicity dummy and interacting it with the home country variable. This raises the possibility that the "home country" effect and the "co-ethnicity" effect may not coincide. In particular, an "international diaspora" may exist which is not associated to any brain gain for the country of origin, and yet play a role at the level of worldwide flows, very much like what observed for trade of heterogeneous goods between countries hosting sizeable ethnic communities (Felbermayr et al., 2010; Rauch and Trindade, 2002).

3.2 Data

3.2.1 Patent and inventor data

Our data results from matching names and surnames of inventors in the EP-INV inventor database (Tarasconi and Coffano, 2014) with information on their countries of origin obtained by Global Name Recognition, a name search technology produced by IBM (from now on, IBM-GNR).

The database contains information on uniquely identified inventors listed on patent applications filed at the EPO from 1978 to around 2010.

Information on inventors includes their home address, as harmonized in the RegPat database (Maraut et al., 2008). Inventor names are disambiguated through a three-step algorithm (see Appendix 1 for a description), followed by a manual check for all citing patents entering our final sample..

Other relevant information are the patents' priority year and technological field, and the identity of their applicants. As for the latter, we adopted the harmonization of applicant names performed by the EEE-PPAT and OECD-HAN projects. Moreover, we also carried out an *ad hoc* work of reconstruction of business groups using Bureau van Dijk Zephyr database on mergers & acquisitions.[7]

---

[6] By inventive activity of international standing we mean one which translates into patents filed at the most important patent offices worldwide, one of which is EPO. Most noticeably, the number of patents filed at SIPO, the Chinese patent office, have literally exploded over the past few years, but this does not translate into an equivalent explosion of patents filed abroad by Chinese applicants.

[7] On OECD-HAN see Thoma et al. (2010) and http://www.oecd.org/sti/inno/43846611.pdf (last visited, May 2015). On EEE-PPAT, see: http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-

IBM-GNR system is a commercial product based upon information collected by the US immigration authorities in the first half of the 1990s. When fed with either a name or a surname, IBM-GNR returns a list of Countries of Association (from now on: CoA) and some statistical information on the strength of the association. As the original dataset included only non-US citizens, the US itself is never listed among the possible CoA.

We treat this information by means of an original algorithm that we describe in Appendix 2. In a nutshell, its purpose is to select one and only one Country of Origin (from now on: CoO), by picking the CoA to whom the inventor's name and surname is most strongly associated. When no CoO can be selected (no association is strong enough) inventors are treated indifferently as locals or foreigners from an unknown CoO.

For the purposes of this paper, we consider all inventors who reside in the US and whose CoO is one of the following: 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 high skilled migrants to the US according to OECD/DIOC data, release 2005/6 (Widmaier and Dumont, 2011). At the same time, none of them has English or Spanish as official languages, which are the most widely spoken languages in the US and our algorithm would find it hard to deal with.

We calibrate our algorithm against a benchmark dataset on the nationality of inventors located in the US, as obtained from PCT patent applications (Miguelez and Fink, 2013). We then retain the results of a "high recall" calibration, one that minimizes false negatives (foreign-origin inventors from the selected CoO mistaken for locals), at the price of low precision. We do so in order to avoid a bias in favour of finding positive co-ethnicity effects in equations (1) and (2).

In Appendix 2 we discuss at length the quality of our ethnic classification. In particular, we compare our results for inventors with census information on US residents by ancestry or country of birth (source: IPUMS-USA). Overall, we find our classification to be more reliable for Asian CoO (with the possible exception of Iran) than for European ones, with Germany being the most problematic case. For this reason, we will test

RA-11-008 (last visited, May 2015). Manual checks are necessary both for companies in different countries and for multinational groups, whose boundaries change over time (so that a patent by company *x* must be assigned to either group *y* or *z* or no group at all, depending on the filing date).

the robustness of our econometric results by using a subset of our data where co-ethnicity is defined according to the inventors' nationality.

Also in Appendix 2 we compare our calculations of foreign inventors' share of patents in the US with those published by Kerr (2008b) for Asian CoO and Russia, over time. The observed trends are very similar. As for values, they are in the same order of magnitude but with our data exhibiting generally lower shares, especially for Russian inventors.

3.2.2 Sampling

We select all patent applications from the EP-INV database, with priority years comprised between 1990 and 2010, and at least one inventor with residence in the US, but a CoO included among the ten of our interest. Our starting sample includes 88,522 inventors and 174,160 patents. We then retain only the applications that have received at least one forward citation from another EPO patent application (either directly, or indirectly, via one or another's patent family).[8]

On this basis, we build two different samples, a "local" and an "international" one, which we will use for investigating the diaspora and brain gain effects, respectively.

For the local sample we retain all cited-citing pairs in which the citing patent comprises among its inventors at least one US-resident. Then, we 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 of interest. 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 under the same IPC groups of the citing patent [9]
3. it comprises among its inventors at least one US-resident

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

[9] Notice that the same patent may be assigned to several IPC groups. Therefore, our matching criteria require the citing patent and its control to be classified under the same number of IPC groups, and to share them all.

This leaves us with 1,043,320 observations, one half of which are cited-citing pairs, the other half cited-control pairs. These are generated by the combination of 89,986 cited patents, 195,595 citing ones and 279,623 controls. Table 1 (part 1) reports details by CoO. As expected, more than half the observations come from the two largest CoO, China and India. The only European country in the same order of magnitude is Germany.

Table 1 HERE

As 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., as above, and, instead of condition 3., the condition of *not* comprising among its inventor any US-resident.[10]

This leaves us with 1,050,236 observations – excluding all the cited-citing pairs (and their respective controls) for which controls cannot be computed. These are generated by the combination of 105,059 cited patents, 266,629 citing ones, and 390,519 controls. Table 1 (part 2) shows that their distribution by CoO of cited patents' inventors is very much the same as that for the local sample.

In the regression setting (which is identical for the two samples), observations are "stacked" and flagged as different by means of the binary variable *Citation* (=1 for cited-citing pairs, =0 for cited-control pairs). Our dependent variable is then the probability of *Citation* being equal to 1, which we estimate by means of a Linear Probability Model (OLS; Logit estimates in Appendix 4).

For all patent pairs in the two samples, we produce the following dummy variables, which enter as independent variables in all the regressions:

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

[10] Notice that, conversely, the cited patent may include, alongside with the US-resident inventor(s), one or more foreign residents. This makes it necessary to control, in our regressions, for the distance between the latter and the inventors of the citing/control patents.

2. *Social distance* S (with S=0,1,2,>3,+∞) : =1 if the geodesic distance between cited patent and the citing (control) is equal to S . Formally: *S = min (Sij)* with S_{ij}=geodesic distance between inventor *i* (*i*=1…I) on the cited patent and inventor *j* (*i*=1…J) on the citing (control) one, as calculated on the entire network of inventors, for all inventors on the cited and the citing (control) patents. Notice that for *i=j* →S=0. If all *i*s and all *j*s belong to disconnected network components then: S=+∞. For each year *t* we calculate a different network of inventors, based on co-inventorship patterns of all patents with priority years (t±5).[11]

3. *Miles*: shortest distance (in miles) between the two patents, based on their inventors' addresses, of which take the log with the addition, in some specifications, of a quadratic term,.[12]

4. Characteristics of the citing (control) patent, as suggested by Marx and Singh (2013), such as: its technological field (OST-30 classification, as from Tarasconi and Coffano, 2014), the number of claims (*claims*), the number of backward citations to prior art (*backward citations*) and to non-patent literature (*NPL citations*), as well as its technological proximity to the cited patent (nr of overlapping IPC-7 codes – *overlap IPCs 7* – and nr of overlapping full IPC codes, out of all codes assigned to the patents).

For the patent pairs in the local 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 we also calculate:

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.[13]

[11] 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).

[12] 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 values at the ZIP code level, the centroid of the city was used (or the county, if the city's was missing, too)

[13] 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 that two inventors in the cited and citing

8. Other measures of country proximity, such as: border-sharing (*Contiguous countries*), *Former colonial relationship*, and language-sharing (*English*, =1 if at least one inventor of the citing (control) patent is located in an English-speaking country; and *Similarity to English*, a language similarity index ranging from 0 to 1, adapted from Miguelez, 2014)
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 2 reports the descriptive statistics for all variables in both samples; for details by country, see tables A3.1-10 in Appendix 3.

Table 2. HERE

Notice that the same cited patent enters our sample as many times as the number of citations it receives. The same applies to each citing patent that cites more than one cited patent. This required correcting for non-independence of errors in regression, which we did by clustering errors by cited patent.

4. Results

4.1 Within-US knowledge flows and the diaspora effect

Table 3 reports the results of five different specifications of equation (1), which do not distinguish by CoO of ethnic inventors. The first specification reproduces Agrawal et al.'s (2008) basic exercise for Indian inventors in the US, which focusses on co-ethnicity and MSA co-location; the second and third ones introduce social

(control) patents are both located outside the US and in the same country, which is not necessarily the CoO of the inventor(s) of the cited patent.

distances between inventors; and the remaining ones add further controls, first for patent characteristics, including technology fixed effects, then for spatial distance.

Table 3 HERE

Estimated coefficients in column (1) have the same sign and are of the same order of magnitude as those in Agrawal et al. (2008): co-ethnicity affects positively the probability to observe 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, which suggests a substitution effect between spatial and ethnic proximity.

When controlling for social distance on the network of inventors (column 2) the estimated coefficients for co-location drops sharply, as social distance affects negatively the probability of citation and it is positively correlated with spatial distance. We also notice that the marginal effect of social distance reduces sharply when the latter increases (the absolute value of coefficients first increases sharply, then less and less). All these results are in accordance with previous findings by Breschi and Lissoni (2009). In addition, we notice that the coefficient for co-ethnicity also shrinks, but not as much (the interaction terms remains unaltered). This suggests that while co-location is strongly correlated to social proximity on the network of inventors, the same does not apply to co-ethnicity. Still, when interacting network-based social distance and co-ethnicity (column 3) we notice that the interaction is positive and significant for social distances higher than 3. This indicates the existence of a substitution effect between network-based social proximity and co-ethnicity, too. Social ties based on ethnicity kick in only when those based on professional experience are lacking. We finally observe that network-based social distance is generally associated to larger marginal effects than co-location or co-ethnicity.

Controlling for the characteristics of patents (column 4) does not alter much the coefficients of interest, which we take as a sign that the original sampling scheme was valid. Adding controls for spatial distance

further alters the estimated co-efficient of co-location (column 5), but neither those for social distance and co-ethnicity.

In table 4 we allow for the estimated coefficient of co-ethnicity to vary across CoO, first without interaction with MSA co-location (column 1), then with interaction (column 2). The importance of co-ethnicity for the probability of citation varies by CoO. Its estimated coefficient is clearly positive and significant only for Asian countries (although unstable across the two specifications for Japan and Iran), Russia, and Germany (although unstable). Marginal effects appear to be higher for Russia followed, in descending order, by China, Iran, India, South Korea, Japan, and, at some distance, 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 results on substitution effects between physical and ethnical proximity are entirely driven by Chinese and Indian inventors. The coefficients for social distance and other controls (unreported) do not differ much from those in table 3.

Table 4 – HERE

Cross-CoO differences in the size and significance of the co-ethnicity coefficients may depend either from the demographic composition of social groups from the same CoO (share of first vs second-generation migrants and/or long established ethnic minorities) or from their social structure (cohesiveness of the social group). These characteristics depend, in turn, on how well we calibrate our algorithm for each specific CoO. The lower its precision, the more likely it is that we mix first generation migrants with established communities (e.g. Italian young PhDs with Italian Americans in New Jersey) or migrants from different CoO, but with the same 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 residents' ancestry. We find measurement errors to be most likely for German inventors, followed at considerable distance by Italians and, at further distance, French and Polish ones.

One way to assess the relative weight of substantive factors vs measurement errors is to make use of a different definition of foreign-origin inventor. In table 5 we exploit information on inventors' nationality, which is a more stringent one. We retrieve such information from PCT patents, for all inventors in our original sample who had at least one patent in the WIPO-PCT database by Miguelez and Fink (2013) (see again Appendix 2). This reduces the sample to around one fifth of the initial one. We then run two sets of regressions: in the first set 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 similar specifications (column 1 to 3, and column 2 to 4) we notice that, in general, the co-nationality one is slightly larger. This suggests that our definition of foreign-origin inventors may comprise late-generation migrants or ethnic communities whose mutual bonds are not as strong as those between first-generation migrants. Still, our results do not change substantially. Coefficients for Poland remain negative, and 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 they do not differ much. Overall, this suggests that, with the exception of Germany, no European countries among those we considered exhibits a diaspora effect, and this is not just a statistical artefact due to a measurement error problem.

Table 5 –HERE

Logit regressions equivalent to OLS ones in tables 3 to 5 can be found in tables A4.2 to A4.4 in Appendix 4. The sign and significance of estimated coefficients do not change, nor the order of magnitude of marginal effects (as estimated at means).

To further probe into the robustness of our results, we estimate separate regressions for different technological classes (of the cited patents). We expect the ratio between first- and late-generation migrants to be higher in science-based technologies, whose inventors are more likely to be PhD holders and possibly

academics, two social categories in which non US-born residents are over-represented (Auriol, 2010; Scellato et al., 2015). [14]

Table 6 shows that Pharmaceuticals & Biotech is the technological class with the most instances of a positive and strongly significant coefficient for co-ethnicity (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. At the same time, these results are not much in contrast with our general ones, as France and Italy never exhibit positive and significant co-ethnicity coefficients while, at the opposite end, this is the case in five instances for China and three for India and Russia, two for South Korea and Germany, and one for Iran, Japan and Poland).

Table 6 – HERE

Appendix 5 reports the results of more robustness checks. First, we test whether our results depend exclusively on the most important high-tech clusters within the US, which are likely to attract a disproportionate number of highly skilled migrants (tables A5.2 and A5.3). Second, we consider the possibility of cohort effects, with different generations of migrant inventors having different propensities to share knowledge with members of their communities (table A5.4). In both cases our main results remain unchanged.

Third, in table A5.5 we consider the possibility that the high significance of several coefficients in tables 3 to 5 depends exclusively on the very large size of our sample. We apply the bootstrap techniques described by Greene (2008, p.596) and Wooldridge (2002, p.378) to specifications (1) and (2) in table 4. While standard errors increase, estimated coefficients stay significant for India and China, as well as for Russia (with one exception only).

[14] Science-based technological classes are those with a high average ratio between backward citations to prior art (pre-existing patents) and to non-patent literature (which is largely made of scientific publications; Callaert et al., 2006), as well as with a high share of academic patents (Lissoni. 2012).

4.2 *International knowledge flows and the brain gain effect*

Table 7 reports the distribution of the international sample according to the values of explanatory variables *Home country* and *Co-ethnicity*, by CoO. Column (5) shows the percentage of observations (patent pairs) with inventors of the citing/control patent who both come from and reside in the same CoO of the cited one (*Co-ethnicity*=1 and *Home country*=1). All developed countries exhibit values over 70%, as native inventors are disproportionately active at home, rather in foreign countries (except the US). On the contrary, BRIC countries have much lower values (from 30% for Russia to 48% for China), due to a more limited inventive activity at home and a large presence of migrant inventors also outside the US. This makes particularly interesting to test for the existence of an international diaspora effect, as mentioned in section 3.1. As for Iran and Poland, the number and percentages of observations with at least *Home country* and *Co-ethnicity* =1 is negligible (columns (2) to (4) and (7)), due both to a feeble inventive activity in the home country and a limited presence of migrant inventors in countries other than the US. [15]

Table 7 and 8 – HERE

Table 8 reports the distribution according to *Home country* and *Same company*, by CoO. Columns (5) to (7) are calculated as in table 7. Column (5) shows, for advanced CoO, a high share of patent pairs in which the company is the same and the citing/control patent is located in the cited inventor's home country (around 60% for Germany and Japan, 40% for France, 13-14% for South Korea and Italy). For the same countries, column (7) shows that a large share of citing/control patents either come from the cited inventor's home country or belong to the same company or both (around 30% for Germany and Japan, 14% for France, 6% for South Korea and Italy). Both statistics are explained with advanced countries hosting more large, R&D intensive firms, than less advanced ones.

[15] Column (6) reports the percentage of citing/control patents by inventors for whom the home country and the CoO coincide. With the exception of the two outliers, Iran and Poland, figures are always over 80%, which we interpret (following Kerr, 2008) as a sign of precision of our algorithm.

Table 2 / part 2 reports the descriptive statistics for the international sample for the regressions that follow. As we do not expect them to produce meaningful results for Iran and Poland, we drop the related observations from the analysis. This reduces our sample to 1,004,950 observations.

Columns (1) in table 9 reports the results of our baseline regression. We observe a positive and significant coefficient of *Home country* for two BRIC countries (China and Russia) as well South Korea and France. However, when we interact *Home country* with *Same company* (column 2) we observe that the positive effect of *Home country* for France occurs mainly via self-citations at the company level (the coefficient of the interaction term is positive and significant). A similar pattern can be detected for other advanced countries, such as Italy and Japan, but not for South Korea (where the interaction term is negative) nor, more interestingly, for any BRIC country.

In column (3) we replace *Home country* with *Co-ethnicity*. Results do not change for China and Russia, nor for the advanced countries (with the partial exception of France, where the interaction term now loses significance). More interestingly, the size and significance of the co-efficient for India changes noticeably, which we take as a suggestion that, for this country, an "international diaspora" may exist, but not to the benefit of the home country.

Table 9– HERE

In order to dig further in this direction, table 10 reports the results of a regression exercise restricted to the BRIC countries in our sample. We allow for the contemporary presence, among the regressors, of *Home country* and *Co-ethnicity*, plus their interaction. We observe that, for China and Russia, the coefficients for both variables persist positive and significant, while for India it is only *Co-ethnicity* that seems to matter. The interaction terms are never significant. We interpret these results as further evidence that for China and Russia the international diaspora and inventors at home benefit of migrants' knowledge feedbacks, while for India no brain gain is detectable. Notice that this result is very close to Agrawal et al.'s (2011), and suggests the latter to be India-specific.

Table 10– HERE

Two further remarks are due. First, we observe a positive and significant coefficient for *Returnee*. However, descriptive statistics in table 2 make clear that returnee inventors are very few (they account for one 0.1% of all observations, as opposed to 3% for *Same company*), so they are an unlikely channel for massive knowledge feedbacks. Second, we observe that Germany does not behave as France and the other advanced countries, nor like BRICs and South Korea. That is, neither the inventors nor the companies active at home, nor the international diaspora seem to have privileged access to knowledge produced by migrant inventors in the US. We have explored the possibility that this result was due to measurement errors, caused by the presence of many German inventors in Swiss companies and/or confusion between German, Swiss, and Austrian inventors caused by the our algorithm. But apparently this is not the case.[16]

Table 11 reports the results of a robustness check run by replacing, when available, *Co-nationality* to either *Home country* or *Co-ethnicity*. Although the much smaller sample size causes several coefficients to lose significance, the main results obtained so far do not change. In general, the same applies to regressions by technology, especially science-based ones. We notice that results for China, Russia and South Korea hold across several technologies (the most notable exceptions being *Instruments* for China and *Pharma & Biotech* for Russia). Results for France seem mostly due to *Electronics* and *Instruments*, and not to hold for *Pharma & Biotech* (results available on request).

Table 11– HERE

[16] We have re-run regressions in table 9 by extending value =1 for Home country to all cases of inventors located in Austria and Switzerland. We have also restricted the regressions to the case of Pharma & Biotech technologies, in which Swiss companies are over-represented. Always to no avail. The same applies to robustness checks based on nationality, which we discuss below.

5. Discussion and conclusions

By means of patent and inventor data, we have investigated whether social ties binding migrants from the same country of origin help diffusing technical knowledge both among migrants (diaspora effect) and towards the country of origin (brain gain effect). We have focussed on the US as a destination country and on five Asian and as many European countries of origin, which we selected among the most important sources of highly skilled migration.

Our empirical exercise has made use of a large and entirely novel sample of patents filed by inventors of foreign origin in the US, which we identified by means of linguistic analysis of names and surnames. We conducted robustness checks based upon inventors' nationality for a sizeable subsample.

We find evidence of a diaspora effect to exist for all Asian countries in our sample (China, India, South Korea, and, to a lesser extent, Japan and Iran) and for one European country (Russia). However, the marginal effect of co-ethnicity is secondary to the effect of proximity in the physical space (co-location at the city or State level) and on the network of inventors. In addition, co-ethnicity ties appear to act as substitute of co-inventorship ties and chains, that is to kick-in between network-distant inventors. The same holds, but only for China and India, for spatial proximity, as already found (for India only) by Agrawal et al. (2008).

As for the brain gain effect, we find that ethnic ties do not necessarily imply a knowledge transfer to the home country. In particular, we see none of this for one of the most important inventor diasporas in the US, namely the Indian one. This may have to do more with the absorptive capacities of the country of origin, than with the international dimension of the diffusion process under consideration. In fact, for both India and other BRIC countries in our sample, we find evidence of an international diaspora effect, which presents some analogy with findings in the trade literature (Felbermayr et al., 2010). By contrast, we find nothing of the kind for advanced countries such as France, Italy, and Japan, whose brain gain effect is mostly mediated by companies' self-citations. Returnee inventors exhibit a high probability to keep using the knowledge produced in their countries of destination, but are very few.

While our results point at important differences across migrants' countries of origin, we can only speculate on the sources of such differences. Despite imperfections in our name-based method for identifying

migrants, our results appear robust enough to rule out the exclusive effect of measurement errors. Still, the cohort composition of migrant communities (the different mix of first- and further-generation migrants between recent Asian migration wavers, and older European ones) play a role, as well as the composition by migration channel (with migration from BRIC possibly occurring more often through the higher education system, and that from advanced countries through multinationals). We will dedicate future research to assess the validity of these intuitions and to investigate policy implications, which we cannot yet deliver on the basis of the present findings.

Our future research plans also include investigating the role of ethnic ties in the formation of networks of inventors, so to reconsider their role in determining collaboration-based social proximity. Besides, we plan to extend the analysis conducted in this paper to Europe, instead of the US, as the focal destination region. This extension will contribute, among other things, to casting light on a policy–sensitive topic such as 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., 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., McHale, J., 2008. How Do Spatial and Social Proximity Influence Knowledge Flows? Evidence from Patent Data. Journal of Urban Economics, 64(2), 258-269.

Agrawal, A., Kapur, D., McHale, J., 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.

Alcacer, J., 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.

Almeida, P., Phene, A., Li, S., 2014. The Influence of Ethnic Community Knowledge on Indian Inventor Innovativeness. Organization Science.

Alnuaimi, T., Opsahl, T., 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, OECD Publishing, Paris.

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

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

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

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

Breschi, S., 2011, The geography of knowledge flows. In: P. Cooke, B.T. Asheim, R. Boschma, R. Martin, D. Schwartz, F. Tödtling (Eds.). Handbook of Regional Innovation and Growth. Edward Elgar Publishing.

Breschi, S., Lissoni, F., 2005a. "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., Lissoni, F., 2005b, Knowledge networks from patent data. In: H.F. Moed, W. Glänzel, U. Schmoch (Eds.). Handbook of quantitative science and technology research. Springer Science+Business Media, Berlin, pp. 613-643.

Breschi, S., 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., Thijs, B., 2006. Traces of prior art: An analysis of non-patent references found in patent documents. Scientometrics, 69(1), 3-20.

Cerna, L., 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., 2015. 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, lbv025.

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

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

Du Plessis, M., Van Looy, B., Song, X., Magerman, T., 2009. Data production methods for harmonized patent indicators: Assignee sector allocation, Luxembourg.

Ellison, G., Glaeser, E.L., Kerr, W., 2007. What causes industry agglomeration? Evidence from coagglomeration patterns, National Bureau of Economic Research.

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

Foley, C.F., 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., Png, I.P.L., 2015. Engineer/Scientist Careers: Patents, Online Profiles, and Misclassification Bias. Strategic Management Journal, (forthcoming).

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., Trajtenberg, M., 2005. Market value and patent citations. RAND Journal of economics, 16-38.

Hall, B.H., Jaffe, A.B., Trajtenberg, M., 2001. The NBER patent citation data file: Lessons, insights and methodological tools, National Bureau of Economic Research.

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

Henderson, R., Jaffe, A., 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., 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., 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, 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. 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., 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.

Maraut, S., Dernis, H., Webb, C., Spiezia, V., 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., 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., Fleming, L., 2009. Mobility, skills, and the Michigan non-compete experiment. Management Science, 55(6), 875-889.

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

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

Miguelez, E., 2014, **Inventor diasporas and internationalization of technology** Cahiers du GREthA, Université de Bordeaux.

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

Mountford, A., 1997. Can a brain drain be good for growth in the source economy? Journal of development economics, 53(2), 287-303.

Nanda, R., 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.

Niebuhr, A., 2010. Migration and Innovation: Does Cultural Diversity Matter for Regional R&D Activity? Papers in Regional Science, 89(3), 563-585.

Peeters, B., Song, X., Callaert, J., Grouwels, J., Van Looy, B., 2010. Harmonizing harmonized patentee names: an exploratory assessment of top patentees, Luxembourg.

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

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

Rauch, J.E., 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. Harvard University Press.

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

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

Stark, O., Wang, Y., 2002. Inducing human capital formation: migration as a substitute for subsidies. Journal of Public Economics, 86(1), 29-46.

Tarasconi, G., Coffano, M., 2014, Crios-Patstat Database: Sources, Contents and Access Rules. Center for Research on Innovation, Organization and Strategy, CRIOS. http://ssrn.com/abstract=2404344 or http://dx.doi.org/10.2139/ssrn.2404344.

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

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., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: A reassessment. American Economic Review, 450-460.

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).

Veugelers, R., 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., Gereffi, G., 2007a. Education, Entrepreneurship and Immigration: America's New Immigrant Entrepreneurs, Part II. Part II (June 11, 2007).

Wadhwa, V., Saxenian, A., Rissing, B., Gereffi, G., 2007b. America's new Immigrant entrepreneurs: Part I. Duke Science, Technology & Innovation Paper(23).

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.

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

Foreign inventors in the US: Testing for Diaspora and Brain Gain Effects

Stefano Breschi , Francesco Lissoni, Ernest Miguelez

This version: *27 July 2015*

APPENDIXES

Appendix 1 – Inventor names' disambiguation

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). [17]

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 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.

[17] 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 2010 (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)

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, our disambiguation algorithm (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.[18]

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 seen in the paper's descriptive statistics, we find very few cases. Whether true or false positives, they are unlikely to affect our findings.

[18] 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 CoAs and two main scores:[19]

- "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.[20]

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.[21]

III. For each inventor, we consider three indicators:
   a. The frequency of her first name(s) in English- and Spanish-speaking CoA [22]
   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.

[19] 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).

[20] 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.

[21] 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).

[22] 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. [23]

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).[24]

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.

[23] 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.

[24] 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

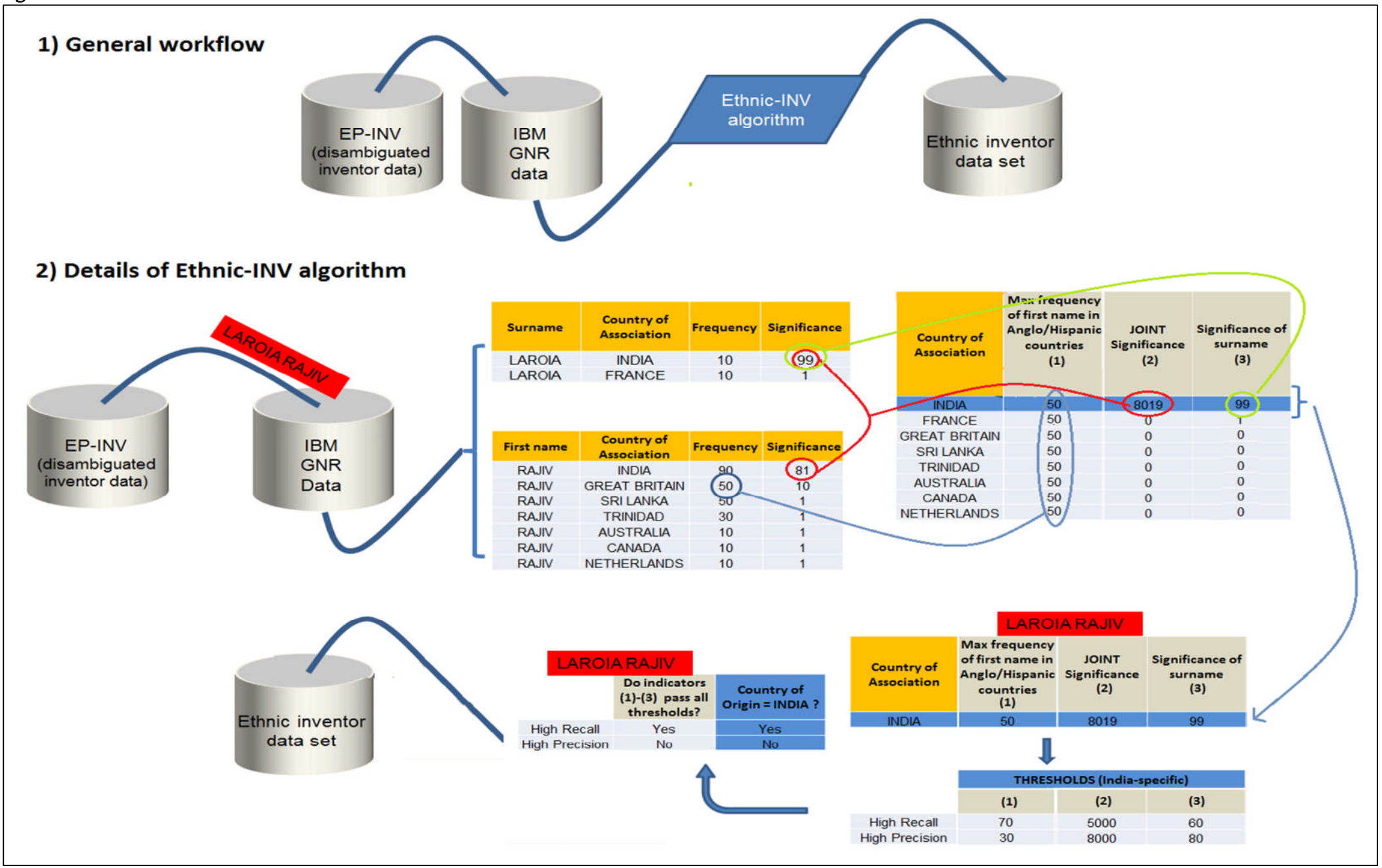


Figure A2.2 - Ethnic-INV algorithm calibration results: China and Italy

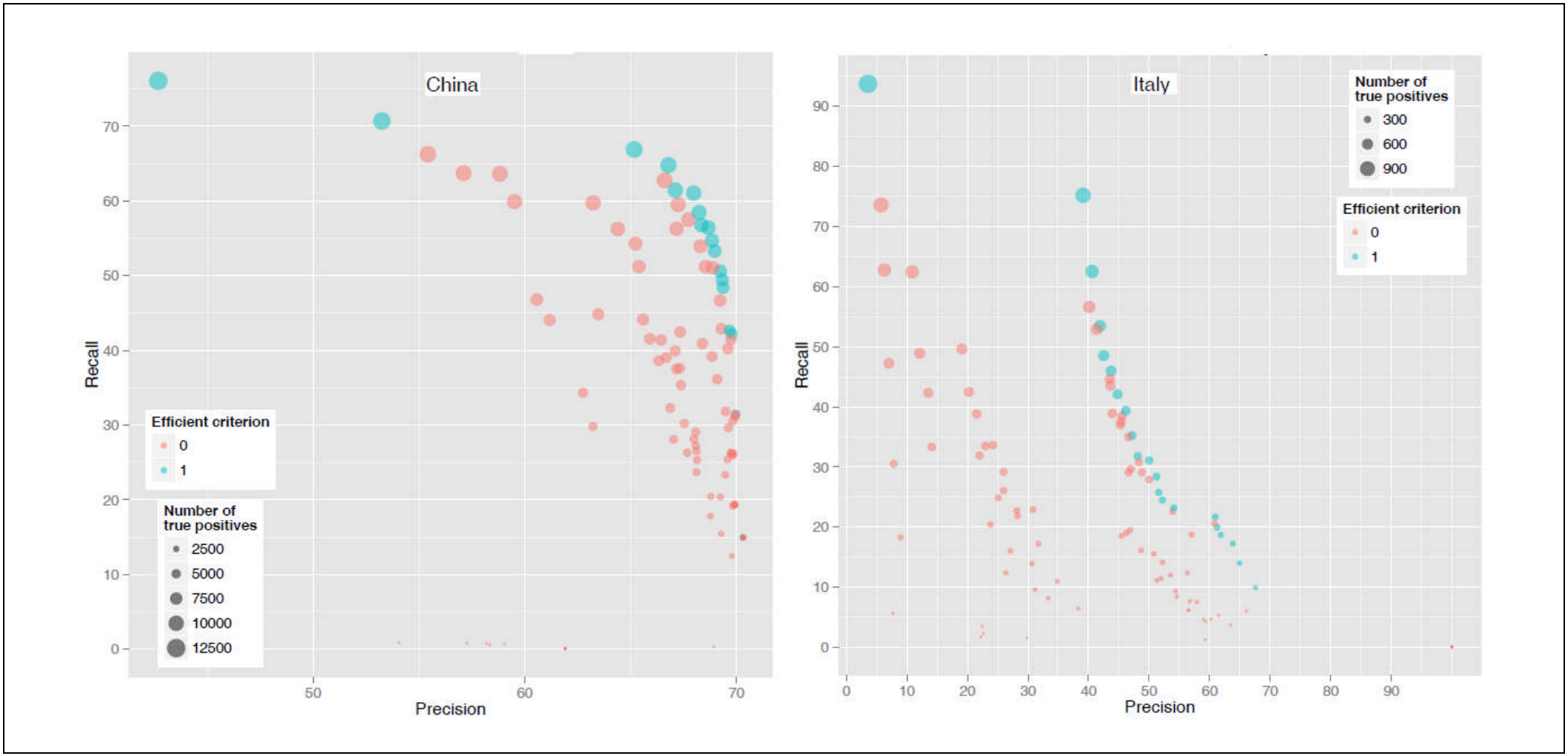
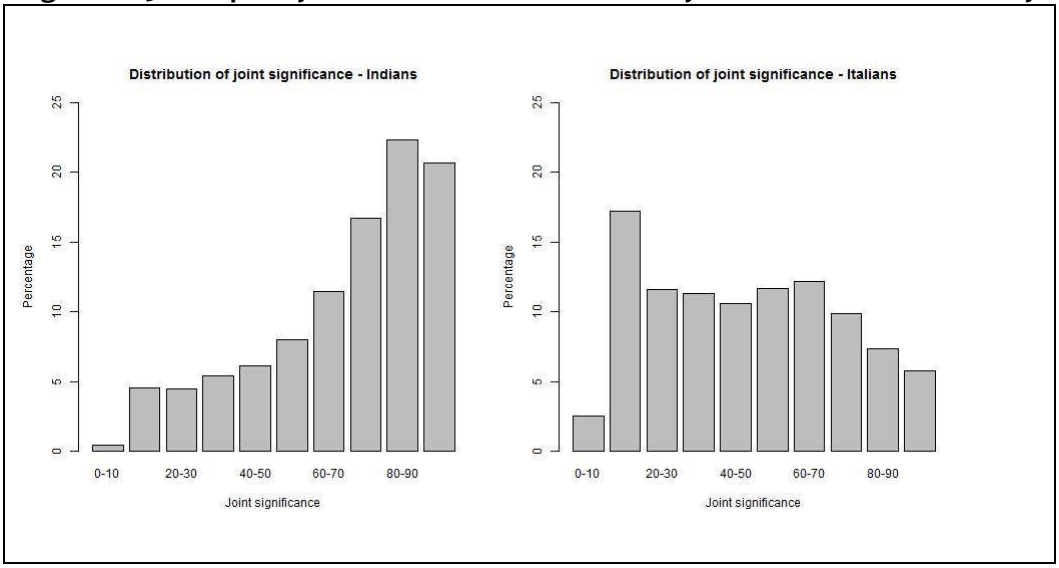


Figure A2.3 - Frequency distribution of values taken by indicator n.2: India vs. Italy



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.[25]

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

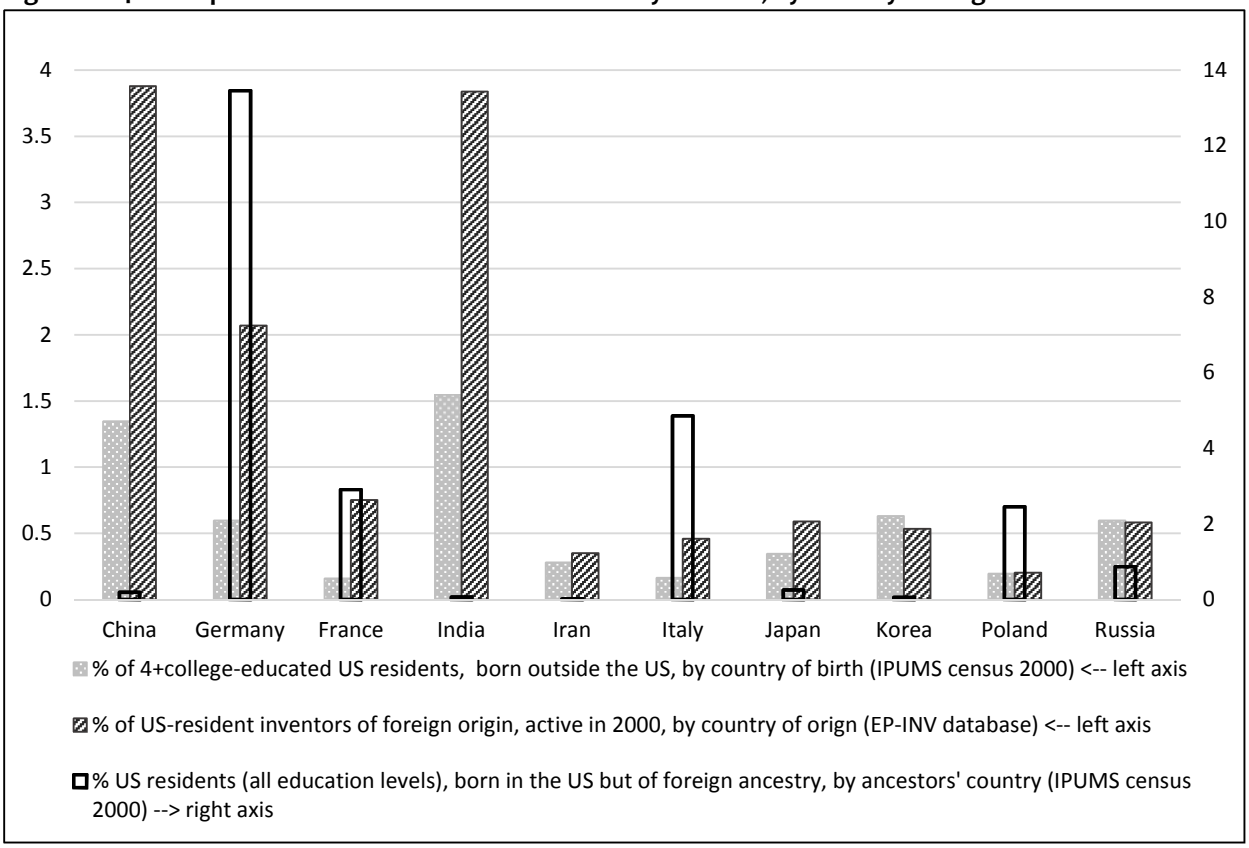
Table A2.1 – Comparison of EP-INV and censual data for year 2000; by Country of Origin

	% 4+college-educated US residents, born outside the US, by country of birth (1)	% US residents (all education levels), born in the US, by ancestors' country (1)	% US-resident inventors of foreign origin, active in 2000, by country of orign (2)
China	1.346	0.189	3.879
Germany	0.598	13.457	2.07
France	0.159	2.912	0.752
India	1.547	0.067	3.839
Iran	0.28	0.016	0.351
Italy	0.164	4.861	0.459
Japan	0.345	0.252	0.589
Korea	0.631	0.059	0.534
Poland	0.196	2.452	0.202

(1) source: IPUMS-USA census data
(2) source: EP-INV database

[25] 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. 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. 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

	% US-resident inventors of foreign nationality, 1995-2005 ; by nationality (1)	% US-resident inventors of foreign origin, active in 2000, by country of origin (2)
China	3.673	3.879
Germany	1.038	2.07
France	0.589	0.752
India	2.984	3.839
Iran	0.110	0.351
Italy	0.228	0.459
Japan	0.483	0.589
Korea	0.482	0.534
Poland	0.111	0.202
Russia	0.469	0.582

(1) source: WIPO-PCT dataset (see Miguelez and Fink, 2013).
(2) source: EP-INV database

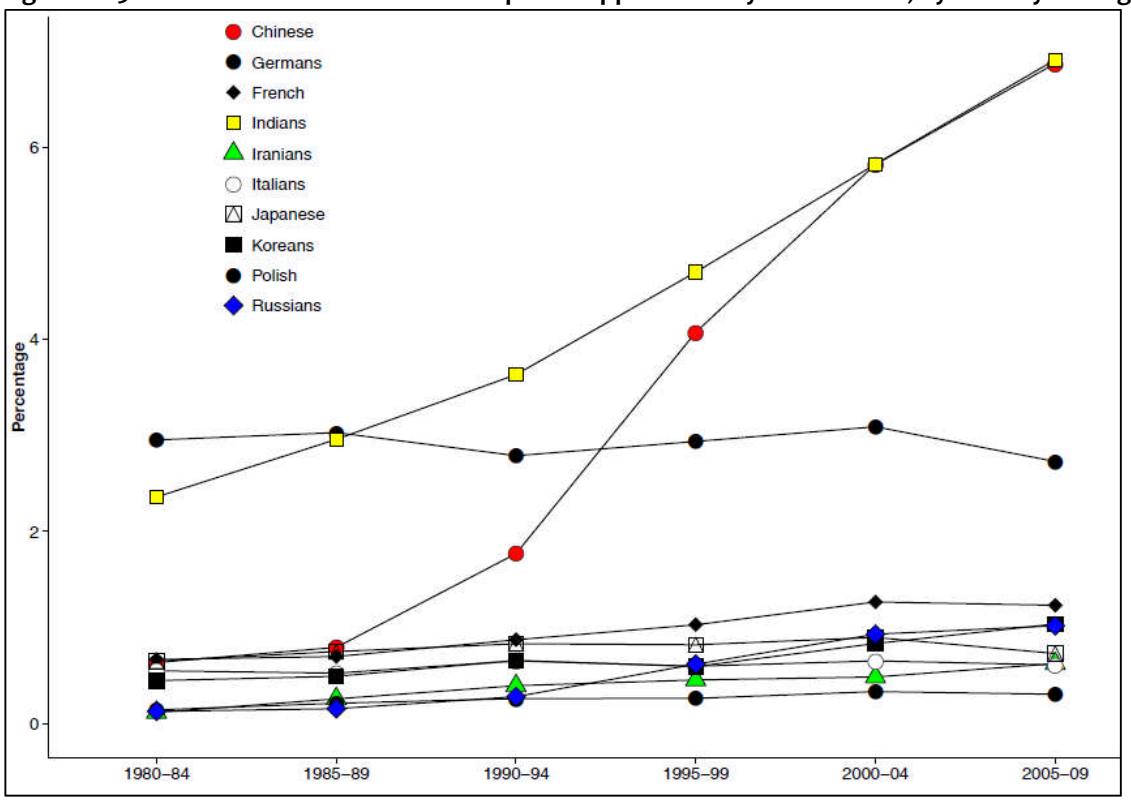
Still, we observe cross-country variations that may be due to lack of precision in the Ethnic-INV algorithm. In particular, we notice larger differences, in relative terms, for Germany, Italy, and Poland, where the share of foreign nationals is about double the share of foreign-origin inventors. But 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.[26] 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).

[26] 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.

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



Appendix 3 – Descriptive statistics: additional tables

Table A3.1 Local and international samples: descriptive statistics. China

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	249348	0.500	0.500	0	1
Co-ethnicity	249348	0.228	0.419	0	1
Same MSA	249348	0.143	0.350	0	1
Same State	249348	0.226	0.418	0	1
Miles	249348	937.615	890.430	0	5081.5
Soc. Dist. 0	249348	0.010	0.099	0	1
Soc. Dist. 1	249348	0.010	0.099	0	1
Soc. Dist. 2	249348	0.007	0.086	0	1
Soc. Dist. 3	249348	0.010	0.100	0	1
Soc. Dist. >3	249348	0.297	0.457	0	1
Soc. Dist. ∞	249348	0.666	0.472	0	1
#claims	249348	7.812	12.641	0	235
backward citations	249348	4.516	3.182	0	87
NPL citations	249348	1.556	2.632	0	57
overlap IPCs 7 digits	249348	1.250	1.680	0	27
overlap IPCs 7 digits / all IPCs	249348	0.269	0.270	0	1
overlap IPCs	249348	0.893	1.773	0	53
2. International sample (citations from outside the US)					
Citation	256244	0.500	0.500	0	1
Co-ethnicity	256244	0.046	0.209	0	1
Home country	256244	0.025	0.157	0	1
Same company	256244	0.025	0.155	0	1
Returnee	256244	0.000	0.014	0	1
Contiguous countries	256244	0.036	0.186	0	1
Former colonial relationship	256244	0.207	0.405	0	1
Same country	256244	0.022	0.146	0	1
English	256244	0.181	0.385	0	1
Similarity to English	256244	0.242	0.259	0	1
Miles	256244	4605.959	1822.230	0	11498.1
Soc. Dist. 0	256244	0.003	0.056	0	1
Soc. Dist. 1	256244	0.005	0.069	0	1
Soc. Dist. 2	256244	0.004	0.062	0	1
Soc. Dist. 3	256244	0.005	0.071	0	1
Soc. Dist. >3	256244	0.248	0.432	0	1
Soc. Dist. ∞	256244	0.736	0.441	0	1
#claims	256244	9.713	12.008	0	383
backward citations	256244	3.966	3.260	0	98
backward NPL citations	256244	1.187	2.254	0	76
overlap IPCs 7 digits	256244	1.168	1.401	0	31
overlap IPCs 7 digits / all IPCs	256244	0.303	0.288	0	1
overlap IPCs	256244	0.819	1.471	0	49

Table A3.2 Local and international samples: descriptive statistics. Germany

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	175570	0.500	0.500	0	1
Co-ethnicity	175570	0.076	0.265	0	1
Same MSA	175570	0.131	0.337	0	1
Same State	175570	0.212	0.409	0	1.0
Miles	175570	909.569	850.374	0	5085.412
Soc. Dist. 0	175570	0.008	0.091	0	1
Soc. Dist. 1	175570	0.009	0.095	0	1
Soc. Dist. 2	175570	0.006	0.077	0	1
Soc. Dist. 3	175570	0.007	0.082	0	1
Soc. Dist. >3	175570	0.210	0.408	0	1
Soc. Dist. ∞	175570	0.760	0.427	0	1
#claims	175570	8.958	12.860	0	259
backward citations	175570	4.726	3.133	0	68
NPL citations	175570	1.152	2.321	0	49
overlap IPCs 7 digits	175570	1.096	1.385	0	23
overlap IPCs 7 digits / all IPCs	175570	0.298	0.296	0	1
overlap IPCs	175570	0.817	1.473	0	28
2. International sample (citations from outside the US)					
Citation	177564	0.500	0.500	0	1
Co-ethnicity	177564	0.307	0.461	0	1
Home country	177564	0.302	0.459	0	1
Same company	177564	0.038	0.190	0	1
Returnee	177564	0.001	0.037	0	1
Contiguous countries	177564	0.029	0.168	0	1
Former colonial relationship	177564	0.191	0.393	0	1
Same country	177564	0.072	0.259	0	1
English	177564	0.163	0.369	0	1
Similarity to English	177564	0.272	0.261	0	1
Miles	177564	4160.425	2128.488	0	11083.11
Soc. Dist. 0	177564	0.004	0.064	0	1
Soc. Dist. 1	177564	0.010	0.098	0	1
Soc. Dist. 2	177564	0.007	0.083	0	1
Soc. Dist. 3	177564	0.006	0.079	0	1
Soc. Dist. >3	177564	0.167	0.373	0	1
Soc. Dist. ∞	177564	0.806	0.395	0	1
#claims	177564	9.791	11.535	0	442
backward citations	177564	4.114	3.195	0	98
backward NPL citations	177564	0.803	1.914	0	76
overlap IPCs 7 digits	177564	1.050	1.188	0	19
overlap IPCs 7 digits / all IPCs	177564	0.329	0.307	0	1
overlap IPCs	177564	0.760	1.267	0	27

Table A3.3 Local and international samples: descriptive statistics. France

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	66170	0.500	0.500	0	1
Co-ethnicity	66170	0.038	0.192	0	1
Same MSA	66170	0.146	0.353	0	1
Same State	66170	0.230	0.421	0	1
Miles	66170	923.263	878.811	0	5024.3
Soc. Dist. 0	66170	0.010	0.098	0	1
Soc. Dist. 1	66170	0.008	0.088	0	1
Soc. Dist. 2	66170	0.006	0.079	0	1
Soc. Dist. 3	66170	0.008	0.089	0	1
Soc. Dist. >3	66170	0.234	0.424	0	1
Soc. Dist. ∞	66170	0.734	0.442	0	1
#claims	66170	8.364	12.562	0	197
backward citations	66170	4.634	3.170	0	64
NPL citations	66170	1.262	2.392	0	50
overlap IPCs 7 digits	66170	1.191	1.580	0	24
overlap IPCs 7 digits / all IPCs	66170	0.294	0.290	0	1
overlap IPCs	66170	0.884	1.731	0	40
2. International sample (citations from outside the US)					
Citation	68100	0.500	0.500	0	1
Co-ethnicity	68100	0.125	0.331	0	1
Home country	68100	0.116	0.320	0	1
Same company	68100	0.036	0.185	0	1
Returnee	68100	0.001	0.031	0	1
Contiguous countries	68100	0.037	0.189	0	1
Former colonial relationship	68100	0.220	0.414	0	1
Same country	68100	0.055	0.228	0	1
English	68100	0.184	0.387	0	1
Similarity to English	68100	0.248	0.257	0	0.67
Miles	68100	4110.604	2181.692	0	11045.67
Soc. Dist. 0	68100	0.006	0.078	0	1
Soc. Dist. 1	68100	0.007	0.084	0	1
Soc. Dist. 2	68100	0.006	0.075	0	1
Soc. Dist. 3	68100	0.006	0.075	0	1
Soc. Dist. >3	68100	0.185	0.388	0	1
Soc. Dist. ∞	68100	0.791	0.407	0	1
#claims	68100	9.804	11.702	0	292
backward citations	68100	4.010	3.140	0	55
backward NPL citations	68100	0.991	2.078	0	33
overlap IPCs 7 digits	68100	1.144	1.392	0	22
overlap IPCs 7 digits / all IPCs	68100	0.328	0.302	0	1
overlap IPCs	68100	0.834	1.509	0	41

Table A3.4 Local and international samples: descriptive statistics. India

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	324034	0.500	0.500	0	1
Co-ethnicity	324034	0.170	0.376	0	1
Same MSA	324034	0.137	0.344	0	1
Same State	324034	0.210	0.408	0	1.0
Miles	324034	928.019	870.285	0	5082.868
Soc. Dist. 0	324034	0.007	0.085	0	1
Soc. Dist. 1	324034	0.007	0.086	0	1
Soc. Dist. 2	324034	0.006	0.078	0	1
Soc. Dist. 3	324034	0.008	0.086	0	1
Soc. Dist. >3	324034	0.227	0.419	0	1
Soc. Dist. ∞	324034	0.745	0.436	0	1
#claims	324034	8.730	12.966	0	235
backward citations	324034	4.533	3.135	0	87
NPL citations	324034	1.252	2.342	0	53
overlap IPCs 7 digits	324034	1.071	1.357	0	26
overlap IPCs 7 digits / all IPCs	324034	0.281	0.284	0	1
overlap IPCs	324034	0.789	1.475	0	47
2. International sample (citations from outside the US)					
Citation	316466	0.500	0.500	0	1
Co-ethnicity	316466	0.022	0.146	0	1
Home country	316466	0.009	0.094	0	1
Same company	316466	0.023	0.149	0	1
Returnee	316466	0.000	0.007	0	1
Contiguous countries	316466	0.038	0.191	0	1
Former colonial relationship	316466	0.198	0.399	0	1
Same country	316466	0.018	0.135	0	1
English	316466	0.176	0.381	0	1
Similarity to English	316466	0.241	0.257	0	1
Miles	316466	4617.670	1797.330	0	11047.56
Soc. Dist. 0	316466	0.003	0.053	0	1
Soc. Dist. 1	316466	0.003	0.056	0	1
Soc. Dist. 2	316466	0.003	0.057	0	1
Soc. Dist. 3	316466	0.004	0.063	0	1
Soc. Dist. >3	316466	0.202	0.401	0	1
Soc. Dist. ∞	316466	0.785	0.411	0	1
#claims	316466	10.078	11.815	0	383
backward citations	316466	3.953	3.164	0	98
backward NPL citations	316466	0.964	1.970	0	76
overlap IPCs 7 digits	316466	1.051	1.214	0	22
overlap IPCs 7 digits / all IPCs	316466	0.313	0.297	0	1
overlap IPCs	316466	0.765	1.340	0	41

Table A3.5 Local and international samples: descriptive statistics. Iran

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	29044	0.500	0.500	0	1
Co-ethnicity	29044	0.016	0.124	0	1
Same MSA	29044	0.159	0.365	0	1
Same State	29044	0.269	0.444	0	1.0
Miles	29044	1001.786	909.522	0	5073.808
Soc. Dist. 0	29044	0.008	0.089	0	1
Soc. Dist. 1	29044	0.008	0.091	0	1
Soc. Dist. 2	29044	0.007	0.082	0	1
Soc. Dist. 3	29044	0.007	0.081	0	1
Soc. Dist. >3	29044	0.193	0.395	0	1
Soc. Dist. ∞	29044	0.777	0.416	0	1
#claims	29044	8.572	12.275	0	227
backward citations	29044	4.618	2.997	0	50
NPL citations	29044	0.930	1.920	0	26
overlap IPCs 7 digits	29044	0.940	1.098	0	26
overlap IPCs 7 digits / all IPCs	29044	0.308	0.303	0	1
overlap IPCs	29044	0.716	1.271	0	39

Table A3.6 Local and international samples: descriptive statistics. Italy

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	46664	0.500	0.500	0	1
Co-ethnicity	46664	0.020	0.141	0	1
Same MSA	46664	0.127	0.333	0	1
Same State	46664	0.208	0.406	0	1.0
Miles	46664	947.658	882.231	0	4929.127
Soc. Dist. 0	46664	0.008	0.089	0	1
Soc. Dist. 1	46664	0.008	0.087	0	1
Soc. Dist. 2	46664	0.008	0.088	0	1
Soc. Dist. 3	46664	0.007	0.084	0	1
Soc. Dist. >3	46664	0.206	0.405	0	1
Soc. Dist. ∞	46664	0.763	0.425	0	1
#claims	46664	9.110	12.927	0	235
backward citations	46664	4.569	3.088	0	44
NPL citations	46664	1.354	2.508	0	57
overlap IPCs 7 digits	46664	1.183	1.519	0	27
overlap IPCs 7 digits / all IPCs	46664	0.301	0.289	0	1
overlap IPCs	46664	0.876	1.629	0	43
2. International sample (citations from outside the US)					
Citation	46228	0.500	0.500	0	1
Co-ethnicity	46228	0.052	0.223	0	1
Home country	46228	0.043	0.203	0	1
Same company	46228	0.026	0.158	0	1
Returnee	46228	0.001	0.023	0	1
Contiguous countries	46228	0.035	0.183	0	1
Former colonial relationship	46228	0.208	0.406	0	1
Same country	46228	0.027	0.161	0	1
English	46228	0.178	0.382	0	1
Similarity to English	46228	0.254	0.257	0	1
Miles	46228	4386.379	1969.206	0	11270.66
Soc. Dist. 0	46228	0.004	0.064	0	1
Soc. Dist. 1	46228	0.006	0.077	0	1
Soc. Dist. 2	46228	0.005	0.070	0	1
Soc. Dist. 3	46228	0.004	0.063	0	1
Soc. Dist. >3	46228	0.187	0.390	0	1
Soc. Dist. ∞	46228	0.794	0.405	0	1
#claims	46228	10.136	11.752	0	383
backward citations	46228	3.931	3.058	0	69
backward NPL citations	46228	0.998	2.095	0	39
overlap IPCs 7 digits	46228	1.089	1.228	0	18
overlap IPCs 7 digits / all IPCs	46228	0.325	0.301	0	1
overlap IPCs	46228	0.812	1.401	0	24

Table A3.7 Local and international samples: descriptive statistics. Japan

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	48172	0.500	0.500	0	1
Co-ethnicity	48172	0.028	0.165	0	1
Same MSA	48172	0.137	0.344	0	1
Same State	48172	0.229	0.420	0	1.0
Miles	48172	995.913	912.139	0	5085.159
Soc. Dist. 0	48172	0.006	0.080	0	1
Soc. Dist. 1	48172	0.006	0.080	0	1
Soc. Dist. 2	48172	0.004	0.066	0	1
Soc. Dist. 3	48172	0.005	0.069	0	1
Soc. Dist. >3	48172	0.213	0.410	0	1
Soc. Dist. ∞	48172	0.765	0.424	0	1
#claims	48172	8.989	13.184	0	247
backward citations	48172	4.503	3.203	0	64
NPL citations	48172	1.625	2.814	0	45
overlap IPCs 7 digits	48172	1.191	1.463	0	27
overlap IPCs 7 digits / all IPCs	48172	0.287	0.278	0	1
overlap IPCs	48172	0.872	1.523	0	32
2. International sample (citations from outside the US)					
Citation	53150	0.500	0.500	0	1
Co-ethnicity	53150	0.284	0.451	0	1
Home country	53150	0.284	0.451	0	1
Same company	53150	0.047	0.212	0	1
Returnee	53150	0.002	0.047	0	1
Contiguous countries	53150	0.032	0.176	0	1
Former colonial relationship	53150	0.189	0.392	0	1
Same country	53150	0.124	0.330	0	1
English	53150	0.163	0.369	0	1
Similarity to English	53150	0.232	0.260	0	1
Miles	53150	4030.211	2148.872	0	11046.71
Soc. Dist. 0	53150	0.004	0.063	0	1
Soc. Dist. 1	53150	0.008	0.091	0	1
Soc. Dist. 2	53150	0.005	0.073	0	1
Soc. Dist. 3	53150	0.004	0.067	0	1
Soc. Dist. >3	53150	0.171	0.376	0	1
Soc. Dist. ∞	53150	0.807	0.395	0	1
#claims	53150	10.231	12.029	0	442
backward citations	53150	4.003	3.214	0	79
backward NPL citations	53150	1.072	2.176	0	41
overlap IPCs 7 digits	53150	1.139	1.282	0	19
overlap IPCs 7 digits / all IPCs	53150	0.313	0.290	0	1
overlap IPCs	53150	0.832	1.387	0	29

Table A3.8 Local and international samples: descriptive statistics. Korea

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	51774	0.500	0.500	0	1
Co-ethnicity	51774	0.031	0.174	0	1
Same MSA	51774	0.139	0.346	0	1
Same State	51774	0.224	0.417	0	1.0
Miles	51774	930.206	893.884	0	4841.666
Soc. Dist. 0	51774	0.009	0.095	0	1
Soc. Dist. 1	51774	0.008	0.088	0	1
Soc. Dist. 2	51774	0.007	0.082	0	1
Soc. Dist. 3	51774	0.008	0.091	0	1
Soc. Dist. >3	51774	0.237	0.425	0	1
Soc. Dist. ∞	51774	0.731	0.443	0	1
#claims	51774	8.502	12.772	0	197
backward citations	51774	4.587	3.148	0	58
NPL citations	51774	1.314	2.436	0	50
overlap IPCs 7 digits	51774	1.117	1.407	0	22
overlap IPCs 7 digits / all IPCs	51774	0.273	0.281	0	1
overlap IPCs	51774	0.787	1.449	0	25
2. International sample (citations from outside the US)					
Citation	49024	0.500	0.500	0	1
Co-ethnicity	49024	0.048	0.214	0	1
Home country	49024	0.047	0.211	0	1
Same company	49024	0.022	0.147	0	1
Returnee	49024	0.000	0.018	0	1
Contiguous countries	49024	0.031	0.172	0	1
Former colonial relationship	49024	0.204	0.403	0	1
Same country	49024	0.026	0.158	0	1
English	49024	0.174	0.379	0	1
Similarity to English	49024	0.241	0.259	0	0.67
Miles	49024	4593.657	1834.148	0	11043.31
Soc. Dist. 0	49024	0.003	0.056	0	1
Soc. Dist. 1	49024	0.005	0.071	0	1
Soc. Dist. 2	49024	0.004	0.060	0	1
Soc. Dist. 3	49024	0.004	0.065	0	1
Soc. Dist. >3	49024	0.197	0.397	0	1
Soc. Dist. ∞	49024	0.787	0.409	0	1
#claims	49024	9.977	11.768	0	240
backward citations	49024	4.070	3.316	0	98
backward NPL citations	49024	0.994	2.051	0	58
overlap IPCs 7 digits	49024	1.078	1.258	0	32
overlap IPCs 7 digits / all IPCs	49024	0.300	0.290	0	1
overlap IPCs	49024	0.775	1.347	0	54

Table A3.9 Local and international samples: descriptive statistics. Poland

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	16064	0.500	0.500	0	1
Co-ethnicity	16064	0.008	0.090	0	1
Same MSA	16064	0.116	0.320	0	1
Same State	16064	0.166	0.372	0	1.0
Miles	16064	923.182	842.585	0	4849.524
Soc. Dist. 0	16064	0.013	0.111	0	1
Soc. Dist. 1	16064	0.010	0.099	0	1
Soc. Dist. 2	16064	0.005	0.068	0	1
Soc. Dist. 3	16064	0.006	0.079	0	1
Soc. Dist. >3	16064	0.213	0.409	0	1
Soc. Dist. ∞	16064	0.754	0.431	0	1
#claims	16064	8.527	12.621	0	209
backward citations	16064	4.650	3.134	0	64
NPL citations	16064	1.282	2.517	0	49
overlap IPCs 7 digits	16064	1.171	1.587	0	18
overlap IPCs 7 digits / all IPCs	16064	0.304	0.299	0	1
overlap IPCs	16064	0.835	1.544	0	19

Table A3.10 Local and international samples: descriptive statistics. Russia

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	36480	0.500	0.500	0	1
Co-ethnicity	36480	0.031	0.173	0	1
Same MSA	36480	0.139	0.346	0	1
Same State	36480	0.223	0.416	0	1.0
Miles	36480	948.014	889.423	0	5080.685
Soc. Dist. 0	36480	0.012	0.107	0	1
Soc. Dist. 1	36480	0.008	0.088	0	1
Soc. Dist. 2	36480	0.005	0.071	0	1
Soc. Dist. 3	36480	0.005	0.073	0	1
Soc. Dist. >3	36480	0.206	0.405	0	1
Soc. Dist. ∞	36480	0.764	0.425	0	1
#claims	36480	7.762	12.169	0	195
backward citations	36480	4.723	3.140	0	64
NPL citations	36480	1.255	2.460	0	45
overlap IPCs 7 digits	36480	0.947	1.149	0	21
overlap IPCs 7 digits / all IPCs	36480	0.292	0.301	0	1
overlap IPCs	36480	0.685	1.263	0	27
2. International sample (citations from outside the US)					
Citation	38174	0.500	0.500	0	1
Co-ethnicity	38174	0.015	0.122	0	1
Home country	38174	0.005	0.071	0	1
Same company	38174	0.020	0.141	0	1
Returnee	38174	0.000	0.014	0	1
Contiguous countries	38174	0.034	0.182	0	1
Former colonial relationship	38174	0.181	0.385	0	1
Same country	38174	0.026	0.160	0	1
English	38174	0.164	0.370	0	1
Similarity to English	38174	0.255	0.261	0	0.67
Miles	38174	4507.309	1838.656	0	11053.67
Soc. Dist. 0	38174	0.005	0.067	0	1
Soc. Dist. 1	38174	0.004	0.063	0	1
Soc. Dist. 2	38174	0.003	0.057	0	1
Soc. Dist. 3	38174	0.003	0.057	0	1
Soc. Dist. >3	38174	0.186	0.389	0	1
Soc. Dist. ∞	38174	0.799	0.401	0	1
#claims	38174	9.564	12.032	0	383
backward citations	38174	4.034	3.148	0	79
backward NPL citations	38174	0.921	1.976	0	24
overlap IPCs 7 digits	38174	0.961	1.070	0	20
overlap IPCs 7 digits / all IPCs	38174	0.321	0.305	0	1
overlap IPCs	38174	0.699	1.210	0	28

Appendix 4 – Regression analysis: Logit estimates

Table A4.1 – Probability of citation from within the US, as a function of co-ethnicity, spatial & social distance, and controls -- Logit regression

	(1)	(2)	(3)	(4)	(5)
Same MSA	0.547***	0.368***	0.367***	0.340***	0.0280*
	(0.00641)	(0.00991)	(0.00991)	(0.0104)	(0.0169)
Co-ethnic	0.204***	0.169***	-0.351**	-0.345**	-0.329*
	(0.00611)	(0.00782)	(0.171)	(0.175)	(0.176)
Co-ethnic * MSA	-0.0546***	-0.0502**	-0.0446**	-0.0451**	-0.0602***
	(0.0161)	(0.0196)	(0.0197)	(0.0207)	(0.0207)
Same State					0.0945***
					(0.0121)
ln(Miles)					-0.157***
					(0.00977)
ln(Miles)^2					0.0121***
					(0.000876)
Soc. Dist. 1		-1.185***	-1.231***	-1.254***	-1.136***
		(0.0852)	(0.0932)	(0.0962)	(0.0962)
Soc. Dist. 2		-2.042***	-2.086***	-2.111***	-1.972***
		(0.0856)	(0.0932)	(0.0967)	(0.0970)
Soc. Dist. 3		-2.543***	-2.580***	-2.622***	-2.470***
		(0.0825)	(0.0897)	(0.0931)	(0.0932)
Soc. Dist. >3		-3.275***	-3.371***	-3.363***	-3.183***
		(0.0770)	(0.0835)	(0.0856)	(0.0858)
Soc. Dist. ∞		-3.420***	-3.497***	-3.434***	-3.249***
		(0.0768)	(0.0832)	(0.0853)	(0.0856)
Co-ethnic *Soc. Dist. 1			0.333*	0.357*	0.330
			(0.196)	(0.200)	(0.201)
Co-ethnic *Soc. Dist. 2			0.351*	0.329*	0.310
			(0.189)	(0.195)	(0.196)
Co-ethnic *Soc. Dist. 3			0.351*	0.364**	0.351*
			(0.181)	(0.186)	(0.186)
Co-ethnic *Soc. Dist. >3			0.602***	0.605***	0.585***
			(0.171)	(0.175)	(0.176)
Co-ethnic *Soc. Dist. ∞			0.482***	0.501***	0.483***
			(0.171)	(0.175)	(0.176)
ln(#claims)				0.00509**	0.00532***
				(0.00204)	(0.00204)
ln(1 + backward citations)				0.363***	0.363***
				(0.00500)	(0.00500)
ln(1 + NPL citations)				-0.0261***	-0.0269***
				(0.00458)	(0.00459)
ln(1 + overlap IPCs 7 digits)				0.922***	0.920***
				(0.00641)	(0.00642)
OST-30 F.E.	no	no	yes	yes	yes
Constant	-0.0991***	3.273***	3.354***	2.381***	2.685***
	(0.00112)	(0.0769)	(0.0833)	(0.0861)	(0.0881)
Observations	1,043,320	1,043,320	1,043,320	1,043,320	1,043,320
Chi2	9750	9601	9742	32758	32890
Log-lik.	-718116	-709371	-709314	-678154	-677638
Pseudo-R2	0.00700	0.0191	0.0192	0.0623	0.0630

The table reports estimated parameters (βs)
Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table A4.2 – Probability of citation from within the US, as a function of co-ethnicity by Country of Origin, spatial & social distance, and controls -– Logit regression

	(1)	(2)	(2-cont.)	(2-cont.)
Same MSA	0.140***	0.148***		
	(0.0152)	(0.0154)		
Same State	0.0908***	0.0906***		
	(0.0121)	(0.0121)		
ln(Miles)	-0.0268***	-0.0268***		
	(0.00292)	(0.00292)	*Co-ethnicity* Same MSA*	
China co-ethnic	0.253***	0.268***	China * Same MSA	-0.0992***
	(0.0108)	(0.0119)		(0.0301)
Germany co-ethnic	0.0450**	0.0356	Germany * Same MSA	0.0647
	(0.0221)	(0.0242)		(0.0624)
France co-ethnic	-0.0425	-0.0166	France * Same MSA	-0.147
	(0.0487)	(0.0544)		(0.126)
India co-ethnic	0.151***	0.156***	India * Same MSA	-0.0397
	(0.0108)	(0.0120)		(0.0300)
Iran co-ethnic	0.224**	0.164	Iran * Same MSA	0.251
	(0.107)	(0.125)		(0.256)
Italy co-ethnic	0.0484	0.0780	Italy * Same MSA	-0.189
	(0.155)	(0.176)		(0.246)
Japan co-ethnic	0.126**	0.151**	Japan * Same MSA	-0.157
	(0.0629)	(0.0701)		(0.174)
Korea co-ethnic	0.162***	0.183***	Korea * Same MSA	-0.115
	(0.0592)	(0.0642)		(0.163)
Poland co-ethnic	-0.204	-0.223	Poland * Same MSA	0.116
	(0.185)	(0.213)		(0.412)
Russia co-ethnic	0.310***	0.254***	Russia * Same MSA	0.360*
	(0.0695)	(0.0770)		(0.209)
Co-ethnicity Same MSA*	*No*	*Yes (see right)*		
Constant	2.417***	2.415***		
	(0.0804)	(0.0804)		
Social distance dummies	yes	yes		
Citing patent characteristics	yes	yes		
OST-30 FE	yes	yes		
Observations	1,043,320	1,043,320		
Chi2	32918	33049		
Log-lik.	-677825	-677811		
Pseudo-R2	0.0627	0.0627		

The table reports estimated parameters (βs)
Clustered standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A4.3 – Probability of citation from within the US, as a function of co-ethnicity or co-nationality -- Logit regression

	CO-ETHNICITY		CO-NATIONALITY	
	(1)	(2)	(3)	(4)
Same MSA	0.139***	0.138***	0.138***	0.138***
	(0.0292)	(0.0292)	(0.0293)	(0.0293)
Same State	0.153***	0.154***	0.153***	0.154***
	(0.0209)	(0.0209)	(0.0209)	(0.0209)
ln(Miles)	-0.0107**	-0.0107**	-0.0111**	-0.0110**
	(0.00537)	(0.00536)	(0.00537)	(0.00537)
Co-ethnicity/ Co-nationality §	0.247***		0.282***	
	(0.0131)		(0.0153)	
China §		0.310***		0.334***
		(0.0157)		(0.0184)
Germany §		0.0838*		0.133**
		(0.0491)		(0.0590)
France §		-0.0526		0.0278
		(0.0875)		(0.104)
India §		0.180***		0.224***
		(0.0213)		(0.0276)
Iran §		0.699**		1.236
		(0.351)		(0.785)
Italy §		0.202		0.250
		(0.213)		(0.168)
Japan §		0.251**		0.156
		(0.114)		(0.142)
Korea §		0.145		0.238*
		(0.110)		(0.132)
Poland §		-1.182		-1.514*
		(0.735)		(0.912)
Russia §		0.438***		0.444***
		(0.129)		(0.149)
Social distance dummies	yes	yes	yes	yes
Citing patent characteristics	yes	yes	yes	yes
OST-30	yes	yes	yes	yes
Constant	2.953***	2.959***	2.957***	2.963***
	(0.168)	(0.168)	(0.168)	(0.168)
Observations	237,696	237,696	237,696	237,696
Chi2	10052	10207	9972	10085
Log-lik.	-154586	-154547	-154600	-154579
Pseudo-R2	0.0617	0.0620	0.0617	0.0618

§ Co-ethnicity in columns 1 and 2 ; co-nationality in columns 3 and 4
The table reports estimated parameters (βs)
Clustered robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A4.4 – Probability of citation from within the US, as a function of co-ethnicity, by technological class of cited patents -– Logit regression

	Electrical eng.; Electronics	Instruments	Chemicals; Materials	Pharma & Biotech.	Industrial processes	Mechanical eng.; Transport	Consumer goods; Civil eng.
Same MSA	0.175***	0.187***	0.0640**	0.0709**	0.144***	0.145**	-0.0248
	(0.0231)	(0.0264)	(0.0284)	(0.0276)	(0.0416)	(0.0606)	(0.0775)
Same State	0.00720	0.0528***	0.159***	0.242***	0.0182	-0.0650	-0.0669
	(0.0171)	(0.0197)	(0.0229)	(0.0223)	(0.0309)	(0.0469)	(0.0612)
ln(Miles)	-0.0285***	-0.0335***	-0.0193***	-0.00255	-0.0565***	-0.0839***	-0.0809***
	(0.00481)	(0.00532)	(0.00503)	(0.00491)	(0.00773)	(0.0112)	(0.0152)
China	0.199***	0.113***	0.302***	0.269***	0.120***	0.0709	-0.120
	(0.0207)	(0.0256)	(0.0152)	(0.0141)	(0.0395)	(0.0751)	(0.139)
Germany	-0.0176	0.0273	0.144***	0.101***	0.00168	-0.00500	-0.251**
	(0.0462)	(0.0354)	(0.0422)	(0.0370)	(0.0599)	(0.0791)	(0.111)
France	0.151	-0.190**	-0.115	-0.113	-0.381**	-0.249	-0.202
	(0.0973)	(0.0966)	(0.0782)	(0.0703)	(0.178)	(0.270)	(0.255)
India	0.142***	0.0258	0.222***	0.175***	0.0144	0.103*	-0.192**
	(0.0147)	(0.0253)	(0.0220)	(0.0209)	(0.0377)	(0.0611)	(0.0944)
Iran	0.155	0.308*	0.194	0.629*	0.738*	-0.220	
	(0.134)	(0.183)	(0.324)	(0.337)	(0.431)	(0.366)	
Italy	-0.0881	-0.101	0.0391	0.226	-0.322	0.392	-0.718
	(0.145)	(0.216)	(0.146)	(0.250)	(0.254)	(0.432)	(0.478)
Japan	-0.135	0.123	0.186*	0.240***	0.213	-0.259	0.498
	(0.115)	(0.128)	(0.0971)	(0.0881)	(0.195)	(0.381)	(0.619)
Korea	0.0980	0.317**	0.102	0.0751	0.332*	0.0982	-0.0451
	(0.103)	(0.125)	(0.0912)	(0.0904)	(0.172)	(0.315)	(0.562)
Poland	-0.136	-0.0596	0.0959	-0.611**	1.477*	-1.665*	0.926
	(0.443)	(0.358)	(0.286)	(0.293)	(0.759)	(1.009)	(1.150)
Russia	0.211**	0.254*	0.573***	0.456***	0.163	0.159	0.351
	(0.106)	(0.133)	(0.130)	(0.122)	(0.211)	(0.435)	(0.491)
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	2.331***	2.433***	2.374***	3.046***	1.888***	1.816***	0.992***
	(0.160)	(0.153)	(0.136)	(0.190)	(0.186)	(0.278)	(0.362)
Observations	338,598	314,880	300,338	364,106	118,550	44,796	23,249
Chi2	16548	11653	11430	12179	8778	4293	2155
Log-lik.	-220543	-202585	-190794	-232490	-73830	-28342	-14944
Pseudo-R2	0.0603	0.0718	0.0835	0.0788	0.102	0.0872	0.0727

The table reports estimated parameters (βs)
Clustered standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A4.5–Probability of citation from outside the US, as a function of inventors' country of residence (Home country) and Country of Origin (Co-ethnicity) -– Logit regression

	HOME COUNTRY		CO-ETHNICITY
	(1)	(2)	(3)
Same company	1.130***	1.127***	1.131***
	(0.0289)	(0.0320)	(0.0315)
Home country / Co-ethnicity §:			
China	0.175***	0.173***	0.178***
	(0.0286)	(0.0287)	(0.0236)
Germany	-0.0138	-0.00616	0.00182
	(0.0132)	(0.0135)	(0.0129)
France	0.0570*	0.0304	0.124***
	(0.0306)	(0.0317)	(0.0305)
India	0.0399	0.0326	0.129***
	(0.0462)	(0.0472)	(0.0304)
Italy	-0.0487	-0.0821	-0.0420
	(0.0570)	(0.0575)	(0.0513)
Japan	0.0165	0.00201	0.0181
	(0.0246)	(0.0255)	(0.0261)
Korea	0.409***	0.438***	0.430***
	(0.0519)	(0.0522)	(0.0514)
Russia	0.618***	0.640***	0.542***
	(0.176)	(0.176)	(0.0990)
Home country / Co-ethnicity # Same company §:			
China # Same company		0.151	0.258
		(0.270)	(0.212)
Germany # Same company		-0.161**	-0.166**
		(0.0693)	(0.0702)
France # Same company		0.375***	0.157
		(0.131)	(0.125)
India # Same company		0.250	0.219
		(0.242)	(0.186)
Italy # Same company		0.721**	0.558*
		(0.359)	(0.314)
Japan # Same company		0.248**	0.235**
		(0.112)	(0.113)
Korea # Same company		-0.676***	-0.772***
		(0.205)	(0.202)
Russia # Same company		-0.918	0.506
		(1.096)	(1.130)
Returnee	1.121***	1.110***	1.087***
	(0.215)	(0.213)	(0.213)
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	1.859***	1.861***	1.865***
	(0.171)	(0.171)	(0.171)
Observations	1,004,950	1,004,950	1,004,950
Chi2	60073	60221	60459
Log-lik.	-629238	-629208	-629161
Pseudo-R2	0.0967	0.0967	0.0968

§ « Home country » effect in columns 1 and 2 ; Co-ethnicity in column 3

The table reports estimated parameters (βs)

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

Table A4.6 – 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 – Logit regression

Same company	1.133***
	(0.0368)
Home country :	
China	0.141*
	(0.0788)
India	-0.350
	(0.221)
Russia	0.818
	(0.551)
Co-ethnicity :	
China	0.176***
	(0.0355)
India	0.164***
	(0.0380)
Russia	0.504***
	(0.122)
Home country # Co-ethnicity :	
China	-0.162*
	(0.0911)
India	0.257
	(0.230)
Russia	-0.742
	(0.576)
Returnee	0.702*
	(0.401)
Observations	621,283
Chi	42178
Log-lik.	-390178
Pseudo-R2	0.0939

The table reports estimated parameters (βs)
Clustered robust 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) -– Logit regression

	HOME COUNTRY	CO-ETHNICITY	CO-NATIONALITY
	(1)	(2)	(3)
Same company	1.009***	1.013***	1.008***
	(0.0738)	(0.0714)	(0.0727)
Home country / Co-ethnicity / Nationality [§]:			
China	0.175***	0.159***	0.178***
	(0.0469)	(0.0392)	(0.0430)
Germany	0.0217	0.0219	0.0409
	(0.0475)	(0.0401)	(0.0427)
France	0.0367	0.0739	0.0551
	(0.0629)	(0.0605)	(0.0594)
India	-0.169*	0.0379	-0.0264
	(0.0988)	(0.0635)	(0.0755)
Italy	0.127	0.0467	0.165
	(0.162)	(0.129)	(0.133)
Japan	0.0518	0.0768	0.0718
	(0.0580)	(0.0559)	(0.0560)
Korea	0.542***	0.503***	0.533***
	(0.109)	(0.107)	(0.109)
Russia	0.349	0.473**	0.376
	(0.430)	(0.233)	(0.304)
Home country / Co-ethnicity / Nationality # Same company [§]:			
China # Same company	-0.828**	0.0177	-0.158
	(0.388)	(0.365)	(0.412)
Germany # Same company	-0.219*	-0.221*	-0.204*
	(0.122)	(0.121)	(0.118)
France # Same company	0.318	0.301	0.270
	(0.209)	(0.213)	(0.205)
India # Same company	0.353	-0.0428	-0.0410
	(0.458)	(0.361)	(0.384)
Italy # Same company	0.924*	0.422	0.626
	(0.472)	(0.511)	(0.509)
Japan # Same company	0.590**	0.551**	0.573**
	(0.250)	(0.250)	(0.251)
Korea # Same company	-0.885	-1.007*	-1.025*
	(0.566)	(0.526)	(0.526)
Returnee	1.355***	1.343***	1.340***
	(0.355)	(0.354)	(0.354)
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	2.376***	2.370***	2.371***
	(0.331)	(0.331)	(0.331)
Observations	163,319	163,316	163,319
Chi2	9964	9963	9965
Log-lik.	-104474	-104477	-104477
Pseudo-R2	0.0771	0.0771	0.0771

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

The table reports estimated parameters (βs)

Clustered robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix 5 – 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 A5.1). Notice that the distribution by CoO/Nationality is very similar in the two samples. For results and related comments, see table 5 in the paper.

Table A5.1. Local and international samples based on Country of Origin (full sample) vs Nationality-based samples (for robustness checks); by CoO/Nationality of cited inventors

	Local 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 run separate regressions by macro-technological classes of patents (see table 6 in the paper).

Third, 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 A5.2). We then control for the fixed effects of either the top MSAs or the top corridors (table A5.3). Our main results remain unaltered.

Table A5.2. 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 A5.3. 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) -- OLS 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

We also 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 A5.4). Our main results remain unchanged.

Table A5.4. - 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) -– OLS 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

Finally, 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 4 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 A5.5, 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 A5.5. Probability of citation from within the US, as a function of co-ethnicity, bootstrap regressions -- OLS

	(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 3

[b] Specification as in column 1 of table 4

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

References

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

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.

Foreign inventors in the US:
Testing for Diaspora and Brain Gain Effects

Stefano Breschi , Francesco Lissoni, Ernest Miguelez

This version: *27 July 2015*

TABLES

Table 1. Local 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 (3)
	Nr	%	nr	%	nr	%	nr
1. Local 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 (1)	108,449	100.00%	354,389	100.00%	521,660	100.00%	1,043,320
Total (2)	89,986		195,595		437,737		875,474
2. International sample (citations from outside the US)							
China	31,321	25.86%	88,675	22.97%	128,122	25.50%	256,244
Germany	21,512	17.76%	72,694	18.83%	88,782	17.67%	177,564
France	8,246	6.81%	29,305	7.59%	34,050	6.78%	68,100
India	37,984	31.36%	114,872	29.75%	158,233	31.49%	316,466
Iran	3,311	2.61%	12,042	2.96%	13,358	2.54%	26,716
Italy	5,019	4.14%	19,834	5.14%	23,114	4.60%	46,228
Japan	6,189	5.11%	23,281	6.03%	26,575	5.29%	53,150
Korea	5,957	4.92%	21,072	5.46%	24,512	4.88%	49,024
Poland	2,089	1.65%	7,391	1.82%	8,348	1.59%	16,696
Russia	4,911	4.05%	16,330	4.23%	19,087	3.80%	38,174
Total (1)	126,621	100.00%	406,226	100.00%	525,118	100.00%	1,050,236
Total (2)	105,059		266,629		432,681		865,362

(1) Total = sum of observations by country of origin (same patent may be recorded under >1 country)
(2) Total = sum of distinct observations
*(3) Nr observations per country = Nr cited-citing pairs * 2*

Table 2. Local and international samples: descriptive statistics

	Obs	Mean	Std. Dev.	Min	Max
1. Local sample (citations from within the US)					
Citation	1043320	0.50	0.50	0	1
Co-ethnicity	1043320	0.13	0.33	0	1
Same MSA	1043320	0.14	0.34	0	1
Same State	1043320	0.22	0.41	0	1
Miles	1043320	933.71	877.68	0	5085
Soc. Dist. 0	1043320	0.01	0.09	0	1
Soc. Dist. 1	1043320	0.01	0.09	0	1
Soc. Dist. 2	1043320	0.01	0.08	0	1
Soc. Dist. 3	1043320	0.01	0.09	0	1
Soc. Dist. >3	1043320	0.24	0.43	0	1
Soc. Dist. ∞	1043320	0.73	0.44	0	1
claims	1043320	8.50	12.80	0	259
backward citations	1043320	4.58	3.15	0	87
NPL citations	1043320	1.33	2.45	0	57
overlap IPCs 7 digits	1043320	1.13	1.47	0	27
overlap IPCs 7 digits / all IPCs	1043320	0.28	0.28	0	1
overlap IPCs	1043320	0.83	1.57	0	53
2. International sample (citations from outside the US)					
Citation	1004950	0.50	0.50	0	1
Co-ethnicity	1004950	0.10	0.30	0	1
Home country	1004950	0.09	0.29	0	1
Same company	1004950	0.03	0.17	0	1
Returnee	1004950	0.00	0.02	0	1
Contiguous countries	1004950	0.03	0.18	0	1
Former colonial relationship	1004950	0.20	0.40	0	1
Same country	1004950	0.04	0.19	0	1
English	1004950	0.17	0.38	0	1
Similarity to English	1004950	0.25	0.26	0	1
Miles	1004950	4452.46	1936.59	0	11498.1
Soc. Dist. 0	1004950	0.00	0.06	0	1
Soc. Dist. 1	1004950	0.01	0.07	0	1
Soc. Dist. 2	1004950	0.00	0.07	0	1
Soc. Dist. 3	1004950	0.00	0.07	0	1
Soc. Dist. >3	1004950	0.20	0.40	0	1
Soc. Dist. ∞	1004950	0.78	0.42	0	1
claims	1004950	9.90	11.82	0	442
backward citations	1004950	4.00	3.20	0	98
backward NPL citations	1004950	1.00	2.07	0	76
overlap IPCs 7 digits	1004950	1.09	1.28	0	32
overlap IPCs 7 digits / all IPCs	1004950	0.31	0.30	0	1
overlap IPCs	1004950	0.79	1.38	0	54

Table 3 – Probability of citation from within the US, as a function of co-ethnicity, spatial & social distance, and controls -- OLS regression

	(1)	(2)	(3)	(4)	(5)
Same MSA	0.135***	0.0894***	0.0887***	0.0778***	0.0104***
	(0.00153)	(0.00236)	(0.00238)	(0.00236)	(0.00379)
Co-ethnic	0.0510***	0.0421***	-0.000590	-0.00330	-0.000133
	(0.00152)	(0.00193)	(0.00721)	(0.00882)	(0.00906)
Co-ethnic * MSA	-0.0161***	-0.0165***	-0.0125***	-0.0127***	-0.0160***
	(0.00375)	(0.00436)	(0.00461)	(0.00455)	(0.00454)
Same State					0.0227***
					(0.00280)
ln(Miles)					-0.0314***
					(0.00203)
ln(Miles)^2					0.00239***
					(0.000186)
Soc. Dist. 1		-0.0442***	-0.0431***	-0.0424***	-0.0182***
		(0.00412)	(0.00446)	(0.00551)	(0.00574)
Soc. Dist. 2		-0.137***	-0.134***	-0.130***	-0.101***
		(0.00655)	(0.00716)	(0.00797)	(0.00823)
Soc. Dist. 3		-0.224***	-0.219***	-0.214***	-0.182***
		(0.00688)	(0.00766)	(0.00839)	(0.00861)
Soc. Dist. >3		-0.392***	-0.402***	-0.373***	-0.335***
		(0.00287)	(0.00302)	(0.00394)	(0.00458)
Soc. Dist. ∞		-0.428***	-0.433***	-0.390***	-0.351***
		(0.00274)	(0.00284)	(0.00382)	(0.00449)
Co-ethnic *Soc. Dist. 1			0.00515	0.0107	0.00538
			(0.0101)	(0.0118)	(0.0120)
Co-ethnic *Soc. Dist. 2			0.00394	0.000358	-0.00362
			(0.0140)	(0.0155)	(0.0157)
Co-ethnic *Soc. Dist. 3			0.00275	0.00808	0.00591
			(0.0140)	(0.0149)	(0.0150)
Co-ethnic *Soc. Dist. >3			0.0629***	0.0630***	0.0590***
			(0.00737)	(0.00898)	(0.00921)
Co-ethnic *Soc. Dist. ∞			0.0335***	0.0402***	0.0366***
			(0.00716)	(0.00875)	(0.00899)
ln(claims)				0.00119**	0.00124***
				(0.000468)	(0.000468)
ln(1 + backward citations)				0.0830***	0.0829***
				(0.00111)	(0.00111)
ln(1 + backward NPL cit.)				-0.00613***	-0.00628***
				(0.00104)	(0.00104)
ln(1 + overlap IPCs 7 digits)				0.213***	0.212***
				(0.00142)	(0.00142)
Technology F.E.	no	no	yes	yes	yes
Constant	0.475***	0.892***	0.898***	0.649***	0.708***
	(0.000278)	(0.00289)	(0.00298)	(0.00451)	(0.00595)
Observations	1,043,320	1,043,320	1,043,320	1,043,320	1,043,320
R2	0.010	0.023	0.023	0.080	0.081
F	3457	7879	4991	2590	2295

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

Table 4 – Probability of citation from within the US, as a function of co-ethnicity by Country of Origin, spatial & social distance, and controls, -- OLS 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		
OST-30 FE	yes	yes		
Observations	1,043,320	1,043,320		
R2	0.080	0.080		
F	2139	1601		

Clustered standard errors 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 or co-nationality -– OLS regression

	CO-ETHNICITY		CO-NATIONALITY	
	(1)	(2)	(3)	(4)
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***
		(0.00356)		(0.00412)
Germany		0.0196*		0.0305**
		(0.0111)		(0.0132)
France		-0.0111		0.00673
		(0.0197)		(0.0231)
India		0.0416***		0.0514***
		(0.00490)		(0.00634)
Iran		0.159**		0.265**
		(0.0754)		(0.131)
Italy		0.0485		0.0596
		(0.0477)		(0.0368)
Japan		0.0549**		0.0338
		(0.0260)		(0.0327)
Korea		0.0308		0.0529*
		(0.0243)		(0.0292)
Poland		-0.251**		-0.243
		(0.126)		(0.160)
Russia		0.101***		0.104***
		(0.0290)		(0.0333)
Social distance dummies	yes	yes	yes	yes
Citing patent characteristics	yes	yes	yes	yes
OST FE	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

§ Co-ethnicity in columns 1 and 2 ; co-nationality in columns 3 and 4

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

Table 6 – Probability of citation from within the US, as a function of co-ethnicity, by technological class of cited patents -- OLS 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 7– "International" sample: distribution of observations (patent pairs) by Country of Origin (CoO) and country of residence of the inventors

	Inventor of citing/control patent is:							
CoO of cited inventor	Nor in home country, nor from same CoO	Not in home country, but from same CoO	In home country, from different CoO	In home country, from same CoO	*Tot*	(4)/(2+4)	(4)/(3+4)	(2+3+4)/Tot
	(1)	(2)	(3)	(4)		(5)	(6)	(7)
China	243700	6088	847	5609	*256244*	48%	87%	4.9%
Germany	117421	6607	5678	47858	*177564*	88%	89%	33.9%
France	58178	2056	1389	6477	*68100*	76%	82%	14.6%
India	309428	4216	182	2640	*316466*	39%	94%	2.2%
Iran	26546	84	2	2	*26634*	2%	50%	0.3%
Italy	43582	661	223	1762	*46228*	73%	89%	5.7%
Japan	37829	210	238	14873	*53150*	99%	98%	28.8%
S.Korea	46596	131	60	2237	*49024*	94%	97%	5.0%
Poland	16578	78	6	12	*16674*	13%	67%	0.6%
Russia	37574	406	20	174	*38174*	30%	90%	1.6%

Table 8– "International" sample: distribution of observations (patent pairs) by Country of Origin (CoO) and country of residence of the inventors, and patent ownership

	Inventor of citing/control patent is:							
CoO of cited inventor	Not in home country, diffe-rent company	Not in home country, same company	In home country, diffe-rent company	In home country, same company	*Tot*	(4)/(2+4)	(4)/(3+4)	(2+3+4)/Tot
	(1)	(2)	(3)	(4)		(5)	(6)	(7)
China	243625	6163	6294	162	*256244*	3%	3%	4.9%
Germany	121515	2513	49380	4156	*177564*	62%	8%	31.6%
France	58775	1459	6904	962	*68100*	40%	12%	13.7%
India	306558	7086	2696	126	*316466*	2%	4%	3.1%
Iran	26024	606	4		*26634*			2.3%
Italy	43221	1022	1819	166	*46228*	14%	8%	6.5%
Japan	36973	1066	13669	1442	*53150*	57%	10%	30.4%
S.Korea	45780	947	2159	138	*49024*	13%	6%	6.6%
Poland	16285	371	18		*16674*			2.3%
Russia	37213	767	186	8	*38174*	1%	4%	2.5%

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) -– OLS 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 [§]:			
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 [§]:			
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

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

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

Table 10 – 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 - OLS regression

Same company	0.214***
	(0.00639)
Home country :	
China	0.0316*
	(0.0178)
India	-0.0771
	(0.0471)
Russia	0.160*
	(0.0942)
Co-ethnicity :	
China	0.0359***
	(0.00740)
India	0.0337***
	(0.00810)
Russia	0.110***
	(0.0255)
Home country # Co-ethnicity :	
China	-0.0314
	(0.0203)
India	0.0603
	(0.0491)
Russia	-0.148
	(0.101)
Returnee	0.0734
	(0.0521)
Observations	621,283
R2	0.120
F	2071

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

Table 11 – Probability of citation from outside the US, as a function of "home-country" effect, co-ethnicity or co-nationality (also by Country of Origin) -– OLS 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