Migrants, Ancestors, and Foreign Investments*

Konrad B. Burchardi[†] **Thomas Chaney**[‡] **Tarek A. Hassan**[§]

June 2018

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 doubling the number of residents with ancestry from a given foreign country relative to the mean increases the probability that at least one local firm engages in FDI with that country by 4 percentage points. We present evidence that this effect is primarily driven by a reduction in information frictions, and not by better contract enforcement, taste similarities, or a convergence in factor endowments.

JEL Classification: O11, J61, L14.

Keywords: migrations, foreign direct investment, international trade, networks, social ties.

*We are grateful to Lorenzo Casaburi, Joshua Gottlieb, Richard Hornbeck, Nathan Nunn, Emir Kamenica, Jacopo Ponticelli, Nancy Qian, and David Strömberg for helpful discussions. Comments from the editor, Gita Gopinath, and four anonymous referees have helped to improve this paper. We also thank seminar participants at the Barcelona GSE, Boston University, Boston College, CEPR ERWIT, University of Chicago, Columbia University, Oxford, Georgetown, Harvard, IFN (Stockholm), Imperial, the University of Maryland, MIT, NBER EEG and Culture & Institutions program meetings, Paris School of Economics, Princeton, Singapore Management University, National University of Singapore, Toulouse, UPF, and the University of Zürich for their comments. Chaney and Hassan are grateful for financial support from NSF grant SES-1061622. Chaney is grateful for financial support from ERC grant N°337272–FiNet. Hassan is grateful for financial support from the IGM and the Fama-Miller Center at the University of Chicago. Mathias Iwanowsky, Markus Schwedeler, Lisa Tarquino and Philip Xu provided excellent research assistance. All mistakes remain our own.

[†]Institute for International Economic Studies, BREAD and CEPR, Stockholm University, SE-106 91 Stockholm, Sweden; E-mail: konrad.burchardi@iies.su.se.

[‡]Sciences Po and CEPR, 28 rue des Saints Pères, 75007 Paris, France; E-mail: thomas.chaney@gmail.com.

[§]Boston University, NBER and CEPR, 270 Bay State Road, Boston MA 02215, USA; E-mail: thassan@bu.edu.

International migrations have reached unprecedented levels over the past decades,[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 their economic effects remain incomplete. At the same time, foreign direct investment (FDI) 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 the ethnic connections emanating from migrations reaching back more than a century have a large positive causal effect on the propensity of US firms to engage in FDI with the historical migrants' countries of origin; and this effect appears to be driven primarily by a reduction of information frictions.

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. For example, historical migrations might have occurred between origins and destinations with certain unobserved climatic or other characteristics, and these characteristics might in turn drive FDI today. Similarly, past FDI might itself cause migration, for example because foreign employers second employees to work at their US subsidiaries.

To overcome these challenges 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 was attracting large numbers of migrants. We observe a large population with Italian ancestry in the West today. 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

[1] In 2013, there were 232 million international migrants, 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).

versus early twentieth century Italy) to instrument for the present-day distribution of ancestries. This 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 drops and becomes insignificant when we instrument for ancestry, suggesting that unobservable factors indeed confound simple OLS estimates of these effects.

Our paper makes three main contributions: (i) historical migrations and the ethnic diversity they created have a quantitatively large causal effect on FDI; (ii) this ethnic determinant of FDI appears to operate mainly because it facilitates the flow of information; and (iii) we propose a general method for instrumenting the composition of ancestry across US counties.

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 country, and increases by 7% the number of local jobs at subsidiaries of firms headquartered in that country. These effects persist over generations: even nineteenth century migrations 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 recursively shows 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 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.

A major advantage of this 'leave-out push-pull' approach is that it yields a specific instrument for migrations from each origin to each destination at each point in time, uniquely allowing us to simultaneously control for both origin and destination fixed effects, and to conduct a number of falsification exercises and robustness checks. For example, 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. These specifications, similar to a difference-in-difference, measure how variations in ancestry driven only by the instrumented inflow of defectors from communist countries explain changes in FDI, from zero in 1989 to its current level in 2014.

This flexibility of our instruments also delivers the statistical power to isolate specific channels linking ancestry to FDI: common ancestry may affect FDI because it (i) induces similarities in tastes for consumption, (ii) causes a convergence in factor endowments, facilitating horizontal FDI, (iii) provides social collateral for contract enforcement, substituting for poor institutions, or (iv) reduces information frictions. We find no evidence in support of the first three channels. Common ancestry does not affect FDI in the final goods sector more than in the intermediate goods sector, does not appear to cause a convergence in the sectoral distribution of employment, and has a significantly weaker impact on FDI for countries with weak institutions.

To provide a direct test for the remaining hypothesis that common ancestry affects FDI by reducing information frictions, we construct a novel measure of information demand about foreign countries using data from Google internet searches. Our index reflects variation across US metro areas in the relative frequency of search terms containing the names of each countries' most prominent politicians, actors, athletes, and musicians. We find a large causal effect of common ancestry on this index: residents of US metro areas with more ancestry from a given country systematically acquire more information about the politics and culture of that country. This fact fully accounts for the effect of ancestry on FDI, in the sense that controlling for our index of information demand drives out the significance of common ancestry in predicting FDI.

We also find that the effects of ancestry on FDI and information flow continue to operate long after migration from the origin country ceases, suggesting that immigrants pass traits to their descendants that facilitate economic exchange with their origin countries. As one example, foreign ancestry increases the use of the origin country's language by US-born individuals.

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 be 60,000 more individuals with German and Irish ancestry in Los Angeles today, and FDI between Los Angeles and Germany and Ireland would have increased by around 60%. The effect of ancestry on FDI is thus large and economically important.

Finally, we note one important limitation to our analysis: Our results rely purely on variation in the composition of FDI within the United States, not between countries. Although we believe that, in light of our results, the ethnic diversity of the United States likely also raises FDI for the country as a whole, we cannot exclude the possibility that increases in FDI in one state are partially or fully offset by decreases in others.

Existing literature. A large 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 et al., 2009; Leblang, 2010), 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 are likely to be non-random.

Three recent papers make attempts at identifying a causal effect of migrations on FDI and trade. Javorcik et al. (2011) use the cost of acquiring a passport and the existing stock of migrants from different countries in the United States to instrument for the impact of migrations between foreign countries and the United States on FDI. However, these instruments are most likely correlated with both migrations to the United States *and* FDI flows, and thus likely violate the exclusion restriction. Cohen et al. (2015) use the location of Japanese internment camps during World War II, and Parsons and Vezina (2016) the placement of Vietnamese refugees after the Vietnam War to identify a causal effect of those migrations on contemporary trade flows between locations within the United States, and Japan and Vietnam, respectively. While the exclusion restriction for the instruments in those two papers is plausible, instrumenting for migrations from only one country makes it impossible to control for destination fixed effects, that is, unobserved factors making a US state both a large recipient of migrants and a large importer and exporter.

Burchardi and Hassan (2013) use variation in wartime destruction across West German re-

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

gions to show evidence of a causal effect of social ties on changes in GDP growth and FDI in East Germany after the fall of the Berlin Wall.[5] Redding and Sturm (2008), Juhász (2014), and Steinwender (2014) study the effect of historical shocks on economic interactions across borders.

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 and destination US counties, 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 show that ancestry affects FDI most likely due to its effects on information flow.

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 Nunn et al. (2015) who study the effect of immigration from all origins during the Age of Mass Migration on present-day outcomes. Fulford et al. (2015) study the effect of historical ancestry composition of US counties on local economic growth. We add to this literature by showing a long-term effect of migration on the absolute advantage in conducting FDI of different regions that 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 however is not appropriate in our context, where unobserved and persistent origin-destination specific characteristics (such as the local climate) 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 historical migrations going back to the 19th century to instrument for the current stock of ancestry. This hybrid approach can easily be replicated for other countries, time periods, or any variable where cumulated flows matter, without the need for a rare or even unique historical accident.

[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 et al. (2015a), and Alesina et al. (2015b).

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, conducts robustness checks, and illustrates the quantitative implications of our findings using two thought experiments. Section 4 examines the mechanism underlying the effect of ancestry on FDI.

1 Data

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

Migrations and Ancestry. Our 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. We cannot use data from the 1940, 1950 and 1960 censuses, because these did not collect information on the year of immigration. The original 1890 census files were lost in a fire.

Throughout the paper, we use $t-1$ and t to denote two consecutive census waves, 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 $t-1$ and t. The exception to this rule is the 1880 census (the first in our sample), which also 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, thus covering the two generations of immigrants arriving prior to 1880.[8] Since 1980, respondents have also been asked about their primary ancestry in both the US Census and the American Community Survey, with the option to 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 (recalled) 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

[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's birthplace.

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

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"). When no direct mapping exists (for example, "Basque" or "Lapland"), 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.[11] The resulting dyadic dataset covers 3,141 US counties, 195 foreign countries, and 10 census waves.

Foreign Direct Investment. Our data on FDI is from the US file of the 2014 edition of the Bureau van Dijk ORBIS data set.[12] 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 partially or fully owned international subsidiaries and branches. In our main specification, we treat all equity stakes of any size as constituting a parent-subsidiary link.[13] 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.[14] Our main outcome variable, $\mathbf{1}\left[FDI_{o,d} > 0\right]$, is 1 if at least one firm in d has at least one parent or subsidiary in o, 0 otherwise. It captures both outward FDI (US firms with foreign subsidiaries) and inward FDI (foreign firms with US subsidiaries). We also count the number of FDI linkages between o and d (the number of foreign parents and subsidiaries in o of all firms in d), the number of unique parents and subsidiaries in both o and d, the number of employees working at firms in d with a foreign parent in o (*# of Employees at Subsidiaries in Destination*),[15] and the 2-digit NAICS code of the sector of the US firm. See Appendix A.2 for details. The resulting dataset covers the same 3,141 US counties, 195 foreign countries and 612,495 origin-destination pairs as above.

Other Data. To streamline the exposition, we discuss our measure of information demand in section 4.2. In addition, we use data on aggregate trade flows between US states and foreign

[10] We also aggregate our data to the PUMA level and show that our results are robust.

[11] Appendix Tables 2 and 3 report summary statistics on these data transitions, including the share of affected respondents. Appendix A.1 provides a detailed description of the data transformation.

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

[13] Appendix Table 11 shows that our results are almost completely unchanged when we restrict ourselves to links with an ownership stake larger than 5%, 25% or 50%.

[14] 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.

[15] When information on the number of employees is missing (which is the case for 95% and 58% of subsidiaries in the destination and origin, respectively), we assume the subsidiary employs one person.

countries for the year 2012 from the US Census Bureau.[16] We construct geographic distances, absolute latitude differences, and measures of agricultural similarity between US counties and foreign countries, and collect information on a number of 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 $\times$ 195 origin-destination pairs.[17] 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. Appendix Figure 3 visualises examples of this relationship: for each of the 10 largest US counties in terms of total FDI linkages, it scatters the number of FDI links to o against the number of individuals of ancestry o in d; regression lines are fitted for the top 10% ancestry groups and bottom 90% separately. In almost all cases larger ancestry groups are more likely to have an FDI link, and they have more FDI links. Panel A of Table 1 also shows that about half of the origin-destination pairs have ancestry of zero: most destinations in the United States do not have populations with ancestry from all 195 origin 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 2,852 individuals. Compared to this stock of ancestry, the flow of immigrants between 1990 and 2000 is relatively small, with 23 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 (its inclusion however has no effect on any of our main results).

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

[16] 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 between state level aggregate trade and state population is 93% for imports and 88% for exports respectively. We are therefore confident our trade dataset at the US state $\times$ foreign country level is not subject to severe measurement error.

[17] 53 countries have no FDI links with US firms in our sample.

4 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.[18]

Until World War I, migration to the United States was largely unregulated. European migrants in particular faced few or no restrictions 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 subsequent 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, Italians migrated to the United States in large numbers around 1910, followed by a peak in migrations from Eastern Europe and 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.

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

[18]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.

States over time. For each census from 1880 to 2010, we project the total number of new migrants from all origins to destination d, I_d^t, on destination and year fixed effects to account for general immigration time trends and persistent destination-specific effects. The figure shows 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.

In the next section, we use historical variation in both the timing of migrations from foreign countries to the U.S., and the timing of how attractive US counties are for newly arriving migrants, as the basis of our identification strategy.

3 Ancestry and Foreign Direct Investment

Before presenting formally our econometric model and results, we use a stylized historical example to describe the intuition behind our 'leave-out push-pull' approach to identification.

The purpose of our identification strategy is to isolate variation in the distribution of present-day ancestry which is independent of unobserved factors that could also affect the distribution of FDI. Figure 3 illustrates our approach 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 peak in the 1900-30 period (triggered by the end of feudalism and demographic pressures, and ending with Mussolini's anti-emigration policies).[19] The top-left part shows the relative attractiveness of US destinations for pre-1900 migrants, when German migrations to the United States peaked, as measured by the location choices of non-European migrants. At that time, most of these non-European migrants settled in the Midwest. Accordingly, we expect most German migrants from this initial wave to also 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 non-European migrants had shifted to the West and South. We expect many Italian migrants to have settled in these areas. The

[19] In absolute terms, there were also large migrations from Italy prior to 1900, but arrivals from Germany and Ireland were far more numerous during that earlier period.

bottom-right part shows the distribution of Italian descendants in 2010, with relatively large populations in the West and South.

This is, in a nutshell, how we identify exogenous variation in the distribution of ancestry across US counties. We use the interaction of time-series variation in the relative attractiveness of different destinations within the United States (measured by the destination choices of migrants from other continents) with the staggered arrival of migrants from different origins (measured by the number of migrants from that origin that migrated to other regions within the US) to instrument for the present-day distribution of ancestry: if, coincidentally, large numbers emigrated from a given origin to the United States at the same time as a given location within the United States was attracting migrants from all over the world, then we expect large numbers of people in that US location with ancestry from that foreign country.

Importantly, this interaction of 'leave-out push-pull' factors is independent of plausible confounding factors that could make (or have made) a given destination in the US more attractive for both migrations and FDI from a given origin country. Consider the example of Italian migrants in the 1900-30 period. Suppose many Italian migrants skilled at growing wine settled in US regions favorable to wine growing (e.g. Napa county). The same unobserved factor (a climate favorable to wine-growing) may very well explain why there are both residents with Italian ancestry in Napa (descendants of wine makers), and FDI linkages between Napa and Italy (wine making multinationals), creating a spurious (not causal) correlation between ancestry and FDI. Our instruments remove this spurious correlation by predicting the number of Italians migrating to Napa 1900-30 using only the interaction of the share of non-European immigrants who settled in Napa with the number of Italians who settled outside the West Coast. Thus, if wine making ability were the only true driver of migrations from Italy (or other European countries) to Napa (or other counties on the West Coast), our leave-out push-pull instruments would predict zero Napa residents with Italian ancestry today.[20]

The same is true for migrations induced by reverse causality and most other confounding factors that might induce a spurious correlation between ancestry and FDI. For example, if the true cause of the large Italian presence in Napa was that an Italian car manufacturer randomly decided to invest in a Napa-based plant, and historically sent Italian workers to operate it, then this investment would affect realized migrations between Italy and Napa, but would again have no effect on the number of Napa residents with Italian ancestry predicted by our instruments.

[20]Incidentally, Figure 3 shows very few non-Europeans emigrated to Napa and Sonoma in 1900-30 (bottom left map), while many residents of Napa and Sonoma have Italian ancestry (bottom right map). Our identification strategy will *not* capture those descendants of Italian migrants in Napa and Sonoma. This is desirable, as both Italian ancestry and FDI linkages would likely be correlated with unobserved factors (local climate).

For this approach to fail, a confounding factor that promotes migration and FDI would have to disproportionately cause large groups of migrants from two origins on two *different continents* to systematically migrate to the same destinations across at least two *different regions* in the *same census periods*. One example would be if migrants who are skilled at growing wine tended to emigrate from Algeria (a non-European country suitable for wine) towards Napa precisely at the same points in history when Italians also went to Napa and to the Champlain Valley in upstate New York (a non-West Coast region suitable for wine), and if Algerians in Napa were a large fraction of all migrants to Napa, and Italians in Champlain Valley a large fraction of all Italians arriving in the United States. We would then predict a large number of Napa residents with Italian ancestry, because of an unobserved factor (local climate) that directly affects both migrations and FDI from Italy and Algeria to Napa and Champlain Valley. We argue below that such occurrences are unlikely (Algerians did not in fact migrate to Napa in large numbers 1900-30), and offer a series of tests to gauge this possibility.

3.1 Identifying the Causal Impact of Migrations

To formally 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 gravity equation,[21]

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

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, zero otherwise.[22] $A_{o,d}$ 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, but also show robustness to a wide range of alternative specifications—see section 3.5). $X'_{o,d}$ 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

[21] The gravity structure can be derived in a variety of models (Arkolakis et al., 2012). See Carr et al. (2001), Razin et al. (2003), Head and Ries (2008), and Ramondo (2014) for applications to foreign direct investment.

[22] We use this combined measure of inward and outward FDI because our main results are largely identical when separately considering inward and outward FDI. We report separate results for each direction below.

of origin, and between origin and destination census region.[23] The error term $\varepsilon_{o,d}$ captures all omitted influences, including any deviations from linearity.[24] Standard errors are clustered at the origin-country level, and our results are robust to alternative methods for calculating standard errors (see section 3.5). 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.

Equation (1) will consistently estimate the parameter of interest if $Cov\left(A_{o,d}^{2010}, \varepsilon_{o,d}\right) = 0$. As discussed above, this condition is unlikely to hold despite the inclusion of origin and destination fixed effects. First, origin-destination specific omitted factors might drive both economic transactions and migration flows, affecting both $A_{o,d}$ and $\mathbf{1}\left[FDI_{o,d} > 0\right]$. (Skills and climate favorable to wine growing in our example above.) Second, past origin-destination specific migration flows might be the result of economic transactions such as FDI or trade, not their driver.[25] Third, ancestry might be selectively recalled because of past or present economic interactions. 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 based on a simple dynamic model of migration: ancestry evolves recursively from the addition of new migrants to the existing stock of ancestry. We assume the combination of three forces determines the allocation 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.

The following equation is a simple linear formulation of these assumptions.

$$A_{o,d}^{t} = a_t + a_{o,t} + a_{d,t} + b_t A_{o,d}^{t-1} + I_o^t \left(c_t \frac{I_d^t}{I^t} + d_t \frac{A_{o,d}^{t-1}}{A_o^{t-1}} \right) + \nu_{o,d}^{t}. \tag{2}$$

The stock of residents of ancestry o in destination d at time t, $A_{o,d}^{t}$, depends on four terms. First, the constant terms a_t, $a_{o,t}$, and $a_{d,t}$ control for residual forces, such as demographics, which may vary over time, space, and between different ethnic groups. Second, the term $b_t A_{o,d}^{t-1}$ corresponds to the fact that ancestry is a stock variable that evolves cumulatively, where b_t modulates how ties to one's ancestry are passed from one generation to the next, including attenuation due to

[23] A census region is one of nine groupings of adjacent US states listed in Appendix Table 5.

[24] We use a simple linear probability model, which allows for a straight-forward interpretation of the coefficient. As a robustness check, we also report results from a probit estimator; see footnote 29.

[25] A real-life example of such reverse causality is the large Japanese ancestry in Scott County, Kentucky, which emerged after Toyota seconded Japanese workers to a newly built manufacturing facility in the 1980s.

internal migrations. Third, the term $I_o^t\left(c_t I_d^t/I^t + d_t A_{o,d}^{t-1}/A_o^{t-1}\right)$ is a linear interpretation of our assumption that net migrations are determined by the combination of push, pull and recursive factors. The push factor (the extent to which migrants are driven out of country o) is measured by the total number of migrants from country o entering the United States at time t, I_o^t. The pull factor (the degree to which county d is appealing to migrants at time t) is measured by the fraction of all migrants entering the United States who settle in county d from all origins, I_d^t/I^t. The recursive factor (the propensity of migrants to settle near their countrymen) is measured by the fraction of people with ancestry from o who already live in d, $A_{o,d}^{t-1}/A_o^{t-1}$. Intuitively, we assume that part of the allocation of new migrants from o across counties (the push factor I_o^t) is proportional to the allocation of migrants from all countries (the pull factor I_d^t/I^t), and part is proportional to the allocation of the existing stock of migrants from o (the recursive factor $A_{o,d}^{t-1}/A_o^{t-1}$). The coefficients c_t and d_t control for the relative importance of the pull and recursive factors. If the pull factor is absent ($c_t = 0$), 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 because persistent forces are likely to shape both migrations and FDI, inducing a correlation between $A_{o,d}^{t-1}$ and $\varepsilon_{o,d}$.[26] 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.

We address this challenge by using the recursive structure of equation (2). Given that our data cover the vast majority of migration to the United States (more than 70 million immigrants, including the entire first and second generation of immigrants alive in 1880), we assume the initial condition $A_{o,d}^{1880\text{“}-1\text{”}} = 0, \forall\,(o,d)$ for simplicity. Solving (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. 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, not specific to destination d but to all destinations in the United States, and a pull factor, I_d^t/I^t, not specific to country o but to migrants from all countries, this formulation already rules out most plausible sources of endogeneity. However, our exclusion restriction could still be violated if $I_{o,d}^t$, or migrations from other origins similar to o, potentially correlated with $\varepsilon_{o,d}$, were a large fraction of I_o^t, I_d^t or I^t.

To address these concerns, we exclude migrants to d's census region from the push factor (we

[26] In the example above, a favorable climate for growing wine induces both migrations from Italy to Napa in 1900-30, a high $A_{o,d}^{t-1}$ term, and many Italy-Napa wine making multinationals in 2014, resulting in a high $\varepsilon_{o,d}$.

replace I_o^t by $I_{o,-r(d)}^t$ in (3), where $-r\,(d)$ means outside of d's census region), and migrants from o's continent from the pull factor (we replace I_d^t/I^t by $I_{-c(o),d}^t/I_{-c(o)}^t$ in (3), where $-c\,(o)$ means outside of o's continent). 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 758 higher-order terms $I_{o,-r(d)}^s \cdots I_{o,-r(d)}^t I_{-c(o),d}^t / I_{-c(o)}^t, \forall t < s \leq 2010$.[27]

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 destination d more attractive for FDI from origin o do not simultaneously affect the interaction of the settlement of migrants from other continents with the total number of migrants from o 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 ($\delta_o \times \delta_{r(d)}$). These specifications are, by construction, robust to any confounding factors that operate within an origin-continent - destination-census-region pair. (In our example above, wine makers from Europe migrating to the West Coast.)

3.2 The First-Stage Relationship

Table 2 shows our basic first-stage estimates of (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 $\{I_{o,-r(d)}^t (I_{-c(o),d}^t / I_{-c(o)}^t)\}_t$. To interpret each coefficient as the marginal effect of migrations in a given period, without affecting the fit of the first stage, we sequentially orthogonalize each of the terms with respect to the previous interactions. For example, the coefficient marked $I_{o,-r(d)}^{1910} (I_{-c(o),d}^{1910} / I_{-c(o)}^{1910})$ shows the effect of the residual obtained from a regression of $I_{o,-r(d)}^{1910} (I_{-c(o),d}^{1910} / I_{-c(o)}^{1910})$ on the same interaction in 1880 and 1900.

All nine coefficients shown in column 1 are positive, and seven are statistically significant at

[27] Principal component analysis (eigenvalue decomposition) is simply a means for compactly summarizing the variation contained in the 758 higher-order terms. In our standard specification, the first five components summarize 99.99% of the variation. To the extent that the higher order terms are valid instruments, the first five principal components are valid instruments as well. Our results are robust to adding these terms or not.

the 1% level (Appendix Figure 1 depicts the coefficients graphically; Appendix Figure 4 depicts the fit of each coefficient). Even our earliest (pre-1880) snapshot of the cross-sectional variation in economic attractiveness to new migrants has left a significant imprint on the present-day ancestry composition of US counties. 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 smaller and insignificant coefficient for 1920-30. A likely explanation is the Great Depression, which induced large reverse migrations 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 controls for distance and latitude difference. Columns 3 and 4 add destination × continent-of-origin and origin × destination-census-region fixed effects, respectively. Columns 1-4 estimate (4) under the restriction that the recursive factor is irrelevant ($d_t = 0$ in (2)). We relax this in column 5 and add the principal components of the higher order interaction terms. Column 6 includes third-order polynomials in the distance and latitude difference between o and d. Column 7 includes migration data from the 2005-2010 ACS survey. Column 8 drops migration prior to 1880. Column 9 estimates (4) in levels rather than logs.

Our standard specification is in column 5. The Kleibergen-Papp Wald rk-statistic against the null of weak identification is 162.2, well above the Stock and Yogo critical values.[28] We reject the null that our instruments are jointly irrelevant in the first stage across all specifications.

3.3 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. The dependent variable is a dummy equal to one if either a parent foreign firm from origin country o owns a US subsidiary in destination US county d (inward FDI), or if a US parent in d owns a foreign subsidiary in o (outward FDI). The results are in Table 3.

In column 1 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. Appendix Figure 2 shows the corresponding

[28] 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.

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 FDI with that origin today. The coefficient of interest falls slightly to 0.190 (s.e.=0.024) in column 2 when we add the first five principal components of the higher-order interactions to our set of instruments. 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.[29] This specification includes destination × continent-of-origin fixed effects and origin × destination-census-region fixed effects. 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 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).[30]

Another useful way to gauge the relative importance of ancestry is its partial R^2 relative to the other controls. Taken together, the standard gravity terms, that is, 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 (panel B) raises the R^2 by 9 percentage points, 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).[31]

The remaining columns of Table 3 probe the robustness of this result. Column 4 adds a third-degree polynomial in distance and latitude difference to capture non-linear effects of distance; column 5 adds an interaction term for the contemporaneous 2010 migrations in the first stage

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

[30] Naturally, the OLS coefficient may also be biased downward simply because ancestry is measured with error.

[31] 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 by 2 percentage points. These numbers are a lower bound on the importance of common ancestry for FDI, since it only accounts for the part of the causal effect of ancestry which is picked up by our instruments.

(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. The coefficient estimate remains remarkably stable and highly statistically significant across specifications.

3.4 The Communist Natural Experiment & Alternative Instruments

The main potential challenge to these results is that, despite our best efforts, confounding factors that make a destination more attractive for both migration and FDI from a given origin 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. In this section, we address this challenge using a natural experiment and a set of alternative instrumentation strategies.

Communist Natural Experiment. We begin by combining our IV with a natural experiment that allows us to focus on *changes* in FDI and *changes* in ancestry: the periods of economic isolation between the United States and communist countries during parts of the 20th century. These periods are 1918-90 for the Soviet Union, 1945-80 for China, 1975-96 for Vietnam, and 1945-89 for Eastern Europe (the non-Soviet members of the Warsaw pact). They provide a useful experiment for two reasons. First, we can confidently assume the prospect of FDI, outlawed for political reasons, did not drive migrations during those periods (ruling out reverse causality). Second, the specification is similar to a difference-in-difference, measuring 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.

Table 4 shows estimates of (1) for each of these countries or sets of countries, using as instruments only migration waves that occurred during the period of isolation. For each country, we find a large and significant causal impact of ancestry on FDI that is remarkably similar to our full-sample estimates from Table 3. (For example, 0.185 (s.e.=0.019) for the Soviet Union alone versus 0.187 (s.e.=0.024) in our standard specification from Table 3.) An exception to this rule is Vietnam, which shows a coefficient less than half the size of the other cases (0.089, s.e.=0.036). Plausibly, this lower coefficient reflects the fact that the majority of Vietnamese immigrants who arrived during the 1975-96 period were granted entry for aiding the US cause during the Vietnam war, and the communist government who defeated them is still in power in 2014 and controls FDI. Vietnamese Americans might thus plausibly be in a worse position than descendants of migrants from other countries to generate FDI between the US and their ancestral country. Pooling across all former Communist countries, we again find a coefficient very close to that of our standard specification in Table 3 (0.206, s.e.=0.031). These results strongly suggest

reverse causality does not drive our results, and our exclusion restriction is likely valid.

Alternative Instruments. The main remaining challenge to our approach is that a common unobserved characteristic of destinations in two different census regions may still directly affect FDI, while also disproportionately causing large groups of migrants from two origins on two different continents to systematically migrate at the same time to the same destinations across multiple census regions (e.g. the Italy-Algeria and Napa-Champlain Valley example above).

If such a confounding factor were driving our results, we would expect that excluding from the construction of our instruments countries with either correlated migrations, or correlated ancestry, would have a large effect on the coefficient of interest. We show below they do not.

The first specification in Panel A of Table 5 excludes from the construction of our pull factors foreign countries that tended to send migrants towards the United States at the same time as a given origin country. For each pair of countries, we compute the correlation coefficent over time of aggregate emigration to the US, and exclude from the (o, d, t) pull factor all migrations to d at t from origin country o' if the correlation with o is above 0.5 and significant at the 5% level. Panel A of Appendix Figure 5 shows these correlations graphically: while correlations within continents (along the diagonal) tend to be larger, there is also a strong correlation in the timing of migrations between some African, American, and Asian countries.[32] Using this alternative approach, we exclude on average 65 countries, of which on average 18 are in the same continent (our baseline leave-out category).

The second specification instead excludes from our instruments foreign countries which have a similar distribution of ancestry within the United States in 2010. For each country pair, we compute the correlation across destination counties of the number of residents with a given ancestry, and again exclude origin countries if the correlation coefficient is above 0.5 and significant at the 5% level. Panel B shows these correlations graphically, with a strong correlation in the distribution of ancestry between Asian and European countries. Using this approach, we exclude on average 36 countries, of which on average 13 are in the same continent.

Panel A of Table 5 shows that our coefficient estimates vary little with these alternative sets of instruments: 0.181 (s.e.=0.027) when excluding countries with correlated migrations over time; and 0.217 (s.e.=0.030) when excluding countries with correlated ancestry over space. The coefficient also remains stable when we exclude migrations to all adjacent states, $I^t_{o,-adj(d)}(I^t_{-c(o),d}/I^t_{-c(o)})$, rather than the surrounding census region (0.192, s.e.=0.022). Appendix Tables 6 and 7 show

[32] Most, if not all, forces that induce such correlations in practice are fully innocuous for our identification. For example, changes in US immigration laws or innovations in transportation technologies might simultaneously affect the push factors of several countries. In this sense, there is no general reason to exclude certain countries with high correlation from the construction of our instruments.

that the same is true across a wide range of similar variations in our leave-out categories: our coefficient estimates remain stable across all variations that exclude at least the own origin-destination pair. We conclude that our original approach (excluding the own continent-of-origin and destination-census-region) is valid and that no spurious correlations are driving our results.

Panel B of Table 5 presents results using subsets of our instruments. The first row uses as instruments only the simple interactions from the first half of the time period covered by our migration data (1880-1930), the second row only from the second half (1970-2010), while the third row excludes migrations from the first census (1880) from the set of our instruments, as these might be more related to stocks than flows. The coefficient of interest again remains stable at 0.209 (s.e.=0.037), 0.175 (s.e.=0.021), and 0.187 (s.e.=0.024) respectively.[33] Appendix Table 8 replicates our results using data on FDI from 2007 rather than 2014 and data on ancestry from 2000 rather than 2010, again with little effect on our results. We conclude that our results are not driven by specific vintages of migrations, pre-, post-WWII, or other.

3.5 Additional Robustness Checks & Intensive Margin

Functional Form. In our main specification, we measure our ancestry variable as the log of one plus the number of residents with foreign ancestry, measured in thousands, $A_{o,d}^{2010} = \ln\left(1 + \frac{1}{1000} Ancestry_{o,d}^{2010}\right)$. In Appendix Table 9, we offer a formal test to justify this choice of functional form by performing a non-linear least squares estimation of

$$\mathbf{1}\left[FDI_{o,d} > 0\right] = \delta_o + \delta_d + \beta \ln\left(1 + \pi Ancestry_{o,d}^{2010}\right) + X'_{o,d}\gamma + \varepsilon_{o,d}, \tag{6}$$

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 $\ln\left(1 + Ancestry_{o,t}/1000\right)$ 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 o, (1) means that increasing ancestry by 1,000 increases the probability of at least one FDI link to o by β percentage points; for a large number of residents with ancestry from o, increasing ancestry by 1% increases the probability of at least one FDI link by β percentage points. Appendix Figure 6 plots the average number of FDI links across centiles of the distribution of ancestry, and shows the effect of ancestry on FDI is highly concave.

[33] The first two coefficients are not statistically different; the p-value for a test of equality is 0.202.

In Appendix Table 10, we further sow the robustness of our results to using a range of alternative functional forms, and to using measures of ancestry from the, 1980, 1990 and 2000 censuses, instead of 2010. Appendix Table 11 shows our main results are robust to varying the cutoff for ownership at which we consider a foreign firm to be a subsidiary or parent (from 5% to 50%). Appendix Table 8 shows our results are robust to using FDI in 2007 (instead of 2014).

Standard Errors. Appendix Table 12 shows our standard specification from column 3 of Table 3 using alternative standard errors. It reports robust standard errors; standard errors clustered by origin, destination, state, continent, and state-country cells. Among all these simple analytic standard errors, clustering by origin, as we do throughout the paper, is the most conservative choice. Doing so allows for arbitrary correlation in the error term across multiple destinations for a given origin, including for spatial correlation of errors. The specification we use throughout the paper thus allows for more flexible patterns of spatial correlation than for example the standard error correction as proposed by Conley (1999).

A possible concern is that errors may still be correlated across origin countries. However, standard errors designed to adjust for such correlations (clustering by county or state) are narrower, suggesting that any such patterns in the error structure are – if they were present – absorbed by the rich set of fixed effects and controls contained in our standard specification. Consistent with this view, the table also shows that standard errors double clustered at county-plus-country and state-plus-country level, as well as various block-bootstrapped standard errors are either narrower or only very marginally wider than those in our standard specification. The conclusion that our results are robust to alternative standard error specifications also carries over to the other main results of our paper (see Appendix Table 13).

An alternative approach to detecting 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 assign the interaction between push and pull factors for country o to randomly selected other countries and calculate the t-statistic on the coefficient of interest. Reassuringly, across 1000 random assignments, the t-statistic rejects the null of no treatment effect in favour of the alternative of a positive treatment effect in only 2.7% of the cases.

Placebo Tests. To assess whether our instrument reliably isolates push factors specific to only one country, or is correlated with omitted variables that affect FDI with other countries, Appendix Table 14 assigns the interaction between push and pull factors for a given origin to a quasi-randomly selected other country: its nearest neighbor in alphabetic order (panel A), or its nearest neighbor in alphabetic order in a *different* continent (panel B). The coefficient estimates

are always near zero and statistically insignificant, suggesting our instrument is not picking up any artificial correlation (positive or negative) between the push factors for different countries.

Robustness in Sub-Samples & Heterogeneous Effects. Figure 4 shows results from separate regressions for all 112 origin countries (left panel) and the 100 largest US counties (right panel).[34] Each figure is a funnel plot of the country/country-specific coefficients on ancestry against the reciprocal of their standard errors, where the circles reflect the relative shares in ancestry and US population, respectively. The coefficients are significant at the 5% level for 84 out of 112 countries and 99 out of 100 counties. This further demonstrates that our results are not driven by any specific subsample or outliers.

Inward and Outward FDI. We estimate our standard specification from column 3 of Table 3 separately for inward FDI, 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, 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, statistically significant, and close to our baseline estimates. We find a somewhat stronger impact of ancestry on outward FDI, $\beta_{out} \approx 0.2$, than on inward FDI, $\beta_{in} \approx 0.15$, although the coefficients are not statistically different.

The Intensive Margin of FDI. Figure 5 shows ancestry has a positive and significant impact on the intensive margin of FDI (the number of foreign subsidiaries), with no obvious outliers, for the largest countries and counties in our sample: Germany and Britain (top parts), and LA and Cook counties (bottom parts). Appendix Table 19 systematically estimates the impact of ancestry on the intensive margin of FDI,

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

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.[35] We use various measures for the volume of FDI: in panel A, the total number of FDI relationships; in panel B, the number

[34] Appendix Figure 7 shows a similar figure for six individual sector groups. Appendix Table 15 shows results from separate regressions for the five largest origins (by number of descendants, panel A) and destinations (in total number of foreign ancestry, panel B). Appendix Tables 16, 17, and 18 show the results from separate regressions for all countries, all sectors, six sector groups, and small and large firms.

[35] To correct the selection bias from dropping log(zero)'s, we implement a Heckman (1979) correction, as in Helpman et al. (2008). We first estimate an IV probit regression at the extensive margin,

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

with $A_{o,d}^{2010}$ instrumented as in equation (4). We extract the predicted latent variable that determines non-zero FDI, $\hat{z}_{o,d} = \Phi^{-1}\left(\hat{\rho}_{o,d}\right)$, and include the inverse Mills ratio $\hat{\mu}_{o,d} = \varphi\left(\hat{z}_{o,d}\right) / \Phi\left(\hat{z}_{o,d}\right)$ within our set $X_{o,d}$ of controls.

of firms in d which are a subsidiary of a firm in o, a measure of inward FDI; and in panel C, the total local employment in county d at subsidiaries of firms in o.

Across all specifications, 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 (7), is statistically and economically 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 the number of FDI relationships by 6.5% and local employment at subsidiaries of foreign firms by 7.3%.[36] 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.

Trade versus FDI. In Appendix Table 20, we test whether ancestry has a similar impact on the intensive margin of FDI (panel A) as on trade (exports in panel B, and imports in panel C). We use readily available data on trade flows between US states and foreign countries sourced from the US Census Bureau. We again instrument for the composition of ancestry as in (4), except that all variables are defined at the state, not county, level. We correct for the selection bias due to zero trade using a Heckman (1979) correction as above.

The impact of ancestry on FDI at the state-level is positive, significant, and larger than on trade, once we instrument and include both origin and destination effects (column 3).[37] The effect of ancestry on trade becomes indistinguishable from zero in our preferred specification in column 3. Although we interpret this non-result with due caution due to the limited data available, it contrasts with 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 (2016) for trade with Vietnam, that all find the presence of migrants facilitates exports. Our preferred specification shows no such positive and statistically significant effect. 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 (column 3 panels B and C), we find none.

Ancestry and Immigration. According to our reduced-form model (2), migrations are driven by economic attractiveness (the interaction of our pull and push factors) and the stock of

[36] Using $\hat{\kappa} = 0.326$ in panel C, column 3 of Appendix Table 19 in equation (7), we have: $\frac{Employment_{o,d}[Ancestry_{o,d}=2\times316]}{Employment_{o,d}[Ancestry_{o,d}=316]} - 1 = \exp\left(0.326\left(\ln\left(1+\frac{2\times316}{1000}\right) - \ln\left(1+\frac{316}{1000}\right)\right)\right) - 1 \approx 0.073$.

[37] 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 for final goods and for intermediate inputs.

ancestry (the recursive factor). To provide direct evidence for these two forces, we estimate

$$I^t_{o,d} = \delta_o + \delta_d + \theta\ I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} + \lambda A^{t-1}_{o,d} + X'_{o,d}\gamma + \vartheta_{o,d} \tag{8}$$

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 6, Panel A estimates (8) with immigration $I^t_{o,d}$ in levels, and (reassuringly) gives a coefficient on the interaction of the push and pull factors close to 1. Columns 2 and 3 estimate (8) in logs for two time periods, 1990 and 2000. Across all specifications, both the coefficient on the push $\times$ pull interaction and on lagged ancestry are positive and significant predictors of current migrations.

In Panel B of Table 6, we show recent changes in FDI are predicted by recent changes in ancestry. In column 1, we add a dummy for FDI in 2007 (the earliest year for which we have data) to our standard specification from column 3 of Table 3, so that the coefficient of interest now reflects the effect of ancestry on *changes* in the extensive margin of FDI between 2007 and 2014.[38] As expected, we find a smaller, but nevertheless positive coefficient (0.086, s.e.=0.018). Column 2 shows almost identical results when using only migrations between 1980 and 2000 for identification. Column 3, shows the same effect at the intensive margin (0.142, s.e.=0.049).

3.6 Quantifying the Effect of Ancestry on FDI

We illustrate the quantitative implications of our findings using two thought experiments. First, we estimate how FDI linkages with 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 in Los Angeles 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 long-term effect of immigration policies 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, essentially closing the United States to legal immigration of laborers from China. In 1943, 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 period of "Chinese

[38] The magnitude of this effect should be interpreted with caution as coverage of the ORBIS data has expanded between 2007 and 2014 so that the size of the effect will reflect both changes in FDI and changes in coverage.

Exclusion." How different would the ancestry composition and FDI of US counties be today had it not been for Chinese Exclusion?

To answer this question we require an estimate for the impact of Chinese exclusion on the number of immigrants from China. We use our own data to derive a rough estimate. We aggregate our immigration data at the time × census-region × origin level and estimate $I^t_{o,r} = \delta_{t,r} + \delta_o - \xi \cdot D^t_{China} + \nu_{t,o,r}$, 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 $\hat{\xi}$ can then be interpreted as the estimated average negative impact of the Chinese Exclusion Act on Chinese migrations. We then calculate a hypothetical time path of immigration in the absence of Chinese exclusion as $\tilde{I}^t_{o,r} \equiv I^t_{o,r} + \hat{\xi} \cdot D^t_{China}$. It suggests that the United States would have received 1.8 million additional Chinese immigrants during the period of exclusion.

Given this hypothetical time path of immigration we can then use our estimates from Table 2 to predict the change in ancestry 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 $d\Pr[FDI_{o,d} > 0] \equiv \hat{\beta} \cdot dA_{o,d}$, where $\hat{\beta}$ is the estimated second-stage coefficient in a specification as in column 3 in Table 3, excluding the principal components to be consistent with the above described methodology to predict hypothetical levels of ancestry.

These calculations suggest that the increase in Chinese migration would have been highly unequally distributed, translating into heterogenous 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, $d\Pr[FDI_{China,d} > 0]$. 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 $d\Pr[FDI_{China,d} > 0]$ 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 and Illinois.[39]

A Gold Rush in Los Angeles. To similarly gauge the magnitude of the estimated intensive margin effects, we derive predictions for the intensity of FDI relationships between Los Angeles county and the world under the hypothetical scenario that Los Angeles had experienced a Gold

[39] Although we believe, in light of our results, that additional Chinese immigration likely also raises FDI between China and the United States as a whole, we cannot exclude the possibility that increases in FDI with China in one state are partially or fully offset by decreases in others.

Rush similar to that in San Francisco, a fivefold increase in the pre-1880 number of immigrants to Los Angeles. Appendix Table 21 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. Columns 1 and 2 show the actual number of individuals of each ancestry in Los Angeles County in 2010 and 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 based on the IV specification corresponding to column 2 of Appendix Table 19, again without the principle components as instruments. A Gold Rush in Los Angeles would have resulted in sizeable effects on the intensity of FDI with those countries that were the source of immigration pre-1880: the intensity of FDI between Los Angeles county and Germany and Ireland 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.

4 Understanding the Effect of Ancestry

So far, we have documented a quantitatively large causal effect of common ancestry on FDI. We now turn to the mechanism linking ancestry to FDI. Existing research suggests that migrations and common ancestry may affect FDI either by making the destination more "similar" to the origin in terms of preferences and skill endowments or by generating social capital that creates an absolute advantage for firms to operate in both the origin and destination country.

In the first category, Atkin (2010) and Bronnenberg et al. (2012) suggest that descendants of migrants may share the same tastes for foods and other products as consumers in their origin country. To the extent that these tastes persist over generations, firms that cater to those tastes may serve both markets. Similarly, we might suspect that migrants may bring with them a specific skill-mix or other factors abundant in their origin country, so that firms can more easily outsource production, using the same skill-mix at home and abroad.

In the second category, common ancestry may create an absolute advantage in conducting FDI for local firms, because social ties between populations in the origin and the destination provide social collateral that helps to enforce contracts when the legal system of o or d is imperfect (Greif, 1993; Besley and Coate, 1995). Alternatively, migrants and descendants of migrants from a given origin may have privileged access to information (Varian, 1990; Stiglitz, 1990): a more intimate

knowledge of the business environment in their origin country and social ties or language skills that provide access to information about business opportunities and practices at a lower cost.

The following section presents evidence testing auxiliary predictions of these distinct channels. The collection of these results suggests that ancestry affects FDI primarily because it creates an absolute advantage for local firms by attenuating information frictions.

4.1 Channel Linking Ancestry to FDI

Panels A-C of Table 7 present indirect tests for the 'similarities' hypothesis. Panel A shows the IV coefficient of ancestry on FDI separately for firms producing final goods and for firms producing intermediate inputs.[40] If common tastes were the explanation behind the positive impact of ancestry on FDI, we would expect its impact to be stronger for final goods, for which consumers' tastes matter directly, than for intermediate inputs, for which tastes matter little. We find, on the contrary, that there is no significant difference between final goods and intermediate input producers; if anything, the point estimate is slightly larger for intermediate input producers than for final goods producers. Panel B shows similar results for inward FDI only, where the local tastes of descendants from country o may plausibly matter more.

Panel C of Table 7 shows the IV coefficients of a regression of ancestry on measures of sectoral similarity between the origin and the destination. For each origin-destination pair, we compute the rank and cosine correlation of the shares of employees in 127 manufacturing sectors.[41] Both correlations increase with the similarity of the allocation of employees across sectors between the origin and destination. If skill similarities were the explanation behind the positive impact of ancestry on FDI, we would expect common ancestry to cause an increase in these measures. We find, on the contrary, that ancestry has no discernible impact on sectoral similarity. This non-result—migrations do not cause a convergence in the sectoral composition of employment—is robust to using alternative measures of sectoral similarity, as well as alternative data sources.[42]

Panel D of Table 7 examines the contract enforcement channel. If contract enforcement were the explanation behind the positive impact of ancestry on FDI, we would expect the impact to be stronger for countries where the quality of the local judiciary is weaker, as ethnic ties would

[40] We use the upstreamness index from Antràs et al. (2012) to define sectors producing final versus intermediate goods: a sector is labelled as final goods (intermediate input) if its upstreamness index is below (above) 2.

[41] We use county and country level industry data from the Bureau of Labor Statistics (BLS) and the UN Industrial Development Organization (UNIDO), respectively. Correlations are calculated for 2006, the year with the largest availability of data (28 countries). Using this smaller sample of countries, the coefficient on ancestry in our standard specification linking ancestry to FDI is 0.348 (s.e.=0.046).

[42] Results are unchanged whether we use rank or cosine correlation, or when we repeat the same exercise using the OECD Stan country level industry data.

substitute for weak institutions. We find the opposite result in columns 1 (extensive margin) and 2 (intensive margin), where we add the interaction of ancestry with a measure of the origin country's judicial quality taken from Nunn (2007) to our simple IV specification of column 2 in Table 3. The coefficients show that the effect of ancestry on FDI is significantly larger for countries with good institutions than for countries with bad institutions, suggesting that common ancestry and good institutions are complements rather than substitutes.

We conclude from Table 7 that the data show no evidence for the 'similarities' hypothesis and that our results do not seem driven by ethnic ties substituting for poor contract enforcement.

4.2 Information Demand

To directly test the remaining hypothesis that common ancestry affects FDI by reducing information frictions, we require data on the flow of information from foreign countries to US destinations. Because such data is not readily available we construct a simple index of differential information demand using data from internet searches.[43] The Google Trends portal provides data on the relative popularity of different search terms across 210 US metropolitan areas ("media markets" according to the Nielsen DMA definition).[44] For each search term i and US metro area d, the portal returns an index number that is equal to the normalized share of searches conducted in d that contain the search term i:

$$G(i,d) = \left[100 \frac{share(i,d)}{max_{\delta}\{share(i,\delta)\}} \mathbf{1}[\#(i,d) > T] \right],$$

where $share(i,d)$ is the share of searches in d that contains i and $\mathbf{1}[\#(i,d) > T]$ is an indicator function that is one if the absolute number of searches containing i in d is greater than some threshold number (Stephens-Davidowitz and Varian, 2015; Liang, 2017).[45] Thus, $G(i,d)$ is equal to 100 in the metro area in which the largest share of searches contain i and a positive number smaller than 100 in all other metro areas that have a sufficient number of searches containing i.

To measure the relative demand in a given metro area for information about a given origin country, we compile a list of the five most prominent actors, athletes, musicians, and politicians for each origin country. We automate this process by searching for "notable [country] [category]" and then extracting the five top suggested names from the Google Answer Box, a feature of

[43] We thank Jack Liang for writing his Bachelor Thesis at the University of Chicago on this topic.

[44] For other recent studies using this data source see Da et al. (2011), Stephens-Davidowitz (2014), Kearney and Levine (2015), and Baker and Fradkin (2016).

[45] As a result of this cutoff, our index tends to assign a value of zero to small origin-destination pairs (38% of our sample). For this reason we focus our attention on the 100 largest origin countries by 2015 population and do not attempt to construct it for all present-day origin countries.

Google search that automatically suggests the most often clicked names associated with this kind of query.[46] We then calculate our Information Demand Index as

$$IDI(o,d) = \frac{1}{20}\sum_{p}\sum_{i \in q(o,p)} G(i,d),$$

where $q(o,p)$ is the set of top five names for country o in category $p \in \{\text{actors, athletes,...}\}$. We implement this procedure for the 100 largest foreign countries by 2015 population. To facilitate the interpretation of results, we standardize this measure to a unit standard deviation.

Panel A of Table 8 shows the results of regressions of our differential information demand index on ancestry (instrumented as in (4)), and our standard set of controls at the metropolitan area - country level. Column 1 documents a large causal effect of ancestry on information demand (0.871, s.e.=0.257), where doubling the number of descendants of migrants from a given origin relative to the mean is associated with a 0.19 standard deviation increase in our index of demand for information about prominent actors, athletes, musicians, and politicians in that origin.[47][48] Columns 2 and 3 show that this effect remains positive and statistically highly significant even when we control for the foreign-born population (first-generation immigrants), and when we condition only on ancestry in 1980 (rather than 2010). Taken together, these results suggest that the differential interest in information about the origin country persists among the US-born descendants of first-generation migrants. The remaining columns show that this persistent interest in the origin country is not limited to politics, but is similar across our four sub-indices for demand for information about actors, athletes, musicians, and politicians.

The longevity of the effect of ancestry on differential information demand suggests that immigrants pass traits to their descendants that facilitate or encourage the exchange of information with their origin countries, such as social ties to family or friends, or knowledge of the origin country's language and culture. Although data on such traits is generally hard to come by at the required level of disaggregation, Panel B shows one additional piece of evidence from the US census: the use of foreign languages. The table shows systematic evidence that a larger community in county d with ancestry from country o has a positive and significant impact on the number of residents in d who speak o's language at home (column 1). This effect persists

[46] See Appendix Table 22 for the full list of search terms used for Germany and Italy as examples. Liang (2017) shows evidence that search terms with multiple meanings (e.g. two prominent politicians from two different countries share the same name) do not impact our results, and gives a detailed account demonstrating that the Google Answer Box generally delivers relevant search terms for each country. See Appendix A.4 for details.

[47] Following the same calculation as above we have $0.871\left(\ln\left(1+\frac{632}{1000}\right)-\ln\left(1+\frac{316}{1000}\right)\right)=0.19$.

[48] Complimentary evidence to ours is provided in Bailey et al. (2016), who find that recent immigrations from origin country o to US county d, as well as the composition of ancestry across US counties d, are close correlates of a measure of social ties derived from Facebook friendship links.

if we remove from d's population all foreign-borns, since they 'mechanically' speak the foreign language from their home country (column 2). Columns 4-6 present the results from separate regressions for large non-English languages: Spanish, Arabic, Chinese and Hindi.[49] The effect of ancestry on foreign languages spoken is positive and statistically significant for all four.

4.3 Ancestry, Information Flow, and FDI

We next ask whether these differences in information flows can account for the the link between ancestry and FDI. We get a superficial answer to this question in Appendix Table 24 by running our simple specification relating ancestry to FDI (column 2 in Table 3), while controlling for our information demand index. We find that the coefficient on ancestry drops to close to zero and becomes statistically insignificant (-0.025, s.e.=0.028), while the coefficient on information demand remains positive and highly statistically significant (0.078, s.e.=0.013). By contrast, controlling for sectoral similarity and the various other channels probed above has no effect on the causally identified coefficient on ancestry. Our results thus suggest the effect of ancestry on FDI transits through the information channel. We study next the mechanisms through which information transmission may generate an absolute advantage in conducting FDI.

Network effects of common ancestry. Theoretical models emphasize the role of networks in facilitating the percolation of information across international borders (Arkolakis, 2010; Chaney, 2014). This class of models tends to predict effects that are concave (as all the relevant information is gradually exhausted), weaker if many people from the same or neighboring origins live in the surrounding area (as relevant information is more likely to have already percolated), and stronger for destinations that are more ethnically diverse (hubs open up more paths for information to percolate through). We test each of these reduced-form predictions.

We have already shown in section 3.5 that the relationship between ancestry and FDI is concave (see Appendix Figure 6). Information percolation on a network also suggests negative spillovers from neighboring regions: if many people with ancestry from o live in locations surrounding d, or if many people in o have an ancestry from countries adjacent to o, it is more likely that relevant information about investment opportunities has already reached the firm, so that the marginal impact of ancestry on FDI is mitigated. In Table 9, panel A, column 1, 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 sep-

[49] Appendix Table 23 presents the regression coefficients of ancestry on foreign languages for the 50 largest linguistic groups.

arate set of instruments, such that both endogenous variables are identified. The coefficient on our measure of ancestry at the state level is -0.015 (s.e.=0.004), suggesting a negative and significant spillover. In column 2 we include a measure of the number of descendants from the closest neighboring country, and we again find a negative and highly significant spillover effect.[50] Our findings are thus consistent with the presence of negative spillovers within co-ethnic networks.

In columns 3 and 4, we repeat the same estimation for the intensive margin of FDI, but in this much smaller sample we lack the statistical power to identify a consistent pattern.

An additional prediction of network-based models is the existence of hubs, dense locations through which distant locations can connect. If such hub-effects were at work, we would expect the effect of ancestry on FDI to be larger in more ethnically diverse destinations. We test this prediction in panel B of Table 9. We interact ancestry with a measure of ethnic diversity (measured as 1 minus the Herfindhal index of ancestry shares). The coefficient on this interaction is positive and significant, both at the intensive and the extensive margin (column 2 and 4). By contrast, we find no effect on the interaction between ancestry and the share of the population that are of foreign descent (column 1 and 3). This suggests the effect in columns 2 and 4 is driven by ethnic diversity, not by the size of the population share with foreign descent.

We conclude that the patterns by which ancestry affects FDI are consistent with the auxiliary predictions of models of network effects, where information (or other effects of social capital) are transmitted internationally through networks created by common ancestry.

Cost of Information Transmission. If information frictions indeed were the explanation behind the positive impact of ancestry on FDI, we would also expect this effect to be stronger in relationships that suffer from higher costs of information transmission. For example, information costs may be larger for distant countries or countries that are ethnically diverse.

We confirm this prediction in the Panel C of Table 9. In all specifications, the coefficient on the interaction between ancestry and geographic distance is positive and significant, both for the extensive (columns 1 and 2) and intensive margins of FDI (columns 3 and 4).[51] Columns 2 and 4 also show some evidence that the effect of ancestry on FDI is larger for more ethnically diverse origins (a higher level of ethno-linguistic fractionalization as defined by Alesina et al. (2003)).[52]

[50] We determine the nearest adjacent country by creating country pairs, using a standard optimal non-bipartisan matching algorithm, such that the average distance between centroids of country pairs is minimised.

[51] Once we account for this interaction, the interaction terms on genetic, linguistic, and religious distance, as defined by Spolaore and Wacziarg (2015), are statistically insignificant, suggesting geographic distance effectively summarizes alternative notions of distance in cultural space.

[52] Results are virtually identical when we consider outward FDI by itself.

4.4 Generational Effects

Having already shown in section 3 that historical migrations prior to World War II had causal effects on FDI that persist to the present day, and that historical migrations predict future migrations through a recursive factor, we now ask whether the effect of ancestry on FDI requires a sustained inflow of migrants from the same origin—or if it persists even after migration from that origin ceases. Appendix Table 25 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 a measure of foreign born from a given origin alive in 2010, instrumenting as in equation (4). The effect remains positive and significant (unsurprisingly, as the correlation between the two variables is 0.59). 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. In the IV specification in column 4 the coefficient turns slightly negative and marginally statistically significant. This suggests the presence of the descendants of immigrants continues to predict FDI even after migrations from the origin cease.

Using the number of foreign born in 1970 as a proxy for second-generation immigrants, column 6 compares the marginal effect of second-generation immigrants to that of the average resident with foreign ancestry. The coefficient on second-generation immigrants remains positive (albeit not significant) when we control for descendants of migrants.[53] Although these specifications, disentangling the marginal effects of several endogenous variables, should be interpreted with caution, they suggest the effect of ancestry on FDI develops over long periods of time, and possibly peaks with the second, but not the first generation of immigrants.

This finding is consistent with a set of microeconomic studies that show that only those individuals that advance to managerial positions successfully establish business linkages to their origin countries (Aleksynska and Peri, 2014); and that it tends to be the second and third generations of immigrants, that achieve such advancement (Borjas, 2006; Algan et al., 2010). Simply put, the second generation of immigrants is more likely than the first to be able to act on any privileged access to information about the origin country.

To conclude, we find a collage of evidence that migration, and the distribution of ancestry that results from it, has a positive impact on FDI primarily because it reduces information frictions associated with foreign direct investment. We also find evidence consistent with network effects

[53] These results hold if we drop migrations from Mexico (the largest origin country in recent decades).

and the inter-generational transmission of traits that facilitate the flow of information between the origin country and the US destination on a long-term basis.

Conclusion

The economic effects of migration loom large in public debates about illegal immigration to the United States and the ongoing flow of migrants to Europe from places such as Syria, Afghanistan, Africa, and the Balkans. Much of the academic debate on the subject has focused on relatively short-term consequences on local labor markets and consumer prices (Card, 1990; Cortes, 2008). We contribute to this debate by showing causally identified evidence that migrations, and the ethnic diversity they create, also have a long-term effects: They enhance the propensity of firms based in areas receiving migrants to interact economically with the migrants' origin countries. This ethnic determinant of foreign direct investment is large, persists for generations, and seems to operate primarily through a reduction in information frictions.

With these findings, we shed new light on the economic effects of migrations, and suggest several promising avenues for future research. First, we document that ethnic diversity increases the likelihood and intensity of FDI. Because the effect of foreign ancestry on FDI is strongly concave, receiving migrants from many small ethnic groups generates more FDI than receiving the same number of migrants from a single ethnic group. In this sense, ethnic diversity creates an absolute advantage in conducting FDI. A similar argument could potentially be made for the impact of ethnic diversity on the adoption of foreign technologies or on economic growth.

Second, the effects of historical migrations on economic development are large and long-lasting. Regions within the United States that received more, often poor, migrants from countries like Ireland or China more than a century ago, as a result, enjoy significantly stronger economic ties with these countries today. Taking in migrants today may thus deliver long-term dividends: International investments follow the paths of historical migrants as much as they follow economic fundamentals, such as productivity, wage differentials, and tax breaks.

Finally, our identification strategy has a range of applications beyond FDI. For example, our approach to identification could be used to study the broader impact of migrations on attitudes towards foreigners of various origins or the effects of migration on long-term economic growth. If global migration pressures increase further, e.g. due to global climate change, having a detailed understanding of these impacts is key to devising effective migration policies.

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.

Algan, Y., C. Dustmann, A. Glitz, and A. Manning (2010): "The Economic Situation of First and Second-Generation Immigrants in France, Germany and the United Kingdom," *The Economic Journal*, 120, F4–F30.

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. (2010): "Market Penetration Costs and the New Consumers Margin in International Trade," *Journal of Political Economy*, 118, 1151–99.

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.

Atkin, D. (2010): "Trade, Tastes and Nutrition in India," *Yale University Economic Growth Center Discussion Paper No. 986*.

Bailey, M., R. Cao, T. Kuchler, J. Stroebel, and A. Wong (2016): "Measuring Social Connectedness," Working paper, NYU.

Baker, S. R. and A. Fradkin (2016): "The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data," *The Review of Economics and Statistics*.

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.

Besley, T. and S. Coate (1995): "Group Lending, Repayment Incentives and Social Collat-

eral," *Journal of Development Economics*, 46, 1–18.

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.

——— (2006): "Making it in American: Social Mobility in the Immigrant Population," Working Paper 12088, National Bureau of Economic Research.

Bronnenberg, B. J., J.-P. H. Dubé, and M. Gentzkow (2012): "The evolution of brand preferences: Evidence from consumer migration," *The American Economic Review*, 102, 2472–2508.

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.

Conley, T. (1999): "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 92, 1 – 45.

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.

Da, Z., J. Engelberg, and P. Gao (2011): "In Search of Attention," *The Journal of Finance*, 66, 1461–1499.

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.

Greif, A. (1993): "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition," *The American Economic Review*, 83, 525–548.

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.

Javorcik, B. S., C. Ozdenc, M. Spatareanu, and N. Cristina (2011): "Migrant Networks and Foreign Direct Investment," *Journal of Development Economics*, 94, 231–41.

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," CEP Discussion Paper 1322, Centre for Economic Performance, LSE.

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.

Kearney, M. S. and P. B. Levine (2015): "Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing," *American Economic Review*, 105, 3597–3632.

Leblang, D. (2010): "Familiarity Breeds Investment: Diaspora Networks and International Investment," *American Political Science Review*, 104, 584 – 600.

Liang, J. (2017): "Cultural Similarity – Measurement using Google Trends," *mimeo University of Chicago*.

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 (2016): "Migrant Networks and Trade: The Vietnamese Boat People as a Natural Experiment," *Economic Journal, forthcoming*.

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).

Stephens-Davidowitz, S. (2014): "The cost of racial animus on a black candidate: Evidence using Google search data," *Journal of Public Economics*, 118, 26 – 40.

Stephens-Davidowitz, S. and Varian (2015): "A Hands-on Guide to Google Data," *Working Paper*.

Stiglitz, J. E. (1990): "Peer Monitoring and Credit Markets," *The World Bank Economic Review*, 4, 351–366.

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

Varian, H. R. (1990): "Monitoring Agents With Other Agents," *Journal of Institutional and Theoretical Economics (Zeitschrift fur die gesamte Staatswissenschaft)*, 146, 153–174.

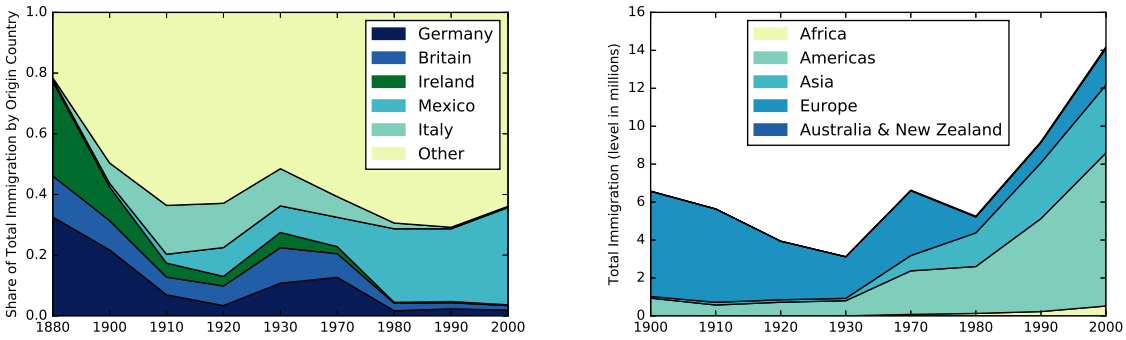


FIGURE 1: ORIGINS OF IMMIGRANTS TO THE UNITED STATES, PRE-1880 TO 2000

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.

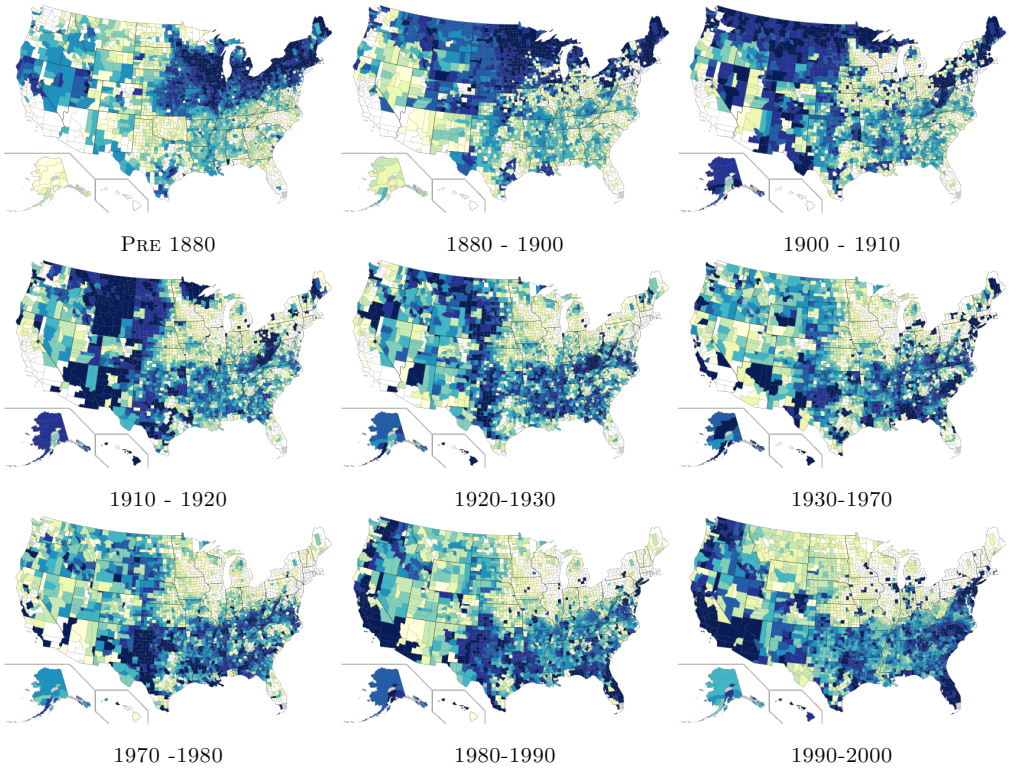
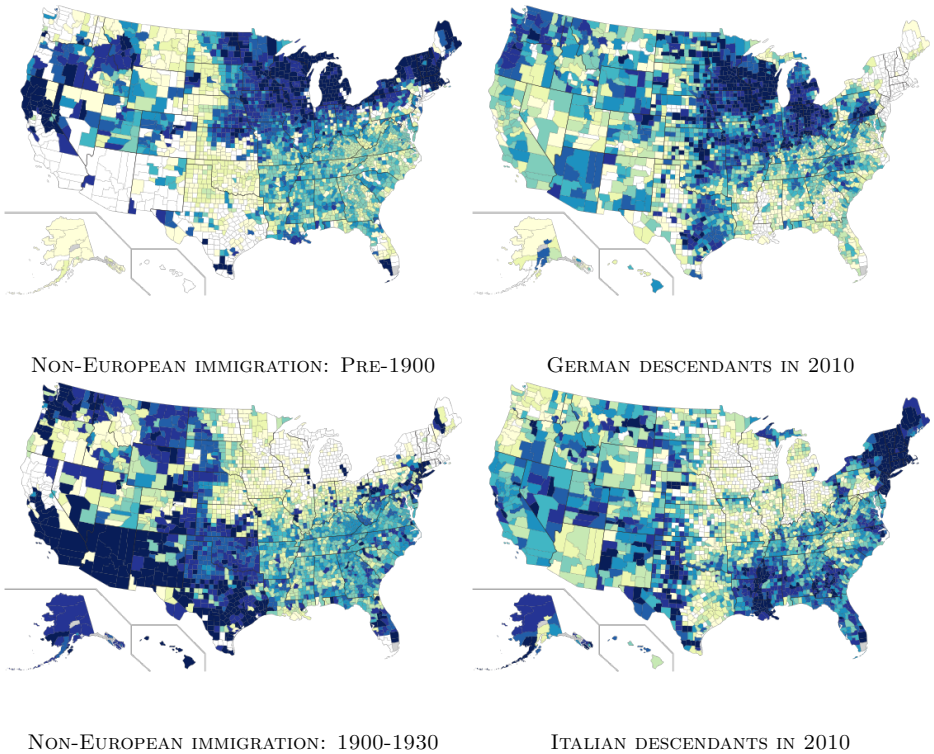


FIGURE 2: DESTINATIONS OF IMMIGRANTS TO THE UNITED STATES, PRE-1880 TO 2000

Notes: This figure maps immigration flows into 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 across counties and within census periods. Darker colors indicate a higher decile.



Non-European immigration: Pre-1900 | German descendants in 2010

Non-European immigration: 1900-1930 | Italian descendants in 2010

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

Notes: This figure contrasts Italian and German ancestry in 2010 (right panels), and non-European immigration patterns pre-1900 and 1910-1930 (left panel). The left two panels are created as in Figure 2, restricted to non-European immigration, and the periods 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.

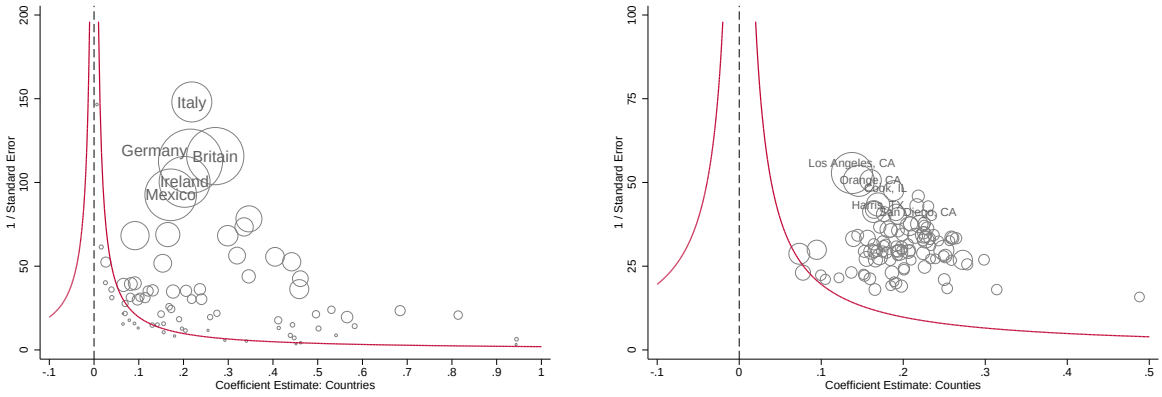


Figure 4: Heterogeneous Effects across Countries and Counties

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 (left) and destination US counties (right). 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. x-axis: estimated coefficients. y-axis: reciprocal of estimated standard errors on ancestry. Circle sizes are proportional to country ancestry (left) and county population (right). Circles above the $y = \pm 1.96/x$ curve indicate statistically significant coefficients. See section 3.5 for details.

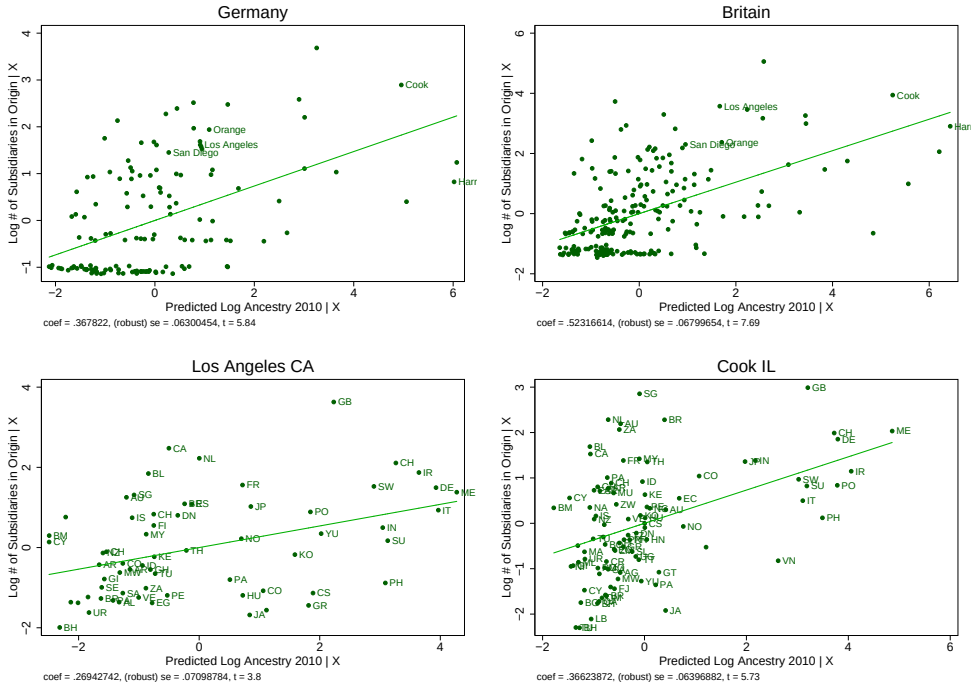


FIGURE 5: ANCESTRY AND FDI: GERMANY AND BRITAIN; LOS ANGELES AND COOK COUNTIES

Notes: The figure shows conditional scatterplots from regressions corresponding to column 3 of Appendix Table 19, restricted to one origin country (top parts: Germany and Britain) or one destination county (bottom parts: LA and Cook counties). The solid line depicts the fitted regression line, controlling for distance and latitude difference. x-axis: predicted log 2010 ancestry. y-axis: log # of subsidiaries in Germany (top left) and Britain (to right) for firms in each US county, log # of subsidiaries of LA (bottom left) or Cook county (bottom right) based firms in each origin country.

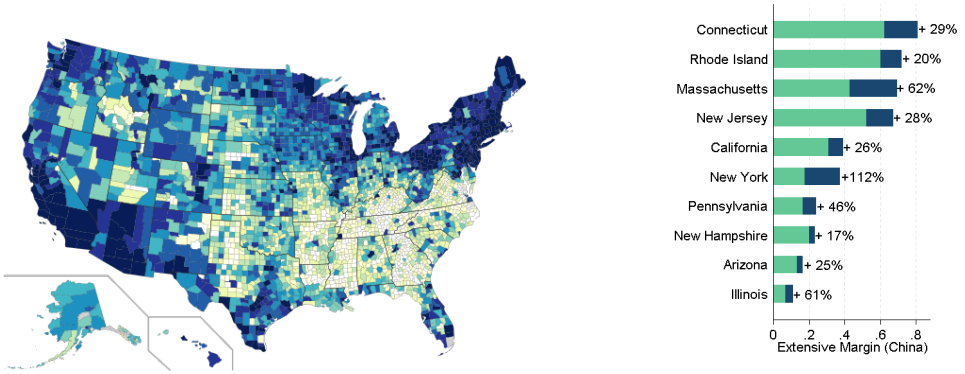


FIGURE 6: THOUGHT EXPERIMENT: REMOVING THE CHINESE EXCLUSION ACT

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. 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 $d\Pr[FDI_{China,d} > 0]$ across counties for the ten US states with the largest change. We also provide the size of this increase relative to the actual fraction in percentage terms. The details of this calculation are section 3.6.

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	0.351	0.028	1.620
	(5.486)	(7.396)	(1.461)	(16.294)
# 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	0.123
	(0.399)	(0.537)	(0.103)	(1.175)
# 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)	0.050	0.088	0.027	0.392
	(2.798)	(3.743)	(2.098)	(7.895)
Information Demand Index (standardized)*	0.599	0.735	0.197	1.489
	(1.000)	(1.092)	(0.443)	(1.448)
N	19,110	14,583	770	4,527
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. Variables in Panel B refer to our sample of countries. Variables in Panel C refer to our sample of counties. 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. *The data is at the metropolitan area level.

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)
$I^{1880}_{o,-r(d)} \times \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$	0.057***	0.056***	0.045***	0.035***	0.056***	0.056***	0.056***		2.145***
	(0.013)	(0.012)	(0.011)	(0.009)	(0.007)	(0.007)	(0.008)		(0.290)
$I^{1900}_{o,-r(d)} \times \frac{I^{1900}_{-c(o),d}}{I^{1900}_{-c(o)}}$	0.097***	0.095***	0.067**	0.068***	0.099***	0.099***	0.100***	0.106**	3.634**
	(0.032)	(0.031)	(0.028)	(0.019)	(0.033)	(0.033)	(0.036)	(0.043)	(1.425)
$I^{1910}_{o,-r(d)} \times \frac{I^{1910}_{-c(o),d}}{I^{1910}_{-c(o)}}$	0.192***	0.193***	0.145***	0.123***	0.137***	0.137***	0.132**	0.123**	5.646**
	(0.039)	(0.039)	(0.042)	(0.031)	(0.050)	(0.050)	(0.054)	(0.049)	(2.345)
$I^{1920}_{o,-r(d)} \times \frac{I^{1920}_{-c(o),d}}{I^{1920}_{-c(o)}}$	0.205***	0.209***	0.176***	0.174***	0.283***	0.283***	0.249***	0.276***	14.726***
	(0.070)	(0.070)	(0.061)	(0.052)	(0.045)	(0.045)	(0.047)	(0.041)	(3.012)
$I^{1930}_{o,-r(d)} \times \frac{I^{1930}_{-c(o),d}}{I^{1930}_{-c(o)}}$	0.062	0.061	0.061	0.035	0.079	0.079	0.065*	0.078	11.812***
	(0.056)	(0.056)	(0.056)	(0.048)	(0.051)	(0.051)	(0.034)	(0.051)	(2.855)
$I^{1970}_{o,-r(d)} \times \frac{I^{1970}_{-c(o),d}}{I^{1970}_{-c(o)}}$	0.183***	0.184***	0.163***	0.149***	0.149***	0.148***	0.150***	0.151***	6.256***
	(0.038)	(0.038)	(0.036)	(0.031)	(0.028)	(0.029)	(0.027)	(0.029)	(0.669)
$I^{1980}_{o,-r(d)} \times \frac{I^{1980}_{-c(o),d}}{I^{1980}_{-c(o)}}$	0.173***	0.173***	0.174***	0.169***	0.214***	0.214***	0.205**	0.213***	18.694***
	(0.066)	(0.066)	(0.064)	(0.061)	(0.076)	(0.077)	(0.080)	(0.075)	(2.390)
$I^{1990}_{o,-r(d)} \times \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$	0.123***	0.124***	0.124***	0.111**	0.101**	0.101**	0.115**	0.102**	10.786***
	(0.048)	(0.048)	(0.048)	(0.044)	(0.045)	(0.045)	(0.048)	(0.044)	(3.675)
$I^{2000}_{o,-r(d)} \times \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$	0.020	0.019	0.026	0.026*	0.046***	0.046***	0.039**	0.046***	5.194***
	(0.017)	(0.017)	(0.016)	(0.015)	(0.017)	(0.017)	(0.016)	(0.017)	(1.148)
$I^{2010}_{o,-r(d)} \times \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$							0.317***		
							(0.089)		
Kleibergen Wald rk statistic	10.608	10.958	8.327	9.607	162.194	158.125	195.423	142.800	910.331
Stock-Yogo 5% critical values	20.53	20.53	20.53	20.53	21.18	21.18	21.23	21.10	21.18
Stock-Yogo 10% critical values	11.46	11.46	11.46	11.46	11.52	11.52	11.51	11.52	11.52
R^2	0.56	0.56	0.66	0.72	0.73	0.73	0.73	0.73	0.49
N	612,495	612,495	612,495	612,495	612,495	612,495	612,495	612,495	612,495
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
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	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)	(7)
Log Ancestry 2010	0.231***	0.190***	0.187***	0.187***	0.189***	0.198***	0.191***
	(0.023)	(0.024)	(0.024)	(0.024)	(0.030)	(0.023)	(0.024)
Log Distance	0.007	0.004	0.024	0.009	0.030	0.026	-0.027
	(0.010)	(0.009)	(0.029)	(0.033)	(0.040)	(0.030)	(0.027)
N	612495	612495	612495	612495	459150	612495	612300
Panel B: OLS	*FDI 2014 (Dummy)*						
Log Ancestry 2010	0.173***	0.173***	0.149***	0.149***	0.145***	0.149***	0.161***
	(0.016)	(0.016)	(0.018)	(0.018)	(0.019)	(0.018)	(0.019)
R^2	0.2967	0.2967	0.3635	0.3635	0.3920	0.3635	0.3930
N	612495	612495	612495	612495	459150	612495	612495
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	Yes	No	No
Agricultural Similarity (Cosine)	No	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	No	Yes	No
Origin × State FE	No	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 $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ as excluded instruments. Columns 3-7 also include the first five principal components of the higher-order interactions of push and pull factors as instruments. To avoid re-calculating the principal components across specifications, we include the 2010 wave in their calculation for all columns; however, the inclusion of the 2010 wave has essentially no effect on our results. For example, the standard specification in column 3 without including the 2010 wave in the calculation of principal components yields 0.187 (0.024). 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 COMMUNIST NATURAL EXPERIMENT

	FDI 2014 (Dummy)				
	(1)	(2)	(3)	(4)	(5)
Log Ancestry 2010	0.185***	0.237***	0.089**	0.172***	0.209***
	(0.019)	(0.024)	(0.036)	(0.021)	(0.032)
N	3,141	3,141	3,141	18,846	28,269
Destination FE	No	No	No	No	Yes
Origin FE	No	No	No	Yes	Yes
Countries considered	Soviet Union	China	Vietnam	Eastern Europe	All communist countries
Period of ec. isolation	1918-1990	1945-1980	1975-1996	1945-1989	
Instruments from	1920-1990	1970-1980	1980-1990	1970-1980	

Notes: The table presents coefficient estimates from IV regressions of equation (1) at the country-county level. The estimates correspond to column1 of Table 3 except in the inclusion of fixed effects, which is specified. 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. In all specifications the instruments are constructed as in column 3 of Table 3; however, we only include as instruments 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: Alternative Instruments

Panel A: Variations of leave-out categories	*FDI 2014 (Dummy)*
Excluding origins with correlated migration flows: $I^t_{o,-r(d)} \times (I^t_{-s^1(o),d}/I^t_{-s^1(o)})$	0.181***
	(0.027)
Excluding origins with correlated 2010 ancestry stock: $I^t_{o,-r(d)} \times (I^t_{-s^2(o),d}/I^t_{-s^2(o)})$	0.217***
	(0.030)
Excluding states adjacent to the destination: $I^t_{o,-adj(d)} \times (I^t_{-c(o),d}/I^t_{-c(o)})$	0.192***
	(0.022)
Panel B: Using subsets of instruments for identification	*FDI 2014 (Dummy)*
Only migrations $1880 - 1930$	0.209***
	(0.037)
Only migrations $1970 - 2000$	0.175***
	(0.021)
Only migrations $1900 - 2000$	0.187***
	(0.024)

Notes: This table presents coefficient estimates from instrumental variable regressions that are variations of our standard specification (column 3 of Table 3). Each row lists the coefficient estimate on *Log Ancestry 2010*. Panel A shows alternative specifications of our leave-out instrument: we exclude from the pull factor all countries for which the time correlation of total migration to the US with o's migration to the US is greater than .5 and significant at the 5% level; we exclude from the pull factor all countries for which the correlation of 2010 ancestry across the US with o's ancestry across the US is greater than .5 and significant at the 5% level; and in the push factor of o we exclude migrations to any state adjacent to the state of d including the state itself, respectively. In Panel B we use throughout the interacted instrument of our standard specification but each specification in this panel uses as instruments only the simple interaction terms from a subset of the full time period covered by our data. Standard errors are given in parenthesis and clustered at the country level. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 6: The Effect of Ancestry on Immigration Flows and FDI Flows

	(1)	(2)	(3)
Panel A	*Immigration 1990-2000*	*Log immigration 1990-2000*	*Log immigration 1980-1990*
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
Panel B	*FDI 2014 (Dummy)*		*Log total # of FDI relationships 2014*
Log Ancestry 2010	0.086***	0.082***	0.142***
	(0.018)	(0.016)	(0.049)
FDI 2007 (Dummy)	0.542***	0.544***	
	(0.031)	(0.032)	
Log Total # of FDI Relationships 2007			0.685***
			(0.020)
N	612,495	612,495	10,851

Notes: Panel A of the table presents the coefficient estimates from IV regressions of equation (8) 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×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered at the origin country level. Panel B: Columns 1 and 2 present coefficient estimates from IV regressions of equation (1) at the country-county level analogous to Table 3 column 3. The dependent variable is a dummy indicating an FDI relationship between origin o and destination d in 2014. Column 1 of panel B includes all double-interactions as instruments ,$\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$, while columns 2 and 3 include only double-interactions of the push and pull factors from the 1990 and 2000 censuses, $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1990,2000}$. Panel B column 3 presents the IV/GMM estimate of equation (7). This specification controls for log distance, latitude difference, origin, and destination fixed effects and does not apply the Heckman Correction. All specifications in panel B include destination, origin, destination × continent, and origin × census region 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 7: The "Similarities" Hypothesis and Contract Enforcement

	(1)	(2)
Panel A: Final vs. Intermediate Goods	*FDI 2014 (Dummy)*	
Log Ancestry 2010	0.156***	0.169***
	(0.026)	(0.024)
N	612,495	612,495
Sample	Final goods	Intermediate goods
Panel B: Final vs. Intermediate Goods	*Inward FDI 2014 (Dummy)*	
Log Ancestry 2010	0.108***	0.117***
	(0.033)	(0.032)
N	612,495	612,495
Sample	Final goods	Intermediate goods
Panel C: Sector similarity	*Rank Correlation*	*Cosine correlation*
Log Ancestry 2010	0.011	0.010
	(0.015)	(0.013)
N	21,518	21,518
Panel D: Judicial Quality	*FDI 2014 (Dummy)*	*Log # of FDI relationships*
Log Ancestry × Judicial Quality	0.180*	1.414***
	(0.094)	(0.243)
N	452,304	10,089

Notes: The table presents coefficient estimates from IV regressions at the country-county level. In Panel A, the outcome variable is the FDI dummy; we restrict our sample to firms producing final goods or intermediate inputs, respectively. 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 number of country-county pairs that have an (non-zero) FDI link in the corresponding sector is 4,201 and 5,842 in columns 1 and 2, respectively. In Panel B, we replicate the same regressions, except that the outcome variable indicates only the existence of any inward FDI. In Panel C, the outcome variable is the rank and cosine correlation of the share of employees in 127 manufacturing sectors within a given origin-destination pair, respectively. The relatively low number of observations is due to data availability in the industry share of employment: When calculating the correlation between industries' share of employment in county d and country o, the correlation coefficient is missing for those country-county pairs that have at least one missing share of employment. In Panel D, the outcome is the extensive (FDI dummy) and intensive (log # of FDI relationships) margin, and the measure of judicial quality is from Nunn (2007). 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 instrumental variables. All specifications control for log distance, latitude difference, and origin and destination fixed effects. In Panels A-C we additionally control for origin×destination-census-region, and destination×continent-of-origin fixed effects. Standard errors are given in parentheses and are clustered by origin country.

Table 8: The Effect of Ancestry on Differential Information Demand and Language

Panel A: Google Trends	Information Demand Index (standardized)			Actors (standardized)	Athletes (standardized)	Musicians (standardized)	Politicians (standardized)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log Ancestry 2010	0.871***	1.717**		0.659**	0.948***	0.531***	0.688***
	(0.257)	(0.783)		(0.257)	(0.359)	(0.076)	(0.183)
Log Foreign-born 2010		-1.108					
		(0.809)					
Log Ancestry 1980			0.801***				
			(0.186)				
N	19,110	19,110	19,110	19,110	19,110	19,110	19,110
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Language	# of residents in d speaking language of o at home	# of US-born in d that speak the language of o at home					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ancestry 2010	2.226***	0.942***		1.170***	1.257***	0.085***	0.015***
	(0.717)	(0.301)		(0.042)	(0.230)	(0.006)	(0.003)
Ancestry 1980			1.241**				
			(0.504)				
N	454,812	454,813	454,813	65,877	78,376	3,137	3,137
Non-English language	Any	Any	Any	Spanish	Arabic	Chinese	Hindi
Destination FE	Yes	Yes	Yes	No	No	No	No
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents coefficient estimates from IV regressions at the country-DMA (Panel A) and country-county (Panel B) level. In Panel A, all dependent variables are based on a count of Google searches in the category specified above the panel. The Information Demand Index in columns 1-3 is a simple average of the other four categories. Each of the four categories is the average of Google Trends value $G(i, d)$, which measures the (normalized) fraction of queries that include search term i relative to the total number of queries of Designated Market Area (DMA) d. For the search terms i we use the first five terms from Google's Answer Box when we search for "notable [foreign country d] [category]". All outcome variables in Panel A are standardized by their standard deviation. In Panel B, the dependent variable in column 1 is the number of residents in d that speak the language of o at home, excluding English; in column 2 and 3, it is the number of US-born residents in d that speak the language of o at home, excluding English; and in columns 3-6, it is the number of US-born residents in d that speak the language indicated in the respective column. Spanish is the official language in twenty-one countries; Arabic is the official language in twenty-five countries; Chinese and Hindi each are the official language in only one country. Ancestry 2010, Ancestry 1980, Log Ancestry 2010, Log Foreign-born 2010, and Log Ancestry 1980 are instrumented as in column 2 of Table 3. All specifications control for log distance and latitude difference. Standard errors are given in parentheses and clustered at the country level (all of Panel A and columns 1-2 of Panel B) or state level (columns 3-6 of Panel B). *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

TABLE 9: NETWORK EFFECTS

	FDI 2014 (Dummy)		*Log Total # of FDI relationships*	
	(1)	(2)	(3)	(4)
PANEL A: SPILLOVERS				
Log Ancestry 2010	0.237***	0.166***	1.027***	0.179***
	(0.020)	(0.026)	(0.168)	(0.052)
Log Ancestry 2010, State Level	-0.015***		-0.315***	
	(0.004)		(0.052)	
Log Ancestry 2010 of Nearest Origin Country		-0.047**		0.037
		(0.019)		(0.176)
N	612,495	612,495	10,851	10,851
PANEL B: DIVERSITY				
Log Ancestry 2010	0.197***	0.123***	0.661***	0.641***
	(0.061)	(0.030)	(0.245)	(0.089)
Log Ancestry × Foreign Share	1.388		3.548	
	(3.103)		(7.803)	
Log Ancestry × Ethnic Diversity		1.270***		3.692***
		(0.204)		(1.009)
N	611,910	612,495	10,851	10,851
PANEL C: FRACTIONALIZATION				
Log Ancestry 2010	0.269***	0.334***	0.914***	1.247***
	(0.037)	(0.081)	(0.149)	(0.109)
Log Ancestry × Geographic Distance	0.101***	0.170**	0.414***	0.864***
	(0.036)	(0.076)	(0.128)	(0.156)
Log Ancestry × Judicial Quality		0.373**		2.375***
		(0.187)		(0.494)
Log Ancestry × Fractionalization		0.470		3.087***
		(0.324)		(0.831)
N	446,022	446,022	10,089	10,089

Notes: The table presents coefficient estimates from IV regressions at the country-county level. The dependent variable in columns 1 and 2 is the dummy for FDI in 2014. The dependent variable in columns 3 and 4 is the log of the number of FDI links in 2014. We use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$ 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. In columns 2 and 4 of Panel A, destination × continent fixed effects are used to control for the increased substitutability of immigrants from a common continent. Foreign Share, Ethnic Diversity, Distance, Judicial Quality, and Fractionalization are demeaned. 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. Judicial quality in the origin is from Nunn (2007); genetic distance is from Spolaore and Wacziarg (2015); and Ethnic Fractionalization refers to 1 minus the Herfindahl index of ethnicities in the origin country calculated using the data in Alesina et al. (2003).

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.[54].

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.[55] 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 that 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.

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

[55] 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%".[56]

[56] 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 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.[57] 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.[58] 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.[59] A higher index for *"Judicial Quality"* corresponds to a higher-quality judicial system. *"Ethnic Diversity"* is a measure of a country's ethnolinguistic fractionalization.[60]

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.[61] *"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.

[57] 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.

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

[59] 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.

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

[61] 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

A.4 Details on the construction of information demand indices

The Information Demand Index is based on data gathered from Google and created in three steps. In the first step we identify five prominent individuals from country o in category p, where $p \in \{\text{actors, athletes, musicians, politicians}\}$. In the second step we utilise Google Trends to obtain data on the spatial variation in the relative frequency of search queries related to these individuals. In the last step we construct indices of the search intensity related to country o in destination d.

Step 1: To identify the top five prominent individuals from o in each category p, we utilise a tool called Google's featured snippet box. Google's featured snippet box is a response to a search query that is generated by Google and pushed to the top of the result list. Google generates these answers by scraping its top results and using an algorithm to provide what it determines to be the most relevant answer.[62] For our purposes we record the top five names in Google's featured snippet box in response to the query "notable [country] [p]", where [country] is one of the 100 largest countries by 2015 population. For example, searching for "notable Belgium actors" yields Google's featured snippet box with an ordered list of Belgian actors. We save the top five names from left to right as the set of search queries $q(o, p)$. If Google's featured snippet box does not give a response for a country, we record a missing entry.[63]

Step 2: Google Trends provides historical and cross-sectional information about the relative importance of a search query. For the United States, the cross-sectional information with the highest granularity is at the level of a Designated Market Area (DMA).[64] Google Trends expresses the relative importance of a search query in a given DMA as an integer value from 0 to 100. This integer value is calculated as follows. First, find the number of searches for the query at hand relative to the total number of searches, and define the maximum search market share of any DMA to 100. Second, divide each search market share by the maximum, and express it as a rounded percentage. If the result does not exceed an unreported threshold, set it to zero (Liang, 2017; Stephens-Davidowitz and Varian, 2015). Formally,

$$G(i,d) = \left\lfloor 100 \frac{share_{i,d}}{\max_{\delta}\{share_{i,\delta}\}} \mathbf{1}[\#(i,d) \geq T] \right\rceil$$

where $\lfloor x \rceil$ is the integer round function, $share_{i,d}$ is the search market share of search query i in DMA d, and T is the unreported search volume threshold. Note that T is defined on the absolute number of searches, rather than the search market share. This implies that DMAs with a larger population will tend to report more data than those with smaller populations. Note also that in addition to $G(i,d)$ being reported as zero for some i and d, we set its value equal to zero if there is no search result from Google's featured snippet box, or if there is no result from Google Trend.

[62] See https://support.google.com/webmasters/answer/6229325?hl=en

[63] This is the case for about 4-10% of our sample, depending on the category.

[64] Google Trends also breaks the information down by major city; however, we would lose non-city data.

Step 3: We define the p-specific Index for each DMA-country pair as

$$I(p,o,d) = \frac{1}{5} \sum_{i \in q(o,p)} G(i,d)$$

We define the Information Demand Index as the average over the p-specific indices:

$$\text{IDI}_{o,d} = \frac{1}{5} \sum_{p} I(p,o,d).$$

to the query "notable [country] [p]", where [country] is one of the 100 largest countries in 2015.

A.5 Details on the construction of crop suitability measures

The crop suitability index for each origin country o and destination county d is taken from the Food and Agriculture Organization of the United Nations Global Agro-Ecological Zones (FAO-GAEZ) data. We estimate potential agricultural similarity between each origin country o and county d by constructing a distance measure based on the difference in crop suitability of the country and county for a select group of crops. The following outlines the steps taken in order to create a crop suitability distance measure for each country-county pair.

Step 1: To identify the crops to be used in constructing the crop suitability distance measure, we compare data from FAOSTAT on the top crops produced by the U.S. in 2014 with data available from FAO-GAEZ on crop suitability. We then select the top 10 crops, based on value of agricultural production, for which there is data in FAO-GAEZ were used: rice, maize, wheat, soybeans, tomatoes, white potatoes, sugar cane, cotton, yams, and cassava. For each crop, we extract the crop suitability data by selecting the total production capacity data (located within the attainable yield/agro-ecological suitability data) and setting the water-supply to rainfed, input level to high, and time period to baseline (1961-1990).

Step 2: To take the global crop suitability data for each crop and define crop suitability for each county and country, we utilise ArcMap software. For counties, the U.S. 1990 counties border map is used. For countries, we extract data on 1990 country borders from a dataset including all country borders for the period post-WWII to 2015.[65] Then, for each crop, we utilise the ArcMap software to calculate the average of crop suitability for each county and country.[66]

Step 3: We rescale the crop suitability data to a 0 to 1 scale and then calculate the crop suitability distance measure for each pair, country o and county d, as

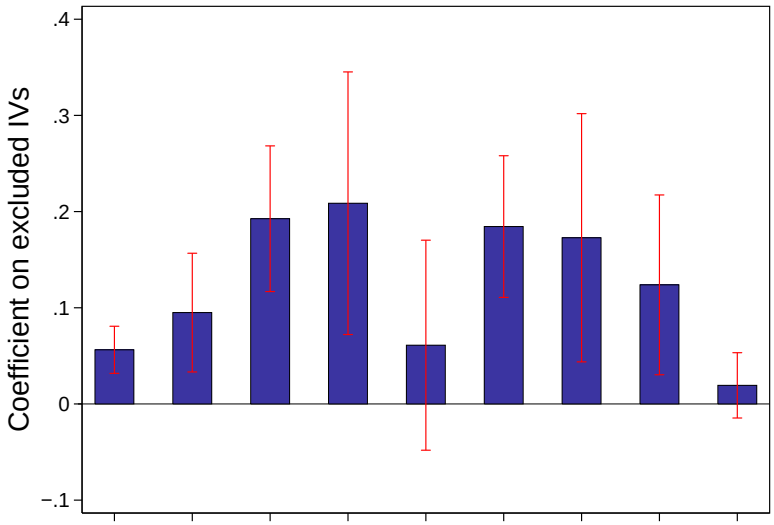
$$DistanceCS_{o,d} = \frac{\sum_{n=1}^{10} s_o^n s_d^n}{\sqrt{\sum_{n=1}^{10} (s_o^n)^2} \sqrt{\sum_{n=1}^{10} (s_d^n)^2}}$$

[65] The original shapefile is version 0.6 (updated November 30, 2016) from Weidmann, Nils B., Doreen Kuse, and Kristian Skrede Gleditsch. "The Geography of the International System: The CShapes Dataset." International Interactions 36, no. 1 (2010).

[66] Data is missing for 23 counties and 35 countries due to issues with overlapping polygons as well as missing 1990 boundaries data for certain countries.

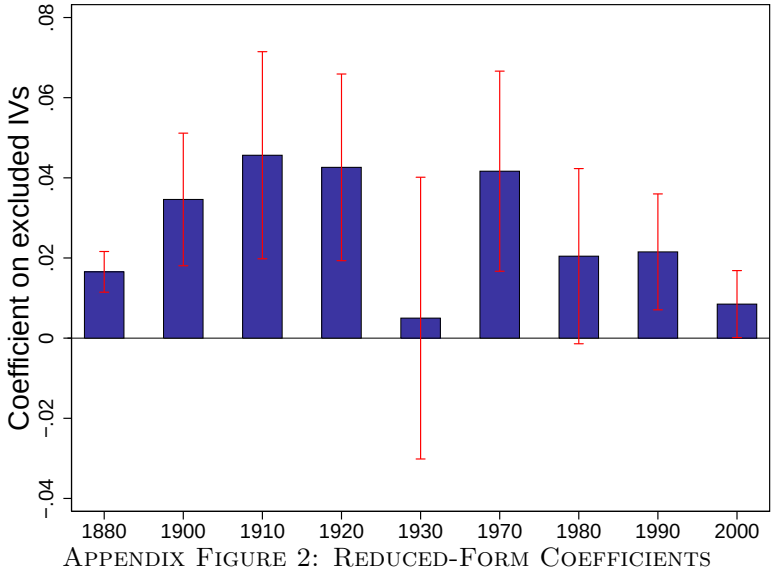
where s^n is the measure of crop suitability for crop n.

B Additional figures and tables



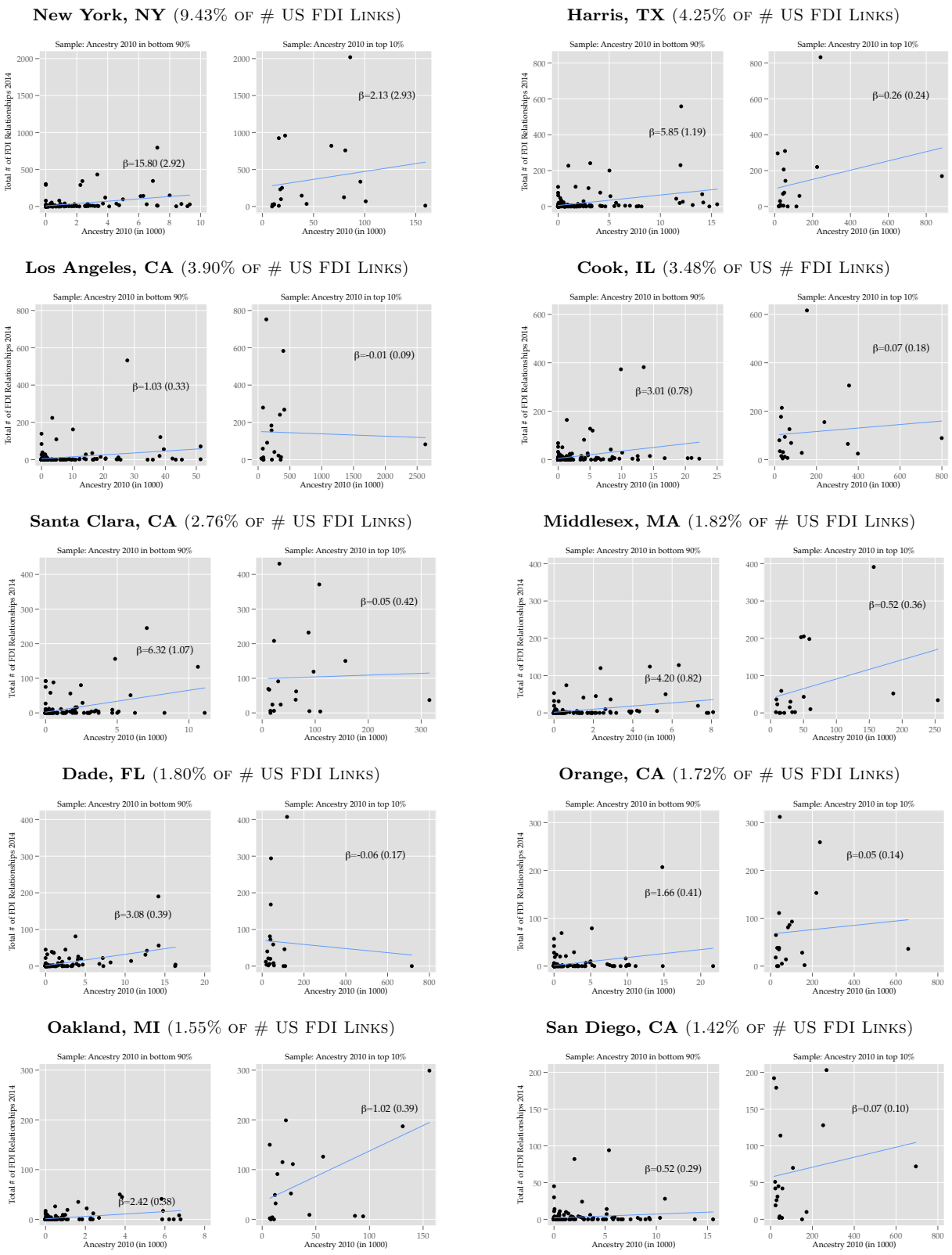
Appendix Figure 1: First-Stage Coefficients

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.



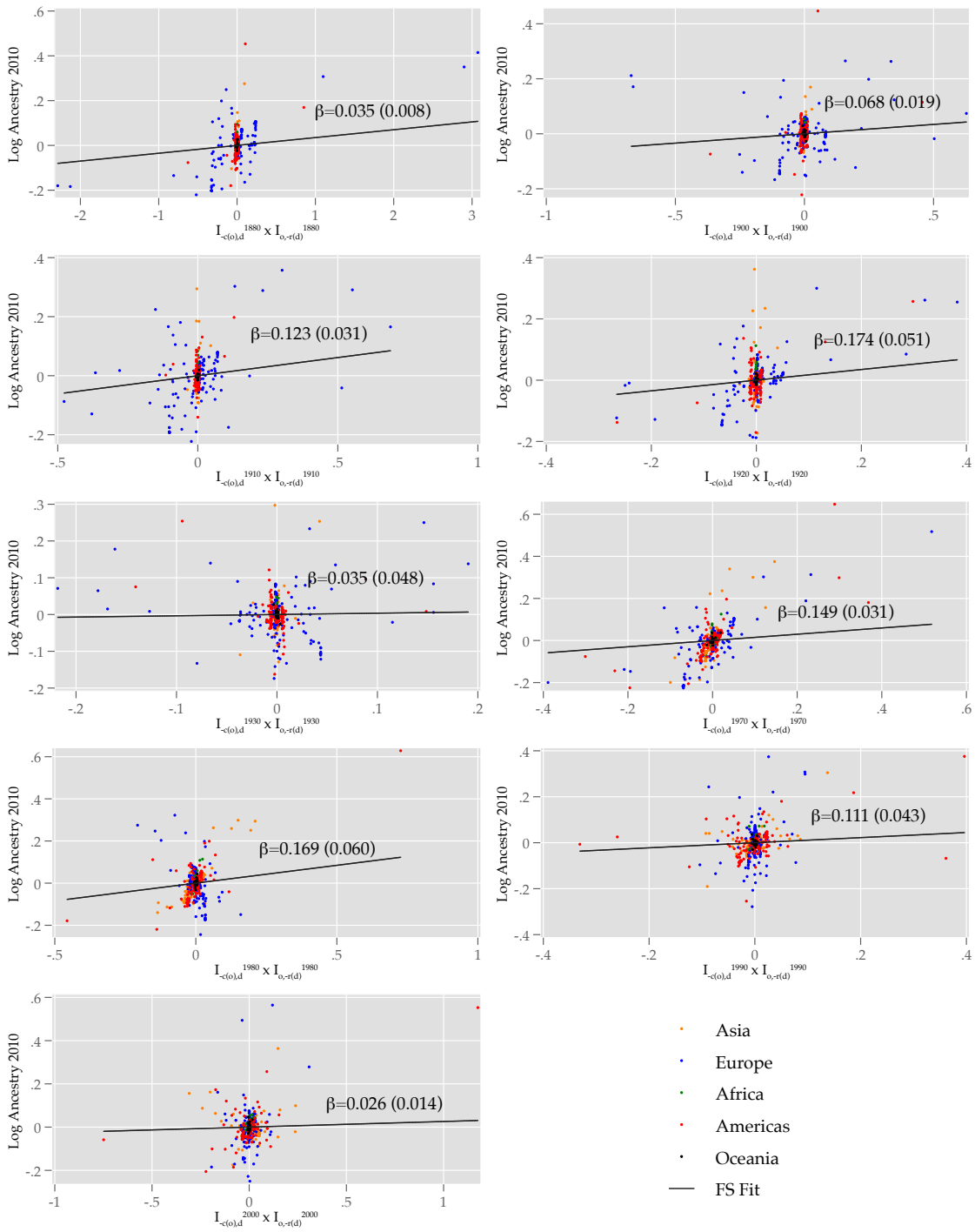
Appendix Figure 2: Reduced-Form Coefficients

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.



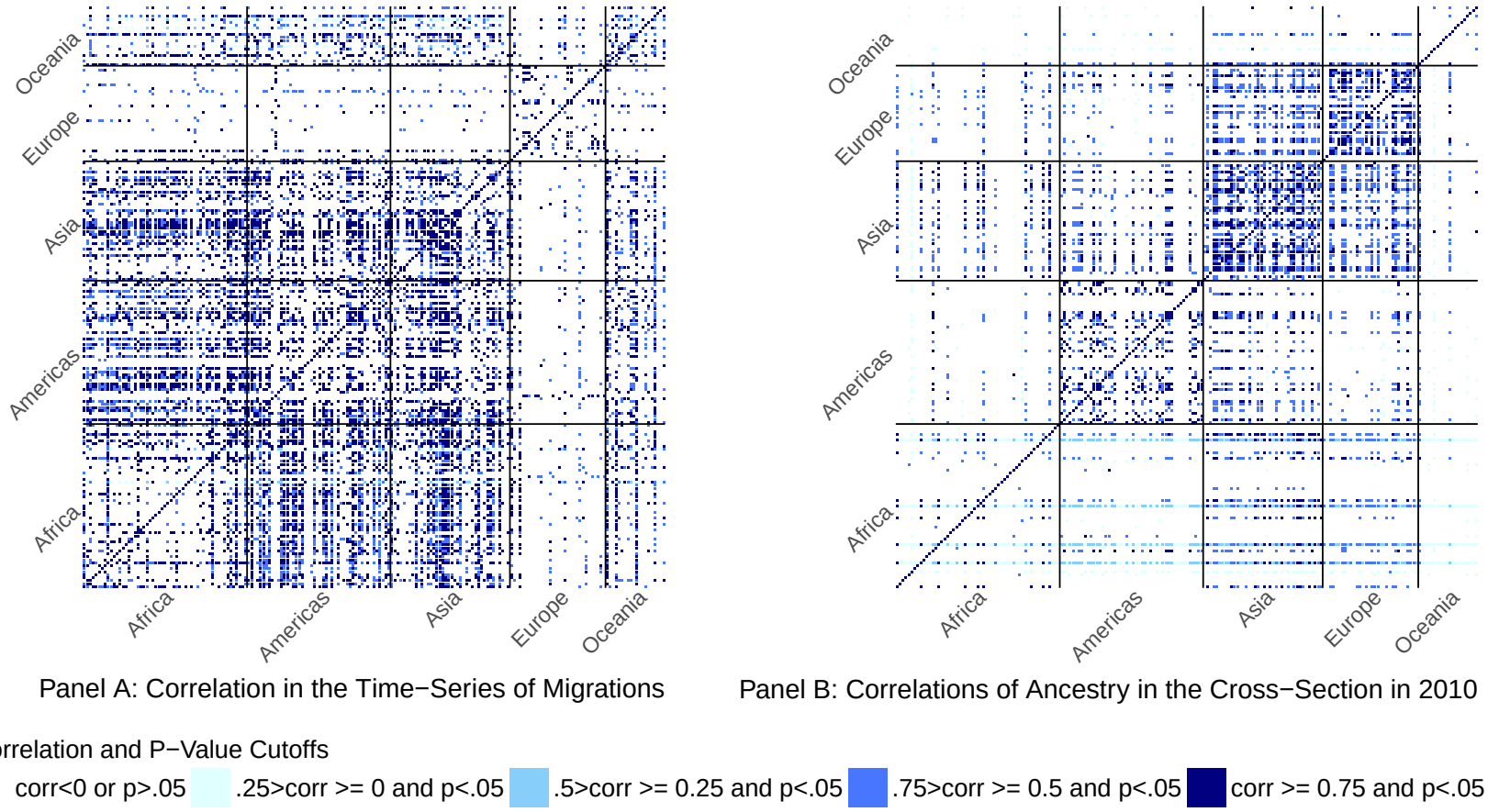
Appendix Figure 3: Ancestry and Total # of FDI relationships (Raw Data)

Notes: The figure presents scatter plots of the raw data for *Ancestry 2010* and *Total # of FDI relationships 2014* for the 10 largest US counties in terms of # of FDI relationships (counties' share of total US FDI relationships indicated in title). For each county, data is shown separately by origins with ancestry share in the bottom 90% of ancestries in d, and in the top 10% of ancestries in d. Linear regressions are fitted separately for each subfigure; coefficient estimates and standard errors (in parentheses) are provided.



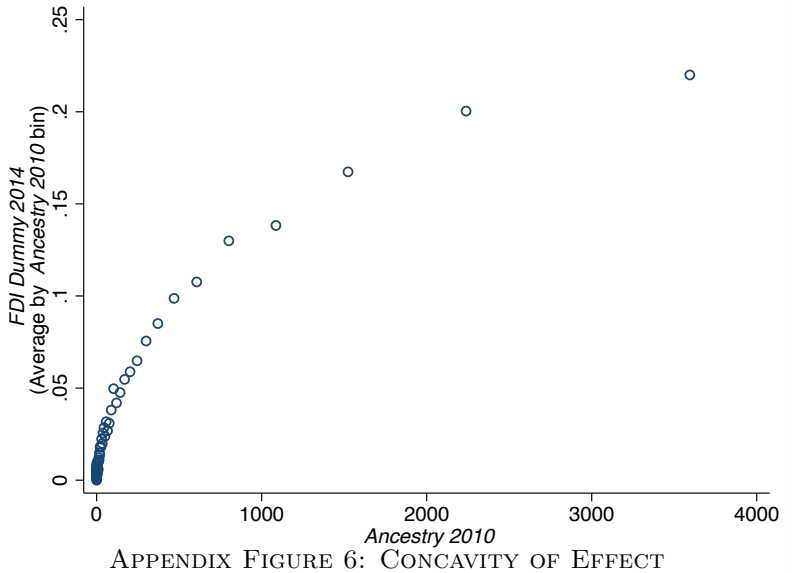
Appendix Figure 4: First Stage Fit

Notes: The figure shows conditional scatter plots of *Log Ancestry 2010* versus each of the interacted instruments $I^t_{-c(o),d} \times I^t_{o,-r(d)}$. Each subfigure is constructed as follows: both *Log Ancestry 2010* and $I^t_{-c(o),d} \times I^t_{o,-r(d)}$ are regressed on destination × continent-of-origin fixed effects, origin × destination-census-region fixed effects, distance, and latitude difference, as well as the interacted instruments for all time periods except t; for visual clarity, residuals of both regressions are binned, separately for each origin o, by quintiles of the residuals of the former regression; the binned data is scattered, colour-coded by the continent of o. Note that each graph shows the partial correlation that identifies the nine coefficients of interest in our standard specification of the first stage in column 4 of Table 2. The first stage – corresponding to a linear least squares fit of the data before binning – is shown as black line, and the respective first stage coefficient estimates (and standard errors in brackets) are shown.



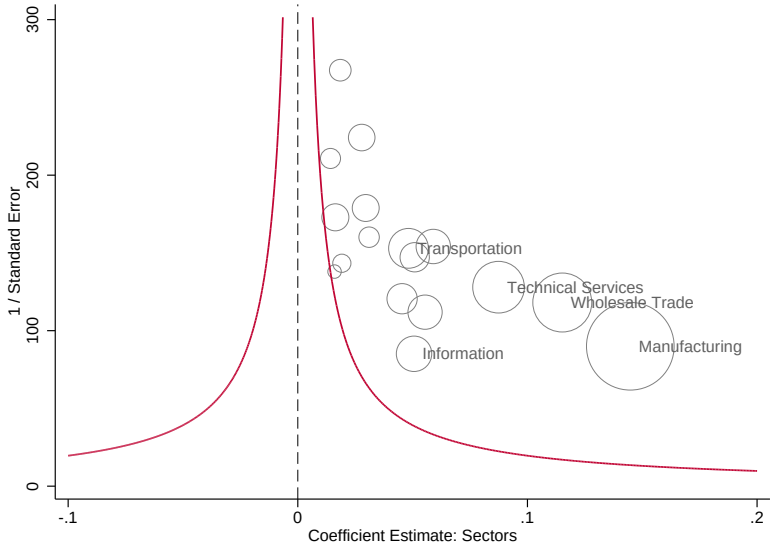
Appendix Figure 5: Correlation Between Countries of Immigration Waves to the US and 2010 Ancestry Across the US

Notes: These correlation plots display the magnitude of the time-series correlation of total migration to the US between each pair of countries (Panel A) and the magnitude of the correlation of ancestry in 2010 across the US counties between each pair of origin countries (Panel B). A pair of origin countries for which the correlation is negative or not significant at the 5% level is displayed with a white dot. For country pairs with positively correlated total migration significant at the 5% level, a darker shaded dot indicates a higher correlation value.



Appendix Figure 6: Concavity of Effect

Notes: This figure plots of the mean of *FDI Dummy 2014* within bins of *Ancestry 2010*. The *Ancestry 2010* bins are constructed as centiles of the conditional distribution of *Ancestry 2010*|*Ancestry 2010*>0. The lowest bin corresponds to *Ancestry 2010*$=0$. We do not plot the mean of *FDI Dummy 2014* in the 99th and 100th centile *Ancestry 2010* bin for visual clarity; the overall concave pattern extends to these observations.



Appendix Figure 7: Heterogeneous Effects across 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 sector. 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 the sector. 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 3.5 for details.

Appendix Table 4: Summary Statistics on the Intensive Margin of FDI

Origin-destination pairs	(1)	(2)	(3)
Ancestry 2010 (in thousands)	10.038	16.502	10.861
	(40.989)	(62.950)	(43.593)
# of FDI Relationships	11.043		
	(39.738)		
# of Parents in Destination		2.282	
		(4.336)	
# of Parents in Origin		8.063	
		(26.132)	
# of Workers Employed at Subsidiary in Destination (in thousands)		6.328	
		(32.695)	
# of Subsidiaries in Origin			1.797
			(10.671)
# of Parents in Destination			0.761
			(3.105)
# of Workers Employed at Subsidiary in Origin (in thousands)			2.435
			(40.360)
N	10851	4065	9082

Notes: The table presents means (and standard deviations). Variables refer to our sample of country-county pairs used in Appendix Table 19. 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 5: 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 6: Alternative Instruments based on Immigration and Ancestry Correlation

	FDI 2014 (Dummy)			
	(1)	(2)	(3)	(4)
Panel A: Time-Series Correlation of Total Migration to the US				
Log Ancestry 2010	0.176***	0.177***	0.177***	0.181***
	(0.027)	(0.028)	(0.027)	(0.027)
Log Distance	0.021	0.021	0.021	0.022
	(0.029)	(0.028)	(0.028)	(0.028)
N	612495	612495	612495	612495
Panel B: Cross-Section Correlation of 2010 Ancestry Across the US				
Log Ancestry 2010	0.217***	0.186***	0.217***	0.217***
	(0.031)	(0.028)	(0.031)	(0.031)
Log Distance	0.031	0.023	0.031	0.031
	(0.031)	(0.029)	(0.031)	(0.031)
N	612495	612495	612495	612495
Correlation Cutoff	.5	.75	.5	.5
Significance Cutoff	NA	NA	.1	.05

Notes: The table displays estimates of the specification from column 3 of Panel A in Table 3 removing or not alternative sets of migrations from the construction of pull factors. In Panel A, we exclude from the pull factor all countries for which the time-series correlation of total migration to the US with o's migration to the US is greater than .5, greater than .75, greater than .5 and significant at the 10% level, and greater than .5 and significant at the 5% level, respectively. In Panel B, we exclude from the pull factor all countries for which the correlation of 2010 ancestry across destination counties with o's ancestry across destination counties is greater than .5, greater than .75, greater than .5 and significant at the 10% level, and greater than .5 and significant at the 5% level, respectively.

Appendix Table 7: The Effect of Ancestry on FDI: Variations of Leave-Out Instrument

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FDI Dummy (2014)						
Panel A: baseline specification	$\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded						
Log Ancestry 2010	0.231***	0.190***	0.187***	0.187***	0.189***	0.198***	0.191***
	(0.023)	(0.024)	(0.024)	(0.024)	(0.030)	(0.023)	(0.024)
N	612495	612495	612495	612495	459150	612495	612300
Panel B: no leave-out	$\{I^t_o(I^t_d/I^t)\}$ excluded						
Log Ancestry 2010	0.204***	0.202***	0.174***	0.174***	0.173***	0.183***	0.215***
	(0.020)	(0.019)	(0.022)	(0.022)	(0.027)	(0.022)	(0.017)
N	612495	612495	612495	612495	459150	612495	612300
Panel C: single country/county leave-out	$\{I^t_{o,-d}(I^t_{-o,d}/I^t_{-o})\}$ excluded						
Log Ancestry 2010	0.212***	0.204***	0.172***	0.171***	0.173***	0.185***	0.216***
	(0.020)	(0.019)	(0.024)	(0.024)	(0.030)	(0.024)	(0.017)
N	612495	612495	612495	612495	459150	612495	612300
Panel D: county/continent leave-out	$\{I^t_{o,-d}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded						
Log Ancestry 2010	0.223***	0.217***	0.183***	0.183***	0.186***	0.200***	0.227***
	(0.022)	(0.021)	(0.024)	(0.024)	(0.030)	(0.024)	(0.018)
N	612495	612495	612495	612495	459150	612495	612300
Panel E: adjacent state leave-out	$\{I^t_{o,-adj(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}$ excluded						
Log Ancestry 2010	0.232***	0.204***	0.192***	0.192***	0.181***	0.206***	0.237***
	(0.024)	(0.022)	(0.022)	(0.022)	(0.027)	(0.021)	(0.019)
N	640764	640764	640764	640764	459150	640764	640560
Panel F: correlated migrations leave-out	$\{I^t_{o,-r(d)}(I^t_{-s^1(o),d})(I^t_{-s^1(o)})\}$ excluded						
Log Ancestry 2010	0.229***	0.188***	0.181***	0.181***	0.173***	0.182***	0.182***
	(0.023)	(0.024)	(0.027)	(0.027)	(0.031)	(0.026)	(0.027)
N	612495	612495	612495	612495	459150	612495	612300
Panel G: correlated ancestry leave-out	$\{I^t_{o,-r(d)}(I^t_{-s^2(o),d})(I^t_{-s^2(o)})\}$ excluded						
Log Ancestry 2010	0.268***	0.200***	0.217***	0.217***	0.222***	0.221***	0.220***
	(0.031)	(0.030)	(0.030)	(0.030)	(0.036)	(0.024)	(0.029)
N	612495	612495	612495	612495	459150	612495	612300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	Yes	No	No
Agricultural Similarity (Cosine)	No	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	No	Yes

Notes: The table repeats the estimates from Panel A in Table 3 in Panel A and then shows variations in the following panels, 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, $adj\,(d)$ refers to the adjacent states for the state of county d; thus we exclude from the push factor of o migrations to any state adjacent to the state of d, including the state itself. In Panel E and Panel F, "s" refers to similar countries; that is, in Panel E we exclude from a given pull factor of o to d all countries for which the time correlation of total migration to the US with o's migration to the US is greater than .5 and significant at the 5% level while for Panel F we exclude from a given pull factor of o to d all countries for which the correlation of 2010 ancestry across the US with o's ancestry across the US is greater than .5 and significant at the 5% level.

Appendix Table 8: 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***	0.184***	0.182***	0.182***	0.188***	0.184***
	(0.018)	(0.020)	(0.020)	(0.020)	(0.019)	(0.021)
KP F-stat on excluded IV's	12.06	10.32	167.32	165.46	156.29	189.21
Stock-Yogo 5% critical values	20.25	20.25	21.10	21.10	21.18	21.10
Stock-Yogo 10% critical values	11.39	11.39	11.52	11.52	11.52	11.52
N	612,495	612,495	612,495	612,495	612,495	612,300
Panel B: OLS			*FDI 2007 (Dummy)*			
Log Ancestry 2000	0.216***	0.184***	0.184***	0.184***	0.184***	0.200***
	(0.015)	(0.018)	(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
Destination × Continent FE	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	Yes	Yes	Yes	Yes	Yes
Principal Components	No	No	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2000}_{o,-r(d)}(I^{2000}_{-c(o),d}/I^{2000}_{-c(o)})$	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 $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,1990}$ as excluded instruments. Columns 3-6 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 1990-2000 wave. 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.

Appendix Table 9: Nonlinear Least Squares Estimation

β	π
0.1683***	0.0010***
(0.0012)	(0.0000)

Notes: The table presents coefficient estimates from a nonlinear least squares regression at the country-county level. The dependent variable is the dummy for FDI in 2014. It shows (un-adjusted) NLS standard errors. We obtain the optimal β and π by solving the nonlinear least squares problem in equation (6), excluding the fixed effects. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 10: Alternative Functional Forms

	FDI 2014 (Dummy)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ancestry 2010	0.002***					
	(0.001)					
Log Ancestry 2010 (-1 for $-\infty$)		0.186**				
		(0.080)				
(Ancestry 2010)$^{1/3}$			0.191***			
			(0.022)			
Log Ancestry 1980				0.218***		
				(0.034)		
Log Ancestry 1990					0.203***	
					(0.028)	
Log Ancestry 2000						0.193***
						(0.024)
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 column 3 of Table 3, except that principal components are excluded. 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 11: Varying ownership cutoffs

Panel A: FDI dummy on ancestry (IV)	*FDI 2014 (Dummy)*			
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.189***	0.190***	0.190***	0.157***
	(0.024)	(0.024)	(0.024)	(0.029)
R^2	0.352	0.352	0.352	0.318
N	612495	612495	612495	612495
Panel B: # of FDI relationships on ancestry (IV)	*Log Total # of FDI relationships*			
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.408***	0.394***	0.402***	0.075
	(0.042)	(0.045)	(0.046)	(0.062)
R^2	0.750	0.749	0.749	0.770
N	10445	10393	10365	6981
Ownership cutoff	keep $\geq 5\%$	keep $\geq 25\%$	keep $\geq 50\%$	keep $< 50\%$
Destination × Continent FE	Yes	Yes	Yes	Yes
Origin × Census Region FE	Yes	Yes	Yes	Yes
Principal Components	Yes	Yes	Yes	Yes

Notes: This table presents coefficient estimates from variations of the IV regression in column 3 of Table 3 (Panel A) and in column 2 of Appendix Table 19 (Panel A). We vary the ownership cutoff across columns: In columns 1, 2, and 3 we keep all shareholder-subsidiary pairs with ownership $\geq 5\%$, $\geq 25\%$, $\geq 50\%$, respectively. The number of origin-destination pairs with any FDI under these cutoffs are 10445, 10393, and 10365. In column 4 we keep all shareholder-subsidiary pairs with ownership $< 50\%$, which results in 6981 origin-destination pairs with any FDI. 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 12: Alternative Standard Errors: Main Specification

Panel A: Analytical	
Robust	0.0092
Cluster by county	0.0171
Cluster by country†	0.0243
Cluster by county and country	0.0280
Cluster by state and country	0.0285
Cluster by state	0.0189
Cluster by continent	0.0070
Cluster by state*country	0.0114
Panel B: Bootstrap	
Robust	0.0090
Cluster by county	0.0152
Cluster by country	0.0284

Notes: This table shows various standard errors on Log Ancestry 2010 based on our standard specification (column 3 of Table 3). The bootstrapped standard errors in Panel B are obtained using 1,000 draws with replacement. † denotes our standard specification.

Appendix Table 13: Alternative Standard Errors: Other Specifications

	Standard specification	Communist natural experiment	Intensive margin	Immigration 1990-2000
	(1)	(2)	(3)	(4)
Outcome variable	FDI Dummy (2014)		Log total # of FDI Relationships	Immigration 1990-2000
Panel A: Clustered by Country (Standard)				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.024)	(0.032)	(0.056)	
Log Ancestry 1990				9.662**
				(4.455)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.358)
Panel B: S.E. Clustered by County				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.017)	(0.032)	(0.077)	
Log Ancestry 1990				9.662**
				(4.327)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.230)
Panel C: S.E. Clustered by State				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.019)	(0.029)	(0.071)	
Log Ancestry 1990				9.662**
				(3.942)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.267)
Panel D: Clustered by County and Country				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.028)	(0.051)	(0.120)	
Log Ancestry 1990				9.662**
				(4.800)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.105)
Panel E: Clustered by State and Country				
Log Ancestry 2010	0.187***	0.209***	0.356***	
	(0.028)	(0.048)	(0.116)	
Log Ancestry 1990				9.662**
				(4.308)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$				1.082***
				(0.044)

Notes: This table shows variations based on four main regressions: standard specification (column 3 in Table 3), communist natural experiment (column 5 in Table 4), intensive margin (based on column 2 in Panel A of Appendix Table 19), and immigration 1990-2000 (column 1 in Table 6). In Panel A, we reproduce the standard error clustering in our main tables; in Panel B, we cluster by county; in Panel C, we cluster by country, in Panel D we double cluster by county and country, and in Panel E we double cluster by state and country.

Appendix Table 14: Placebo Regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FDI 2014 (Dummy)						
Panel A	*Assign to alphabet neighbor*						
Log Ancestry 2010	-0.012	-0.007	0.009	0.009	0.006	0.010	0.012
	(0.019)	(0.014)	(0.027)	(0.027)	(0.033)	(0.027)	(0.030)
N	612495	612495	612495	612495	459150	612495	612300
Panel B	*Assign to alphabet neighbor on a different continent*						
Log Ancestry 2010	-0.025	-0.020	0.010	0.010	0.004	0.014	0.013
	(0.021)	(0.014)	(0.033)	(0.033)	(0.038)	(0.037)	(0.037)
N	612495	612495	612495	612495	459150	612495	612300
Destination FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	Yes	Yes	No	No
Agricultural Similarity (Cosine)	No	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)}(I^{2010}_{-c(o),d}/I^{2010}_{-c(o)})$	No	No	No	No	No	Yes	No
Origin × State FE	No	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 15: The Effect of Ancestry on FDI: Five 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 all 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. Robust standard errors are reported. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 16: 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
Grenada	n/a	n/a	0
Sierra Leone	n/a	n/a	0
Somalia	n/a	n/a	0
Iran	n/a	n/a	0
Tonga	n/a	n/a	0
Cape Verde	n/a	n/a	0
Mauritania	n/a	n/a	0
Nepal	n/a	n/a	0
Greenland	n/a	n/a	0
Yemen	n/a	n/a	0
Equatorial Guinea	n/a	n/a	0
Mongolia	n/a	n/a	0
State of Palestine	n/a	n/a	0
Lao	n/a	n/a	0

Notes: The table is an extension of Appendix Table 15 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 17: 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 18: Heterogeneous Effects across Sectors and Firms

FDI 2014 (Dummy)	*Log Ancestry 2010* (1)	*FDI 2014 (Dummy)* > 0 (2)
Panel A: Individual Sectors		
Manufacturing	0.165***	
	(0.024)	
Trade	0.152***	
	(0.026)	
Information, Finance, Management, and other Services	0.143***	
	(0.024)	
Construction, Real Estate, Accomodation, Recreation	0.125***	
	(0.021)	
Health, Education, Utilities, and other Public Services	0.042**	
	(0.019)	
Natural Resources	0.035***	
	(0.009)	
Panel B: Small vs. Large Firm Size		
Above Median	0.112***	1,840
	(0.018)	
Below Median	0.051**	723
	(0.024)	
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) and for small versus large firms (panel B). 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.

Appendix Table 19: The Effect of Ancestry on the Intensive Margin of FDI

	OLS	IV/GMM	IV/GMM	IV/GMM
	(1)	(2)	(3)	(4)
Panel A	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.245***	0.356***	0.292***	0.147***
	(0.048)	(0.056)	(0.021)	(0.031)
N	10,851	10,851	10,851	10,851
Panel B	*Log # of subsidiaries in destination with shareholders in origin*			
Log Ancestry 2010	0.275***	0.339***	0.288***	0.242***
	(0.050)	(0.059)	(0.016)	(0.045)
N	9,082	9,082	9,082	9,082
Panel C	*Log # of workers employed at subsidiaries in destination*			
Log Ancestry 2010	0.304*	0.077	0.326***	0.192
	(0.175)	(0.236)	(0.051)	(0.139)
N	9,082	9,082	9,082	9,082
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
Principal Components	No	Yes	Yes	Yes
Heckman Correction	No	No	No	Yes

Notes: The table presents OLS (column 1) and IV/GMM (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 controls 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.

Appendix Table 20: The Effect of Ancestry on the Intensive Margin of Trade (State Level)

	OLS	IV	IV
	(1)	(2)	(3)
Panel A	*Log Total # of FDI relationships*		
Log Ancestry 2010	1.001***	1.374***	0.079***
	(0.077)	(0.183)	(0.025)
R^2	0.659	0.626	0.847
N	2,208	2,202	2,191
Panel B	*Log Aggregate Exports*		
Log Ancestry 2010	1.519***	2.993***	-0.149
	(0.173)	(0.357)	(0.138)
R^2	0.416	0.374	0.665
N	4,799	4,783	4,739
Panel C	*Log Aggregate Imports*		
Log Ancestry 2010	1.927***	3.447***	0.003
	(0.148)	(0.497)	(0.150)
R^2	0.419	0.360	0.576
N	3,823	3,764	3,815
Origin FE	Yes	Yes	Yes
Destination FE	No	No	Yes
Heckman Correction	Yes	Yes	Yes
Panel D	*Log Exports to Vietnam*		
Log Ancestry 2010	1.169***	1.230***	
	(0.124)	(0.124)	
R^2	0.680	0.678	
N	51	51	
Panel E	*Log Exports to Japan*		
Log Ancestry 2010	0.898***	1.107***	
	(0.197)	(0.128)	
R^2	0.442	0.419	
N	51	51	
Origin FE	Yes	Yes	
Destination FE	No	No	

Notes: The table presents OLS and IV estimates of equation (7) at the state level for FDI and trade. The dependent variables are the log number of total FDI links in 2014 (Panel A), the log of aggregate exports (from the US state) (Panel B), aggregate imports (Panel C), exports to Vietnam (Panel D), and exports to Japan (Panel E). Exports and imports are measured in US dollars in 2011. In all columns, we use $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,...,2000}$ 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 double clustered at the destination state and origin country. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Appendix Table 21: Thought Experiment: A Gold Rush in Los Angeles in 1880

			Predicted Counterfactual Change	
	Ancestry 2010	FDI #	*Ancestry 2010*	*FDI # (in %, IV)*
	(1)	(2)	(3)	(4)
Germany	343,276	241	+65,344	+62.55
Ireland	256,621	40	+61,701	+58.21
UK	396,439	582	+26,645	+21.91
Norway	39,515	55	+4,657	+3.52
Sweden	51,395	71	+4,010	+3.03
France	77,372	278	+3,293	+2.48
Canada	27,722	531	+3,132	+2.36
Switzerland	10,156	162	+2,456	+1.84
Czechoslovakia	17,905	4	+2,140	+1.60
Netherlands	38,392	121	+1,638	+1.23

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 as in reality (columns 3 and 4). Column 3 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 principal components. Column 4 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 $\{I^t_{o,-r(d)}(I^t_{-c(o),d}/I^t_{-c(o)})\}_{t=1880,\ldots,2000}$, similar to column 2 of Appendix Table 19 without the principal components as instruments. All regressions control for log distance and latitude difference and include a origin × destination-census-region, and destination × continent-of-origin 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 thought experiment are presented in section 3.6.

Appendix Table 22: Search terms for Germany and Italy

Germany	Italy
Politicians	
Angela Merkel	Aldo Moro
Helmut Kohl	Benito Mussolini
Willy Brandt	Alessandra Mussolini
Joseph Goebbels	Amintore Fanfani
Karl Marx	Angelino Alfano
Actors	
Jürgen Prochnow	Isabella Rossellini
Til Schweiger	Robert De Niro
Franka Potente	John Turturro
Udo Kier	Roberto Rossellini
Daniel Brühl	Roberto Benigni
Athletes	
Katarina Witt	Mario Andretti
Dirk Nowitzki	Armin Zoggeler
Boris Becker	Roberto Baggio
Steffi Graf	Andrea Barzagli
Franz Beckenbauer	Gerhard Plankensteiner
Musicians	
Ludwig van Beethoven	Antonio Vivaldi
Nena	Gioachino Rossini
Johann Sebastian Bach	Giacomo Puccini
Nina Hagen	Ennio Morricone
Felix Mendelssohn	Luciano Pavarotti

This table shows the top five results from Google's Answer Box for each category for the countries Germany and Italy when typing "notable [country] [category]" into Google.

Appendix Table 23: The Effect of Ancestry on Language: Language Specific Effects

	Point Estimate	Standard Error	N	# of US-born in d that speak o at home in 2010
Aleut	1.608***	(0.028)	3,137	116
Malay	1.376	(0.897)	9,411	176
Arabic	1.222***	(0.171)	78,376	45,953
Spanish	1.172***	(0.380)	65,877	65,877
French	0.212***	(0.018)	87,836	87,416
Haitian Creole	0.198***	(0.015)	3,137	595
Greek	0.196***	(0.033)	6,273	1,849
Vietnamese	0.188***	(0.006)	3,137	1,739
Portuguese	0.174***	(0.030)	28,233	13,095
Korean	0.170***	(0.041)	6,274	3,040
Mon-Khmer	0.167***	(0.005)	3,136	316
Urdu	0.159***	(0.020)	3,137	499
Bengali	0.153***	(0.015)	3,137	191
Japanese	0.142***	(0.007)	3,137	1,972
Persian	0.102***	(0.005)	3,137	538
Chinese	0.085***	(0.005)	3,137	1,745
Thai	0.077***	(0.010)	3,137	619
Polish	0.061***	(0.015)	3,137	1,670
Filipino	0.059***	(0.002)	3,137	1,229
Laotian	0.050***	(0.009)	3,137	592
Albanian	0.049***	(0.014)	3,137	181
Italian	0.041***	(0.005)	6,274	5,068
Samoan	0.036	(0.031)	3,137	261
Amharic	0.035**	(0.015)	3,137	123
Tongan	0.034	(0.027)	3,137	115
Russian	0.027***	(0.008)	3,137	53
German	0.022***	(0.002)	15,685	15,513
Hindi	0.015***	(0.003)	3,137	558
Rumanian	0.014**	(0.007)	3,137	417
Turkish	0.012**	(0.005)	3,137	263
Croatian	0.012**	(0.006)	3,137	65
Swahili	0.006	(0.005)	6,274	606
Finnish	0.005	(0.012)	3,137	373
Magyar	0.004**	(0.002)	3,137	578
Indonesian	0.002	(0.003)	3,137	130
Swedish	0.002**	(0.001)	3,137	888
Dutch	0.002**	(0.001)	9,411	4,746
Norwegian	0.002**	(0.001)	3,137	965
Pashto	0.002	(0.002)	3,137	26
Czech	0.001***	(0.000)	3,137	367
Burmese	0.001	(0.001)	3,137	19
Sinhalese	0.000	(0.001)	3,137	9
Danish	0.000***	(0.000)	6,274	1,200
Irish	0.000***	(0.000)	3,137	459
Afrikaans	-0.000***	(0.000)	3,137	6
Nepali	-0.000***	(0.000)	3,137	52

Bulgarian	-0.000***	(0.000)	3,137	85
Creole	n/a	n/a	3,136	2,252
Bantu	n/a	n/a	9,411	432

Notes: The table is an extension of Table 8, where we only show the results for a set of selected languages. The table is sorted on the size of the point estimate. The last column shows the # of US-born residents in d that speak the language of o at home.

Appendix Table 24: Accounting For the Effect of Ancestry

	FDI Dummy (2014)				
	(1)	(2)	(3)	(4)	(5)
Log Ancestry 2010	0.222***	0.213***	0.212***	0.213***	-0.025
	(0.021)	(0.068)	(0.068)	(0.025)	(0.028)
Sector Similarity (Rank Correlation)		0.012			
		(0.020)			
Sector Similarity (Cosine Correlation)			0.022		
			(0.023)		
Log # of residents in d that speak language of o at home				0.005	
				(0.006)	
Information Demand Index (standardized)					0.078***
					(0.013)
N	612,495	23,708	23,708	454,812	19,110
Destination FE	Yes	Yes	Yes	Yes	Yes
Origin FE	Yes	Yes	Yes	Yes	Yes
Principal Components	Yes	Yes	Yes	Yes	Yes

Notes: This table shows IV regressions at the county-country (columns 1-4) and DMA-country (column 5) level. Each column is a variation of the simple specification (column 2 in Table 3) that has origin and destination fixed effects. All variables are defined as in the previous tables. The relatively low number of observations in columns 2 and 3 is due to data availability in the industry share of employment: When calculating the correlation between industries' share of employment in county d and country o, the correlation coefficient is missing for those country-county pairs that have at least one missing share of employment. 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 25: Generational Effects

	FDI 2014 (Dummy)					
	IV	IV	OLS	IV	IV	IV
	(1)	(2)	(3)	(4)	(5)	(6)
Log Ancestry 2010	0.187***		0.155***	0.242***		0.163***
	(0.024)		(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)
N	612,495	612,495	612,495	612,495	612,495	612,495

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 Kleinbergen-Paap rk LM statistic on the excluded instruments is 18.211 with a p-value of 0.150. We are thus unable to reject the null that our instruments do not induce differential variation in the two endogenous variables, and interpret any difference in the coefficient estimates with caution. The Kleinbergen-Paap rk LM statistic on the excluded instruments is 29.04 with a p-value of 0.007. We thus have sufficient power to detect differences between the coefficient estimates on the two endogenous variables.