NBER WORKING PAPER SERIES

MIGRANTS, ANCESTORS, AND INVESTMENTS

Konrad B. Burchardi
Thomas Chaney
Tarek A. Hassan

Working Paper 21847
http://www.nber.org/papers/w21847

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2016

We are grateful to Richard Hornbeck, Emir Kamenica, Jacopo Ponticelli, Nancy Qian, and David Stromberg for helpful discussions. We also thank seminar participants at the University of Chicago, MIT, Princeton, Harvard, Boston University, Boston College, Oxford, Imperial, the University of Maryland, CEPR ERWIT, NBER ENT, Paris School of Economics, Singapore Management University, National University of Singapore, UPF, Barcelona GSE, and IFN (Stockholm), for their comments. Chaney is grateful for financial support from ERC grant N 337272FiNet. Hassan is grateful for financial support from the IGM and the Fama-Miller Center at the University of Chicago. Philip Xu provided excellent research assistance. All mistakes remain our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Migrants, Ancestors, and Investments
Konrad B. Burchardi, Thomas Chaney, and Tarek A. Hassan
NBER Working Paper No. 21847
January 2016, Revised August 2016
JEL No. F21,G15,J61,L14,N3,O11

ABSTRACT

We use 130 years of data on historical migrations to the United States to show a causal effect of the ancestry composition of US counties on foreign direct investment (FDI) sent and received by local firms. To isolate the causal effect of ancestry on FDI, we build a simple reduced-form model of migrations: migrations from a foreign country to a US county at a given time depend on (i) a push factor, causing emigration from that foreign country to the entire United States, and (ii) a pull factor, causing immigration from all origins into that US county. The interaction between time-series variation in origin-specific push factors and destination-specific pull factors generates quasi-random variation in the allocation of migrants across US counties. We find that a doubling of the number of residents with ancestry from a given foreign country relative to the mean increases by 4 percentage points the probability that at least one local firm engages in FDI with that country, and increases by 29% the number of employees at domestic recipients of FDI from that country. This effect operates mainly through the descendants of migrants rather than migrants themselves and increases in size with the ethnic diversity of the local population, the distance to the origin country, and the quality of its institutions.

Konrad B. Burchardi
Institute for International Economic Studies
Stockholm University
SE-106 91 Stockholm
Sweden
konrad.burchardi@iies.su.se

Thomas Chaney
Department of Economics
Toulouse School of Economics
21 Allee de Brienne
31000 Toulouse
France
thomas.chaney@gmail.com

Tarek A. Hassan
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
tarek.hassan@chicagobooth.edu

Over the past decades, international migrations have reached unprecedented levels,[1] shaping an increasingly ethnically diverse and socially connected world. The economic consequences of these migrations are at the heart of fierce political debates on immigration policy, yet our understanding of the economic effects of migrations remains incomplete. At the same time, foreign direct investment (FDI) undertaken by multinational firms has become a defining feature of international production.[2] Local policymakers see attracting and retaining FDI as a major goal, and technology transfers through FDI are both a conduit for technological progress abroad and a source of revenue for US firms.[3] Migrations and FDI create two parallel global networks, one of ethnic connections, one of parent-subsidiary linkages. How do these two networks affect each other? In this paper, we estimate the long-term effect of immigration on the patterns of FDI sent and received by US firms, and shed light on the mechanism behind this effect. We show that immigration and FDI are intimately related: The ethnic diversity created by migrations reaching back more than a century has a large positive causal effect on the propensity of US firms to engage in FDI with the historical migrants' countries of origin.

Evaluating the causal impact of migrations on FDI requires a rigorous identification strategy, as unobserved factors may simultaneously affect migrations, ancestry, and FDI, creating a spurious correlation between them. We construct a set of instrumental variables (IV) for the present-day ancestry composition of US counties, best explained by the examples of migrations from Germany and Italy. German migrations peaked at the end of the nineteenth century when the Midwest was booming and attracting large numbers of migrants. We observe a large population with German ancestry in the Midwest today. Italian migrations peaked a few decades later, at the beginning of the twentieth century when the West were attracting large numbers of migrants. We observe a large population with Italian ancestry in the West. We use this interaction of time-series variation in the relative attractiveness of different destinations within the United States (e.g. end of nineteenth century Midwest versus early twentieth century West) with the staggered arrival of migrants from different origins (e.g. end of nineteenth century Germany versus early twentieth century Italy) to instrument for the present-day distribution of ancestries. Our formal IV strategy is essential. For instance, while the effect of ancestry on FDI is positive in both ordinary least squares (OLS) and IV specifications, its effect on international trade becomes insignificant or even negative when we instrument for ancestry, suggesting that

[1] The number of international migrants worldwide reached 232 million in 2013, an all time high (UN Population Facts No. 2013/2).

[2] In 2009, 55% of all US exports emanated from US multinationals that operated subsidiaries abroad. These firms employ 23 million Americans, while US subsidiaries of foreign firms employ another 5 million. Source: Office of the United States Trade Representative, Fact Sheet on International Investment.

[3] See McGrattan and Prescott (2010) and Holmes et al. (2015).

unobservable factors indeed confound simple OLS estimates of these effects.

Our paper makes three main contributions: (i) foreign ancestry has a positive causal effect on FDI that increases with the ethnic diversity of the local population, the distance to the origin country, and the quality of its institutions; (ii) this ethnic determinant of FDI has deep historical roots and operates mainly through the descendants of migrants rather than migrants themselves; and (iii) we propose a general method for instrumenting the composition of ancestry.

Before describing the related literature, we summarize our main empirical results.

We find that, for an average US county, doubling the number of individuals with ancestry from a given origin country increases by 4 percentage points the probability that at least one firm from this US county engages in FDI with that origin country, and increases by 29% the number of local jobs at subsidiaries of firms headquartered in that origin country. The presence of second and third generation descendants of immigrants has a significantly larger effect than first-generation immigrants. Even the earliest migrations in the nineteenth century for which we have data significantly affect the patterns of FDI today.

To arrive at those findings on the *causal* impact of foreign ancestry on the patterns of FDI, we follow an IV strategy. We motivate our approach using a simple reduced-form dynamic model of migrations. Migrations from a given origin country o to a given US destination county d in period t depend on the total number of migrants arriving in the United States from o (a push factor), the relative economic attractiveness of d to migrants arriving in t (a pull factor), and the size of the pre-existing local population of ancestry o in d at t, allowing for the fact that migrants tend to prefer settling near others of their own ethnicity (a recursive factor). Solving the model shows that the number of residents in d today who are descendants of migrants from o is a function of simple and higher-order interactions of the sequence of pull and push factors.

To construct valid instruments from this sequence of interactions, we isolate variation in the pull and push factors that is plausibly independent of any unobservables that may make a given destination within the US differentially more attractive for both settlement and FDI from a given origin country. To that end, we measure the pull factor from country o to county d as the fraction of migrants coming from anywhere in the world who settle in d at time t, excluding migrants from the same continent as o. The pull from o towards d thus depends only on the destination choices of migrants arriving at the same time from other continents. Similarly, we measure the push factor as the total number of migrants arriving in the United States from o at time t, excluding migrants from o who settled in the same region as d. We then instrument for the present-day number of residents in county d with ancestry from country o using the full set of simple and higher-order interactions of these pull and push factors. Using the entire series of interactions going back to 1880 maximizes the statistical power of our IV strategy.

Flexibly applying our instrumentation strategy to the entire set of origins and destinations allows us to corroborate our approach in various ways and to guard against a range of potentially confounding factors. We are able simultaneously to control for both origin and destination fixed effects, thus controlling for all origin- and destination-specific factors, such as differences in size, market access, and productivity. Regional interactions of the origin and destination fixed effects also allow us to address a wide range of threats to our identifying assumption.

In addition, we conduct a range of falsification exercises and robustness checks. Importantly, we obtain quantitatively very similar effects of ancestry on FDI when we combine our IV strategy with a natural experiment surrounding the rise and fall of communism. Making use of the periods of economic isolation between the United States and former communist countries, these specifications (similar to a difference-in-difference) measure how cross sectional variations in ancestry driven only by the inflow of migrants over a period of exclusion explain changes in FDI, from zero during the exclusion period to its current level in 2014.

Our approach delivers the statistical power to generate three main additional findings. First, while the effect of ancestry on FDI is almost always positive, significant, and of similar magnitude across individual US counties, origin countries, and sectors, this effect increases with the ancestry diversity of the US county, as well as the geographic distance from, the judicial quality of, and the ethnic diversity of the foreign country. Once these characteristics are accounted for, other measures of genetic, linguistic, and religious distance do not significantly affect the size of the effect of ancestry on FDI. We also find the effect is stronger for firms producing intermediates than final goods, suggesting local tastes do not explain our results. Second, the effect of immigration on FDI is significantly smaller for first-generation immigrants, so that the full effect of ancestry on FDI unfolds over multiple generations. Third, we find evidence of negative spillovers, where, for instance, a large Polish contingent in a given US county has a lower effect on FDI if nearby counties also host large contingents from Poland, or if the county also hosts a large contingent from the nearby Czech Republic. Overall, our results suggest that a small minority of second- or third- generation immigrants from a distant part of the world not otherwise represented in the local ethnic mix has the largest marginal impact on FDI.

To illustrate the quantitative implications of our results, we conduct two thought experiments. In the first, we calculate the effect of Chinese exclusion – the effective ban on Chinese immigration between 1882 and 1965. Absent this ban, we predict the fraction of counties in the Northeast with FDI links to China would have increased substantially (e.g. doubled in New York state).

In the second, we calculate the effect of a hypothetical "L.A. gold rush" – an early population growth in Los Angeles before 1880 similar to the experience of San Francisco. We predict there would have been 60,000 more individuals with German and Irish ancestry in Los Angeles, and FDI between Los Angeles and Germany and Ireland would have increased by around 60%.

Finally, we note two important limitations to our analysis. First, our results rely purely on cross-sectional variation in FDI within the United States. Although we believe that, in light of our results, the ethnic diversity of the United States likely also raises the extent of FDI for the United States as a whole, we cannot exclude the possibility that increases in FDI in one state are partially or fully offset by decreases in other states. Second, our results appear most consistent with a model in which common ancestry affects FDI by diminishing information frictions rather than through shared tastes or the provision of social collateral, but we provide only indirect evidence on the nature of the mechanism linking ancestry to FDI. Some of our earlier work provides more direct evidence. Burchardi and Hassan (2013) use microeconomic data to show that social ties to East Germans provided West German households and firms with valuable information after the fall of the Berlin Wall. Chaney (2014) structurally estimates a model in which the percolation of information about foreign trading partners drives firm-level exports.

Existing literature. A large body of literature shows that measures of affinity between regions, such as common ancestry, social ties, trust, and telephone volume, correlate strongly with aggregate economic outcomes, such as foreign direct investment (Guiso, Sapienza, and Zingales, 2009), international asset flows (Portes and Rey, 2005), and trade flows (Gould, 1994; Rauch and Trindade, 2002).[4] How much of this association should be interpreted as causal, however, remains an open question because these measures of affinity, and ancestry in particular, are likely to be nonrandom across regions. Two recent papers use historical decisions by the US government on the location of Japanese internment camps during World War II (Cohen, Gurun, and Malloy, 2015) and the placement of Vietnamese refugees after the Vietnam War (Parsons and Vezina, 2014) to identify a causal effect of concentrations of descendants of these migrants on contemporary trade flows between locations within the United States and Japan and Vietnam, respectively. Burchardi and Hassan (2013) use variation in wartime destruction across West German regions in 1945, a time when millions of refugees were arriving from East Germany, as an instrument for the share of the population with social ties to the East, and show evidence of a causal effect of these social ties on GDP growth and FDI in East Germany after the fall of

[4] Also see Head and Ries (1998), Combes, Lafourcade, and Mayer (2005), Garmendia, Llano, Minondo, and Requena (2012), and Aleksynska and Peri (2014) for the relationship between common ancestry and trade and Bhattacharya and Groznik (2008) for its relationship with FDI.

the Berlin Wall.[5] Other papers studying the effect of historical shocks on economic interactions across borders include Redding and Sturm (2008), Juhász (2014), and Steinwender (2014).

We contribute to this literature in several ways. First, we identify a causal effect of ancestry on FDI in a setting with a high degree of external validity directly relevant for assessing, for example, the long-term effects of immigration policy. Second, because our identification strategy can be applied to all origin countries, we are able to guard against a wide range of possible confounding factors and to relate to the previous literature by employing a gravity equation with both destination and origin fixed effects. Third, we identify the determinants of the heterogeneity in the effect of ancestry on FDI across origins, destinations, and sectors, and show evidence of negative geographic spillovers. Fourth, our results suggest the causal effects of ancestry on FDI and trade flows may be very different, although they appear similar in OLS regressions.

Our paper also contributes to the debate on the costs and benefits of immigration. Much of the existing literature has focused on the effects of migration on local labor markets, mostly in the short run.[6] A more recent literature focuses on the effect of cultural, ethnic, and birthplace diversity on economic development and growth.[7] Most closely related are Fulford, Petkov, and Schiantarelli (2015) who study the effect of ancestry composition of US counties on GDP growth over time and Nunn, Qian, and Sequeira (2015) who study the effect of immigration from all origins during the Age of Mass Migration on present-day outcomes. We add to this literature by examining the effect of migration on the pattern of economic exchange and employment. By looking at the composition of US residents with *foreign ancestry*, as opposed to just *foreign-born* residents, we are able to separate the short-term and long-term effects of migration. In this sense, our results show a long-term effect of migration on the comparative advantage in FDI of different regions that operates across multiple generations and may explain part of the association between diversity and long-term growth found in other studies.

Our approach to identification is related to Card (2001) who instruments immigration flows from origin o to destination d with the interaction of the total immigration from o to the United States (the push factor) and the spatial distribution of previous migrants from o in the United States (the recursive factor). This strategy has been widely used in the literature to instrument for changes in labour supply caused by immigration. However, it is not appropriate in our context, where unobserved and persistent origin-destination specific characteristics (such as factor

[5] See Fuchs-Schündeln and Hassan (2015) and Chaney (2016) for surveys of this literature.

[6] See for example Card (1990), Card and Di Nardo (2000), Friedberg (2001), Borjas (2003), and Cortes (2008). Borjas (1994) provides an early survey.

[7] See Ottaviano and Peri (2006), Putterman and Weil (2010), Peri (2012), Ashraf and Galor (2013), Ager and Brückner (2013), Alesina, Harnoss, and Rapoport (2015a), and Alesina, Michalopoulos, and Papaioannou (2015b).

endowments) may drive both the spatial distribution of previous migrants and FDI. Our approach instead combines a push-pull model similar to that of Card (2001) with a two-dimensional version of the leave-out approach of Bartik (1991) and Katz and Murphy (1992), and uses multiple subsequent waves of historical migrations going back to the 19th century to instrument for the current stock of ancestry. This hybrid approach, based on a simple recursive model of migrations, can easily be replicated for other countries, other time periods, or variables other than migrations where cumulated flows matter, without the need for a rare or even unique historical accident.

The remainder of this paper is structured as follows. Section 1 introduces our data. Section 2 gives a brief overview of the history of migration to the United States. Section 3 identifies the causal effect of ancestry composition on FDI at the extensive margin, discusses various challenges to our identifying assumption, and conducts a range of robustness checks and falsification exercises. Section 4 probes the effect of ancestry composition on FDI at the intensive margin and other outcomes, and illustrates the quantitative implications of our findings using two thought experiments. Section 5 examines the mechanism underlying the effect of ancestry on FDI by by probing the heterogeneity of the effect across countries, counties, and sectors and by testing for spill over effects. Section 6 concludes.

1 Data

We collect data on migrations and ancestry, on foreign direct investment and trade, and on origin and destination characteristics. Below is a description of our data, along with their source. Further details are given in Appendix A.

Migrations and ancestry. All migration and ancestry data are constructed from the individual files of the Integrated Public Use Microdata Series (IPUMS) samples of the 1880, 1900, 1910, 1920, 1930, 1970, 1980, 1990, and 2000 waves of the US census, and the 2006-2010 five-year sample of the American Community Survey. We weigh observations using the personal weights provided by these data sources. Appendix Table 1 summarizes specific samples and weights used.

Throughout the paper, we use $t-1$ to denote the census wave just before t, o for the foreign country of origin, and d for the US destination county. We construct the number of migrants from origin o to destination d at time t, $I^t_{o,d}$, by counting the number of respondents who live in d, were born in o, and emigrated to the United States between t and $t-1$. The exception to this rule is 1880 census, which did not record the year of immigration. The variable $I^{1880}_{o,d}$ instead measures the number of residents who were either born in o or whose parents were born in o.[8] Since 1980, respondents have also been asked about their primary ancestry in both the US Census and the American Community Survey, and they can provide multiple answers. $Ancestry^t_{o,d}$ corresponds to the number of individuals residing in d at time t who report o as first ancestry. Note that this measure captures self-reported and hence recalled ancestry, which may potentially be more relevant for economic exchange than genetic (factual) ancestry.[9]

The respondents' residence is recorded at the level of historic counties, and at the level of historic county groups or PUMAs from 1970 onwards. Whenever necessary we use contemporaneous population weights to transition data from the historic county group or PUMA level to the historic county, and then use area weights to transition data from the historic county level to the 1990 US county level.[10] The respondents' stated ancestry (birthplace) often, but not always, directly corresponds to foreign countries in their 1990 borders (for example, "Spanish" and "Denmark"). However, in other cases no direct mapping exists (for example, "Basque" or "Lapland"). For these cases, we construct transition matrices that map data from the answer level to the 1990 foreign country level, using approximate population weights where possible and approximate area weights otherwise. In the few cases when answers are imprecisely specific or such a mapping cannot be constructed (for example, "European" or "born at sea"), we omit the data. Appendix Tables 2 and 3 report summary statistics on these data transitions, including the share of affected respondents. See Appendix A.1 for details.

Foreign direct investment. Our data on FDI is from the US file of the 2014 edition of the Bureau van Dijk ORBIS data set.[11] For each US firm, the database lists the location of its (operational) headquarters, the addresses of its foreign parent entities, and the addresses of its international subsidiaries and branches. We use a 5% ownership threshold to define a parent-subsidiary link. Altogether we have information on 36,108 US firms that have at least one foreign parent or subsidiary. Collectively, these firms have 102,618 foreign parents and 176,332 foreign subsidiaries in 142 countries (in their 1990 borders).[12] We then aggregate this information to the county level. Our main outcome variable, *FDI Dummy*, is 1 if at least one firm within

[8] If the own birthplace is in the United States, imprecisely specific (e.g., a continent), or missing, we instead use the parents' birthplace, assigning equal weights to each parent' birthplace.

[9] See Duncan and Trejo (2016) for recent evidence on recalled versus factual ancestry in CPS data.

[10] We also aggregate our data to the PUMA level and show that our results are robust (available on request).

[11] In robustness checks we show that our results do not change when we instead use data from the 2007 file.

[12] Although Bureau van Dijk cross checks the data on international subsidiaries and branches using both US and foreign data sources, we cannot exclude the possibility that coverage may be better for some countries than for others. However, all of our specifications control for country fixed effects such that any such variation in coverage at the country level would not affect our results.

a given destination county has at least one parent or subsidiary in the origin country. For each destination county, we also count the total number of FDI linkages (the total number of foreign parents and subsidiaries of all firms within the county), and the total number of unique parents and subsidiaries in both the origin and the destination. We also count the total number of employees working at firms with a foreign parent in a given destination (*# of Employees at Subsidiaries in Destination*) and the total number of employees working at subsidiaries of US firms in a given origin country (*# of Employees at Subsidiaries in Origin*).[13] The ORBIS database also gives the 2007 NAICS code of the sector of the US firm, allowing us to disaggregate these data by 2-digit sector.[14] See Appendix A.2 for details.

Other data. We use data on aggregate trade flows between US states and foreign countries for the years 2002 to 2011 from the US Census Bureau. When we aggregate our dataset across US states, the correlation with aggregate trade between the entire US and foreign countries from the NBER bilateral trade dataset is 99.9% for imports and 99.7% for exports respectively (in 2008). When we aggregate our data across foreign countries, the correlation of between state level aggregate trade and state population is 93% for imports and 88% for exports respectively. We are therefore confident our trade dataset disaggregated at the US state x foreign country level is not subject to severe measurement error. To further guard against measurement error, we also use data for the manufacturing sector only, final goods only, or intermediate inputs only.

We also construct bilateral distances and latitude differences between US counties and foreign countries, and characteristics for countries, counties, and sectors. See Appendix A.3 for details.

Summary statistics. Panel A of Table 1 gives summary statistics on our sample of 3,141 × 195 origin-destination pairs.[15] Column 1 shows means and standard deviations for all observations. Columns 3-4 show the same statistics for the subsamples of origin-destination pairs containing only observations with non-zero ancestry, and ancestry in the bottom and top quintile, respectively. The table shows that a lot of the variation both in ancestry and FDI is at the extensive margin. Only 1.8% of origin-destination pairs have an FDI link. Conditional on the US county having any population with origins in the foreign country, 3.1% have an FDI link. The larger this population, the larger the probability of finding an FDI link, with 12.8% of the origin-destination pairs in the top quintile having an FDI link. Similarly, about half of the

[13] When information on the number of employees is missing (which is the case for 6% and 25% of subsidiaries in the destination and origin, respectively), we assume the subsidiary employs one person.

[14] Appendix Table 4 provides a list of sectors and sector groups.

[15] 53 countries thus have no FDI links with US firms in our sample.

origin-destination pairs have ancestry of zero, reflecting the fact that most destinations in the United States do not have populations with ancestry from all 195 countries. The mean number of individuals with ancestry from a given origin is 316, but is highly skewed, with a mean in the top quintile of 2.9 million individuals. Compared to this stock of ancestry, the flow of immigrants between 1990 and 2000 is relatively small, with 25 on average across the sample. The summary statistics also show that the number of first-generation immigrants (foreign born) measured in the 2010 American Communities Survey appears somewhat understated (69 on average). This fact is known in the literature and appears to affect only the measurement of immigration flows but not the stock of ancestry (Jensen et al., 2015). For this reason, we exclude the 2000-2010 wave of migrations from our standard specification and instead rely on the pre-2000 census numbers.

Panels B and C show summary statistics following the same format for destination counties and origin countries for variables used in our estimation of heterogenous effects. Appendix Table 5 gives summary statistics on the intensive margin of FDI.

2 Historical background

The 1880 US census counted 50 million residents, 10 million of which were first- or second-generation immigrants from 195 countries. The censuses taken since 1880 counted an additional 67 million immigrants. Our sample period thus covers the vast majority of migrations.[16]

During the first part of this period, up until World War I, migration to the United States was largely unregulated. European migrants in particular faced few or no restrictions to entry and came in large numbers. Figure 1 shows the extent and the changing composition of migration over time. Although the peak of British migration was passed before the beginning of our sample, the numbers for 1880 clearly show the effect of the potato famines and the subsequently large inflow of Irish migrants. The second big wave of migration in our sample is that of Germans in the aftermath of the failed revolutions of 1848 and the consolidation of the German empire under Prussian control in 1871. Similarly disrupted by political changes and an economic crisis in the South, Italian migrants began flocking to the United States in large numbers around 1910, followed by a peak in migrations from Eastern Europe, and in particular from Russia, in the years after the October Revolution. The inflow of migrants overall dropped dramatically during World War I, falling below 4 million during the period between 1910 and 1930.

Although economic and political factors in the origin countries dominated the timing of

[16] The historical information in this section is from Daniels (2002) and Thernstrom (1980). Also see Goldin (1994) for the political economy of US immigration policy.

these earlier European migrations, US immigration policies became relatively more important during the 1920s. The first important step toward regulating the inflow of migrants was the Chinese exclusion act of 1882 that ended the migration of laborers, first from China, and then in following incarnations from almost all of Asia. These restrictions were followed by literacy and various other requirements that came into effect after 1917, culminating in the establishment of a quota system in 1921. The quota system limited the overall number of immigrants, reduced the flow of migrants from Southern and Eastern Europe, and effectively shut out Africans, Asians, and Arabs. Combined with the effects of the Great Depression, these new regulations led to negative net migration in the early 1930s and then a stabilization at relatively low levels of immigration. The quota system was abolished in 1965 in favor of a system based on skills and family relationships, leading both to a large increase in the total number of migrants and a shift in composition toward migrants from Asia and the Americas, in particular from Mexico.

Figure 2 maps the spatial settlement pattern of newly arrived immigrants in the United States over time. For each census from 1880 to 2010, we compute the total number of new migrants to destination d, I_d^t,[17] projected on destination and year fixed effects to account for general immigration time trends and persistent destination-specific effects. We show only the residuals from this projection, color coded by decile. Migrants initially settled on the East Coast of the United States (in the mid-19th century), and then the frontier for migrants moved to the Midwest (in the late-19th century), to the West (1900-30), and to the South (in the 1980s). Starting in the 1970s, we can also see graphically the increased settlement of migrants in urban centers, with a series of dark dots appearing around large urban areas.

Below we use the interaction of this time-series variation in the relative attractiveness of different destinations within the United States with the staggered arrival of migrants from different origins as the basis of our identification strategy.

3 Ancestry and Foreign Direct Investment

To evaluate the effect of the presence of descendants of migrants from a given origin on the probability that at least one firm within a given destination has an FDI link with a firm based in the origin country (inward or outward), we estimate the structural equation,

$$\mathbf{1}\left[FDI_{o,d} > 0\right] = \delta_o + \delta_d + \beta A_{o,d}^{2010} + X_{o,d}'\gamma + \varepsilon_{o,d}, \tag{1}$$

[17] Note we treat our first (1880) census differently: because we have no previous census, we define the number of migrants for 1880 as all residents in 1880 who are either foreign born or whose parents are foreign born.

where $\mathbf{1}\left[FDI_{o,d}>0\right]$ is a dummy variable equal to 1 if any firm headquartered in destination d is either the parent or the subsidiary of any firm headquartered in origin o in 2014. $A_{o,d}$ is a measure of common ancestry, usually calculated as the log of 1 plus the number of residents in d that report having ancestors in origin o in 2010, measured in thousands. (We choose this functional form in anticipation of non-parametric results discussed below, but also show robustness to a wide range of alternative specifications in section 3.4). $X'_{o,d}$ is a vector of control variables that always includes the geographic distance between o and d, and the difference in latitude between o and d. δ_o and δ_d represent a full set of origin and destination fixed effects, augmented in most of our specifications by fixed effects for the interaction between destination and continent of origin, and between origin and destination census region.[18] The coefficient of interest is β, which measures the effect of ancestry on the probability that an FDI relationship exists between firms in o and d. The error term $\varepsilon_{o,d}$ captures all omitted influences, including any deviations from linearity.[19] Throughout the main text, we report heteroskedasticity-robust standard errors clustered at the origin-country level. In the appendix, we report standard errors calculated using alternative methods for the main results of the paper, and show our results are robust.

Equation (1) takes the form of a gravity equation, widely used in the empirical literature describing the pattern of international trade and FDI. We maintain the same form for consistency with this literature. Moreover, the gravity form is appealing on theoretical grounds because it can be derived in a variety of models.[20] The destination and origin fixed effects absorb all differences in productivity, market size, and market access between origins and destinations that systematically affect prices. We may thus interpret the coefficient β as the effect of ancestry controlling for the large set of conventional economic forces shaping international exchanges.

Equation (1) will consistently estimate the parameter of interest if $Cov\left(A_{o,d}^{2010},\varepsilon_{o,d}\right)=0$. This condition is unlikely to hold in our data, despite the inclusion of origin and destination fixed effects. First, past origin-destination specific migration flows might be the result of economic transactions such as FDI or trade, not their driver.[21] Second, origin-destination specific omitted factors might drive both economic transactions and migration flows, affecting both $A_{o,d}$ and $\mathbf{1}\left[FDI_{o,d}>0\right]$. An example of such an endogenous assignment of migrants and FDI would be

[18] A census region is one of nine groupings of adjacent US states listed in Appendix Table 6.

[19] We use a simple linear probability model, which allows for a straight-forward interpretation of the coefficient. In robustness checks, we also report results from a probit estimator.

[20] See Arkolakis et al. (2012) for a derivation of the gravity structure of international trade in a variety of theoretical settings. See Carr et al. (2001), Razin et al. (2003), Head and Ries (2008), and Ramondo (2014) for an application of the gravity structure to foreign direct investment.

[21] An example of such reverse causality is the strong concentration of Japanese in Scott County, Kentucky, which emerged after Toyota seconded Japanese workers to a newly built manufacturing facility in the 1980s.

a simple extension of the Heckscher-Ohlin model: High skill migrants from high skill abundant countries migrate to low skill abundant US counties, to benefit from a higher skill premium, while firms in high skill abundant countries tend to engage in FDI towards high skill abundant US counties, to use their technology with the same skill mix. So migrations are driven by factor endowment *differences*, while FDI flows are driven by factor endowment *similarities*, inducing a downward bias in the OLS estimate of β. Third, ancestry might be selectively recalled where economic transactions exist.[22] These challenges are not unique to our data, but are likely concerns with any data where ethnic linkages and economic transactions are simultaneously observed.

To address these concerns, we devise an instrumental variables (IV) strategy. This strategy is guided by a simple dynamic model of migration, which helps to understand the variation we are using and relate our approach to the existing literature. The stock of residents of ancestry from origin o in destination d at time t, $A^t_{o,d}$, depends on the past stock of residents with ancestry from o and the newly arrived migrants from o who settle in d. The combination of three forces determines the number of new migrants: A country-specific *push factor* drives migrants out of country o into the United States; a *pull factor* attracts migrants entering the United States to county d, irrespective of their origin; and a *recursive factor* corresponds to the tendency of newly arrived migrants to settle in communities where people with the same ancestry already live.

Formally, the stock of residents in d with ancestry from o at time t evolves according to

$$A^t_{o,d} = a_t + a_{o,t} + a_{d,t} + b_t A^{t-1}_{o,d} + I^t_o \left(c_t \frac{I^t_d}{I^t} + d_t \frac{A^{t-1}_{o,d}}{A^{t-1}_o} \right) + \nu^t_{o,d}. \tag{2}$$

The constant terms a_t, $a_{o,t}$, and $a_{d,t}$ control for residual forces, such as demographics, which may vary over time, over space, and between different ethnic groups. The term $b_t A^{t-1}_{o,d}$ corresponds to the fact that ancestry is a stock variable that evolves cumulatively. The constant b_t controls for demographics and for how ties to one's ancestry are passed from one generation to the next. The term I^t_o, the total number of migrants from country o entering the United States, measures the strength of the push factor, the fact that migrants are driven out of country o. The fraction of all migrants entering the United States who settle in county d from all origins, I^t_d/I^t, measures the strength of the economic pull factor, the fact that county d is particularly appealing to migrants at time t. The fraction of people with ancestry from country o who already live in county d, $A^{t-1}_{o,d}/A^{t-1}_o$, measures the strength of the recursive factor, the propensity of migrants to settle near their countrymen. The coefficients c_t and d_t control the relative importance of the pull and recursive factors. If the pull factor is absent, and only the recursive factor affects the allocation

[22] See Duncan and Trejo (2016) for recent evidence on selective recall in self-reported ancestry data.

of newly arrived migrants, $c_t = 0$, our model collapses exactly to the Card (2001) model. Finally, $\nu_{o,d}^t$ is a sequence of error terms that are potentially correlated with $\varepsilon_{o,d}$.

Equation (2) is not a suitable first stage if persistent forces shape both the settlement of migrants and FDI, i.e. if $\nu_{o,d}^{t-1}$ and $\varepsilon_{o,d}$ are correlated. In this case $A_{o,d}^{t-1}$ and $\varepsilon_{o,d}$ would be correlated. Therefore an IV strategy following Card (2001), using variations in I_o^t and $A_{o,d}^{t-1}$ as instruments, would not be suitable in our setting.

We address this challenge by noting that equation (2) is recursive, both because ancestry is passed down from generation to generation (the first $A_{o,d}^{t-1}$ term) and because newly arrived migrants' decision of where to settle depends on where past migrants have settled (the second $A_{o,d}^{t-1}$ term). Given that our data cover the vast majority of migration to the United States (a total of 70 million immigrants), we assume the initial condition $A_{o,d}^{1880``-1"} = 0, \forall\,(o,d)$ for simplicity.[23] Solving equation (2) recursively, we get,

$$A_{o,d}^{2010} = \sum_{t=1880}^{2010} \left(a_t + a_{o,t} + a_{d,t} + c_t I_o^t \frac{I_d^t}{I^t} + \nu_{o,d}^t \right) \prod_{s=t+1}^{2010} (b_s + d_{o,s} I_o^s), \tag{3}$$

where the constant $d_{o,s}$ only contains information on total migrations from o in previous periods.

The reduced-form equation (3) highlights how ancestry is the result of a sequence of migration waves and their subsequent cumulative effect. In each period t, the interaction of the contemporaneous push factor (I_o^t) and economic pull factor (I_d^t/I^t) determines the flow of migration from o to d. Demographic factors (the b_s's) and the recursive factor (the $d_{o,s}$'s) then amplify these initial waves of migrants. This simple specification is flexible, allowing for cases in which no migrants from a given origin country exist at some initial period of time. In the absence of a recursive factor, $d_t = 0$, ancestry only depends on the interactions of the push and pull factors.

This specification suggests plausibly exogenous variation in $I_o^t\,(I_d^t/I^t)$ would allow the construction of an instrument for $A_{o,d}^{2010}$. By interacting a push factor, I_o^t, which is not specific to destination d, but common to all destinations in the United States, and a pull factor, I_d^t/I^t, which is not specific to country o but to migrants from all countries, we rule out most plausible sources of endogeneity. However, our exclusion restriction could still be violated since $I_{o,d}^t$ is mechanically a component of I_o^t, I_d^t and I^t and potentially related to $\epsilon_{o,d}$. This would be a concern if at some point in time, migrants from o to d represent a large fraction of all migrants from o ($I_{o,d}^t$ a large fraction of I_o^t), or a large fraction of all migrants to d ($I_{o,d}^t$ a large fraction of I_d^t), or if migrants

[23] Remember we treat our first 1880 Census differently, and count all first- *and* second-generation immigrants in the I^{1880} terms. So assuming zero stock of ancestry pre-1880 is a good approximation: $A^{1880``-1"} = 0$, where the superscript 1880"-1" denotes pre-1880 ancestry.

from other origins with unobserved similarities to o represent a large fraction of all migrants.

To address these concerns, we exclude from the push factor migrants from o going to all destinations in d's census region, and from the economic pull factor, migrants from all origins in the same continent as o. We replace I_o^t by $I_{o,-r(d)}^t$, the migrants from o who settle in destinations *not* in the same census region as d; and I_d^t/I^t by $I_{-c(o),d}^t/I_{-c(o)}^t$, the fraction of migrants *not* coming from origins in the same continent as o who settle in county d. $-r\left(d\right)$ stands for all destinations outside of d's census region, and $-c\left(o\right)$ stands for all origins outside of o's continent.

Replacing the $I_o^t \frac{I_d^t}{I^t}$ terms by $I_{o,-r(d)}^t \frac{I_{-c(o),d}^t}{I_{-c(o)}^t}$ in (3), our first-stage specification is thus

$$A_{o,d}^{2010} = \delta_o + \delta_d + \sum_{t=1880}^{2000} \alpha_t I_{o,-r(d)}^t \frac{I_{-c(o),d}^t}{I_{-c(o)}^t} + \sum_{n=1}^{5} \delta_n PC_n + X'_{o,d}\gamma + \eta_{o,d}, \tag{4}$$

where $\sum_{n=1}^{5} \delta_n PC_n$ stands for the first five principal components summarizing the information contained in the the series $I_{o,-r(d)}^s \cdots I_{o,-r(d)}^t \frac{I_{-c(o),d}^t}{I_{-c(o)}^t}, \forall t < s \leq 2010$. We prefer summarizing the higher-order interactions in (3) as principle components because it avoids adding an excessive number of highly co-linear regressors. Our results are robust to including these terms or not.

Our key identifying assumption is

$$Cov\left(I_{o,-r(d)}^t \frac{I_{-c(o),d}^t}{I_{-c(o)}^t}, \varepsilon_{o,d} | controls\right) = 0. \tag{5}$$

It requires that any confounding factors that make a given destination more attractive for both migration and FDI from a given origin country do not simultanously affect the interaction of the settlement of migrants from other continents with the total number of migrants arriving from the same origin but settling in a different census region.

To further relax this assumption, most of our specifications also control for interactions of fixed effects that are symmetric to the construction of our instruments: The interaction between destination and continent-of-origin fixed effects ($\delta_d \times \delta_{c(o)}$) and the interaction between origin and destination-census-region fixed effects ($\delta_o \times \delta_{r(d)}$). In these regressions we restrict ourselves to using variation across origin countries from the same continent, holding the destination constant, and variation across destinations within the same census region, holding the origin constant. These specifications are by construction robust to any confounding factors that are origin-census region or continent-destination specific. We further corroborate the robustness of our approach below using a series of falsification exercises and placebo treatments.

3.1 The First-Stage Relationship

Table 2 shows our basic first-stage regressions, estimates of equation (4). Column 1 is the most parsimonious specification regressing our measure of ancestry on origin and destination fixed effects and the nine simple interaction terms $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_t$. To facilitate the interpretation of the results, we sequentially orthogonalize each of the terms with respect to the interaction terms from the previous censuses. For example, the coefficient marked $I^{1900}_{o,-r(d)}(I^{1900}_{-c(o),d}/I^{1900}_{-c(o)})$ shows the effect of the residual obtained from a regression of $I^{1900}_{o,-r(d)}(I^{1900}_{-c(o),d}/I^{1900}_{-c(o)})$ on the same interaction in 1880, the coefficient marked 1910 shows the effect of the residual from a regression of the 1910 interaction on the interactions from the previous two censuses, and so on. Although this procedure has no effect on the fit and predictive power of the first stage as a whole, we find it useful because it allows us to interpret each coefficient as the marginal effect of the innovation in the migration pattern of the period reported with respect to the previous periods.

All nine coefficients shown in column 1 are positive, and seven are statistically significant at the 5% level. Figure 4 depicts the coefficients graphically. The first main insight from this figure is that even our earliest (pre-1880) snapshot of the cross-sectional variation in economic attractiveness to new migrants has left its imprint on the present-day ancestry composition of US counties: the destinations that were relatively more attractive to the typical migrant arriving pre-1880 continue to the present day to house significantly larger numbers of residents of the ethnic groups that arrived in large numbers prior to 1880. The overall pattern of coefficients suggests a hump-shape, where very recent waves of migrants have a smaller impact on current ancestry than migrations a few decades back, but the effect of past migrations eventually fades after about one century. An exception to the general pattern is the coefficient for 1920-30, which is smaller and insignificant. A likely explanation is the Great Depression, which induced large reverse migrations from the United States of recently arrived migrants, demonstrating our model is less well suited for periods with negative net migration.

Taken together, the nine simple interactions incrementally increase the R^2 of the regression by 4 percentage points and explain about 9% of the variation in ancestry not explained by origin and destination fixed effects. Column 2 adds controls for distance and latitude difference. Both columns 1 and 2 estimate equation (4) under the restriction that the recursive factor is irrelevant ($d_s = 0$ in (2)). Column 3 relaxes this restriction and adds the higher-order interaction terms, raising the R^2 by another three percentage points. Columns 4 and 5 add destination per continent-of-origin fixed effects and origin per destination-census-region fixed effects, respectively.

Column 5 is our standard specification. The F-statistic against the null that the excluded instruments are irrelevant in this specification is 161.7.[24] Column 6 includes third-order polynomials in the distance and latitude difference between o and d. Columns 7 through 9 successively show variations of our instrumentation strategy: column 7 includes migration data from the 2005-2010 ACS survey, column 8 drops migration prior to 1880, and column 9 estimates our standard specification in levels rather than logs. Throughout all of these variations, we can comfortably reject the null that our instruments are jointly irrelevant in the first stage.

We illustrate our first-stage identification with a stylized graphical illustration in Figure 3, using two specific examples: That of migrations from Germany, with a migration peak in the pre-1900 period (corresponding to the failed 1848 revolution and the consolidation of the German empire under Prussian control), and that of Italy, with a migration peak in the 1900-30 period (triggered by the end of feudalism and demographic pressures, accelerated with the institutional support for emigration in 1901, and ending with the quota system in the U.S. in 1921 and Mussolini's anti-emigration policies). The top-left part shows the relative attractiveness of US destinations for pre-1900 migrants, when German migrations to the United States peaked, where we exclude migrations from Europe – analogously to our regression specification. At that time, most non-European migrants settled in the Midwest. We expect most German migrants from this initial wave to have settled in the Midwest. The top-right part shows the distribution of US residents with German ancestry in 2010, with disproportionately many in the Midwest. The bottom-left part shows the relative attractiveness of US destinations for non-European migrants during the 1900-30 period, when Italian migrations to the United States peaked. At that time, the preferred destination for migrants had shifted to the West and South. We expect many Italians migrants to have settled in the West and South. The bottom-right part shows the distribution of Italian descendants in 2010, with large populations in the West and South.

3.2 Instrumental Variables Results

In our IV estimation, we explicitly test the hypothesis that an increase in the number of descendants from a given origin increases the probability that at least one local firm engages in FDI with that country. In column 1 of Table 3, we estimate equation (1) while instrumenting (the log of) ancestry in 2010 with the simple interaction terms $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_t$ and controlling for origin and destination fixed effects, distance, and latitude difference. The coefficient estimate on ancestry is 0.231 (s.e.=0.023), statistically significant at the 1% level. The coefficient on distance is not statistically distinguishable from zero, perhaps reflecting the fact that US counties

[24] The Hansen J test statistic is 15.891 with a p-value of 0.255. We thus fail to reject the null that our instruments are uncorrelated with the error term and correctly excluded from the second-stage regression.

do not differ much in their distance to most foreign countries, and that these smaller differences are irrelevant once we control for the effect of the distance between the United States as a whole and the country in question (absorbed in the country fixed effect). By contrast, the difference in latitude is positive and significant, showing that, all else being equal, firms tend to engage in FDI with origin countries that are climatically different from their own location. Appendix Figure 1 presents the corresponding reduced form results graphically. All nine coefficients are greater than zero, and seven of them are statistically significant at the 5% level. Destinations that received an (exogenous) increase in the number of migrants from a given origin in any of the nine consecutive waves of immigration thus tend to have a significantly higher probability of engaging in bilateral FDI with these origin countries today. In column 2 of Table 3, we add the five principal components of the higher-order interactions to our set of instruments, resulting in a slight fall in the coefficient of interest to 0.190 (s.e.=0.024).

Column 3 shows our standard specification. The estimate, 0.187 (s.e.=0.024), implies that doubling the number of residents with ancestry from a given origin relative to the sample mean (from 316 to 632) increases by 4 percentage points the probability that at least one firm engages in FDI with that origin.[25] This specification includes destination per continent-of-origin fixed effects and origin per destination-census-region fixed effects. For a given origin country, this demanding specification uses only variation across different destinations within the same census region while controlling for the fact that each destination may have a high or low idiosyncratic propensity to interact with the continent containing the origin country, and symmetrically for destinations. Reassuringly, adding these 17,460 fixed effects has almost no effect on our coefficient of interest (0.187, s.e.=0.024 versus 0.190, s.e.=0.024). Comparing this estimate with the same column in panel B shows that it is about 25% larger than the corresponding OLS coefficient. The endogenous assignment of migrants to destinations within the United States thus appears to induce a downward bias in the OLS coefficient, consistent with a simple extension of the Heckscher-Ohlin model: Migrations tend to be driven by *differences* in factor endowments (creating differences in wages between origin country and destination county), while FDI flows are driven by *similarities* in factor endowments (as firms use FDI to export their technology to countries with a similar mix of factor endowments).

Another useful way to gauge the relative importance of ancestry is its partial R^2 relative to the controls included in the specification. Taken together, the standard gravity terms, that is,

[25] Using $\hat{\beta} = 0.187$ from column 3 in Table 3 in equation (1), we have: $\mathbf{1}\left[FDI_{o,d} > 0 | Ancestry_{o.d} = 632\right] - \mathbf{1}\left[FDI_{o,d} > 0 | Ancestry_{o.d} = 316\right] = 0.187\left(\ln\left(1 + \frac{632}{1000}\right) - \ln\left(1 + \frac{316}{1000}\right)\right) \approx 0.0402$. An IV probit estimate of the same specification yields a marginal effect of Log Ancestry 2010 on $\Pr\left[FDI > 0\right]$ of 0.104 (s.e.=0.037).

the origin and destination fixed effects, distance, and latitude difference, explain 20.3% of the variation in the FDI Dummy. Adding ancestry to these variables in a simple OLS specification (shown in panel B) raises the R^2 by another 9 percentage points, accounting for about about half as much variation as the combined explanatory power of the economic fundamentals reflected in the gravity terms (although this effect is not necessarily causal). Instead adding our nine simple interactions to the standard gravity terms, thus running the most parsimonious reduced form, raises the R^2 by 1.5 percentage points, and adding them in combination with the five principal components raises the R^2 by 2 percentage points. These numbers are a lower bound on the importance of common ancestry for FDI, because it only accounts for part of the causal effect.

The remaining columns of Table 3 probe the robustness of this result. The coefficient estimate remains remarkably stable and highly statistically significant across specifications. Column 4 adds a third-degree polynomial in distance and latitude difference to capture a potentially non-linear effect of distance; column 5 adds an interaction term for the contemporaneous 2010 migrations in the first stage (as in column 7 of Table 2); and column 6 adds a more stringent set of origin×destination-state fixed effects, exploiting only variation within US states. All of these variations leave our coefficient of interest virtually unchanged.

3.3 The Communist Natural Experiment

The main potential challenge to our approach is that, despite our efforts, confounding factors that make a given destination more attractive for both migration and FDI from a given origin country may still, in some complicated way, be correlated with our instruments, although they only use information about migrations from other continents and to other census regions. To address this concern, we use a natural experiment and focus on changes in FDI and changes in ancestry, similar to a difference-in-difference approach: The periods of economic isolation between the United States and communist countries during parts of the 20th century.

These periods of isolation are the Soviet Union from 1918 to 1990, China from 1945 to 1980, Vietnam from 1975 to 1996, and Eastern Europe (the non-Soviet members of the Warsaw pact) from 1945 to 1989. At the end of each of these periods of isolation, practically no FDI existed between the United States and each of these countries.[26] Moreover, any migrants arriving in the United States during the period of isolation would plausibly not have expected to be able to conduct FDI or otherwise interact economically with their countries of origin.

Table 4 shows estimates of (1) for each of these countries or sets of countries, using as

[26] See the UNCTAD time series for the stock of FDI at www.unctadstat.unctad.org.

instruments only migration waves that occurred during the period of isolation. This specification offers two advantages. First, we can confidently assume the prospect of FDI, outlawed for political reasons, did not drive migrations during those periods. Second, our estimate is similar to a difference-in-difference: We measure how cross sectional variations in ancestry driven only by the inflow of migrants over a period of exclusion explain changes in FDI, from zero during the exclusion period to its current level in 2014. For all countries, we find a large causal impact of ancestry on FDI. The magnitude of the estimated impact of ancestry is broadly similar to the one we estimated for all countries in Table 3, and the estimated coefficients are statistically significant for the Soviet Union and China, and Eastern Europe. The coefficient is not statistically significant for Vietnam, most likely because most migration from Vietnam occurred before or after a relatively short 20 year period of exclusion. Pooling all countries, we find a coefficient very close to that of our standard specification (0.234, s.e.=0.098).

The fact that we find similar results in these more restrictive natural experiments as in our baseline specification in Table 3 bolsters our confidence that our exclusion restriction is valid, and that neither reverse causality nor omitted variables drive our baseline results.

3.4 Robustness checks

To corroborate the validity of our exclusion restriction, we run a series of robustness checks.

Functional form exploration. In our main specification, we measure our ancestry variable, $A^t_{o,d}$, as the log of one plus the number of residents with foreign ancestry, measured in thousands. Our results are robust to a wide range of alternative functional form specifications, and we offer a formal justification for our main specification using a non-linear estimation.

In panel A of Appendix Table 7, we show our results also hold when we use a non-parametric specification for ancestry. We divide the absolute numbers of individuals of a given ancestry, $Ancestry^{2010}_{o,d}$, into quantiles, including the same covariates as in our simple specification from column 2 in Table 3. We experiment with different numbers of quantiles (4, 5 and 6). Across all specifications, the effect on ancestry on FDI is monotonically increasing and concave.[27]

In panel B of Appendix Table 7, we offer a formal test to justify the functional form $A^{2010}_{o,d} = \ln\left(1 + \frac{1}{1000} Ancestry^{2010}_{o,d}\right)$. To that end, we perform a non-linear least squares estimation of

$$\mathbf{1}\left[FDI_{o,d} > 0\right] = \delta_o + \delta_d + \beta \ln\left(1 + \pi Ancestry^{2010}_{o,d}\right) + X'_{o,d}\gamma + \varepsilon_{o,d},$$

[27] For instance, with sextiles in column 3, the increment in the probability of positive FDI of adding the same 1000 more descendants gradually decreases from quantile to quantile: +0.129 from quantile 1 to 2; +0.061 from 2 to 3; +0.050 from 3 to 4; and +0.001 from the 4 to 5.

again including the same covariates as in our simple specification from column 2 in Table 3. We find a point estimate of $\beta = 0.1683$ and $\pi = 0.0010$. This finding forms the basis for our choice of functional form applied throughout the paper. This functional form is convenient because it offers a compact way to model the non-linear impact of ancestry. For small ancestry ($Ancestry_{o,d} \ll 1000$), the function $\ln(1 + Ancestry_{o,t}/1000)$ is approximately linear in $Ancestry_{o,d}$. For large ancestry ($Ancestry_{o,d} \gg 1000$), it behaves approximatively like $\ln(Ancestry_{o,d})$. So for a small number of residents with foreign ancestry, the coefficient β in (1) measures the proportional impact of ancestry on the extensive margin of FDI; for a large number of residents with foreign ancestry, β is simply the elasticity of the extensive margin of FDI with respect to ancestry.

In Appendix Table 8, we further explore the robustness of our results to alternative functional forms. In column 1, we simply measure ancestry in levels, and find a positive and significant effect. In column 2, we use $\ln\left(Ancestry_{o,d}^{2010}\right)$, and use the value -1 instead of $-\infty$ for $Ancestry_{o,d}^{2010} = 0$, and find a result similar to our baseline specification. In column 3, we use $\left(Ancestry_{o,d}^{2010}\right)^{1/3}$ as an alternative concave function, and find again a robust positive and significant effect of ancestry. In columns 4, 5, and 6, we replicate our results using measures of ancestry from the, 1980, 1990 and 2000 censuses, instead of 2010, and change the dates for our IV interaction terms accordingly. The estimated impact of ancestry on FDI varies little when we move the cutoff date.

Placebo Test. Our next robustness check uses a placebo treatment to assess whether our instrument reliably isolates push factors that are specific only to one country, or is correlated with omitted variables that affect FDI with other countries in a systematic fashion.

The results are presented in Appendix Table 9. In panel A, we assign the interaction between push and pull factors for a given origin to a quasi-randomly selected other country: Its nearest neighbor in alphabetic order. To further check whether the same push factors might affect two countries in different continents, panel B assigns the interaction between push and pull factors for a given country to its nearest neighbor in alphabetic order in a *different* continent. Across all specifications, our placebo treatment is always statistically insignificant, and the point estimates are near zero. We conclude from this placebo test that our instrument is not picking up any artificial correlation (positive or negative) between the push factors for different countries.

Standard Errors. Appendix Table 10 shows our standard specification from column 3 of Table 3 using alternative standard errors clustered by origin, destination, state, state-country, state-continent, and various block-bootstrapped standard errors. Across all these specifications, clustering by origin, as we do throughout the paper, is the most conservative choice.

An alternative way to detect any tendency to over-reject the null is to reassign the "treatment" to a different set of outcome observations, in the spirit of Fisher's randomization inference procedure. We repeatedly assign the interaction between push and pull factors for country o to randomly selected other countries. Appendix Figure 2 shows the histogram of t-statistics on the estimated coefficient on ancestry across 200 random assignments. The t-statistics across random assignments are centered around zero, with no noticeable tendency for positive or negative estimates. Reassuringly, the rates of false positives and negatives are 1.5% and 11.5%.

Ancestry and immigration. According to our reduced-form model of migration, the number of migrants arriving at a given destination is a function of the economic attractiveness of the destination at the time (measured by the interaction of our pull and push factors) and the stock of descendants of migrants from the same origin (the recursive factor). To provide direct evidence these two forces are at work (the push × pull and recursive factors), we estimate the specification

$$I^t_{o,d} = \delta_o + \delta_d + \theta \; I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} + \lambda A^{t-1}_{o,d} + X'_{o,d}\gamma + \vartheta_{o,d} \tag{6}$$

for $t = 2000, 1990$ (the census years for which we have information on lagged ancestry), where we again instrument for $A^{t-1}_{o,d}$ using (4).

Column 1 of table 5 estimates (6) with immigration $I^t_{o,d}$ in levels, and gives a coefficient on the interaction of the push and pull factors close to 1. This finding is what we would expect if newly arrived migrants were distributed uniformly on average. Columns 2 and 3 estimate (6) in logs for two time periods, 1990 and 2000. Across all specifications, both the coefficient on the interaction and on lagged ancestry are positive and significant predictors of current migrations.

Overview of additional robustness checks. Appendix Table 11 shows results from separate regressions for the five largest origins (by number of descendants), and destinations (in total number of foreign ancestry).[28] The impact of ancestry on FDI is similar across specifications.

Appendix Table 14 shows plausible variations of our leave-out instrument, removing or not different sets of migrants from the $I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})$ interaction terms. In panel A, we do not remove any migrants from o to d when computing our push and pull factors, using directly the $I^t_o(I^t_d/I^t)$ interaction terms. In panel B, we only remove migrants from o in the pull factor and migrants to d in the push factor, that is, the interactions $I^t_{o,-d}(I^t_{-o,d}/I^t_{-o})$. In panel C, we additionally remove migrants from the same continent from the pull factor, $I^t_{o,-d}(I^t_{-c(o),d}/I^t_{-c(o)})$.

[28] Appendix Tables 12 and 13 shows the results from separate regressions for all countries and sectors, respectively.

In panel D, we go further than in our standard specification, additionally removing migrants from all adjacent states in calculating the pull factor. Across these variations, the coefficient of interest in our standard specification (column 3) is stable between 0.172 (s.e.=0.024) in panel B and 0.192 (s.e.=0.022) in panel D, where as expected, less aggressive leave-out instruments produce estimates that are marginally closer to the OLS coefficient (0.149, s.e.=0.018).

In Appendix Table 15, we replicate our results using a different time period, 2007 FDI and 2000 ancestry instead of 2014 FDI and 2010 ancestry, and find similar estimates. Further we replicate our results using a different level of geographic aggregation, Public Use Microdata Areas (PUMAs) instead of US counties, and find similar estimates (results available upon request).

4 The Effect of Ancestry on Other Outcomes

4.1 Inward and outward FDI

We first distinguish between inward and outward FDI. To do so, we estimate our standard specification of (1) from column 3 of Table 3 separately for inward FDI, where the outcome variable is a dummy equal to 1 if at least one firm in US county d is a subsidiary of a parent in foreign country o, and for outward FDI, where the outcome variable is a dummy equal to 1 if at least one firm in US county d is the parent of a subsidiary in foreign country o. The coefficients for both outward and inward FDI are positive and statistically significant. We find a stronger impact of ancestry on outward FDI, $\beta_{out} \approx 0.2$, than on inward FDI, $\beta_{in} \approx 0.15$, although both coefficients are not statistically distinguishable from each other.

4.2 The intensive margin of FDI

So far, we have studied the impact of ancestry on the extensive margin of FDI, the probability that at least one firm engages in FDI. We now turn to the impact of ancestry on the intensive margin of FDI: Conditional on being positive, how large are FDI flows for a given size of the local population with a given foreign ancestry?

In Table 6, we estimate by IV various specifications of

$$\ln FDI_{o,d} = \delta_o + \delta_d + \kappa A_{o,d}^{2010} + X_{o,d}'\gamma + \zeta_{o,d}. \tag{7}$$

where $FDI_{o,d}$ corresponds to various measures of the volume of FDI between o and d and where we instrument $A_{o,d}^{2010}$ with the same first-stage equation (4) as earlier. Because of the log specification, cases of zero FDI will automatically be dropped from our sample. This creates a selection problem, as counties with non-zero FDI are likely to be systematically different from those with zero FDI. To correct for this potential selection bias, we implement a simple Heckman (1979) correction. We first estimate an IV probit regression for the extensive margin of FDI

$$\rho_{o,d} = \Pr\left(FDI_{o,d} > 0 | observables\right) = \Phi\left(\delta_o^{pr} + \delta_d^{pr} + \beta^{pr} A_{o,d}^{2010} + X_{o,d}' \gamma^{pr}\right), \tag{8}$$

where $A_{o,d}^{2010}$ is again instrumented as in equation (4). We extract an estimate for $\hat{z}_{o,d} = \Phi^{-1}\left(\hat{\rho}_{o,d}\right)$, the predicted latent variable that determines non-zero FDI. We then include an inverse Mills ratio term, $\hat{\mu}_{o,d} = \varphi\left(\hat{z}_{o,d}\right)/\Phi\left(\hat{z}_{o,d}\right)$, within our set $X_{o,d}$ of controls in the intensive margin equation (7), where φ and Φ denote respectively the p.d.f. and c.d.f. of the normal distribution. This correction for selection, the extensive margin of FDI, is similar to the procedure in Helpman et al. (2008) for international trade.[29]

We use various measures for the volume of FDI. In panel A of Table 6, we count the total number of FDI relationships, that is, the sum of the number of firms in d which are either parent or subsidiary of a firm in o and the number of firms in o which are parent or subsidiary of a firm in d. In panel B, we only count the number of firms in d which are a subsidiary of a firm in o, a measure of inward FDI. In panel C, we only count the number of firms in o which are parent of a firm in d, an alternative measure of inward FDI. In panel D, we measure the total local employment in county d at subsidiaries of firms in o, giving us a measure of the impact of inward FDI on local employment. Panels E, F, and G use similar measures for outward FDI: the number of foreign subsidiaries of local firms in E, the number of local parents of foreign subsidiaries in F, and the number of foreign workers employed by subsidiaries of local firms in panel G.

Across most specifications, and for most measures of the intensive margin of FDI, we find a positive impact of ancestry on the volume of FDI. The effect of ancestry on the intensive margin of FDI, the coefficient κ in equation (7), is large and significant across most specifications. Doubling the number of residents in county d who report ancestry from country o (from the mean, 316, to 632) increases local employment at subsidiaries of foreign firms by 29%.[30] Alternatively,

[29] Note Helpman et al. (2008) correct for both the selected presence of zeros, as well as for the unobserved selection of which firm engages in foreign activities, export in their case. We only correct for the presence of zeros, not for the selection of firms, for three reasons. First, we are not interested in how ancestry affects the volume of FDI of one individual firm but rather in how ancestry affects the *total* volume of FDI between a US county and a foreign country, unlike Helpman et al. (2008) who are interested in how various covariates affect the export of one individual firm. Second, we directly use firm-level data, so that we do not require an explicit correction for firm selection. Finally, at the very fine level of geographic disaggregation we use – US counties as opposed to entire countries in Helpman et al. (2008) – the simple structural model they use to motivate their correction for firm selection is unlikely to be appropriate.

[30] Using $\hat{\kappa} = 1.168$ in panel D, column 2 of Table 6 in equation (7), we have: $\frac{Employment_{o,d}[Ancestry_{o,d}=2\times 316]}{Employment_{o,d}[Ancestry_{o,d}=316]} - 1 = \exp\left(0.401\left(\ln\left(1+\frac{2\times 316}{1000}\right) - \ln\left(1+\frac{316}{1000}\right)\right)\right) - 1 \approx 0.29$.

increasing the number of residents in d with ancestry from o by one standard deviation from the mean increases local employment by 16% of a standard deviation.

With all measures of the volume of FDI, the estimated impact of ancestry is larger in our IV specification (column 2) than in the OLS specification (column 1). This downward bias of the OLS estimate is similar to our result for the extensive margin of FDI, and again consistent with a simple extension of the Heckscher-Ohlin model: Migrations tend to increase with factor endowment *differences*, while the intensity of FDI tends to decrease with factor endowment *similarities*. For all measures of the volume of FDI, the impact of ancestry is larger when we include our complete set of interacted fixed effects (column 2) than when we use a simple gravity specification without interacted fixed effects (column 3). Finally, correcting for selection using a Heckman type procedure always leads to a lower estimated impact of ancestry on the volume of FDI (column 4).[31] For our measures of outward FDI (panels E, F and G), the impact of ancestry becomes insignificant, with slightly negative point estimates. Except for those three cases, correcting for selection does not change our results substantially.

Figure 5 illustrates these results graphically by estimating equation (7) using data only for Germany and Britain (top parts), and LA and Cook counties (bottom parts). Each graph shows a conditional scatterplot of the number of subsidiaries as a function of ancestry. They all show a positive and significant slope close to the corresponding full-sample estimate in column 3 of Table 6 and no obvious outliers.

The conclusion from Table 6 is that foreign ancestry affects both the extensive and intensive margins of FDI. More descendants of foreign migrants increases the likelihood that local firms engage in FDI, the number of firms that do so, and the local employment by foreign-owned firms.

4.3 International trade

To compare our methodology to the existing literature, we apply our methodology to the effect of common ancestry on international trade.

We do not have access to micro data on international trade at the level of individual US counties. Instead, we use readily available trade data from the US Census Bureau at the level of US states. Our instrument for the composition of ancestry is the same as in equation (4) for FDI, except that all variables are defined at the state level, and not at the county level.

[31] Note that the number of interacted fixed effects in column 2 is too large for a probit estimation of the extensive margin of FDI to be computationally feasible. Moreover, Greene et al. (2002) show that that probit regressions tend to give biased estimates in the presence of a large number of fixed effects. For both reasons, in column 4, we implement the Heckman-type correction for selection in the simple specification with fixed effects only for origins and destinations.

To compare the magnitude of the effect of ancestry on international trade, we focus our attention on the intensive rather than extensive margin of trade, as most of the literature has done. Table 7 presents the results of the estimation of various specifications of equation (7), where the dependent variable is now total exports or imports in 2011 between US state d and foreign country o. Again, we allow for a Heckman-type correction for the selection bias due to zero trade. Panel A shows the results for the intensive margin of FDI aggregated to the state level for comparison. Panels B and C show the corresponding results for exports and imports. Panels D and E show results for Japan and Vietnam only. The main finding emerging from the table is that when we properly instrument for ancestry, and include a full set of origin country and destination US state fixed effects, we continue to find a robustly positive and significant effect of ancestry on FDI, but not on international trade of US states. This result can be seen in columns 3 and 4 of Table 7, where the effect of ancestry on trade is insignificant or negative.

This finding is in stark contrast to earlier findings in the literature, started by the seminal contributions of Gould (1994) and Rauch and Trindade (2002) (using OLS), and the recent IV results of Cohen et al. (2015) for trade with Japan, and Parsons and Vezina (2014) for trade with Vietnam, that all find the presence of migrants facilitates both exports and imports. We do not find any such positive impact of ethnic ties (ancestry) on international trade. A closer look at the data suggests two important features are essential in reaching this negative conclusion: When either a formal identification is missing (OLS in column 1), or no control for destination –US state– fixed effect is included (column 2), we erroneously find a positive and significant estimated impact of ancestry on trade. But when both are present (columns 3 and 4), we find none. This upward bias of the OLS estimate (column 1) compared to the IV (column 3) is again consistent with a simple Heckscher-Ohlin model, where both migrations and trade are positively affected by differences in factor endowments. The fact that the semi-elasticity of trade w.r.t. ancestry is close to unity in IV regressions without state fixed effects (column 2) is consistent with a simple scaling effect, where larger US states have proportionately more trade and larger populations.

In unreported robustness checks, we find similar results for other years, or when restricting our analysis only to trade in manufacturing goods, where determining the final destination (origin) of an import (export) is less subject to measurement error, as well as for separate regressions on final goods and on intermediate inputs.

4.4 Quantitative implications

Having estimated the impact of ancestry on a range of different outcomes, we illustrate the quantitative implications of our findings using two thought experiments. First, we estimate how investment relations between US counties and China might have evolved if Chinese migrants had not been effectively barred from entering the United States between 1882 and 1965. Second, we report how FDI relationships between Los Angeles and the world might have evolved if Los Angeles had had an influx of migrants in the 1800s similar to that resulting from the San Francisco Gold Rush. These thought experiments are not meant as formal counterfactuals, but merely as illustrations of the magnitude of the effect of ancestry on FDI implied by our estimates.

The effect of Chinese exclusion. The US government passed the Chinese Exclusion Act into law in 1882 in response to increased immigration from China. It essentially closed the United States to legal immigration of laborers from China. It was in force until 1943, when it was replaced by the Magnuson Act, which allocated a quota of 105 immigrants per year from China, and was in effect until 1965, when the removal of the quota system allowed for large-scale Chinese immigration for the first time. We refer to the entire period from 1882 through to 1965 as the "Chinese Exclusion." How different would the ancestry composition and FDI of US counties be today had it not been for Chinese Exclusion?

We first derive a rough prediction for the time path of Chinese migration to the United States in this scenario. We aggregate our immigration data at the time × census-region × origin level to run a regression of the form $I^t_{o,r} = \delta_{t,r} + \delta_o - \xi \cdot D^t_{China} + \nu_{t,o,r}$, where D^t_{China} is a dummy equal to 1 if $o = China$ and $t \in [1882, 1965]$, and $\delta_{t,r}$ and δ_o are time×census region and origin fixed effects, respectively. The coefficient of interest, ξ, estimates the average negative impact of the Chinese Exclusion Act on immigration from China. Defining the hypothetical time path of immigration as $\tilde{I}^t_{o,r} \equiv I^t_{o,r} + \xi \cdot D^t_{China}$, we then predict the change in ancestry using the estimates from our standard first-stage regression as $dA_{o,d} \equiv \sum_t \hat{\alpha}_t \cdot \left(\tilde{I}^t_{o,-r(d)} - I^t_{o,-r(d)}\right) \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$, where $\hat{\alpha}_t$ are the estimated first-stage coefficients. The hypothetical change of FDI relations with China at the county level is $d\Pr\left[FDI_{o,d} > 0\right] \equiv \hat{\beta} \cdot dA_{o,d}$, where $\hat{\beta}$ is the estimated second-stage coefficient from a specification similar to column 3 in Table 3 but without the principal components.[32]

Our estimates suggest that in the absence of the Chinese Exclusion Act, the United States would have received 1.8 million additional Chinese immigrants during the period of exclusion. This increase would have been highly unequally distributed, translating into heterogenous

[32] Dropping the principal components from this specification has only a negligible effect on the coefficient of interest which rises only sightly from 0.187 (s.e.=0.024) to to 0.188 (s.e.=0.023).

changes in the incidence of FDI relationships with China. The map in Figure 6 depicts the expected change in the probability of positive FDI with China, $d\Pr\left[FDI_{China,d} > 0\right]$. The absence of Chinese exclusion would have resulted in substantially stronger FDI ties with the Northeast, the Midwest and the Southwest. The bar graph depicts the fraction of counties within a state which have positive FDI with China in 2014, and the predicted change in this measure of the extensive margin of FDI linkages, i.e. the unweighted average of $d\Pr\left[FDI_{China,d} > 0\right]$ across counties within the state. To save space, the graph shows only the ten states with the highest predicted change. For example, we predict that in the absence of Chinese exclusion, the proportion of counties with an FDI link to China would have doubled in New York, and increased by 60% in Massachusetts or Illinois.

Los Angeles Gold Rush. To similarly gauge the size of the estimated intensive margin effects, we derive predictions on the intensity of FDI relationships between Los Angeles county and the world under the hypothetical scenario that Los Angeles had experienced a Gold Rush similar to that in San Francisco. In particular, we derive predictions on the intensity of FDI relationships with the world if the number of immigrants pre-1880 had been fivefold the actual number of immigrants to Los Angeles. Table 8 presents the results of this thought experiment for the 10 foreign countries with the largest predicted change in their ancestry group in Los Angeles in 2010. Column 1 presents the actual number of individuals of each ancestry in Los Angeles County in 2010. Column 2 presents the total number of FDI links recorded in our data between Los Angeles County and the respective origin countries. Columns 3 and 4 present the predictions of our thought experiment. The calculations are based on the IV specification corresponding to column 2 of Table 6 without the principle components as instruments. A Gold Rush in Los Angeles would have resulted in sizeable effects on the intensity of foreign direct investment relations with those countries that were the source of immigration pre-1880: The intensity of foreign direct investment between Los Angeles County on the one side and Germany and Ireland on the other side would have increased by around 60%. Column 4 presents the predicted absolute change in the size of the ancestry groups, based on a reduced form regression analogous to column 9 of Table 2 with *Ancestry 2010* (in levels) as outcome variable, again excluding the principle components. It suggests that the population of Irish and German descent living in Los Angeles County today would each be counting about 60,000 more individuals.

5 Understanding the Effect of Ancestry

A clear advantage of the fact that our instrumentation strategy can be flexibly applied to the entire set of time periods, origin countries and destination counties, is that it delivers the statistical power and a sufficiently large number of instruments to probe the nature of this effect.

5.1 First-generation immigrants

Having shown the historical stock of ancestry predicts subsequent migrations, we ask whether the effect of ancestry on FDI requires a sustained inflow of migrants from the same origin. To address this question, Table 9 compares the (causally identified) effect of ancestry to that of foreign born, that is, first-generation immigrants. Column 1 replicates our standard specification for comparison. Column 2 replaces our measure of ancestry in equation (1) with the log of 1 plus the number of foreign born from a given origin alive in 2010 (measured in thousands, using the same functional form as for our measure of ancestry), instrumenting as in equation (4). As expected, we obtain a positive and statistically significant coefficient on foreign born (the correlation between the two variables is 0.59). However, when we simultaneously include both endogenous variables in the specification, the coefficient on ancestry remains positive and statistically significant at the 1% level, whereas the coefficient on foreign born in 2010 is close to zero and insignificant in the OLS specification in column 3 and turns negative in the IV specification in column 4.[33] Because each foreign born also increases the number of individuals with foreign ancestry and the coefficient on foreign born is smaller than the one on ancestry in absolute terms, we may interpret this result as stating that foreign born have a positive effect on the probability of bilateral FDI, but their effect is smaller than that of their descendants. The results are similar for 2000 ancestry and immigration in column 7.

Using the number of foreign born in 1970 as a proxy for second-generation immigrants, columns 5 and 6 shows that, by contrast, the effect of second-generation immigrants is larger than for first-generation migrants, and it remains positive (albeit not significant) when we control for descendants of migrants with a foreign ancestry.[34] These results suggest first-generation immigrants have a smaller effect on FDI than their descendants and that the effect of ancestry on FDI fully develops only over long periods of time, consistent with the temporal pattern of reduced-form coefficients shown in Figure 1.

[33] The Kleinbergen-Paap statistic on the excluded instruments is 18.211, not rejecting the null that our instruments do not induce differential variation in the two endogenous variables with a p-value of 0.15.

[34] These results continue to hold when we drop migrations from Mexico (the largest origin country in recent decades) from the sample.

5.2 Heterogenous Effects

We next explore how the effect of ancestry on FDI varies across origins, destinations, and sectors.

Heterogeneous effect across origin countries. We begin by dropping the destination fixed effects from equation (1) and running 112 separate IV regressions of the FDI dummy on ancestry for each origin country that has at least one FDI link with the United States. The top part in Figure 7 plots the coefficients on ancestry against the reciprocal of the standard error, yielding a funnel plot, where the size of the circles is proportional to the share of each origin country in the total foreign ancestry of the United States. All coefficients to the northeast of the red line are statistically significant at the 5% level. Of the 112 coefficients, 84 are positive and statistically significant at this level. For easier readability, the plot excludes coefficient estimates larger than 1 and with a reciprocal of the estimated standard error exceeding 150 (the full set of results is in Appendix Table 12). The plot shows the estimates for larger origin countries tend to be more precise and clustered in a tight band. The coefficients for the five largest origin countries range from 0.171 (s.e.=0.011) for Mexico to 0.271 (s.e.=0.009) for the Britain. The plot also suggests some heterogeneity may exist in the size of the effect.

In panel A of Table 10, we explore the heterogeneous effect of ancestry on the extensive margin of FDI across countries. We use the simple IV specification of column 2 in Table 3 and add interactions of ancestry with measures of distance and various country characteristics (while of course always controlling and instrumenting for the main effect of ancestry). Columns 1-3 show the interaction of ancestry with the geographic distance and measures of genetic, linguistic, and religious distance between the United States and the origin country as defined by Spolaore and Wacziarg (2015). The results show a consistently positive and statistically significant effect on the interaction between ancestry and geographic distance. Once we account for this interaction, the interactions with the three other measures of distance are statistically insignificant, suggesting the effect of ancestry on FDI increases with geographic distance but not other measures of cultural distance. The remaining columns also show a consistently positive effect on the interaction between ancestry and judicial quality as defined by Nunn (2007), suggesting ancestry has a relatively larger effect on FDI when the origin country has good institutions.

Panel B shows broadly similar results for the intensive margin of FDI. The only qualitative difference is the positive and significant effect of the interaction with ethnic diversity in the origin country (as provided by Alesina et al. (2003)) once we control for the interactions with geographic distance and judicial quality, suggesting ancestry from a given origin has a larger effect at the intensive margin when the origin country is more ethnically diverse.

Heterogeneous effect across destination US counties. The middle part in Figure 7 shows coefficients and standard errors from 100 separate IV regressions of the FDI dummy on ancestry for the 100 US counties with the largest population in 2010. The circle sizes are proportional to the size of the total population of the county. The coefficient is positive and statistically significant for 99 of these 100 regressions. The coefficient varies from 0.073 (s.e.=0.034) for Kings county in New York to 0.488 (s.e.=0.063) for Orleans county in Louisiana.

Table 11 probes the heterogeneity of this effect in more detail by again reverting to our simple IV specification from column 2 in Table 3 and interacting our measure of ancestry with the share of the county's population that are of any foreign ancestry (column 1), the diversity of ancestries within the destination county, measured as 1 minus the Herfindhal index of ancestry shares (column 2), or both (column 3). Only the interaction with ethnic diversity is positive and significant: Ancestry has a stronger impact on FDI flows to and from US counties with more diverse ancestries, while the overall share of residents with foreign ancestry matters little. These results suggest US counties with more diverse populations may act as hubs for FDI, where the presence of descendants from a wide variety of origins enhances the effect of ancestry.

Heterogeneous effect across sectors. To explore the heterogeneity of the effect across sectors, we reconstruct our data set separately for all 20 2-digit NAICS sectors, considering in each case only FDI links sent and received by US firms in that sector. The bottom part in Figure 7 shows coefficients and standard errors from 20 separate IV regressions of the FDI dummy on ancestry for each sector, where the size of the circles is proportional to the total number of non-zero county-country FDI links in the sector (the full set of results is in Appendix Table 13). We find a statistically significant positive effect of ancestry on FDI in 18 of the 20 sectors. However, because the number of US firms in some of these sectors is less than 500, interpreting these results may be difficult. Panel A of Table 12 instead aggregates sectors into five groups with a more comparable number of firms. It shows the effect is largest in manufacturing (0.165, s.e.=0.024) and smallest in sectors dealing in natural resources (0.007, s.e.=0.003). The overall pattern of results appears consistent with the view that the effect of ancestry on FDI may be larger in sectors where production involves more differentiated inputs (Nunn (2007)), but we do not have detailed data on sufficiently many sectors to test this hypothesis formally.

Panel B of Table 12 presents the IV coefficient of ancestry on FDI separately for firms producing final consumption goods and firms producing intermediate inputs.[35] The impact of ancestry on FDI is larger for intermediates than for final goods. This suggests shared tastes do not play a major role in the impact of ancestry on the patterns of FDI.

Panel C of Table 12 presents the IV coefficients of ancestry on FDI separately for the subset of large versus small firms (with a number of employees above and below the median, respectively). We find the impact of ancestry on FDI is positive for both categories of firms, but it is significantly larger (about twice the size) for large than small firms (although this latter result is harder to interpret because firm size may well be endogenous to FDI).

5.3 Spillovers

In Table 13, we test for the presence of spillovers within states and between migrants from proximate origins. In column 1 of panel A, we use our simple specification from column 2 in Table 3, but add the total number of descendants of ancestry o at the state level. We are able to identify the effect of this spillover at the state level by aggregating our instruments from equation (4) to the state level and including them as a separate set of instruments in the specification, such that both endogenous variables are identified. The coefficient on our measure of ancestry at the state level is -0.020 (s.e.=0.010), suggesting a negative and significant spillover from a larger presence of descendants from the same origin in the state on the effect of ancestry on FDI at the county level. In column 2, we instead include (and instrument for) the number residents in the nearest adjacent county with ancestry from the same origin country, where we find a negative, albeit insignificant, effect.[36] Column 3 includes the number of descendants from origins within the same continent, where we find a positive but again insignificant effect. However, when we include in column 4 a measure for the number of descendants from the closest neighboring country, we find a negative and highly significant effect.

Overall, the evidence thus points to the presence of negative spillovers. The effect of ancestry on the extensive margin of FDI falls with the population of migrants from the same origin in the state as a whole, or with origin from neighboring countries. For example, a large Polish contingent in a given county has a lower effect on the probability of FDI with Poland if the state overall contains a large Polish contingent. Similarly, if the destination county also hosts a large number of descendants from a nearby origin, such as the Czech Republic, the Polish contingent

[35] To separate firms into final-goods producers and intermediate-goods producers, we use the upstreamness index from Antràs et al. (2012). A sector is labelled as final goods (intermediate input) if its upstreamness index is below (above) 2.

[36] We determine the nearest adjacent county (country) such that the average distance within adjacent county (country) pairs is minimised using a standard optimal non-bipartisan matching algorithm.

has a smaller marginal effect on FDI.

In panel B, we repeat the same estimation for the intensive margin of FDI, but appear to lack the statistical power to identify significant spillovers.

6 Conclusion

The economic effects of migration loom large in public debates about illegal immigration to the U.S. and the ongoing flow of migrants to Europe from places such as Syria, Afghanistan, and the Balkans. Much of the academic debate on the subject has focused on the relatively short-term consequences, identifying effects of immigration on local labor markets and consumer prices (Card, 1990; Cortes, 2008). In this paper, we add to this debate by showing causally identified evidence of an effect of migration, and the ethnic diversity resulting from it, on the propensity of firms based in the areas receiving migrants to interact economically with the migrants' origin countries. This effect of ancestry on FDI operates over long periods of time, spanning generations rather than decades, and explains an economically large share of the variation in patterns of FDI across US counties and states.

Our identification strategy uses 130 years of census data to isolate variation in today's ancestry composition of US counties that derives solely from the interaction of time-series variation in the relative attractiveness of different destinations within the United States with the staggered timing of factors that drove out-migration from the migrants' countries of origin. This approach allows us to generate four main insights.

First, we are able to causally identify and quantify the effect of ancestry on FDI in a setting with a high degree of external validity while guarding against a wide range of possible confounding factors, including unobserved origin and destination effects. We find that a doubling of a US county's residents with ancestry from a given foreign country relative to the mean increases by 4 percentage points the probability that at least one local firm engages in FDI with that country.

Second, the presence of descendants of first-generation immigrants rather than first-generation immigrants themselves generates the majority of the effect of ancestry on FDI. The effect of ancestry on FDI is thus long lasting and appears to unfold over generations rather than years, where even the earliest migrations for which we have data going back to the 19th century significantly affect the pattern of FDI today.

Third, the effect of ancestry on FDI increases with the geographic distance to the origin country and the quality of its institutions. Once these factors are controlled for, other measures of genetic, linguistic, and religious distance do not significantly affect the size of the effect of ancestry on FDI. The effect also does not appear to vary systematically between firms producing final goods versus intermediate inputs.

Fourth, we find a range of results that show a positive effect of ethnic diversity on FDI. The most obvious of these findings is the strong indication of concavity in the number of descendants of migrants from a given origin, such that a more ethnically diverse population, combining many smaller communities from different origins, should generate more FDI than one large community of foreign descent. Further, we find negative spillovers both within states and between migrants from geographically proximate countries, such that a larger community of the same ethnic descent in surrounding counties or a larger community of descent from a neighboring country decreases the effect of ancestry on FDI. In addition, the effect of ancestry on FDI significantly increases with the diversity of the community of residents with foreign ancestry. All three findings taken together suggest ethnic diversity may be a quantitatively important driver of FDI.

Taken together, our results suggest that receiving migration from a foreign country has a positive long-term effect on the ability of local firms to interact economically with the migrants' country of origin. This effect increases with the institutional quality of the origin country, suggesting that, for example, receiving migrants from a war-torn country may have larger positive effects once the country stabilizes. The collage of our results also appears more consistent with a model in which common ancestry mitigates informational frictions, but does not operate though contract enforcement or common tastes. In the presence of informational frictions, common ancestry may act as a conduit for transmitting information. This information channel is more relevant for remote countries (the positive interaction with distance). Moreover, information transmission tends to follow the shortest path, so that a small increment in the number of residents with common ancestry matters more when few residents have information about a country than when many do (the concave impact of ancestry and the negative spillover effects). On the other hand, common ancestry has a smaller impact on FDI in weak judicial environments (the positive interaction with judicial quality), and thus does not appear to predominantly act as a substitute for contract enforcement. Finally, the fact that the effect of ancestry on FDI is weaker for firms producing final goods than intermediate inputs appears to exclude mechanisms that rely on common tastes between descendants of migrants and their countries of origin. However, we note the caveat that all of our evidence on the mechanism through which ancestry facilitates FDI is indirect and should thus be interpreted with caution. We leave a thorough study of this mechanism for future research.

References

Ager, P. and M. Brückner (2013): "Cultural divsersity and economic growth: Evidence from the US during the age of mass migration," *European Economic Review*, 64, 76–97.

Aleksynska, M. and G. Peri (2014): "Isolating the Network Effect of Immigrants on Trade," *The World Economy*, 37, 434–45.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8, 155–194.

Alesina, A., J. Harnoss, and H. Rapoport (2015a): "Birthplace Diversity and Economic Prosperity," Working paper.

Alesina, A., S. Michalopoulos, and E. Papaioannou (2015b): "Ethnic Inequality," *Journal of Political Economy*, forthcoming.

Antràs, P., D. Chor, T. Fally, and R. Hillberry (2012): "Measuring the Upstreamness of Production and Trade Flows," *American Economic Review Papers and Proceedings*, 102, 412–416.

Arkolakis, C., A. Costinot, and A. Rodríguez-Clare (2012): "New Trade Models, Same Old Gains?" *American Economic Review*, 102, 94–130.

Ashraf, Q. and O. Galor (2013): "The "Out of Africa" Hypothesis Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, 103, 1–46.

Bartik, T. J. (1991): *Who benefits from state and local economic development policies?*, no. wbsle in Books from Upjohn Press, W.E. Upjohn Institute for Employment Research.

Bhattacharya, U. and P. Groznik (2008): "Melting Pot or Salad Bowl: Some Evidence from US Investments Abroad," *Journal of Financial Markets*, 11, 228–258.

Borjas, G. J. (1994): "The Economics of Immigration," *Journal of Economic Literature*, XXXII, 1667–1717.

——— (2003): "The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market," *The Quarterly Journal of Economics*, 118, 1335–1374.

Burchardi, K. B. and T. A. Hassan (2013): "The Economic Impact of Social Ties: Evidence from German Reunification," *Quarterly Journal of Economics*, 128, 1219–1271.

Card, D. (1990): "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245–257.

——— (2001): "Immigrant Inflows, Native outflows, and the Local Labor Market Impacts of Higher Immigration," *Journal of Labor Economics*, 19, 22–64.

Card, D. and J. Di Nardo (2000): "Do Immigrant Inflows Lead to Native Outflows?" *American Economic Review*, 90, 360–367.

Carr, D. L., J. R. Markusen, and K. E. Maskus (2001): "Estimating the Knowledge-Capital Model of the Multinational Enterprise," *American Economic Review*, 91, 693–708.

Chaney, T. (2014): "The network structure of international trade," *The American Economic Review*, 104, 3600–3634.

——— (2016): "Networks in International Trade," in *Oxford Handbook of the Economics of Networks*, ed. by Y. Bramoulle, A. Galleoti, and B. Rogers, Oxford University Press.

Cohen, L., U. Gurun, and C. Malloy (2015): "Resident Networks and Firm Value," *The Journal of Finance*, forthcoming.

Combes, P., M. Lafourcade, and T. Mayer (2005): "The trade-creating effects of business and social networks: Evidence from France." *Journal of International Economics*, 66 (1), 1–29.

Cortes, P. (2008): "The Effect of Low-Skilled Immigration on U.S. Prices: Evidence from CPI

Data," *Journal of Political Economy*, 116, pp. 381–422.

Daniels, R. (2002): *Coming to America*, HarperCollins Publishers.

Duncan, B. and S. J. Trejo (2016): "The Complexity of Immigrant Generations: Implications for Assessing the Socioeconomic Integration of Hispanics and Asians," Working Paper 21982, National Bureau of Economic Research.

Friedberg, R. (2001): "The impact of mass migration on the Israeli labor market." *Quarterly Journal of Economics*, 116 (4), 1373–1408.

Fuchs-Schündeln, N. and T. A. Hassan (2015): "Natural Experiments in Macroeconomics," Working Paper 21228, National Bureau of Economic Research.

Fulford, S. L., I. Petkov, and F. Schiantarelli (2015): "Does It Matter Where You Came From? Ancestry Composition and Economic Performance of U.S. Counties, 1850-2010," Institute for the Study of Labor (IZA) Discussion Paper No. 9060.

Garmendia, A., C. Llano, A. Minondo, and F. Requena (2012): "Networks and the disappearance of the intranational home bias," *Economics Letters*, 116, 178–182.

Goldin, C. (1994): "The Political Economy of Immigration Restriction in the United States, 1890 to 1921," in *The Regulated Economy: A Historical Approach to Political Economy*, ed. by C. Goldin and G. D. Libecap, University of Chicago Press, 223–258.

Gould, D. M. (1994): "Immigrant links to the home country: Empirical implications for U.S. bilateral trade flows." *The Review of Economics and Statistics*, 76, 302–316.

Greene, W., C. Han, and P. Schmidt (2002): "The bias of the fixed effects estimator in nonlinear models," Unpublished Manuscript, Stern School of Business, NYU.

Guiso, L., P. Sapienza, and L. Zingales (2009): "Cultural biases in economic exchange." *The Quarterly Journal of Economics*, 124, 1095–1131.

Head, K. and J. Ries (1998): "Immigration and Trade Creation: Econometric Evidence from Canada," *Canadian Journal of Economics*, 31, 47–62.

——— (2008): "FDI as an Outcome of the Market for Corporate Control: Theory and Evidence," *Journal of International Economics*, 74, 2–20.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating Trade Flows: Trading Partners and Trading Volumes," *Quarterly Journal of Economics*, 123, 441–487.

Holmes, T. J., E. R. McGrattan, and E. C. Prescott (2015): "Quid Pro Quo: Technology Capital Transfers for Market Access in China," *The Review of Economic Studies*, 82, 1154–1193.

Jensen, E. B., R. Bhaskar, and M. Scopilliti (2015): "Demographic Analysis 2010: Estimates of Coverage of the Foreign-Born Population in the American Community Survey," Tech. rep., U.S. Census.

Juhász, R. (2014): "Temporary Protection and Technology Adoption: Evidence from the Napoleonic Blockade," .

Katz, L. F. and K. M. Murphy (1992): "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *Quarterly Journal of Economics*, 107, 35–78.

Kaufmann, D., A. Kraay, and M. Mastruzzi (2003): "Governance Matters III: Governance Indicators for 1996–2002," Working Paper No. 3106, World Bank.

McGrattan, E. R. and E. C. Prescott (2010): "Technology Capital and the US Current Account," *American Economic Review*, 100, 1493–1522.

Nunn, N. (2007): "Relationship-Specificity, Incomplete Contracts, and the Pattern of Trade," *Quarterly Journal of Economics*, 122, 569–600.

Nunn, N., N. Qian, and S. Sequeira (2015): "Migrants and the Making of America," *Working Paper*.

Ottaviano, G. I. and G. Peri (2006): "The economic value of cultural diversity: evidence from US cities," *Journal of Economic Geography*, 6, 9–44.

Parsons, C. and P.-L. Vezina (2014): "Migrant Networks and Trade: The Vietnamese Boat People as a Natural Experiment," Mimeo University of Oxford.

Peri, G. (2012): "The Effect of Immigration on Productivity: Evidence from U.S. States," *The Review of Economics and Statistics*, 94, 348–358.

Portes, R. and H. Rey (2005): "The Determinants of Cross-Border Equity Flows," *Journal of International Economics*, 65, 269–296.

Putterman, L. and D. N. Weil (2010): "Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality," *The Quarterly Journal of Economics*, 125, 1627–1682.

Ramondo, N. (2014): "A Quantitative Approach to Multinational Production," *Journal of International Economics*, 93, 108–122.

Rauch, J. and V. Trindade (2002): "Ethnic Chinese Networks In International Trade," *The Review of Economics and Statistics*, 84, 116–130.

Razin, A., Y. Rubinstein, and E. Sadka (2003): "Which countries export FDI, and how much?" Tech. rep., National Bureau of Economic Research.

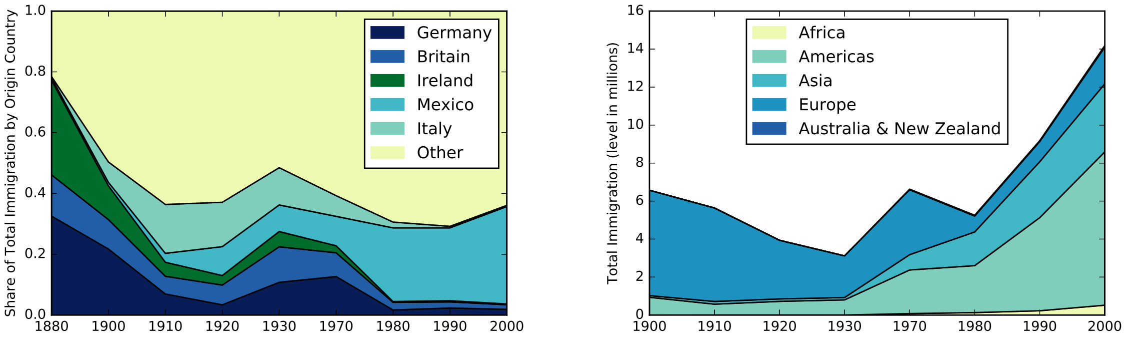
Redding, S. and D. Sturm (2008): "The Cost of Remoteness: Evidence from German Division and Reunification," *The American Economic Review*, 98, 1766–1797.

Spolaore, E. and R. Wacziarg (2015): "War and Relatedness," *The Review of Economics and Statistics*, forthcoming.

Steinwender, C. (2014): "Information Frictions and the Law of One Price: "When the States and the Kingdom became United"," Working Papers 190, Oesterreichische Nationalbank (Austrian Central Bank).

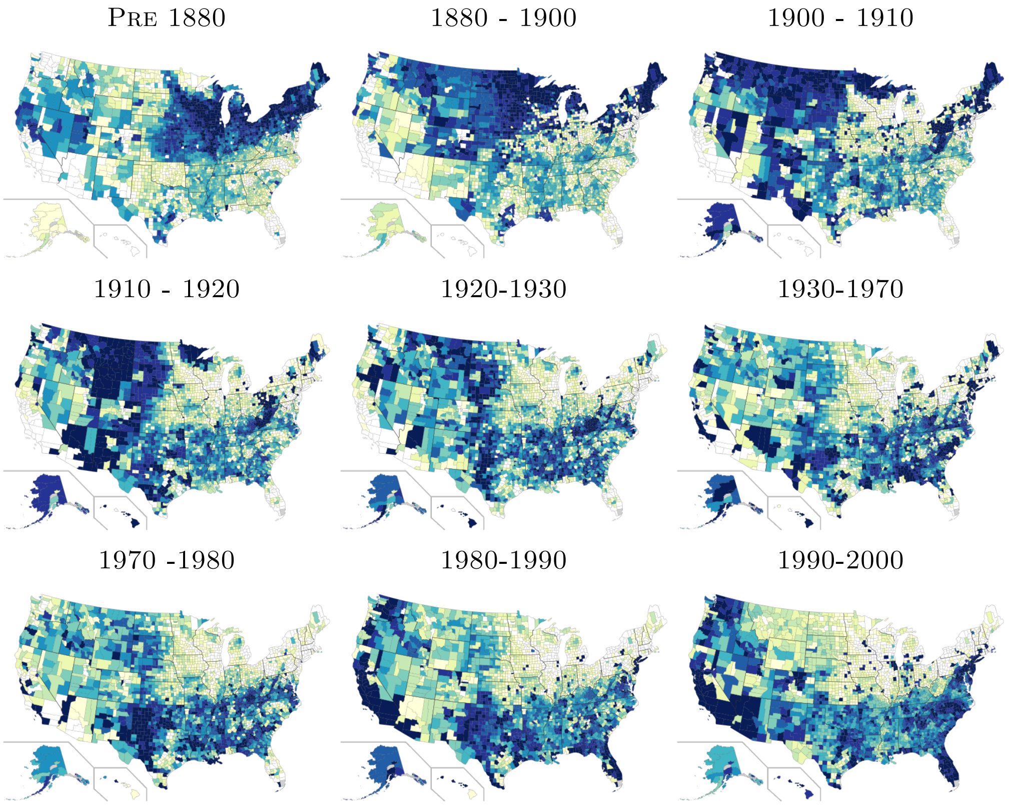
Thernstrom, S. (1980): *Havard encyclopedia of American ethnic groups*, Harvard University Press, Cambridge MA.

Figure 1: Origins of Immigrants to the United States, pre-1880 to 2000



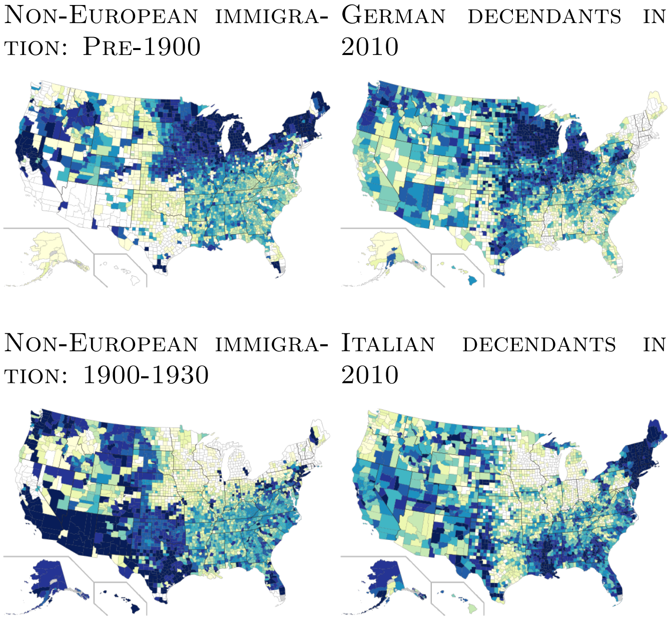
Notes: The left side depicts the share of total immigration to the United States in each census period for the largest five origin countries of US residents that claim foreign ancestry in the 2010 census: Germany, Britain, Ireland, Mexico, and Italy. The right side shows the the number of migrants (in millions) by continent of origin. See section 1 of the main text and appendix A.1 for details.

Figure 2: Destinations of Immigrants to the United States, pre-1880 to 2000



Notes: This figure maps immigration flows into the US counties by census period. We regress the number of immigrants into US county d at time t, I_d^t, on destination county d and year t fixed effects, and calculate the residuals. The maps' color coding depicts the residuals' decile in the distribution of residuals across counties and within census periods. Darker colors indicate a higher decile.

Figure 3: Migrants and Ancestors: The Cases of Germany Pre-1900 and Italy 1900-1910


Non-European immigration: Pre-1900

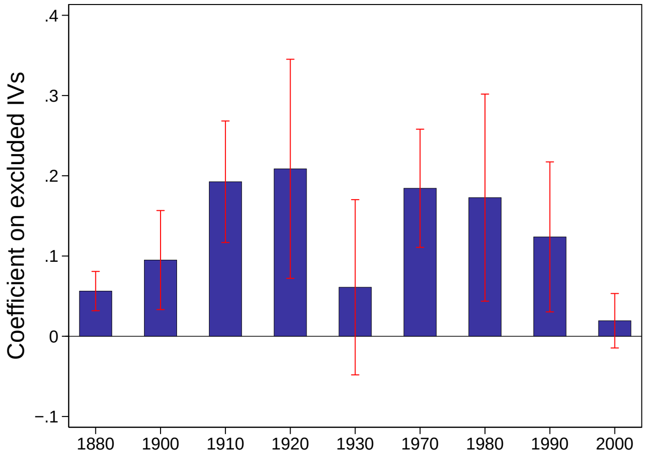
German decendants in 2010

Non-European immigration: 1900-1930

Italian decendants in 2010

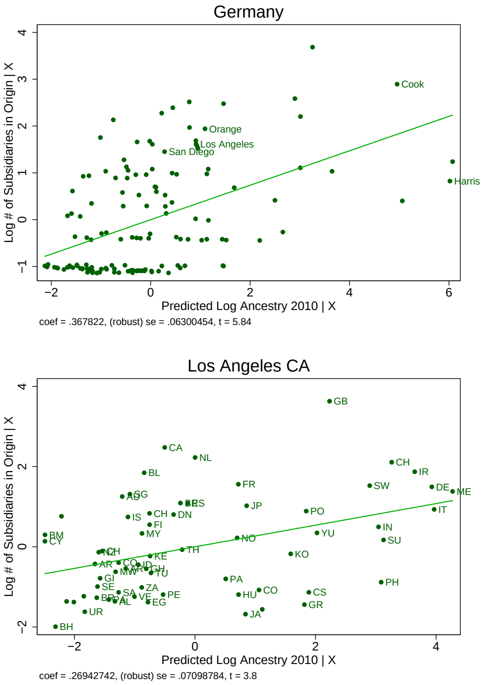
Notes: This figure contrasts Italian and German ancestry in 2010 (right panels), and non-European immigration patterns in pre-1900 and 1910-1930 (left panel). The left two panels are created as the maps in Figure 2, restricted to non-European immigration and the two periods we consider (pre-1900 and 1910-1930). The right two panels plot the county level residuals from a regression of log ancestry in 2010 on county, Italy and Germany fixed effects on the sample of European countries. The maps' color coding depicts the residuals' decile in the distribution of residuals across counties. Darker colors indicate a higher decile.

Figure 4: First-Stage Coefficients



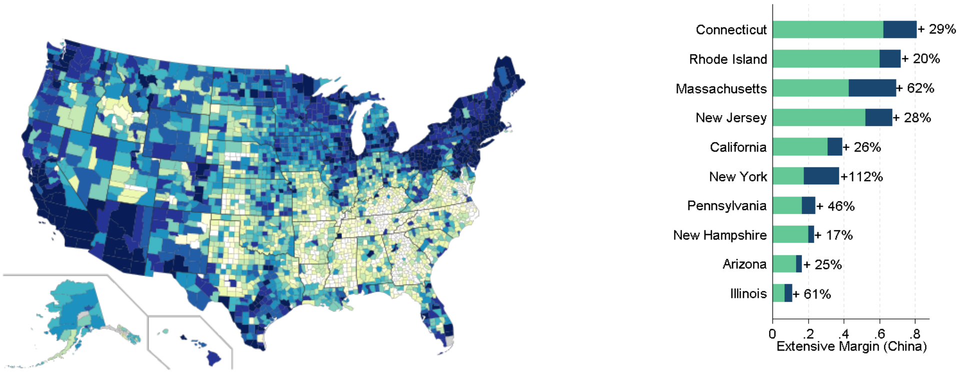
Notes: Coefficient estimates (bars) and 95% confidence intervals (lines) on the excluded instruments $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ from Table 2, column 2. The dependent variable is Log Ancestry 2010. Robust standard errors are clustered at the origin country level.

Figure 5: Ancestry and FDI: Germany and Britain; Los Angeles and Cook counties



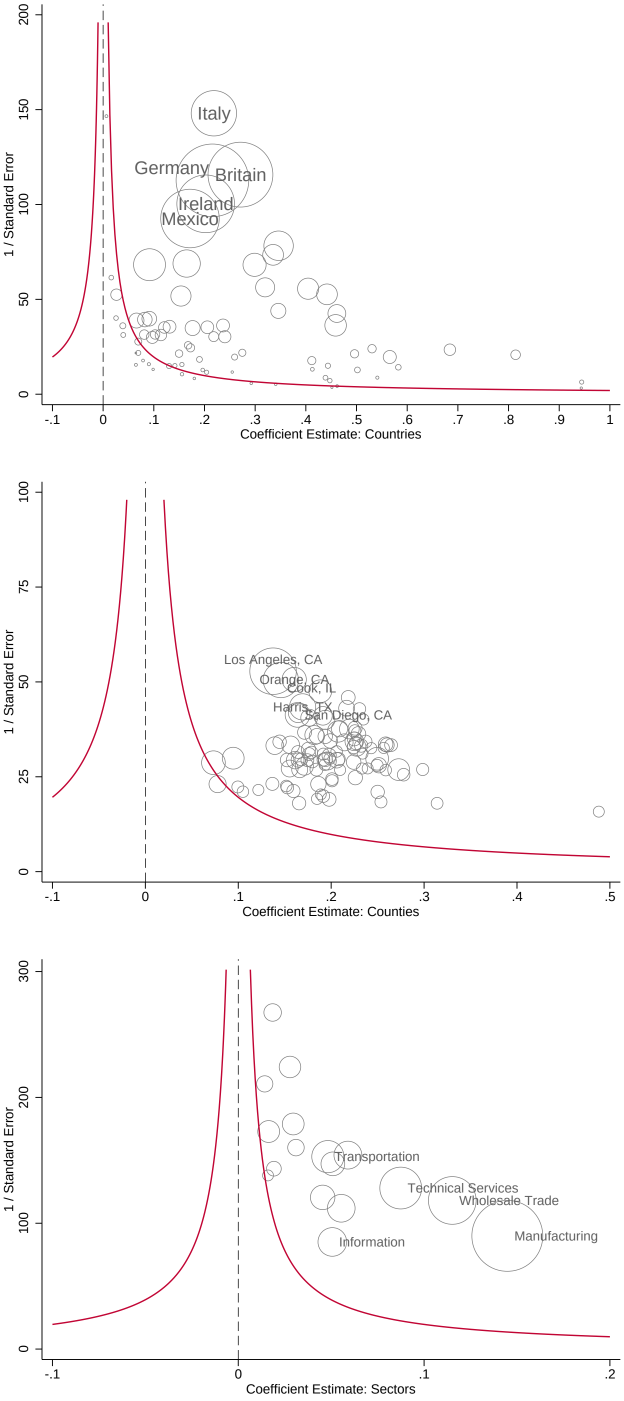
Notes: The figure shows conditional scatterplots from regressions corresponding to column 3 of Table 6, but restricting the sample to one origin country only (top parts: Germany and Britain) or one destination county only (bottom parts: LA and Cook counties). The solid line depicts the fitted regression line. In each case, the regression controls for distance and latitude difference. The upper-left panel shows a conditional scatterplot from a regression of log # of subsidiaries in Germany for firms in each US county on predicted log 2010 German ancestry. The upper-right panel shows a similar scatterplot for British ancestry and the log # of subsidiaries in Britain. The bottom-left panel shows a conditional scatterplot from the corresponding regression for LA county, CA and the log # of subsidiaries of LA based firms in each origin country. The bottom-right part shows the same plot for Cook county, IL.

Figure 6: Thought Experiment: Removing the Chinese Exclusion Act



Notes: The map on the left depicts for each US county the predicted increase in the probability of having positive FDI relations with China in a counterfactual world where the "Chinese Exclusion" Act of 1882 had never been passed, that is, if Chinese immigration to the United States had not been discriminated against between 1882 and 1965. Darker colors indicate larger increases. The bar graph on the right shows the fraction of counties within each state with FDI relations with China (light color) and the predicted increment in the fraction of counties with FDI relations with China (dark color), which we calculate as the unweighted average of $d\Pr\left[FDI_{China,d} > 0\right]$ across counties in a given state. We also provide the size of this increase relative to the actual fraction in percentage terms. The histogram only depicts the ten US states with the largest change. Interpretation: If Chinese immigration to the United States had been free, the fraction of counties in Massachusetts with FDI relations to China would have increased from 43% to 69%, a 62% increase. The details of this calculation are section 4.4.

Figure 7: Heterogeneous Estimates across Countries, Counties, and Sectors



Notes: This figure shows funnel plots of the estimated coefficients and standard errors from separate IV regressions of the FDI dummy on Log 2010 Ancestry for each origin country (top), destination US counties (middle), and sectors (bottom). In all regressions, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$ and principal components as excluded instruments, and control for log distance as well as latitude difference. We plot the estimated coefficients (x axis) against the reciprocal of estimated standard errors on ancestry. The size of the circle is proportional to the size of country ancestry (top), the size of county population (middle), and the size of the sector (bottom). The imposed curve is $y = 1.96/x$ for positive x region and $y = -1.96/x$ for negative x region. Circles above the curve indicate statistically significant coefficients. See section 5.2 for details.

Table 1: Summary Statistics

	All	*Ancestry > 0*		
		All	*Bottom Quintile*	*Top Quintile*
	(1)	(2)	(3)	(4)
Panel A: Origin-destination pairs				
FDI Dummy	0.018	0.031	0.003	0.127
	(0.132)	(0.173)	(0.052)	(0.333)
Ancestry 2010 (in thousands)	0.316	0.575	0.000	2.852
	(5.962)	(8.036)	(0.000)	(17.790)
Immigrants between 1990-2000 (in thousands)	0.023	0.042	0.000	0.199
	(1.070)	(1.443)	(0.001)	(3.221)
Immigrants between 2000-2010 (in thousands)	0.020	0.036	0.000	0.173
	(0.665)	(0.898)	(0.002)	(1.999)
Foreign-born 2010 (in thousands)	0.069	0.125	0.000	0.594
	(2.749)	(3.708)	(0.004)	(8.267)
Geographic Distance (km)	9,122.393	8,397.379	9,142.553	7,463.619
	(3,802.105)	(3,763.718)	(4,299.572)	(2,986.233)
Latitude Difference (degree)	19.440	16.319	18.915	13.750
	(11.312)	(10.902)	(11.388)	(8.807)
# of FDI Relationships	0.196	0.352	0.028	1.622
	(5.490)	(7.401)	(1.461)	(16.307)
# of Subsidiaries in Origin	0.033	0.060	0.003	0.270
	(1.345)	(1.813)	(0.281)	(3.844)
# of Parents in Destination	0.015	0.027	0.001	0.125
	(0.407)	(0.548)	(0.103)	(1.201)
# of Workers Employed at Subsidiary in Origin (in thousands)	0.039	0.069	0.010	0.319
	(4.941)	(6.661)	(1.298)	(14.750)
# of Subsidiaries in Destination	0.068	0.122	0.011	0.562
	(1.903)	(2.565)	(0.546)	(5.667)
# of Parents in Origin	0.079	0.143	0.012	0.664
	(2.282)	(3.077)	(0.580)	(6.811)
# of Workers Employed at Subsidiary in Destination (in thousands)	1.873	3.398	0.087	16.512
	(86.649)	(116.896)	(5.658)	(260.739)
N	612,495	336,380	67,276	67,267
Panel B: Countries				
Genetic Distance	0.103	0.084	0.106	0.066
	(0.053)	(0.041)	(0.050)	(0.036)
N	155	119	18	25
Linguistic Distance	0.950	0.937	0.990	0.920
	(0.110)	(0.121)	(0.010)	(0.114)
N	132	103	8	26
Religious Distance	0.820	0.807	0.923	0.732
	(0.129)	(0.137)	(0.050)	(0.128)
N	131	101	8	25
Judicial Quality	0.503	0.537	0.546	0.661
	(0.208)	(0.214)	(0.224)	(0.202)
N	144	115	15	26
2010 Country Diversity	0.442	0.405	0.433	0.239
	(0.269)	(0.256)	(0.246)	(0.197)
N	162	122	20	27
Panel C: Counties				
2010 Share of Population with Foreign Ancestry	0.577	0.577	0.560	0.648
	(0.188)	(0.187)	(0.223)	(0.137)
2010 Diversity of Ancestries	0.790	0.789	0.764	0.838
	(0.075)	(0.075)	(0.071)	(0.077)
N	3,141	3,137	628	627

Notes: The table presents means (and standard deviations). Variables in Panel A refer to our sample of (country-county) pairs used in Tables 2, 3, 4, 5, 6, 9, 13, and Appendix Table 9. Variables in Panel B refer to our sample of counties used in Table 10. Variables in Panel C refer to our sample of counties used in Table 11. Column 1 shows data for all observations. Columns 2 to 4 show all, the bottom quintile, and the top quintile of observations with positive ancestry, respectively. In Panel A, the FDI dummy is a dummy variable equal to 1 if the destination county has either subsidiaries or shareholders in the origin country. The details of variables in Panel B are given in the Data Appendix. The ancestry-diversity variable is computed as 1 minus the Herfindahl index of ancestry group shares in each county.

Table 2: First-stage: The Effect of Historical Migrations on Ancestry

	Log Ancestry 2010								*Ancestry 2010*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$	0.057***	0.056***	0.089***	0.068***	0.056***	0.056***	0.056***		2.145***
	(0.013)	(0.012)	(0.010)	(0.009)	(0.007)	(0.007)	(0.008)		(0.290)
$I^{1900}_{o,-r(d)} \frac{I^{1900}_{-c(o),d}}{I^{1900}_{-c(o)}}$	0.097***	0.095***	0.162***	0.115***	0.099***	0.099***	0.100***	0.159***	3.634**
	(0.032)	(0.031)	(0.036)	(0.035)	(0.033)	(0.033)	(0.036)	(0.027)	(1.425)
$I^{1910}_{o,-r(d)} \frac{I^{1910}_{-c(o),d}}{I^{1910}_{-c(o)}}$	0.192***	0.193***	0.205***	0.159***	0.137***	0.137***	0.132**	0.090**	5.646**
	(0.039)	(0.039)	(0.036)	(0.038)	(0.050)	(0.050)	(0.054)	(0.040)	(2.345)
$I^{1920}_{o,-r(d)} \frac{I^{1920}_{-c(o),d}}{I^{1920}_{-c(o)}}$	0.205***	0.209***	0.296***	0.261***	0.283***	0.283***	0.249***	0.319***	14.726***
	(0.070)	(0.070)	(0.063)	(0.057)	(0.045)	(0.045)	(0.047)	(0.064)	(3.012)
$I^{1930}_{o,-r(d)} \frac{I^{1930}_{-c(o),d}}{I^{1930}_{-c(o)}}$	0.062	0.061	0.074	0.071	0.079	0.079	0.065*	0.097	11.812***
	(0.056)	(0.056)	(0.059)	(0.059)	(0.051)	(0.051)	(0.034)	(0.074)	(2.855)
$I^{1970}_{o,-r(d)} \frac{I^{1970}_{-c(o),d}}{I^{1970}_{-c(o)}}$	0.183***	0.184***	0.180***	0.158***	0.149***	0.148***	0.150***	0.190***	6.256***
	(0.038)	(0.038)	(0.033)	(0.032)	(0.028)	(0.029)	(0.027)	(0.033)	(0.669)
$I^{1980}_{o,-r(d)} \frac{I^{1980}_{-c(o),d}}{I^{1980}_{-c(o)}}$	0.173***	0.173***	0.223***	0.216***	0.214***	0.214***	0.205**	0.331***	18.694***
	(0.066)	(0.066)	(0.083)	(0.081)	(0.076)	(0.077)	(0.080)	(0.122)	(2.390)
$I^{1990}_{o,-r(d)} \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$	0.123***	0.124***	0.117***	0.117***	0.101**	0.101**	0.115**	0.127**	10.786***
	(0.048)	(0.048)	(0.044)	(0.045)	(0.045)	(0.045)	(0.048)	(0.063)	(3.675)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$	0.020	0.019	0.037*	0.039**	0.046***	0.046***	0.039**	0.087**	5.194***
	(0.017)	(0.017)	(0.021)	(0.017)	(0.017)	(0.017)	(0.016)	(0.042)	(1.148)
$I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$							0.317***		
							(0.089)		
R^2	0.56	0.56	0.59	0.68	0.73	0.73	0.73	0.73	0.49
F Stat on excluded IVs	10.604	10.954	2447.187	340.940	161.729	157.672	194.863	99.464	907.720
p-value on F Stat	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Distance	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude Difference	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	No	No	Yes	No	No	No

Notes: The table presents coefficient estimates of our first stage equation (4) at the country-county level. All specifications control for origin and destination fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. In columns 1-8 the dependent variable is the log of 1 plus the number of residents of the county in 2010 that report having ancestors in the origin country, measured in thousands *(Log Ancestry 2010)*. In column 9 the dependent variable is the level of ancestry in 2010 (again in thousands). The excluded instruments are, for each census period, interactions of pull and push factors in migration, $I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})$, where $I^t_{o,-r(d)}$ stands for the number of migrants from o who settle in destinations *not* in the same census region as d in period t and $I^t_{-c(o),d}/I^t_{-c(o)}$ for the fraction of migrants *not* coming from origins in the same continent as o who settle in county d. Columns 3-9 also include the first five principal components of higher-order interactions of these factors. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 3: Second-stage: The effect of Ancestry on FDI

Panel A: IV	*FDI 2014 (Dummy)*					
	(1)	(2)	(3)	(4)	(5)	(6)
Log Ancestry 2010	0.231***	0.190***	0.187***	0.187***	0.198***	0.191***
	(0.023)	(0.024)	(0.024)	(0.024)	(0.023)	(0.024)
Log Distance	0.007	0.004	0.024	0.009	0.026	-0.027
	(0.010)	(0.009)	(0.029)	(0.033)	(0.030)	(0.027)
Latitude Difference	0.006**	0.005**	0.006*	-0.000	0.006*	0.003
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)
N	612495	612495	612495	612495	612495	612300
Panel B: OLS	*FDI 2014 (Dummy)*					
Log Ancestry 2010	0.173***	0.173***	0.149***	0.149***	0.149***	0.161***
	(0.016)	(0.016)	(0.018)	(0.018)	(0.018)	(0.019)
R^2	0.2967	0.2967	0.3635	0.3635	0.3635	0.3930
N	612495	612495	612495	612495	612495	612495
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from IV (Panel A) and OLS (Panel B) regressions of equation (1) at the country-county level. The dependent variable in all panels is a dummy indicating an FDI relationship between origin o and destination d in 2014. The main variable of interest is *Log Ancestry 2010*, instrumented using various specifications of equation (4). In all columns in Panel A, we include $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,...,2000}$ as excluded instruments. Columns 2-6 also include the first five principal components of the higher-order interactions of push and pull factors as instruments. Column 5 also includes the interaction of the push and pull factor constructed using data from the 2006-2010 American Community Survey. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. (We also run an IV probit regression using the specification in column 2 yielding a marginal effect evaluated at the mean of *Log Ancestry 2010* on FDI equal to 0.104***(0.037).)

Table 4: The Effect of Ancestry on FDI: The Case of Communist Countries

	FDI 2014 (Dummy)				
	(1)	(2)	(3)	(4)	(5)
Log Ancestry 2010	0.197***	0.380***	0.075	0.242**	0.234**
	(0.066)	(0.053)	(0.054)	(0.104)	(0.098)
N	3141	3141	3141	18846	28269
F Stat on excluded IVs	1.877	5.874	10.537	4.825	3.545
Destination FE	No	No	No	No	Yes
Countries considered	Soviet Union	China	Vietnam	Eastern Europe	All communist countries
Years excluded	1918-1990	1949-1980	1975-1996	1945-1989	

Notes: The table presents coefficient estimates from IV regressions of equation (1) at the country-county level. Each column uses data from a subset of origin countries: Soviet Union (column 1), China (column 2), Vietnam (column 3), as well as Albania, Bulgaria, Czechoslovakia, Hungary, Poland, and Romania (column 4). The dependent variable in all columns is a dummy indicating an FDI relationship between origin country o and destination county d in 2014. All specifications use the same set of instruments as the one in column 3 of Table 3, but only exclude the interaction terms containing measures of pull and push factors in migrations that occur during the years of economic isolation from the United States indicated above; the remaining variables are included as controls. All specifications control for log distance, latitude difference, and origin fixed effects. Standard errors are given in parentheses and are robust. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 5: The Effect of Ancestry on Immigration

	Immigration 1990-2000	*Log immigration 1990-2000*	*Log immigration 1980-1990*
	(1)	(2)	(3)
Log Ancestry 1990	9.662**	0.556***	
	(4.455)	(0.075)	
Log Ancestry 1980			0.447***
			(0.076)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$	1.082***	0.033**	
	(0.358)	(0.015)	
$I^{1990}_{o,-r(d)} \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$			0.061***
			(0.015)
N	612,495	612,495	612,495
F-stat on excluded IVs	8.899	8.899	11.407

Notes: The table presents the coefficient estimates from IV regressions of equation (6) at the country-county level. The dependent variable is the immigration flow from 1990 to 2000 in columns 1-2 and the immigration flow from 1980 to 1990 in column 3. In all columns, we instrument for *Log Ancestry* with the double-interactions of pull and push factors from prior censuses, $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,1980}$. All specifications control for log distance, latitude difference, origin-region, and destination-continent fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 6: The Effect of Ancestry on the Intensive Margin of FDI

	OLS	IV	IV	IV
	Log Total # of FDI relationships			
Panel A	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.245***	0.373***	0.288***	0.146***
	(0.051)	(0.041)	(0.019)	(0.028)
N	10851	10851	10851	10851
	Log # of subsidiaries in destination			
Panel B	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.275***	0.351***	0.279***	0.242***
	(0.053)	(0.047)	(0.014)	(0.041)
N	9082	9082	9082	9082
	Log # of parents in origin			
Panel C	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.267***	0.302***	0.279***	0.238***
	(0.056)	(0.064)	(0.014)	(0.041)
N	9082	9082	9082	9082
	Log # of workers employed at subsidiaries in destination			
Panel D	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.597***	1.168***	0.439***	0.371
	(0.146)	(0.244)	(0.094)	(0.237)
N	9082	9082	9082	9082
	Log # of subsidiaries in origin			
Panel E	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.098**	0.401***	0.187***	-0.055
	(0.041)	(0.045)	(0.019)	(0.048)
N	4065	4065	4065	4065
	Log # of parents in destination			
Panel F	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.119***	0.421***	0.185***	-0.052*
	(0.036)	(0.031)	(0.016)	(0.031)
N	4065	4065	4065	4065
	Log # of workers employed at subsidiaries in origin			
Panel G	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.200	0.903***	0.351***	-0.211
	(0.152)	(0.090)	(0.055)	(0.182)
N	4065	4065	4065	4065
Destination FE	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes
Destination × Continent FE	Yes	Yes	No	No
Origin × Census Region FE	Yes	Yes	No	No
Heckman Correction	No	No	No	Yes

Notes: The table presents the OLS (column1) and IV (columns 2-4) estimates of equation (7). The dependent variables are specified for each panel in the table. The main variable of interest is *Log Ancestry 2010.* All IV columns use as instruments the same set of variables as column 3 of Table 3. All specifications control for log distance, latitude difference, origin, and destination fixed effects. The coefficient estimates on these specifications are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 7: The Effect of Ancestry on the Intensive Margin of Trade (State Level)

	OLS	IV	IV	IV
	Log Total # of FDI relationships			
Panel A	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.181***	0.978***	0.275***	0.089*
	(0.036)	(0.071)	(0.076)	(0.048)
R^2	0.8481	0.6624	0.8467	0.8521
N	2384	2384	2384	2216
F Stat on excluded IVs		4471.146	87.376	44.860
	Log Aggregate Export			
Panel B	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.084**	1.033***	-0.186***	-0.170*
	(0.033)	(0.080)	(0.047)	(0.091)
R^2	0.8372	0.7012	0.8351	0.6703
N	7902	7902	7904	4762
F Stat on excluded IVs		4747.006	280.702	243.618
	Log Aggregate Import			
Panel C	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.308***	1.335***	-0.269***	-0.021
	(0.051)	(0.100)	(0.081)	(0.155)
R^2	0.7749	0.6746	0.7673	0.5789
N	6210	6210	6210	3821
F Stat on excluded IVs		6349.321	266.488	28.722
	Log Exports to Vietnam			
Panel D	(1)	(2)		
Log Ancestry 2010	1.169***	1.230***		
	(0.124)	(0.124)		
R^2	0.6799	0.6783		
N	51	51		
F Stat on excluded IVs		5.938		
	Log Exports to Japan			
Panel E	(1)	(2)		
Log Ancestry 2010	0.898***	1.107***		
	(0.197)	(0.128)		
R^2	0.4417	0.4188		
N	51	51		
F Stat on excluded IVs		67.220		
Origin FE	Yes	Yes	Yes	Yes
Destination FE	Yes	No	Yes	Yes
Heckman Correction	No	No	No	Yes

Notes: The table presents the OLS (column 1) and IV (columns 2-4) estimates of equation (7) at the state level for FDI and trade. The main variable of interest is *Log Ancestry 2010*. The dependent variables are the log number of total FDI links in 2014, the log of aggregate exports (from the US state), aggregate imports, exports to Vietnam, and exports to Japan in panels A, B, C, D, and E, respectively. Exports and imports are measured in US dollars in 2011. In all columns, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as excluded instruments. All specifications control for log distance, latitude difference, and origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 8: Thought Experiment: A Gold Rush in Los Angeles in 1880

	Ancestry 2010	FDI #	*Predicted Counterfactual Change* *FDI # (in %, IV)*	 *Ancestry 2010*
	(1)	(2)	(3)	(4)
Germany	343,276	241	+62.70	+65,344
Ireland	256,621	40	+58.34	+61,701
UK	396,439	582	+21.95	+26,645
Norway	39,515	55	+3.53	+4,657
Sweden	51,395	71	+3.03	+4,010
France	77,372	278	+2.48	+3,293
Canada	27,722	531	+2.36	+3,132
Switzerland	10,156	162	+1.85	+2,456
Czechoslovakia	17,905	4	+1.61	+2,140
Netherlands	38,392	121	+1.23	+1,638

Notes: The table presents the number of individuals of selected ancestries living in Los Angeles County (column 1), the number of FDI links between Los Angeles County and the countries of origin (column 2), and the predicted changes in these variables under a counterfactual scenario where the pre-1880 pull factor of Los Angeles is 5 times as large the true size (columns 3 and 4). Column 3 shows the predicted change of *Total # of FDI relationships* (in percent) based on the IV regression of *Log Total # of FDI relationships* on *Log Ancestry 2010*, instrumented for by $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880..2000}$, similar to column 2 of Table 6 without the principal components as instruments. Column 4 shows the predicted absolute change in ancestry, based on a regression analogous to column 9 of Table 2 with *Ancestry 2010* (in levels) as dependent variable, again excluding the principle components. All three regressions control for log distance and latitude difference and include a origin × census region, and destination × continent fixed effects. Only the 10 countries with the highest absolute change in ancestry are shown in the interest of space. The details for the construction of this counterfactual are presented in section 4.4.

Table 9: The effect of Ancestry versus Foreign-born on FDI

	FDI 2014 (Dummy) IV	 IV	 OLS	 IV	 IV	 IV	 IV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log Ancestry 2010	0.184***		0.155***	0.242***		0.163***	
	(0.010)		(0.022)	(0.043)		(0.014)	
Log Foreign-born 2010		0.207***	-0.012	-0.082*			
		(0.014)	(0.031)	(0.049)			
Log Foreign-born 1970					0.286***	0.046	
					(0.025)	(0.034)	
Log Ancestry 2000							0.236***
							(0.037)
Log Foreign-born 2000							-0.076*
							(0.045)
N	612495	612495	612495	612495	612495	612495	612495

Notes: The table presents the OLS (column 3) and IV (all other columns) estimates of equation (1), contrasting the effect of ancestry and first-generation immigrants (foreign-born) on FDI. The dependent variable is the dummy for FDI in 2014. All IV columns use as instruments the same set of variables as column 3 of Table 3. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. The coefficient estimates on these control variables are not reported in the interest of space. Standard errors are given in parentheses and clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. For column 4, the Kleibergen-Paap LM statistic on the excluded instruments is 18.211 with p-value 0.1497. For column 7, the Kleibergen-Paap LM statistic on the excluded instruments is 19.336 with p-value 0.113.

Table 10: Heterogeneous Effects across Origin Countries

	FDI 2014 (Dummy)						*Log Total # of FDI relationships*					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Log Ancestry × Geographic Distance	0.099***	0.107***	0.135**			0.170**	0.404***	0.571***	0.479*			0.865***
	(0.030)	(0.034)	(0.065)			(0.076)	(0.077)	(0.094)	(0.253)			(0.156)
Log Ancestry × Genetic Distance	-1.105						-6.159					
	(1.005)						(4.195)					
Log Ancestry × Linguistic Distance		-0.417						-3.973***				
		(0.510)						(1.465)				
Log Ancestry × Religious Distance			-0.869						-1.330			
			(0.559)						(1.581)			
Log Ancestry × Judicial Quality				0.180*		0.373**				1.414***		2.376***
				(0.094)		(0.187)				(0.243)		(0.494)
Log Ancestry × Fractionalization					-0.240**	0.470					-1.486***	3.091***
					(0.095)	(0.324)					(0.328)	(0.829)
N	486855	414612	411471	452304	508842	446022	9970	9345	9221	10089	10166	10089

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable for Panel A is the dummy for FDI in 2014. The dependent variable for Panel B is the log of the number of FDI links in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,...,2000}$ and principal components as instruments. All specifications control for log distance, latitude difference, origin, and destination fixed effects, as in column 2 of Table 3. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The measures of genetic, linguistic, and religious distance between the United States and the origin country are from Spolaore and Wacziarg (2015). The measure of judicial quality is from Nunn (2007). Ethnic Diversity refers to 1 minus the Herfindahl index of ethnicities in the origin country calculated using the data in Alesina et al. (2003).

Table 11: Heterogeneous Effects across Destination Counties

	FDI 2014 (Dummy)			*Log Total # of FDI relationships*		
	(1)	(2)	(3)	(4)	(5)	(6)
Log Ancestry × Foreign Share	1.388		2.996	3.538		-0.007
	(3.103)		(3.096)	(7.806)		(6.463)
Log Ancestry × Ethnic Diversity		1.270***	1.343***		3.694***	3.694***
		(0.204)	(0.223)		(1.010)	(1.062)
N	611910	612495	611910	10851	10851	10851

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable for Panel A is the dummy for FDI in 2014. The dependent variable for Panel B is the log of the number of FDI links in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,...,2000}$ and principal components as instruments. All specifications control for log distance, latitude difference, origin, and destination fixed effects, as in column 2 of Table 3. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. Foreign Share is the share of the destination county's population that are of any foreign ancestry in 2010. Diversity of Ancestries is measured as 1 minus the Herfindhal index of ancestry shares in the destination county.

Table 12: Heterogeneous Effects across Sectors and Firms

	Log Ancestry 2010	*FDI 2014 (Dummy)* > 0
FDI 2014 (Dummy)		
Panel A: Individual Sectors		
Manufacturing	0.165*** (0.024)	5,549
Trade	0.158*** (0.025)	3,212
Information, Finance, Management, and other Services	0.143*** (0.024)	3,628
Construction, Real Estate, Accomodation, Recreation	0.125*** (0.021)	1,637
Health, Education, Utilities, and other Public Services	0.042*** (0.019)	689
Natural Resources	0.035*** (0.009)	669
Panel B: Final vs. Intermediate Goods		
Intermediate Goods	0.169*** (0.024)	5,842
Final Goods	0.156*** (0.026)	4,201
p-value of χ^2 test, H_0: equality of coefficients	0.000	
Panel C: Small vs. Large Firm Size		
Above Median	0.112*** (0.018)	1,840
Below Median	0.051*** (0.024)	723
p-value of χ^2 test, H_0: equality of coefficients	0.000	

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from IV regressions at the country-county level. Each row of the table corresponds to a separate regression. The dependent variables in all rows are dummy variables that are one if any firm within the indicated subset of firms in destination county d has a parent or subsidiary in origin country o. These subsets of firms are five sector groups (panel A), firms producing final goods versus intermediate inputs (panel B), and for small- versus large firms (panel C). The composition of sector groups in panel A is given in Appendix Table 4. Final goods and intermediate inputs are defined as 4-digit NAICS sectors with upstreamness index below and above 2, respectively, where we use the upstreamness index from Antràs et al. (2012). The cutoff value between small and large firms is the median employee number, which is 1380 for US firms that are subsidiaries and 1057 for US firms that are parents. Throughout, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as intrumental variables. "*FDI 2014 (Dummy)* > 0" refers to the number of country-county pairs that have an (non-zero) FDI link in the corresponding sector. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 13: Spillover Effects

Panel A: IV	*FDI 2014 (Dummy)*			
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.191***	0.190***	0.231***	0.197***
	(0.024)	(0.028)	(0.021)	(0.020)
Log Ancestry 2010, State Level	-0.020**			
	(0.010)			
Log Ancestry 2010 of Nearest County within State		-0.004		
		(0.016)		
Log Ancestry 2010, Continent Level			0.012	
			(0.018)	
Log Ancestry 2010 of Nearest Origin				-0.056***
				(0.020)
N	612495	612495	612495	612495
Panel B: IV	*Log Total # of FDI relationships*			
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.359***	0.200*	0.314***	0.283***
	(0.132)	(0.112)	(0.111)	(0.099)
Log Ancestry 2010, State Level	-0.160			
	(0.102)			
Log Ancestry 2010 of Nearest County within State		0.186*		
		(0.104)		
Log Ancestry 2010, Continent Level			-0.097	
			(0.161)	
Log Ancestry 2010 of Nearest Origin				0.049
				(0.214)
N	10851	10851	10851	10851

Notes: The table presents coefficient estimates from the extensive-margin equation (1) (Panel A) and the intensive-margin equation (7) (Panel B) at the country-county level. The dependent variable for Panel A is the dummy for FDI in 2014. The dependent variable for Panel B is the log of total FDI in 2014. In all columns, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as instrumental variables. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Online Appendix
"Migrants, Ancestors, and Foreign Investments"

Konrad B. Burchardi

Thomas Chaney

Tarek A. Hassan

A Data Appendix

Overview

To construct the migration and ancestry data up until the year 2000, we download the 1880, 1900, 1910, 1920, 1930, 1970, 1980, and 2000 waves of the Integrated Public Use Microdata Series (IPUMS) from `https://usa.ipums.org/usa-action/samples`. For each wave, we select the largest available sample; for example, if a 1% and 10% sample was available for 1880 data, we used the 10% sample. To construct the 2010 data, we used the 2006-2010 American Community Survey (ACS) sample provided on the IPUMS website. For a more detailed overview on the specific waves used, see Appendix Table 1.

For each sample, we obtain the following variables: year, datanum, serial, hhwt, region, statefip, county, cntygp97, cntygp98, puma, gq, pernum, perwt, bpl, mbpl, fbpl, nativity, ancestr1, yrimmig, mtongue, mmtongue, fmtongue, and language.

We construct the number of migrants from origin country o to destination county d in t, $I^t_{o,d}$, as well as the measure of ancestry $A^t_{o,d}$ from 1980 onward. We first aggregate the individual-level census data to counts of respondents at the level of historic US counties (or country groups from 1970 onwards) and foreign countries, and then transform the data into 1990 country-county level using various transition matrices. Details are given in the following sections.

How we create transition matrices

We create a set of transition matrices that transform non-1990 countries to 1990 countries and non-1990 counties/county groups to 1990 counties.

- Birthplace-to-country: The aim is to construct transition matrices that map all the birthplace answers into 1990 countries. In each wave of the US Census, respondents were asked to report their country of birth. All possible answers (across time) are listed here: `https://usa.ipums.org/usa-action/variables/BPL#codes_section`. The censuses from 1850-2012 contain roughly 550 possible different answers to the question of birthplace. In each census data set, they are saved in the variable "bpld." What follows is our procedure for building those matrices:

  1. We start with a transition matrix of zeros, with all possible answers to the 1990 birthplace question as rows and all 1990 countries as columns. A cell in row r and column c of the transition matrix answers the question, "What is the probability that an individual who claims his/her birthplace as r refers to the area that in 1990 is country c?" So all cells contain values in [0,1], and rows sum up to 1.

2. For each row r in the transition matrix, if r with certainty refers to the area that in 1990 is country c, we simply change the entry in cell (r,c) from 0 to 1; if r does refer to an area that in 1990 is in multiple countries, then we search for the 1990 population of each possible country, and assign probabilities in proportion to the population data. We use the population information from the Worldbank database.[37].

   Panel A in Appendix Table 2 lists the distribution of weights that we end up using, and the affected countries and persons.

- Ancestry-to-country: The aim is to construct transition matrices that map all the answers to the ancestry question into 1990 countries. The 1980, 1990, 2000, and 2010 census data provide information on the ancestry (ancestr1, 3-digit version). All possible answers (across time) are listed here: https://usa.ipums.org/usa-action/variables/ANCESTR1/#codes_section. The procedure is the same as in the birthplace-to-country procedure. Panel B in Appendix Table 2 lists the distribution of weights that we end up using, and the affected countries and persons.

- Group-to-county & PUMA-to-county: The aim is to construct transition matrices that map all the county groups/PUMAs into individual counties. For the years 1970 and 1980, the US census data are at the US county group level. A "county group" is an agglomeration of US counties. For the years 2000 and 2010, the census data are at the PUMA level. A "PUMA" is also an agglomeration of US counties.[38] To construct transition matrices from county agglomeration level to county level, we download the corresponding matching files from the IPUMS website. We use data on the population of each county (within each county group/PUMA) to assign a probability that an observation from county group/PUMA g in year t is from county c in year t. This approach gives a transition matrix from year t county groups to year t counties. Appendix Table 3 lists the distribution of weights that we end up using, and the affected counties and persons.

- County-to-county: The aim is to construct transition matrices tthat map all the non-1990 counties into 1990 counties. This step is necessary because the list and boundaries of US counties changed over time. Similarly to the birthplace-to-country and ancestry-to-country procedure, we use one transition matrix per census year (1880, 1900, 1910, 1920, 1930, 1970, 1980, 2000, 2010). Such a transition matrix has as rows all US counties, indexed c, in year t, and as columns all 1990 US counties, indexed m. Each cell of the transition matrix takes a value that answers the question, "Which fraction of the area of the county c in year t is in 1990 part of county m?" Appendix Table 3 lists the distribution of weights that we end up using, and the affected counties and persons. More specifically, we build these matrices as follows:

  1. We download the year-specific map files. For 1880 us counties, we obtain the 503MB GIS file from Atlas: http://publications.newberry.org/ahcbp/downloads/united_states.html and extract the 1880 part. For 1900, 1910, 1920, and 1930 counties, we obtain the maps from IPUMS: https://usa.ipums.org/usa/volii/ICPSR.shtml.

[37] http://data.worldbank.org/indicator/SP.POP.TOTL

[38] Detailed description of "county group" and "PUMA" can be found here: https://usa.ipums.org/usa/volii/tgeotools.shtml.

Finally, for 1970, 1980, and 1990 counties, we obtain the maps from NHGIS: https://data2.nhgis.org/main.

2. We project non-1990 maps onto 1990 counties. We used the intersect command in ArcGIS to map year-specific counties onto 1990 counties based on area. This approach gives a transition matrix from non-1990 counties to 1990 counties.

Appendix Table 1: Description of each IPUMS wave

Wave	Description
1880	We use the 10% sample with oversamples; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1900	We use the 5% sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1910	We use the 1% sample; the sample is unweighted; we use the region identifiers statefip and county.
1920	We use the 1% sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1930	We use the 5% sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and county.
1970	We use the 1% Form 1 Metro sample; the sample is unweighted; we use the region identifiers statefip and cntygp97 (county group 1970); note that only four states can be completely identified because metropolitan areas that straddle state boundaries are not assigned to states; identifies every metropolitan area of 250,000 or more.
1980	We use the 5% State sample; the sample is unweighted; we use the region identifiers statefip and cntygp98 (county group 1980); the sample identifies all states, larger metropolitan areas, and most counties over 100,000 population.
1990	We use the 5% State sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use the region identifiers statefip and puma; the sample identifies all states, and within states, most counties or parts of counties with 100,000 or more population.
2000	We use the 5% Census sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use region identifiers statefip and puma; the sample identifies all states, and within states, most counties or parts of counties with 100,000 or more population.
2010	We use the American Community Service (ACS) 5-Year sample; the sample is weighted, so we use the provided person weights to get to a representative sample; we use region identifiers statefip and puma, which contain at least 100,000 persons; the 2006-2010 data contains all households and persons from the 1% ACS samples for 2006, 2007, 2008, 2009 and 2010, identifiable by year.

Appendix Table 2: Historical birthplace to current country: transition matrices

Panel A: Birthplace		weights $\in (0,1)$	weight $= 1$	weights $= 0$
1880	# of answers	22	258	9
	# of persons	26,301	50,177,184	4,933
	% of persons	0.05%	99.94%	.01%
1900	# of answers	15	131	6
	# of persons	23,345	6,555,140	5,339
	% of persons	0.35%	99.56%	.08%
1910	# of answers	20	99	4
	# of persons	31,072	5,613,136	3,105
	% of persons	0.55%	99.39%	.05%
1920	# of answers	13	174	7
	# of persons	36,070	3,905,455	12,559
	% of persons	0.91%	98.77%	.32%
1930	# of answers	25	194	9
	# of persons	35,930	3,086,341	61,462
	% of persons	1.13%	96.94%	1.93%
1970	# of answers	12	77	3
	# of persons	318,800	6,323,100	230,800
	% of persons	4.64%	92.00%	3.36%
1980	# of answers	32	222	7
	# of persons	491,760	4,774,820	313,300
	% of persons	8.81%	85.57%	5.61%
1990	# of answers	24	209	7
	# of persons	721,595	8,532,585	484,433
	% of persons	7.41%	87.62%	4.97%
2000	# of answers	11	136	0
	# of persons	1,122,532	13,144,632	0
	% of persons	7.87%	92.13%	0%
2010	# of answers	14	137	1
	# of persons	1,302,255	11,131,046	17,148
	% of persons	10.46%	89.40%	.14%
2010*	# of answers	14	188	1
	# of persons	3,512,123	300,415,680	37,469
	% of persons	1.16%	98.83%	.01%
Panel B: Ancestry		weights $\in (0,1)$	weight $= 1$	weights $= 0$
1980	# of answers	29	227	143
	# of persons	924,400	198,525,616	27,412,380
	% of persons	0.41%	87.51%	12.08%
1990	# of answers	29	239	9
	# of persons	2,941,941	217,720,512	27,445,182
	% of persons	1.19%	87.75%	11.06%
2000	# of answers	17	137	22
	# of persons	6,000,639	191,300,704	84,120,558
	% of persons	2.13%	67.98%	29.9%
2010	# of answers	19	142	30
	# of persons	8,454,279	229,211,968	66,299,030
	% of persons	2.78%	75.41%	21.81%

The table reports statistics on the transition of data from the ‘answer’ level to 1990 country level. For each survey wave, and each question – birthplace in Panel A and primary ancestry in Panel B – the table reports the number of answers that can be directly linked to a 1990 country (weight = 1), that are assigned to several 1990 countries using population weights (weights $\in (0,1)$) and that cannot be linked to any modern country with sufficient certainty (weights = 0). The table also reports the number of respondents (scaled from the original data using the person weights provided) in each category. Answers with weights zero essentially consists of ”Not Reported” (e.g. 23, 24, 54 and 30 million respondents for the 1980, 1990, 2000 and 2010 ancestry data, respectively) and ”African-American” (e.g. 26, 22 and 25 million respondents for the 1990, 2000 and 2010 ancestry data, respectively). The remainders are mostly cases such as ”African”, ”Uncodable”, ”Bohemian”, ”Nuevo Mexicano”, ”Other”, etc. In Panel A, all years except 1880 consist of the number of persons that report birthplace since the last Census wave. For the 2010 Census wave the additional entry (denoted by a *) reports the respective numbers for all respondents in that wave.

Appendix Table 3: Historical state-county unit to 1990 state-county unit: transition matrices

Census wave		weights $\in (0, 1)$	weight $= 1$	weights $= 0$
1880	# of counties	658	1854	1
	% of persons (birthplace data)	21.54%	78.45%	.01%
1900	# of counties	2211	7	4
	% of persons (birthplace data)	99.09%	0.87%	.05%
1910	# of counties	1517	5	1
	% of persons (birthplace data)	99.00%	0.94%	.05%
1920	# of counties	1355	7	0
	% of persons (birthplace data)	90.80%	9.20%	0%
1930	# of counties	1801	6	0
	% of persons (birthplace data)	90.61%	9.39%	0%
1970	# of countygroups	310	98	0
	% of persons (birthplace data)	34.07%	65.93%	0%
1980	# of countygroups	580	573	0
	% of persons (birthplace data)	17.96%	82.04%	0%
	% of persons (ancestry data)	40.02%	59.98%	0%
1990	# of PUMAs	541	1185	0
	% of persons (birthplace data)	8.97%	91.03%	0%
	% of persons (ancestry data)	32.15%	67.85%	0%
2000	# of PUMAs	620	1451	0
	% of persons (birthplace data)	10.66%	89.34%	0%
	% of persons (ancestry data)	30.36%	69.64%	0%
2010	# of PUMAs	619	1449	1
	% of persons (birthplace data)	12.31%	87.65%	.03%
	% of persons (ancestry data)	30.13%	69.81%	.05%

The table reports statistics on the transition of data from the 'historical spatial area' level to 1990 US county level. For each Census wave the table reports the number of contemporaneous spatial areas that are a subset of a 1990 US county (weight = 1) and the number of contemporaneous spatial areas whose data is transitioned to 1990 US county level using non-degenerate weights (weights $\in (0, 1)$). For Census waves 1880 to 1930 the share of their contemporaneous county spatial area in each 1990 US county area is used as weight. For waves 1970 to 2010 there are two steps: In step 1 the share of their contemporaneous countygroup (waves 1970 and 1980) or PUMA (waves 1990 to 2010) population in the contemporaneous county population are used as weights; in step 2 the share of their contemporaneous county spatial area in each 1990 US county area is used as weight. The two-step procedure is necessary because the 1970 to 2010 Census waves do not have a county-level identifier (to protect the privacy of the respondents). The table also reports the share of respondents affected by this transition in the birthplace and ancestry data, respectively.

A.1 Details on the construction of migration and ethnicity data

Details calculation of post-1880 flow of immigrants

For each census wave after 1880, we count the number of individuals in each historic US county d who were born in historic country o (as identified by birthplace variable "bpld" in the raw data) that had immigrated to the United States since the last census wave that contains the immigration variable (not always 10 years earlier). Then we transform these data

- from the non-1990 foreign-country ("bpld") level to the 1990 foreign-country level using bpld-to-country transition matrices.
- from the US-county group/puma level to the US-county level using group/puma-to-county transition matrices.
- from the non-1990 US-county level to the 1990 US-county level using county-to-county transition matrices.
- from the post-1990 US-county level to the 1990 US county level. Based on the information from https://www.census.gov/geo/reference/county-changes.html, a new county is either created from part of ONE 1990 county or assigned a new FIPS code after 1990, so we manually change that county's FIPS code to what it was in 1990. A few counties' boundaries have been changed after 1990 but that only involved a tiny change in population, so we ignore these differences.

Details calculation of pre-1880 stock of immigrants

For the year 1880, we calculate for each historic US county d the number of individuals who were born in a historic foreign country o (no matter when they immigrated). We add to those calculations the number of individuals in county d who were born in the United States, but whose parents were born in historic foreign country o. (If the parents were born in different countries, we count the person as half a person from the mother's place of birth, and half a person from the father's place of birth). Then we transform these data

- from the pre-1880 foreign-country ("bpld") level to the 1990 foreign-country level using the pre-1880 country-to-country transition matrix.
- from the pre-1880 US-county level to the 1990 US-county level using the pre-1880 county-to-county transition matrix.

Details calculation of stock of ancestry (1980, 1990, 2000, and 2010)

For the years 1980, 1990, 2000, and 2010, we calculate for each US county group the number of individuals who state as primary ancestry ("ancestr1" variable) some nationality/area. We transform the data

- from the ancestry-answer ("ancestr1") level to the 1990 foreign-country level using ancestry-to-country transition matrices.

- from the US-county group/puma level to the US county-level using group/puma-to-county transition matrices.

- from the non-1990 US-county level to the 1990 US-county level using county-to-county transition matrices.

- from the post-1990 US-county to the 1990 US-county level. Based on the information from https://www.census.gov/geo/reference/county-changes.html, a new county is either created from part of ONE 1990 county or assigned a new FIPS code after 1990, so we manually change that county's FIPS code to what it was in 1990. A few counties' boundaries have been changed after 1990 but that only involved a tiny change in population, so we ignore the difference.

A.2 Details on the construction of FDI data

Our FDI data are from the US file of the Bureau van Dijk ORBIS dataset. For each US firm, the raw data set lists the location of its (operational) headquarters, the addresses of its foreign parent entities, and the addresses of its international subsidiaries and branches. It also provides the number of employees for both US and foreign firms. The steps for building the data follow below.

Clean postcode information

We use firm's postcode as a unique identifier for the county location of the US firm, and then need to ensure that one county uniquely corresponds to one postcode. Vance, NC; Wakulla, FL; Citrus, FL; Rankin, MS; Union, OH; and Du Page, IL share at least one postcode with a neighboring county. In each case we assign that postcode wholly to the county with the larger population (according to Google 2012 population data). In the last step, we hand-coded missing postcodes that we took from main data set. Only one such case existed: 75427 for Dallas.

Build the parent data

We used the following variables from the parent dataset: "Mark" "Company name" "BvD ID number" "Country ISO Code" "City" "Postcode" "NAICS 2007 Core code (4 digits)" "NAICS, text description" "Number of employees 2013" "Shareholder - Name" "Shareholder - BvD ID number" "Shareholder - City" "Shareholder - Postal code" "Shareholder - NAICS 2007, Core code" "Shareholder - NAICS 2007, text description" "Shareholder - Country ISO code" "Shareholder - Direct %" "Shareholder - Total %" "Shareholder - Number of employees". Here "shareholder" is equivalent to "parent" in our context. The key data-building steps are as follows:

1. Assign numerical values to "Shareholder Direct" and "Shareholder Total":

    - When the stake of a shareholder is described by an acronym rather than a number, we replace it with numerical values as follows: MO, majority owned, is replaced by "75%"; JO, jointly owned, is replaced by "50%"; NG, negligent, is replaced by '0%'; BR, branch and WO, wholly owned are both replaced by "100%".[39]

[39] See http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2407845 for reference.

- When the stake of a shareholder is described by the following expressions, we replace it with a numerical value as follows: Values with a “>”, e.g., “ > 25.00” were replaced by the original number plus 10; values with a “<”, e.g., “ < 34.00”, were replaced by the original number minus 10; values with a “±”, e.g. “±25.00”, were replaced by the original number.

2. Postcode matching: We matched both US firms and US parents (foreign parents were ignored in this step), with our postcode data. Besides the original string variable postcode, we generated new variables postcode5digit and postcodeextension and labeled them “Postal code (5 digit)” and “Postal code (extension).” Similarly, shareholders had shareholderpostcodeUS5digit and shareholderpostcodeUSextension (note the spelling postal code in shareholder variables was unified to postcode).

3. Country-code matching: We matched both companies and their parents. Each firm had four country variables: numerical country code, country name, and 2- and 3- digit ISO country code. Then we adjusted those 2014 country codes to 1990 codes based on the information on post-1990 country changes.

Build the subsidiary data

We used the following variables from the subsidiary dataset: “Mark” “Company name” “BvD ID number” “Country ISO Code” “City” “Postcode” “NAICS 2007 Core code (4 digits)” “NAICS, text description” “Number of employees 2007” “Subsidiary - Name” “Subsidiary - BvD ID number” “Subsidiary - Country ISO code” “Subsidiary - City” “Subsidiary - Postal code” “Subsidiary - NAICS 2007, Core code” “Subsidiary - NAICS 2007, text description” “Subsidiary - Number of employees” “Subsidiary - Direct %” ”Subsidiary - Total%” “Branch - Name” ”Branch - BvD ID number” “Branch - Country ISO code” “Branch - City” “Branch - Postcode” “Branch - NAICS 2007, Core code” “Branch - NAICS 2007, text description” “Branch - Number employees”. The data cleaning process is identical to that of the parent data described above, with the exception that we merged subsidiaries with branches and refer to them collectively as “subsidiaries”.

A.3 Details on the construction of our other data

International trade.— The data on trade between US states and foreign countries, both at the aggregate level and at the sectoral level, are from the Commodity Flow Survey for the year 2012. The data are collected by the US Census Bureau. A representative sample of establishments are surveyed every five years, and information on their shipments collected. The value of all shipments crossing the US international border are recorded as international trade, along with their foreign origin/destination country. We only used thes readily available data aggregated at the US state and foreign country level. Although they do not cover all of the US foreign trade (the data com from a representative survey, not from the universe of foreign transactions), they are the only publicly available source of international data disaggregated at a geographic level below that of the entire United States. For each origin country and destination state, $Import_{o,d}$ are aggregate imports (in dollars) from country o to US state d in 2012, and $Export_{o,d}$ are aggregate exports (in dollars) from US state d to country o in 2012, where we keep the convention of using o for foreign countries and d for US administrative units, states or counties.

Bilateral distances and latitude differences.— To compute the distance between US counties or states and foreign countries, we used the coordinates for all postal codes within a county or state, and the coordinates of the main city for foreign countries.[40] We define the latitude and longitude of a US county as the unweighted average of the latitudes and longitudes of all postal codes within the county. We define the latitude and longitude of a US state as the unweighted average of the latitude and longitude of all counties within the state. The distance between foreign country o and a US county or state d, $Distance_{o,d}$, is computed as the great circle distance between the two, measured in kms. The latitude difference between a foreign country o and a US county or state d, $Latitude\,Difference_{o,d}$, is the absolute difference between the latitudes of the two, measured in degrees.

Country characteristics.— To shed light on the mechanism through which the presence of foreign ancestry affects the patterns for foreign investment, we constructed several measures of foreign country and US county characteristics. *"Genetic Distance"* is a measure of the genetic distance between a given foreign country and the United States, normalized to take values between 0 and 1. *"Linguistic Distance"* is a measure of the linguistic distance between a given foreign country and the United States; it measures the probability that a randomly selected person in the United States speaks the same language as a randomly selected person from that country. *"Religious Distance"* measures the religious distance between a given foreign country and the United States, with a similar construction as the linguistic distance.[41] A higher index for *"Genetic Distance"*, *"Linguistic Distance"*, or *"Religious Distance"* corresponds to a greater distance between the United States and that country. *"Judicial Quality"* is a measure of the judicial quality in a given country.[42] A higher index for *"Judicial Quality"* corresponds to a higher-quality judicial system. *"Ethnic Diversity"* is a measure of a country's ethnolinguistic fractionalization.[43]

US county characteristics.— We define three US-county level measures. *"Diversity of Ancestries"* is a measure of the diversity of communities from different ancestries in a given US county.[44] *"Foreign Share"* measures the share of residents in a given county who claim foreign ancestry.

Sectoral characteristics.— We separated sectors into final consumption goods and intermediate inputs. To do so, we use the measure of upstreamness from Antràs et al. (2012). We classified 4-digit NAICS sectors as "final goods" if their upstreamness index is below 2, and as "intermediates" if their upstreamness index is above 2.

[40] The geo-coordinates are downloaded from `www.geonames.org`and `www.cepii.fr`, respectively. When a county has multiple postcodes we randomly select one of them and use the geocoordinates for that randomly selected postcode.

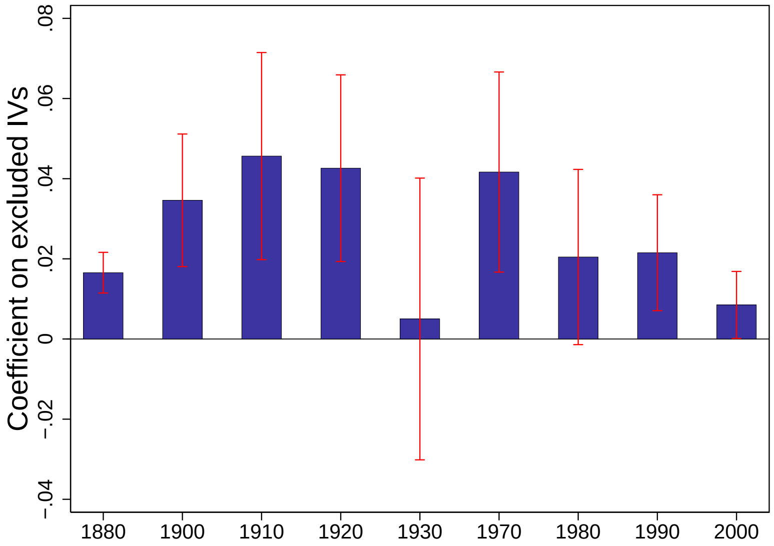
[41] Both genetic and religious distance measures come from Spolaore and Wacziarg (2015).

[42] The measure of judicial quality comes from Kaufmann et al. (2003) and is used in Nunn (2007). It is based on a weighted average of variables measuring perceptions of the effectiveness of the judiciary and the enforcement of contracts.

[43] The measure of fractionalization comes from Alesina et al. (2003). It is equal to 1 minus the Herfindahl index of ethnolinguistic group shares.

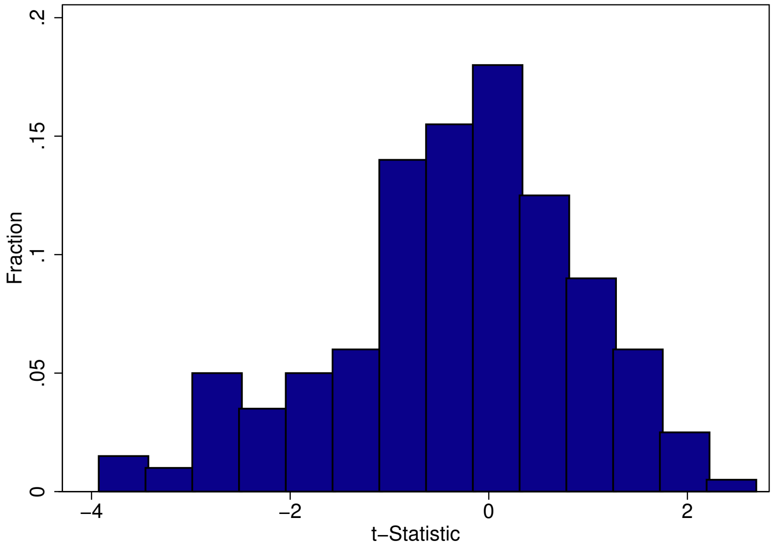
[44] It is equal to 1 minus the Herfindahl index of ancestry, measured as the sum of squared fractions of all possible ancestry among people who report foreign ancestry within that US county

B Additional figures and tables



Appendix Figure 1: Reduced-Form Coefficients

Notes: Coefficient estimates (bars) and 95% confidence intervals (lines) on the excluded instruments $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ from a reduced form regression corresponding to the specification in column 2 of Table 2, using the 2014 FDI dummy as dependent variable. Robust standard errors are clustered at the origin country level. The R^2 of this regression is 0.218.



Appendix Figure 2: Placebo Experiment: Histogram of t-stats

Notes: The figure presents a placebo experiment as an extension to Appendix Table 9 Panel B, where we repeatedly assign the interaction between push and pull factors for country *o* to randomly selected other countries. Each time, we assign each country to some random country on a different continent, run the same specification as in Table 9 Panel B Column 3, and report the t-statistic on the estimated coefficient on *Log Ancestry 2010*. We repeat the procedure 200 times and generate the histogram. Using ±1.96 as cut-off values, the false positive rate is 1.5% and the false negative rate is 11.5%.

Appendix Table 4: Composition of Sector Groups used in Table 12

Group	NAICS Sectors	# of US Firms
Manufacturing	Manufacturing	10009
Trade	Wholesale Trade Retail Trade	7191
Information, Finance, Management, and Other Services	Information Finance and Insurance Professional, Scientific, and Technical Services Management of Companies and Enterprises Administrative and Support and Waste Management and Remediation Services Other Services (Except Public Administration)	10052
Construction, Real Estate, Accomodation, Recreation	Construction Transportation and Warehousing Real Estate and Rental and Leasing Arts, Entertainment, and Recreation Accommodation and Food Services	3039
Health, Education, Utilities, and Other Public Services	Utilities Educational Services Health Care and Social Assistance	1257
Natural Resources	Agriculture, Forestry, Fishing and Hunting Mining, Quarrying, and Oil and Gas Extraction	871

Appendix Table 5: Summary Statistics on the Intensive Margin of FDI

Origin-destination pairs	(1)	(2)	(3)
Ancestry 2010 (in thousands)	10.038	10.861	16.502
	(40.989)	(43.593)	(62.950)
# of FDI Relationships	11.057		
	(39.766)		
# of Subsidiaries in Destination		4.572	
		(14.951)	
# of Parents in Origin		5.354	
		(17.975)	
# of Workers Employed at Subsidiary in Destination (in thousands)		11.872	
		(44.522)	
# of Subsidiaries in Origin			5.018
			(15.739)
# of Parents in Destination			2.318
			(4.431)
# of Workers Employed at Subsidiary in Origin (in thousands)			5.812
			(60.380)
N	10851	9082	4065

Notes: The table presents means (and standard deviations). Variables refer to our sample of (country,county) pairs used in Table 6. Column 1 shows data for observations that have at least one FDI link. Column 2 shows data for observations that have at least one subsidiary in the origin. Column 3 shows data for observations pairs that have at least one subsidiary in the destination.

Appendix Table 6: Assignment of States to Census Regions

Census Region	State Names
New England	Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont
Middle Atlantic	New Jersey, New York, Pennsylvania
East North Central	Illinois, Indiana, Michigan, Ohio, Wisconsin
West North Central	Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota
South Atlantic	Delaware, District Of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia
East South Central	Alabama, Kentucky, Mississippi, Tennessee
West South Central	Arkansas, Louisiana, Oklahoma, Texas
Mountain	Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming
Pacific	Alaska, California, Hawaii, Oregon, Washington

Appendix Table 7: Non-parametric and Nonlinear Least Squares Estimation

	(1)	(2)	(3)
		FDI Dummy (2014)	
Panel A: non-parametric OLS			
Ancestry 2010 Quantile 0		(omitted)	
Ancestry 2010 Quantile 1	0.002	0.002	0.001
	(0.008)	(0.007)	(0.006)
Ancestry 2010 Quantile 2	0.018	0.012	0.013
	(0.017)	(0.013)	(0.011)
Ancestry 2010 Quantile 3	0.124***	0.041*	0.024
	(0.032)	(0.022)	(0.017)
Ancestry 2010 Quantile 4		0.165***	0.061**
		(0.037)	(0.026)
Ancestry 2010 Quantile 5			0.205***
			(0.040)
N	612495	612495	612495

Panel B: Nonlinear Least Squares		
	β	π
Estimates	0.1683***	0.0010***
	(0.0011)	(0.0003)

Notes: The table presents coefficient estimates from non-parametric OLS (Panel A) and nonlinear least squares (Panel B) regressions at the country-county level. The dependent variable in both panels is the dummy for FDI in 2014. The cutoffs for the number of residents in county d with ancestry from country o are $\{0; 50; 145; 655; +\infty\}$ for quartiles (column 1), $\{0; 50; 108; 281; 1116; +\infty\}$ for quintiles (column 2), and $\{0; 50; 91; 186; 452; 1616; +\infty\}$ for sextiles (column 3). All the numbers are in units. Interpretation: between sextiles 1 and 2, the increment in the probability of positive FDI from adding 1000 more descendants (column 3) is +0.129 $\left(1000\,(0.013 - 0.001)\,/\,\left(\frac{187+92}{2} - \frac{92+0}{2}\right) \approx 0.1290\right)$. Standard errors given in parentheses are clustered at the country level in Panel A. Panel B shows (un-adjusted) NLS standard errors. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. For Panel B, we obtain the optimal β and π by solving a nonlinear least squares problem as mentioned in the text.

Appendix Table 8: Alternative Functional Forms

	FDI 2014 (Dummy)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ancestry 2010	0.002***					
	(0.001)					
Log Ancestry 2010 (-1 for $-\infty$)		0.190**				
		(0.080)				
(Ancestry 2010)$^{1/3}$			0.187***			
			(0.021)			
Log Ancestry 1980				0.127***		
				(0.031)		
Log Ancestry 1990					0.128***	
					(0.036)	
Log Ancestry 2000						0.132***
						(0.038)
N	612495	612495	612495	612495	612495	612495

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable is the dummy for FDI in 2014. The main variable of interest in each column is the measure of ancestry indicated by the first column of the table. In the second row, we use Log(Ancestry/1000) instead of Log(1+Ancestry/1000), and replace Log(0) with -1. All specifications are the same as that in Table 3, column 3. Standard errors are given in parentheses and are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 9: Placebo Regressions

	(1)	(2)	(3)	(4)	(5)	(6)
	FDI 2014 (Dummy)					
Panel A	*Assign to alphabet neighbor*					
Log Ancestry 2010	-0.012	-0.007	0.009	0.009	0.010	0.012
	(0.020)	(0.015)	(0.028)	(0.028)	(0.028)	(0.031)
N	612495	612495	612495	612495	612495	612300
Panel B	*Assign to alphabet neighbor on a different continent*					
Log Ancestry 2010	-0.026	-0.021	0.009	0.009	0.013	0.012
	(0.021)	(0.015)	(0.033)	(0.033)	(0.038)	(0.037)
N	612495	612495	612495	612495	612495	612300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from placebo regressions corresponding the the specifications in Table 3. In Panel A, we assign the outcomes (FDI 2014 Dummy) for each origin country to the next country in the alphabet. In Panel B, we assign the outcomes (FDI 2014 Dummy) for each origin country to the next country in the alphabet that is from another continent.

Appendix Table 10: Alternative Standard Errors

	(1)	(2)	(3)
	FDI Dummy (2014)		
Panel A: Bootstrap	Raw	Origin Panel	Destination Panel
Log 2010 Ancestry	0.1875***	0.1875***	0.1875***
	(0.0091)	(0.0281)	(0.0154)
N	612495	612495	612495
Panel B	Robust	County Cluster	State Cluster
Log 2010 Ancestry	0.1875***	0.1875***	0.1875***
	(0.0091)	(0.0171)	(0.0187)
N	612495	612495	612495
Panel C	Country Cluster	State-Continent Cluster	State-Country Cluster
Log 2010 Ancestry	0.1875***	0.1875***	0.1875***
	(0.0239)	(0.0130)	(0.0112)
N	612495	612495	612495

Notes: The table shows alternative standard errors for our standard specification in Table 3, column 3. Panel A shows bootstrapped standard errors constructed using 50 draws with replacement. Column 1 shows a conventional bootstrap; column 2 and 3 block-bootstrap across origins and destinations, respectively. Panels B and C show standard errors clustered in various dimensions.

Appendix Table 11: The Effect of Ancestry on FDI: 5 Largest Countries and Counties

	FDI 2014 (Dummy)
Panel A: Top 5 Ancestries	*Log Ancestry 2010*
Germany	0.216***
	(0.009)
Britain	0.271***
	(0.009)
Mexico	0.171***
	(0.011)
Ireland	0.202***
	(0.010)
Italy	0.219***
	(0.007)
Panel B: Largest 5 Counties	*Log Ancestry 2010*
Los Angeles, California	0.137***
	(0.019)
Cook, Illinois	0.146***
	(0.020)
Harris, Texas	0.169***
	(0.023)
San Diego, California	0.164***
	(0.024)
Orange, California	0.160***
	(0.020)

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable in both panels is the dummy for FDI in 2014. Panel A presents the coefficient on *Log Ancestry 2010* when we run our estimation separately for each of the largest five origin countries. Panel B presents the coefficient on *Log Ancestry 2010* when we run our estimation separately for each of the five US counties with the largest population in 2010. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as IVs. All specifications control for log distance and latitude difference. Standard errors are clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 12: The Effect of Ancestry on FDI: Country Specific Effects

	Point Estimate	Standard Error	*FDI 2014 (Dummy)* > 0
United Arab Emirates	11.875***	(2.712)	60
Kuwait	6.098***	(2.120)	22
Finland	4.113***	(0.513)	180
New Zealand	2.980***	(0.511)	107
Oman	2.481	(1.597)	6
British Virgin Islands	2.467***	(0.604)	100
Australia	2.201***	(0.384)	369
Malaysia	2.005***	(0.406)	90
South Africa	1.832***	(0.247)	80
Tunisia	1.438***	(0.345)	9
Iceland	1.359***	(0.276)	25
Saudi Arabia	1.144***	(0.158)	29
Belgium and Luxembourg	1.086***	(0.087)	354
Puerto Rico	1.034***	(0.240)	26
Israel	0.944***	(0.156)	137
Bahamas	0.943***	(0.308)	44
Switzerland	0.814***	(0.048)	371
Denmark	0.684***	(0.043)	278
Thailand	0.583***	(0.070)	68
Japan	0.566***	(0.051)	575
Uruguay	0.541***	(0.115)	21
Austria	0.531***	(0.042)	148
Chile	0.502***	(0.078)	73
Brazil	0.496***	(0.047)	140
Barbados	0.462**	(0.234)	38
Canada	0.461***	(0.024)	809
Norway	0.459***	(0.028)	239
Malta	0.451	(0.281)	11
Costa Rica	0.447***	(0.140)	30
Turkey	0.444***	(0.067)	48
Netherlands	0.442***	(0.019)	398
Panama	0.439***	(0.115)	44
Indonesia	0.413***	(0.076)	29
Argentina	0.412***	(0.056)	64
Sweden	0.405***	(0.018)	323
Senegal	0.383	(0.314)	2
France	0.346***	(0.013)	528
South Korea	0.346***	(0.023)	155
Liberia	0.341*	(0.190)	6
Spain	0.335***	(0.014)	300
India	0.320***	(0.018)	233
China	0.299***	(0.015)	248
Kenya	0.292*	(0.175)	5
Venezuela	0.275***	(0.046)	32
Britain	0.271***	(0.009)	664
Egypt	0.259***	(0.051)	23
Belize	0.255***	(0.086)	14

Hungary	0.240***	(0.033)	52
Colombia	0.237***	(0.028)	45
Italy	0.219***	(0.007)	489
Peru	0.218***	(0.033)	30
Germany	0.216***	(0.009)	608
Portugal	0.206***	(0.028)	85
Samoa	0.204**	(0.086)	5
Ireland	0.202***	(0.010)	247
Morocco	0.197**	(0.078)	11
Nigeria	0.190***	(0.055)	18
Sri Lanka	0.180	(0.120)	6
Czechoslovakia	0.177***	(0.029)	54
Romania	0.173***	(0.041)	23
Mexico	0.171***	(0.011)	259
Pakistan	0.168***	(0.039)	23
USSR	0.165***	(0.015)	97
Ghana	0.156	(0.095)	6
Bulgaria	0.156**	(0.064)	11
Philippines	0.154***	(0.019)	50
Lebanon	0.150***	(0.047)	20
Bolivia	0.142**	(0.066)	8
Greece	0.131***	(0.028)	42
Trinidad and Tobago	0.130*	(0.067)	15
Socialist Yugoslav	0.121***	(0.028)	29
Jamaica	0.114***	(0.032)	15
Honduras	0.103***	(0.032)	14
Algeria	0.099	(0.076)	3
Guatemala	0.097***	(0.033)	14
Poland	0.092***	(0.015)	63
Viet Nam	0.091***	(0.025)	18
Jordan	0.090	(0.063)	7
Cameroon	0.085	(0.065)	2
Dominican Republic	0.082***	(0.025)	16
Ecuador	0.081**	(0.032)	15
Paraguay	0.079	(0.056)	4
Nicaragua	0.069*	(0.036)	7
Albania	0.069	(0.046)	3
North Korea	0.068	(0.072)	1
El Salvador	0.066**	(0.026)	13
Sudan	0.065	(0.065)	1
Fiji	0.065	(0.046)	5
Bangladesh	0.040	(0.032)	2
Cambodia	0.039	(0.028)	3
Haiti	0.026	(0.019)	2
Ethiopia	0.026	(0.025)	1
Syria	0.016	(0.016)	1
Myanmar	0.007	(0.007)	1
Afghanistan	0.003	(0.003)	1
Guyana	0.002	(0.002)	1

Iraq	0.002	(0.002)	1
Cuba	-0.000***	(0.000)	1
Libya	-0.022	(0.024)	1
Nepal	n/a	n/a	0
Grenada	n/a	n/a	0
State of Palestine	n/a	n/a	0
Sierra Leone	n/a	n/a	0
Yemen	n/a	n/a	0
Equatorial Guinea	n/a	n/a	0
Somalia	n/a	n/a	0
Greenland	n/a	n/a	0
Cape Verde	n/a	n/a	0
Mauritania	n/a	n/a	0
Tonga	n/a	n/a	0
Lao	n/a	n/a	0
Mongolia	n/a	n/a	0
Iran	n/a	n/a	0

Notes: The table is an extension of Table 11 Panel A, where we only show the results for top five ancestries. Results are sorted on the point estimate. The last column shows the number of US counties that have an FDI link with the corresponding country. All countries with ancestry < 1 are discarded.

Appendix Table 13: The Effect of Ancestry on FDI: Sector-Specific Effects

20 Sectors Based on 2007 NAICS code	Point Estimate	Standard Error	*FDI 2014 (Dummy)* > 0
Manufacturing	0.165***	(0.024)	5,549
Wholesale Trade	0.141***	(0.026)	2,513
Professional, Scientific, and Technical Services	0.122***	(0.024)	1,925
Retail Trade	0.085***	(0.020)	846
Information	0.084***	(0.018)	906
Transportation and Warehousing	0.084***	(0.016)	620
Administrative and Support and Waste Management and Remediation Services	0.083***	(0.018)	855
Real Estate and Rental and Leasing	0.077***	(0.020)	662
Finance and Insurance	0.071***	(0.019)	1,143
Other Services (except Public Administration)	0.053***	(0.014)	301
Management of Companies and Enterprises	0.049***	(0.014)	524
Construction	0.040**	(0.016)	510
Accommodation and Food Services	0.035***	(0.010)	239
Arts, Entertainment, and Recreation	0.030***	(0.006)	131
Mining, Quarrying, and Oil and Gas Extraction	0.028***	(0.009)	528
Health Care and Social Assistance	0.024**	(0.012)	291
Utilities	0.022*	(0.012)	338
Educational Services	0.009	(0.006)	111
Agriculture, Forestry, Fishing and Hunting	0.007**	(0.003)	149
Public Administration	0.001	(0.001)	10

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from IV regressions for each of the 20 2-digit NAICS sectors at the country-county level. Each row of the table corresponds to one regression. The dependent variable in each row is a dummy variable for FDI in 2014 in the sector indicated. The last column shows the number of country-county pairs that have an FDI link with the corresponding country. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ and principal components as IVs. All specifications control for log distance, latitude difference, origin×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 14: The Effect of Ancestry on FDI: Variations of Leave-Out Instrument

	(1)	(2)	(3)	(4)	(5)	(6)
	FDI Dummy (2014)					
Panel A	$\{I_o^t(I_d^t/I)\}$ excluded					
Log Ancestry 2010	0.204***	0.202***	0.174***	0.174***	0.183***	0.215***
	(0.020)	(0.019)	(0.022)	(0.022)	(0.022)	(0.017)
N	612495	612495	612495	612495	612495	612300
Panel B	$\{I^t_{o,-d}(I^t_{-o,d}/I^t_{-o})\}$ excluded					
Log Ancestry 2010	0.212***	0.204***	0.172***	0.171***	0.185***	0.216***
	(0.020)	(0.019)	(0.024)	(0.024)	(0.024)	(0.017)
N	612495	612495	612495	612495	612495	612300
Panel C	$\{I^t_{o,-d}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded					
Log Ancestry 2010	0.223***	0.217***	0.183***	0.183***	0.200***	0.227***
	(0.022)	(0.021)	(0.024)	(0.024)	(0.024)	(0.018)
N	612495	612495	612495	612495	612495	612300
Panel D	$\{I^t_{o,-adj(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded					
Log Ancestry 2010	0.232***	0.204***	0.192***	0.192***	0.206***	0.237***
	(0.024)	(0.022)	(0.022)	(0.022)	(0.021)	(0.019)
N	640764	640764	640764	640764	640764	640560
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table shows variations of the estimates from Panel A in Table 3, removing or not different sets of migrants from the interaction of pull and push factors. The construction of the interaction is indicated above each panel. In Panel D, "adj" refers to the adjacent states for the state of county d.

Appendix Table 15: The effect of Ancestry in 2000 on FDI in 2007

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: IV			*FDI 2007 (Dummy)*			
Log Ancestry 2000	0.250***	0.214***	0.182***	0.182***	0.201***	0.184***
	(0.018)	(0.028)	(0.020)	(0.020)	(0.019)	(0.021)
First-stage F on instruments	12.06	3,793.67	167.32	165.46	169.31	189.21
N	612,495	612,495	612,495	612,495	612,495	612,300
Panel B: OLS			*FDI 2007 (Dummy)*			
Log Ancestry 2000	0.216***	0.216***	0.184***	0.184***	0.184***	0.200***
	(0.015)	(0.015)	(0.018)	(0.018)	(0.018)	(0.019)
N	612,495	612,495	612,495	612,495	612,495	612,300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from IV (Panel A) and OLS (Panel B) regressions of equation (1) at the country-county level. The dependent variable in all panels is a dummy indicating an FDI relationship between origin o and destination d in 2007. The main variable of interest is *Log Ancestry 2000*, instrumented using various specifications of equation (4). In all columns in Panel A, we include $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ as excluded instruments. Columns 2-6 also include the first five principal components of the higher-order interactions of push and pull factors as instruments. Column 5 also includes the interaction of the push and pull factor constructed using data from the 2006-2010 American Community Survey. All specifications control for log distance, latitude difference, origin, and destination fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the origin country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. (We also run an IV probit regression using the specification in column 2 yielding a marginal effect evaluated at the mean of *Log Ancestry 2000* on FDI equal to 0.206***(0.044).)