Migrants, Ancestors, and Investment*

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

November, 2015
Preliminary and Incomplete

Abstract

We use 130 years of data on historical migrations to the U.S. to show a causal effect of the ancestry composition of U.S. counties on foreign direct investment (FDI) sent and received by firms within these counties. To isolate the causal effect of ancestry on FDI, we build a simple reduced form model of migrations: migrations from a foreign country towards a U.S. county at a given time depend on (i) a push factor, causing emigration from that foreign country to the whole U.S, and (ii) a pull factor, causing immigration from the whole world into that US county. The interaction between time-series variation in country-specific push factors and time-series variation in the county-level pull factors generates quasi-random variation in the allocation of migrants across U.S. counties. We find that a doubling of the number of residents with ancestry from a given foreign country relative to the mean increases the probability that at least one local firm invests in that country by 4.2 percentage points, and increases the number of employees at domestic recipients of FDI from that country by 31%. The size of these effects increases with the ethnic diversity of the local population, the geographic distance to the origin country, and the ethno-linguistic fractionalization of the origin country.

JEL Classification: O11, J61, L14.

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

*We are grateful to Richard Hornbeck, Emir Kamenica, and Nancy Qian. We also thank seminar participants at the CEPR's summer trade meeting in Paris (ERWIT) and at the University of Chicago for useful comments. All mistakes remain our own.

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

‡Toulouse School of Economics and CEPR, 21 Allée de Brienne 31000 Toulouse, France; E-mail: thomas.chaney@gmail.com.

§University of Chicago, Booth School of Business, NBER and CEPR, 5807 South Woodlawn Avenue, Chicago IL 60637 USA; E-mail: Tarek.Hassan@ChicagoBooth.edu.

1 Introduction

Over the last two centuries, the U.S. has attracted migration from all corners of the earth, making it one of the most ethnically diverse nations on the planet. As of the 2010 census, 63% of U.S. residents could trace their ancestry to a foreign country, including descendants of migrants from virtually all countries in the world. At the same time, U.S. firms send and receive a vast amount of foreign direct investment to these countries. In this paper, we quantify the causal impact of the ancestry composition of U.S. counties, defined as the presence in the U.S. of descendants of foreign migrants, on their patterns of foreign direct investment.

We find a sizeable causal impact of foreign ancestry on foreign direct investment. For an average U.S. county, doubling the number of individuals with ancestry from a given average foreign country, increases the probability that at least one firm from that U.S. county has foreign direct investment relations with that foreign country by 4.2 percentage points (or 237% relative to the mean). At the intensive margin, it increases the total number of firms engaged in such activity by 10%, and increases the number of local jobs at subsidiaries of firms headquartered in the origin country by 31%. This effect appears strongly concave, where going from 1,000 to 10,000 descendants increases the probability of an FDI link by a factor of 3, suggesting that higher diversity of origins may increase the overall level of FDI activity in a given destination.

To assess the *causal* impact of foreign ancestry on the patterns of foreign direct investment, we follow an instrumental variable strategy using the history of migrations into the U.S. We isolate quasi-random variation in the allocation of migrants across destinations within the U.S. that results from the interaction of two facts: First, migrants from different origins tended to arrive in the U.S. at different times. Second, the set of destinations that are most economically attractive to the typical migrant arriving in the U.S. changed over time.

We motivate our approach using a simple reduced-form dynamic model of migrations. Migrations from a given foreign origin country o towards a given destination U.S. county d in period t depend linearly on the total number of migrants arriving in the U.S. from o (a push factor), the relative economic attractiveness of destination county d to migrants (a pull factor), and the size of the pre-existing local population of ancestry o (a recursive factor), allowing for the fact that migrants tend to prefer settling near to others of their own ethnicity. Recursively solving the model shows that the number of residents today who are descendants of migrants from o is a function of simple- and higher-order interactions of the push and pull factors at time t.

To prevent our results from being driven by omitted variables that affect both migrations and FDI, we interact the number of migrants arriving from o at time t who do not settle in the

vicinity of d (origin push factor), with the share of migrants arriving at time t that settle in destination d who are not from the same continent as country o (destination pull factor). We predict the ethnic composition of U.S. counties in 2010, using the entire history of migrations towards the U.S. over 130 years, isolating variations in today's ancestry composition of U.S. counties that derive solely from time-series variations in the relative attractiveness of different destinations within the U.S.

To illustrate our approach, consider the examples of migrations from Ireland and Germany. The peak of Irish migrations to the U.S. occurred in the mid-nineteenth century, when the great potato famine triggered massive migrations from Ireland. At that time, the eastern half of the U.S. was attracting large numbers of foreign migrants, including from non-European countries. We predict a large population with Irish ancestry on the eastern half of the U.S. today, as subsequent waves of Irish migrants were more likely to settle where previous Irish migrants had gone. The peak of German migrations occurred several decades later, at the end of the nineteenth century. At that time, it was the Midwest which was booming and attracting large numbers of migrants. We predict a large population with German ancestry in the Midwest, which persists to this day. It is the difference in the timing of migrations from Ireland and Germany which generates quasi-random variations in the predicted distribution of ancestry across the U.S.

Our main identifying assumption for a causal interpretation of the results (we relax it further in the main text) is that omitted variables making a given location within the U.S. differentially more attractive for migrants from a specific origin for both settlement and FDI do not simultaneously affect large groups of migrants originating from different continents nor do they have large effects outside of the surrounding states of the destination in question.

Flexibly applying our instrumentation strategy to the entire set of origin countries and destinations counties allows us to corroborate our identifying assumption in various ways and to guard against a range of potentially confounding factors using fixed effects. We are able to simultaneously control for both origin and destination fixed effects, thus controlling for all origin and destination-specific factors, such as differences in size, market access, and productivity. Most existing studies based on an instrumental variable approach focus on the U.S. and a single foreign country, and are thus unable to control for destination characteristics. Regional interactions of the origin and destination fixed effects also allow us to address a wide range of threats to our identifying assumption. We are also able to conduct a range of falsification exercises and robustness checks. For example, we obtain quantitatively similar effects of ancestry on FDI if we condition only on migrations from former communist countries that occurred during a period of economic isolation between the origin country and the U.S.

Our approach delivers the statistical power and a sufficiently large number of instruments to reliably estimate heterogeneous effects of ancestry on FDI across origins, destinations and sectors. Across countries, the effect is almost always positive and significant, and of similar size. Among the five largest origin countries, the estimates range from 0.174 (s.e.=0.011) for Mexico to 0.265 (s.e.=0.009) for the UK. The effect increases with the geographic distance and the judicial quality of the origin country, and with its ethnolinguistic fractionalization. Once these characteristics are accounted for, other measures of genetic, linguistic, and religious distance do not significantly affect the size of the effect of ancestry on FDI. The effect is similar across destination counties, but increases with the degree of ethnic diversity. The effect is heterogeneous across sectors, being about five times larger for manufacturing than for natural resources.

We find evidence of negative spill-over. The effect of ancestry on FDI falls with the population of migrants from the same origin in neighboring counties, and from neighboring countries: a large Polish contingent in a given county has a lower effect on FDI if nearby counties also host large contingents from Poland, or if this county hosts a large contingent from the Czech Republic.

We are able to distinguish the effect of first-generation immigrants from the effect of ancestry (that is, the presence of the descendants of first-generation immigrants). Once the effect of ancestry is controlled for, first-generation immigrants have no additional effect on the probability that local firms will engage in FDI with the origin.

By aggregating our dataset to the state level, we are also able to consider the effect of ancestry on trade flows between U.S. states and foreign countries. Although our results with respect to FDI remain largely stable after performing this aggregation, we find no systematic causal impact of ancestry on the patterns of international trade once we control for county fixed effects.

We view our results as convincing evidence that migration, and the ethnic diversity resulting from it, affect the pattern of international investment in a quantitatively important way. This effect appears to unfold and persist over several generations, and to increase in size even after the first generation of immigrants has passed. Our evidence on heterogeneous effects, negative spill-overs, and the concavity of the effect is consistent with a positive impact of regional diversity on FDI: the more diverse and less ethnically homogeneous the local population, the larger the total effect of ancestry on FDI. It is also consistent with the "strength of weak ties" (Granovetter, 1973), where a small minority from a distant part of the world that is is not otherwise represented in the local ethnic mix has the largest marginal impact on FDI.

To illustrate the quantitative implications of our results we conduct two counterfactual calculations.

Finally, we note two important limitations to our analysis. First, our results rely purely on

cross-sectional variation in FDI within the U.S. Although we believe that, in light of our results, it is likely that the ethnic diversity of the U.S. also raises the extent of FDI for the U.S. as a whole, we cannot exclude the possibility that increases in FDI in one state are partially or fully offset by decreases in FDI elsewhere. Second, although our results on the heterogeneity in the effect of ancestry on FDI provide some indirect evidence on the nature of the mechanisms linking ancestry composition to FDI, our results do not offer any direct evidence. We believe that the comparative advantage of our present setup is to establish causality and quantify the aggregate effect of ancestry within a setting with a high degree of external validity. Quantifying the relative roles of underlying microeconomic mechanisms, such as the transmission of information via co-ethnic networks versus the provision of social collateral, and enforceability of contracts, requires data at a lower level of disaggregation than we have available. Moreover, we believe that the answers to these questions are likely to vary by ethnicity and historical context such that they may be best addressed in studies of specific historical episodes, including some of our earlier work (Burchardi and Hassan, 2013).

A large body of literature shows that measures of affinity between regions, such as common ancestry, social ties, trust, and telephone volume, correlate strongly with aggregate economic outcomes, such as foreign direct investment (Guiso et al., 2009), international asset flows (Portes and Rey, 2005), and trade flows (Gould, 1994; Rauch and Trindade, 2002).[1] How much of this association should be interpreted as causal, however, remains an open question as these measures of affinity, and ancestry in particular, are likely to be nonrandom across regions. Two recent papers use historical decisions by the U.S. government on the location of Japanese internment camps during World War II (Cohen et al., 2015) and the placement of Vietnamese refugees after the Vietnam War (Parsons and Vezina, 2014) to identify a causal effect of concentrations of descendants of these migrants on contemporary trade flows between locations within the U.S. and Japan and Vietnam, respectively. Burchardi and Hassan (2013) use variation in wartime destruction across West-German regions in 1945, a time when millions of refugees were arriving from East Germany, as an instrument for the the share of the population with social ties to the East, and show evidence of a causal effect of these social ties on GDP growth, entrepreneurial activity, and FDI in East Germany after the fall of the Berlin Wall.[2]

We contribute to this literature in several ways. First, we are able to identify a causal

[1] Combes et al. (2005) and Garmendia et al. (2012) study the impact of migrations on trade using variation only within France and Spain, respectively. Head and Ries (1998) study the impact of migrations on trade of Canadian provinces. Aleksynska and Peri (2014) distinguish the impact of migrations on trade by migrants' occupation. Docquier et al. (2014) distinguish between potential and actual migrations.

[2] See Fuchs-Schuendeln and Hassan (2015) and Chaney (2014) for surveys of this literature.

effect of ancestry on FDI in a setting with a high degree of external validity. This is directly relevant to assessing, for example, the long-term effects of immigration policy. Second, because our identification strategy can be flexibly applied to the entire set of origin countries, we are able to guard against a wide range of possible confounding factors and to relate directly to the previous literature by employing a gravity equation that features both destination and origin fixed effects. This gravity form also allows us to quantitatively compare the relative contributions of conventional gravity forces and of ancestry in determining the patterns of FDI. Third, the increased statistical power allows us to identify the determinants of heterogeneity in the effect of ancestry on FDI across origins,[3] destinations, and sectors, as well as show evidence of negative geographic spill-overs. Fourth, our results suggest the causal effects of ancestry on FDI and trade flows may be very different, although they appear similar in OLS regressions.

Our paper also contributes to the debate on the costs and benefits of immigration. Much of the existing literature has focused on the effects of migration on local labor markets (Card, 1990; Card and Di Nardo, 2000; Friedberg, 2001; Cortes, 2008). to this literature by examining the effect of migration on the pattern of economic exchange and employment. By looking at the composition of residents of the U.S. with *foreign ancestry*, as opposed to just *foreign born* residents, we are able to separate the short-term and long-term effects of migration. In this sense, our results show a long-term effect of migration on the comparative advantage in FDI of different regions that operates across multiple generations.

On a methodological level, our paper relates to a wider literature that has employed leave-out shift-share instruments in the identification of causal effects (Bartik, 1991; Katz and Murphy, 1992). Our instrumentation strategy, based on a simple recursive model of migrations, can easily be replicated for other countries, other time periods, or even other variables beyond migrations where cumulated flows matter. Although it requires information on flows of migration covering several decades, which may be prohibitively costly in some context, our identification strategy does not require a rare or even unique historical accident. Our approach can therefore be applied to other countries and outcome variables.

The remainder of this paper is structured as follows. Section 2 introduces our data. Section 3 gives a brief overview of the history if migration to the U.S. Section 4 identifies the causal effect of ancestry composition on FDI at the extensive margin, discusses various challenges to our identifying assumption, and conducts a range of robustness checks and falsification exercises. Section 5 probes the effect of ancestry composition on FDI at the intensive margin and other

[3]For instance, using the measure of fractionalization from Alesina et al. (2003), we find ancestry matters more for FDI to and from countries with higher degree of fractionalization.

outcomes, and illustrates the quantitative implications of our findings using two counterfactual calucations. Section 6 examines the mechanism underlying the effect of ancestry on FDI by testing for spill-over effects and by probing the heterogeneity of the effect across countries, counties, and sectors. Section 7 concludes.

2 Data

We collect data on migrations and ancestry, on foreign direct investment and trade, and on country and county characteristics. Below is a description of our data, along with their source.

Migrations and ancestry.— All migration and ancestry data are constructed from the individual files of the Integrated Public Use Microdata Series (IPUMS) samples of the 1880, 1900, 1910, 1920, 1930, 1970, 1980, 1990 and 2000 waves of the US census, and the 2006-2010 5-year sample of the American Community Survey. We aggregate the individual level data to counts of respondents at the level of historic US counties (or county groups from 1970 onwards) and foreign countries.

Throughout the paper, we use the convention $t-1$ to denote the census wave just before t, o for the foreign country of origin, and d for the U.S. destination county. We construct the number of migrants from origin country o to destination country d in t, $I^t_{o,d}$, by counting the number of respondents who live in d, were born in origin country o , and immigrated to the United States between t and $t-1$. The exception to this rule is the 1880 census, which did not record the year of immigration. The variable $I^{1880}_{o,d}$ measures the number of individuals whose birthplace is in origin country o and who live in d at the time of the census.[4]

Since 1980, respondents are also asked about their ancestry in both the US Census and the American Community Survey, and they can provide multiple answers. Our measure *Ancestry (o,d)* is constructed by counting the number of individuals residing in d at time t who report as first ancestry o.

The definition of (i) US counties and changed over time, and the responses to the (ii) birthplace and (iii) ancestry question do not necessarily refer to a modern day country. In order to construct a panel, we transition all data to the 1990 definition of US counties and foreign countries. The t-specific US county transition matrix applies the share of the area of each historic US county in each modern US county as weights. For all data from 1970 onwards, we use population weights to transition data from the period t county group level to the period-t county

[4]If the own birthplace is in the US, imprecisely specific (e.g. a continent) or missing, we instead use the parents birthplace, giving equal weights to father and mother.

level. The t-specific foreign country transition matrix applies as weights, if possible, the share of the population of a historic country that lived in the area of the modern country and otherwise the share of the area of the historic country that is in 1990 part of each modern country. We transition the ancestry data by assigning to each response a probability with which the answer refers to an area in modern country o. See Appendix A.1 for details on these procedures.

Foreign direct investment.— Our data on FDI is from the U.S. file of the 2014 edition of the Bureau van Dijk ORBIS dataset. For each U.S. firm, the database lists the location of its (operational) headquarters, the addresses of its foreign parent entities, and the addresses of its international subsidiaries and branches. We consider a foreign entity to be a parent of the U.S. firm if the foreign entity owns at least 5% of the firm's equity. Similarly, we consider a foreign entity a subsidiary of the U.S. firm if the U.S. firm owns at least 5% foreign entity's equity. Altogether we have information on 36,108 U.S. firms that have at least one foreign parent or subsidiary. Collectively, these firms have 102,618 foreign parents and 176,332 foreign subsidiaries in 142 countries (in their 1990 borders). We then aggregate this information to the county level. Our main outcome variable, *FDI Dummy*, is one if at least one firm within a given destination county has at least one parent or subsidiary in origin country o. For each U.S. county we also count the total number of FDI linkages (the total number of foreign parents and subsidiaries of all firms within the county), and the total number of unique parents and subsidiaries in both the origin and the destination. We also count the total number of employees working at firms with a foreign parent in a given origin country (*# of Employees at Subsidiaries in Destination*) and the total number of employees working at subsidiaries of U.S. firms in a given origin country (*# of Employees at Subsidiaries in Origin*).[5] The ORBIS database also gives the 2007 NAICS code of the sector. Appendix A.2 gives further details on the construction of our FDI 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 5 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 use this readily available data aggregated at the US state and foreign country level. While the data on bilateral trade flows between US states and foreign countries from the Commodity Flow Survey does not cover all of the US foreign trade (it comes

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

from a representative survey, not from the universe of foreign transactions), it is the only publicly available source of international data disaggregated at a geographic level below that of the entire US. This allows us to contrast the patterns of trade between different US states as a function of the ethnic composition of those states.

For each origin country o and destination US state d, $Import_{o,d}$ 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 use the coordinates for all postal codes[6] within a county or state, and the coordinates of the main city for foreign countries.[7] 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.— In order to shed light on the mechanism through which the presence of foreign ancestry affects the patterns for foreign investment, we construct several measures of foreign country and US county characteristics. $Genetic\,Distance_o$ is a measure of the genetic distance between foreign country o and the US, normalized to take values between zero and one. $Linguistic\,Distance_o$ is a measure of the linguistic distance between foreign country o and the US; it measures the probability that a randomly selected person in the US speaks the same language as a randomly selected person form country o. $Religious\,Distance_o$ measures the religious distance between foreign country o and the US, with a similar construction as the linguistic distance.[8] A higher index for $GeneticDistance_o$, $LinguisticDistance_o$ or $ReligiousDistance_o$ corresponds to a higher distance between the US and country o. $Judicial\,Quality_o$ is a measure of the judicial quality in country o.[9] A higher index for $JudicialQuality_o$ corresponds to a higher

[6]The geo-coordinates are downloaded from www.geonames.org.

[7]The geo-coordinates are downloaded from www.cepii.fr.

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

[9]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.

quality judicial system. $Fractionalization_o$ is a measure of the ethnolinguistic fractionalization of country o.[10]

US county characteristics.— We define three US county level measures. $Ethnic\,Diversity_d$ is a measure of the diversity of ethnic groups in county d.[11] $Foreign\,Share_d$ measures the share of residents in county d who claim foreign ancestry. $Population_d$ is the population size of US county d. $IncomePerCapita_d$ is a measure of the per capita income (in thousand dollars) within US county d. Both population and income data are taken from the Bureau of Economic Analysis.

Panel A of Table 1 gives summary statistics on our sample of 3,141 $\times$ 195 origin-destination pairs. Column 1 shows means and standard deviations for all observations. Columns 3-4 show the same statistics for the sub-samples 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 county having any population with origins in the foreign country, 3.1% have an FDI link. The larger this population the larger is the probability of finding an FDI link, with 12.8% of the origin-destination pairs in the top quintile having an FDI link. Similarly, about half of the origin-destination pairs have ancestry of zero, reflecting the fact that most destinations in the U.S. do not have populations with ancestry from all 195 countries. The mean number of individuals with ancestry from a given origin is 318, but is highly skewed, with a mean in the top quintile of 2.8 million individuals. Compared to this stock of ancestry, the flow of immigrants between 1990 and 2000 is relatively small, with 25 on average across the sample. The summary stats also shows that the number of first-generation immigrants measured in the 2010 American Communities Survey (Total Immigrants 2010) is grossly understated (29 on average). This fact is known in the literature and appears to affect only the measurement if immigration flows but not the stock of ancestry (Jensen et al. (2015)). For this reason we exclude the 2000-2010 wave of migrations from our standard specification and instead rely on the pre-2000 census numbers.

Panels B and C show summary statistics following the same format for destination counties and origin countries used in our estimation of heterogenous effects.

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

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

3 Historical background

The U.S. census conducted in 1880 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 100 million immigrants from 195 countries. Our sample period thus covers the vast majority of migration to the U.S.

During the first part of this period up until World War I, migration to the U.S. was largely un-regulated. Particularly European migrants faced little or no restrictions to entry and came in large numbers. Figure FIXME shows the extent and the changing composition of migration over time. While the peak of British migration was passed before the beginning of our sample, the number for 1880 clearly show effect of the potatoe famines and the subsequently large inflow of (predominantly poor) 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, followed by a peak in Italian migration in 1910, and in migrations from Eastern Europe. The inflow of migrants overall dropped dramatically during World War I, falling below 4 million during the period between 1910 and 1920.

While the timing of these earlier European migrations was clearly dominated by economic and political factors in the origin countries, U.S. immigration policies became relatively more important during the 1920s. The first important step towards 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 against Asians were followed by literacy and various other requirements that came into effect after 1917, culminating in the establishment of a quota system in 1921.

Up until 1882, migration into the U.S.

1) Relatively unfettered migration for all non-Asians until 1920s

Irish

Germans

Scandinavians

Eastern Europeans

Chinese Japanese

changing composition over time

Chines exclusion 1882 extended to other asians over time

literacy tests from 1917

quotas established in 1920s

2)

reduction of migration during WWI

negative net migration in great depression

3) liberalization in 1965 Hart-Cellar act

Asians, Americas

describe different destinations

4 Ancestry and Foreign Direct Investment

We first explore the effect of the presence of descendants of migrants from a given origin country on the probability that at least one firm within a given destination county has an FDI link (incoming or outgoing) with a firm based in the origin country. The structural equation of interest is

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

where $\mathbf{1}\left[FDI_{o,d} > 0\right]$ is a dummy variable that is one if any firm headquartered in destination county d is either the parent or the subsidiary of any firm headquartered in origin country o. $A_{o,d}$ is a measure of common ancestry, usually calculated as the log of one plus the number of residents in d that report to have ancestors in origin o measured in thousands. We choose this functional form in anticipation of non-parametric results discussed below, but show robustness to a wide range of alternative specifications. $X'_{o,d}$ 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, respectively. In addition, most of our specifications also control for the interaction between destination county and continent-of-origin fixed effects, and for the interaction between origin country and destination-census-region fixed effects. The coefficient of interest is β, which measures the effect of ancestry on the probability that an FDI relationship exists between firms in o and d. The error term $\varepsilon_{o,d}$ captures all omitted influences, including any deviations from linearity.[12] Throughout the main text we report heteroskedasticity-robust standard errors clustered at the origin country level. In the appendix, we report standard errors calculated using alternative methods for all the main results of the paper, and show our results are robust.

Equation (1) takes the form of a gravity equation, a form that is widely used in the empirical literature describing the pattern of international trade and FDI. We maintain the same form for

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

consistency with this literature. Moreover, the gravity form is appealing on theoretical grounds as it can be derived in a variety of trade models.[13] In these models, the destination and origin fixed effects absorb all differences in productivity, market size, and market access between origins and destinations which systematically affect prices. We may thus interpret the coefficient β as the effect of ancestry controlling for the large set of conventional economic forces shaping international exchanges.

Equation (1) will consistently estimate the parameter of interest if $Cov\left(A_{o,d}^{2010}, \varepsilon_{o,d}\right) = 0$. However, this condition is unlikely to hold in the data due to omitted variables that may simultaneously affect $A_{o,d}$ and $\mathbf{1}\left[FDI_{o,d} > 0\right]$. For example, we might suspect the same omitted factor made Scandinavians prefer to settle in the Minnesota over other destinations in the U.S. and increases FDI flows (for example because the climate is similar in Minnesota and in Scandinavia). Such omitted variable would bias an OLS estimate of β.

Such concerns are aggravated by the fact that migrants appear to prefer to settle near to populations that share their own ancestry (Card and Di Nardo (2000)). As a result, small variations in omitted factors that affected migration early on may result in much larger variation in the ancestry of present-day populations.

To address these concerns, we build a simple dynamic model of migrations. The number of residents of destination county d of ancestry from origin country o at time t, $A_{o,d}^{t}$, depends on the past stock of residents with ancestry from o and the newly arrived migrants from o who settle in the same destination. We assume the number of new migrants from country o who settle in county d is determined by the combination of three forces: a country-specific *push factor* drives migrants out of country o into the U.S.; a *pull factor* attracts migrants entering the U.S. to county d, irrespective of their origin; and a *social network factor* corresponding to the tendency of newly arrived migrants to settle in communities where people with the same ancestry already live.

Formally, the stock of residents in county d with ancestry from country o at time t follows,

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

The constant terms a_t, $a_{o,t}$ and $a_{d,t}$ control for demographic forces, which may vary over time, over space, and between different ethnic groups. The term, $b_t A_{o,d}^{t-1}$, corresponds to the fact that ancestry is a stock variable, which evolves cumulatively as new waves of migrants arrive. The

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

constant b_t controls for demographics and for how ties to one's ancestry are passed from one generation to the next. The term I_o^t, the total number of migrants from country o entering the U.S., measures the strength of the push factor, the fact that migrants are driven out of country o. The term I_d^t/I^t, the fraction of all migrants entering the U.S. who settle in county d, irrespective of their country of origin, measures the strength of the economic pull factor, the fact that county d is particularly appealing to migrants at time t. Finally, the term $A_{o.d}^{t-1}$ measures the strength of the social network factor, the propensity of migrants to settle near their countrymen. The coefficient d_t controls the relative importance of the pull and social network factors.

Equation (2) is recursive, both because ancestry is passed down from generation to generation (the first $A_{o.d}^{t-1}$ term), and because the choice by newly arrived migrants of where to settle depends on where past migrants have settled (the second $A_{o.d}^{t-1}$ term). We simply assume the initial condition $A_{o,d}^{-1880} = 0, \forall\,(o,d)$. As we are treating our very first census of 1880 differently from the other censuses, recording not only foreign borns, but also children of foreign borns, this seems to be a good approximation for the U.S., where the population rapidly increased in the 19th Century. Solving equation (2) recursively, we get,

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

The reduced form equation (3) highlights how ancestry can be understood as the result of a sequence of migration waves, and their subsequent cumulative effect. In each period t, the interaction of the contemporaneous push factor (I_o^t) and economic pull factor (I_d^t/I^t) determine the flow of migration from o to d. These initial waves of migrants are then amplified by demographic factors (the b_s's) and by the social network factor (the $c_s d_s$'s), as new immigrants from o tend to settle where previous waves of migrants from o have already settled. Note that this simple specification is flexible. In particular, cases where no migrants from a given origin country exist for some initial period of time are perfectly admissible. In the absence of a social network factor, $d_t = 0$, ancestry can simply be understood as the sum over all time periods of the interaction of the push and economic pull factor.

Crucially, this specification highlights potentially exogenous sources of variation in $A_{o,d}^{2010}$. It suggests that plausibly exogenous variation in $I_o^t \frac{I_d^t}{I^t}$ would allow to construct an instrument for $A_{o,d}^{2010}$. Any estimate of β in (1) would be biased if $cov\left(I_o^t \frac{I_d^t}{I^t}, \varepsilon_{o,d}\right) \neq 0$. By interacting a push factor, I_o^t, which is not specific to county d, but common to all destinations in the U.S., and an economic pull factor, I_d^t/I^t, which is not specific to country o but to migrants form all countries, we rule out most plausible sources of endogeneity. However, our exclusion restriction could still

be violated if at some point in time, migrants from o to d represent a large fraction of all migrants from o ($I^t_{o,d}$ a large fraction of I^t_o), or a large fraction of all migrants to d ($I^t_{o,d}$ a large fraction of I^t_d), or if migrants from other origins with unobserved similarities to o represent a large fraction of all migrants. To address these concerns, we exclude migrants from o going to all counties in the same Census region (a group of states four to seven adjacent states) as d from the push factor, and migrants from all countries in the same continent as o from the economic pull factor. In other words, we replace I^t_o by $I^t_{o,-r(d)}$, the migrants from o who settle in counties *not* in the same Census region as d; and we replace I^t_d/I^t by $I^t_{-c(o),d}/I^t_{-c(o)}$, the fraction of migrants *not* coming from countries in the same continent as o who settle in county d. $-r\,(d)$ stands for all counties outside of d's Census region, and $-c\,(o)$ stands for all countries outside of o's continent.

Using the functional form derived in (3), and replacing the $I^t_o \frac{I^t_d}{I^t}$ terms by $I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$, our first stage specification is thus

$$A^{2010}_{o,d} = \delta^f_o + \delta^f_d + \sum_{t=1880}^{2000} \alpha^f_t I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} + \sum_{n=1}^{5} \delta_n PC_n + X'_{o,d}\gamma^f + \eta_{o,d}, \tag{4}$$

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

Our key identifying assumptions is

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

Simply put, it requires that any confounding factors that may make a given destination more attractive for both migration and FDI from a given origin country do not simultaneously affect the settlement of the average migrant originating from another continent and the total number of migrants arriving from the same origin but settling in a different census region.

To further relax this assumption, most of our specifications also control for interactions of fixed effect that are symmetric to the construction of our instruments: the interaction between destination county and continent-of-origin fixed effects ($\delta^f_{o,-r(d)}$ terms) and for the interaction between origin country and destination-census-region fixed effects ($\delta^f_{-c(o),d}$ terms). For these specifications the identifying assumption is then merely that any confounding factors do not systematically affect both sides of the interaction across origin-destination pairs. We further corroborate this assumption below using a series of falsification exercises and placebo treatments.

4.1 The First Stage Relationship

We start with a stylized graphical presentation of our first stage identification. Figure 2 shows the change over time of which locations in the U.S. are most attractive to foreign migrants. From each census from 1880 to 2010, we compute the total number of new migrants in each U.S. county decade by decade.[14] To facilitate the interpretation, we then project the number of migrants in a destination for each decade on a destination and a year fixed effect, and show only the residuals from this projection. Figure 2 clearly shows how migrants initially settled on the East coast of the U.S., then the frontier for migrants started moving to the Midwest (in the late 19th Century), then to the West (in the 1930's), and finally to the South (in the 1980's). Starting in the 1970's, we can also see graphically migrants going to urban centers, with a series of dark dots appearing around large urban areas. Overall, this figure clearly shows the time-series variation in the attractiveness of different destinations within the U.S., i.e. variation in our pull factor over time and across destinations.

We then study two specific examples graphically, that of migrations from Ireland, with a migration peak in our first 1880 Census (corresponding to the massive migrations from Ireland to the U.S. in the 1850's), and that of migrations from Germany, with a migration peak in the 1880-1900 period. The top panel of Figure 3 shows the relative attractiveness of U.S. counties for pre-1880 migrants, where we exclude migrations from Europe so that Irish migrants do not drive our results. Not surprisingly, most migrants at that time settled along the Eastern seashore. We therefore expect most Irish migrants from this initial wave settled in the Eastern part of the U.S. The bottom panel of Figure 3 shows the distribution of U.S. residents with Irish ancestry in 2010. The Eastern part of the U.S. still contains disproportionately many Irish descendants. The top panel of Figure 4 shows the relative attractiveness of U.S. countries for non-European migrants during the 1880-1900 period, i.e. the peak of German migrations to the U.S. At that time, the preferred destination for migrants had shifted to the Midwest and the West. We expect many German migrants from this first wave to have settled in the Midwest and the West. The bottom panel of Figure 4 shows the distribution of German descendants in 2010, with large populations in the Midwest. Note that for both both countries, those initial waves do not predict large populations of Irish and German descendants in Florida and Southern California, two areas with both large Irish and German ancestry, and strong FDI ties to Ireland and Germany. Those settlements are either due to some common unobserved factors (precisely what we want to avoid

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

with our instrumental variable approach), to forces outside our simple reduced form model, or to subsequent waves of migration. Figures 3 and 4 suggest a visual example of the variation captured by our instrumental variable approach.

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

All of the nine coefficients shown in column 1 are positive and seven are statistically significant at the 5% level. Figure 5 depicts the coefficients graphically. The first main insight from this figure is that even our earliest (pre 1880) snapshot of the cross-sectional variation in economic attractiveness to new migrants has left its imprint on the present-day ancestry composition of U.S. counties: the destination counties that were relatively more attractive to the typical migrant arriving pre 1880 continue to house significantly larger numbers of residents of the ethnic groups arriving at the U.S. border prior to 1880. The overall pattern of coefficients suggests a hump shape, where very recent waves of migrants have a smaller impact on current ancestry than migrations a few decades back, but the effect of past migrations eventually fades away after about one century. The pattern is consistent with the effects of two opposing forces, where migrants arriving earlier tend to pass their ancestry on to more than one descendant, and migrations reaching further back into history are more likely to be forgotten. A possible explanation for the smaller and insignificant coefficient for 1920-30 is the Great Depression, which induced large reverse migrations from the U.S. Our model is well suited for periods of positive net inflows of migrants, less so for outflows of migrants.

Taken together, the nine simple interactions incrementally increase the R^2 of the regression by 8 percentage points and explain about 16% of the between-variation. Column 2 adds controls for distance and latitude difference. Column 3 adds the principal components, raising the R^2 by another three percentage points. Column 4 shows our standard specification, controlling

in addition, for the interaction of destination and continent-of-origin fixed effects as well as the interaction of origin and destination-census-region fixed effects. The F-Statistic against the null that the excluded instruments as irrelevant in this specification is 433.[15] The remaining columns successively add additional controls and show variations of our instrumentation strategy, including migration data from the 2005-2010 ACS survey, dropping migration prior to 1880, and estimating our standard specification in levels rather than logs. Throughout all of these variations we can comfortably reject the null that our instruments are jointly irrelevant in the first stage.

As a prelude to our instrumental variables estimates, Figure 11 plots the coefficients from a reduced-form regression that relates the FDI Dummy directly to the simple push-pull factor interaction terms $\left\{ I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} \right\}$ and the controls from our most parsimonious specification in column 1 of Table 2. (That is, it controls only for destination and origin fixed effects and omits the instruments consisting of higher-order interactions.) All nine coefficients are greater than zero, and seven of them are statistically significant at the 5% level. Destination counties that received an (exogenous) increase in the number of migrants from a given origin in any of the nine consecutive waves of immigration thus tend to have a significantly higher probability of engaging in bilateral FDI with these origin countries today.

In our instrumental variables estimation we explicitly test the hypothesis that an increase in the number of descendants from a given origin country increases the probability that at least one local firm engages in FDI with that country. In column 1 of Table 3 we estimate (1) while instrumenting (the log of) ancestry in 2010 with the simple interactions $\left\{ I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} \right\}$ and controlling for origin and destination fixed effects, distance, and the difference in latitude between destination and origin. The coefficient estimate on ancestry is 0.237 (s.e.=0.024) and statistically significant at the 1% level. The coefficient on distance is not statistically distinguishable from zero, perhaps reflecting the fact that U.S. counties do not differ much in their distance to most foreign countries, and that these smaller differences are irrelevant once the effect of the distance between the U.S. as a whole and the country in question is controlled for (it is absorbed in the country fixed effect). By contrast, the difference in latitude is positive and significant, showing that, other things equal, firms tend to engage in FDI with origin countries that are climatically different from their own location.

In column 2 we add the five principal components of the higher-order interactions to our set of instruments, resulting in a slight fall in the coefficient of interest to 0.190 (s.e.=0.024). The estimate implies that doubling the number of residents with ancestry from a given origin (from

[15] The J test statistic for overidentifcation is 18.975 with a p-value of 0.124. We thus fail to reject the null that our instruments are uncorrelated with the error term and correctly excluded from the second stage regression.

318 to 636) increases the probability that at least one firm engages in FDI with that origin by 4.1 percentage points.

Another useful way to gauge the relative importance of ancestry is by comparing its partial R^2 to that of the controls included in the specification. Taken together, the standard gravity terms, that is the origin and destination fixed effects as well as distance and latitude difference explain 20.3% of the variation in FDI Dummy. Adding ancestry to these variables in a simple OLS specification (shown in Panel B) accounts for an additional 9 percentage points of the variation — meaning that it accounts for almost half as much as the combined explanatory power of these conventional economic forces driving economic exchange. Instead adding our nine simple interactions to the standard gravity terms raises the R^2 by 1.5 percentage points and adding them in combination with the five principal components raises it by 1.8 percentage points.

Column 3 shows our standard specification that also includes the interaction of destination and continent-of-origin fixed effects as well as the interaction of origin and census region of the destination fixed effects. For a given origin country, this quite demanding specification uses only variation across different destinations within the same census region while controlling for the fact that each destination may have an idiosyncratically high or low propensity to interact with the continent containing the origin country. Symmetrically, for each destination, it uses only variation across different origins within the same continent while controlling for the fact that each origin may have an idiosyncratically high or low propensity to interact with the destination county. Reassuringly, the coefficient estimate is almost identical to that in column 2 (0.197, s.e.=0.030). Comparing this estimate with the same column in Panel B shows that it is about 25% larger than the corresponding OLS coefficient. The endogenous assignment of migrants to destinations within the U.S. thus appears to induce a mild downward bias in the OLS coefficient.

The remaining columns probe the robustness of this result. Column 4 adds a third degree polynomial in distance and latitude difference, to capture a potentially non-linear effect of distance. Our coefficient of interest remains essentially unchanged (0.197). Column 5 adds an interaction term for the contemporaneous 2010 migrations, again leaving our coefficient of interest unaffected (0.190). Column 6 adds a more stringent set of origin-state fixed effects, exploiting only variations with state, leaving our coefficient of interest the same (0.201).

4.2 Validity of the Exclusion Restriction

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

4.2.1 Placebo test

Our first robustness check uses a placebo treatment. We want to assess whether our instrument really picks up push factors that are specific only to one country, or whether these push factors also affect FDI with other countries in a systematic fashion.

The results are presented in Appendix table 16. In panel A, we assign the interaction between push and pull factors for a given country o to a randomly taken other country, its nearest neighbor in the alphabetic order. To further check whether two countries in different continents may be affected by the same push factors, we also assign in panel B the interaction between push and pull factors for a given country to its nearest neighbor in the alphabetic order in a *different* continent. Finally, to assess whether our controls for the principal components in equation (4), the fixed effects in equation (1), or the controls for distance and latitude difference matter, we estimate various specifications with different sets of controls.

Across all specifications, our placebo treatment is always statistically insignificant, and the point estimates are near zero.

We also bootstrap our placebo test, and repeatedly assign the interaction between push and pull factors for country o to randomly taken other countries. Figure 13 shows the distribution (histogram) of t-statistics on the estimated coefficient on ancestry across 200 random assignments. The t-stats across those random assignments are centered around zero, with no noticeable tendency for positive or negative estimates.

We conclude from this placebo test that our instrument is not picking up any artificial correlation (positive or negative) between the push factors for different countries.

4.2.2 The Communist Natural Experiment

Our second robustness check uses a more direct natural experiment approach. We use specific historical episodes during which FDI between the United States and some specific countries was virtually inexistent, and not expected to be allowed. Those countries are the Soviet Union from 1918 to 1990, China from 1945 to 1980, Vietnam from 1975 to 1996, and Eastern Europe (the non Soviet members of the Warsaw pact) from 1945 to 1989. For each of those countries or sets of countries, we only use as instrument data on migrations from those countries to the U.S. during the period of exclusion. We can plausibly assume migrations during those periods were not driven by the prospect for FDI, which was non existent.

The results are presented in table 4. For all countries, we find a large causal impact of ancestry on FDI, even though we are instrumenting ancestry using only periods when FDI with

those countries was non existent. The magnitude of the estimated impact of ancestry is similar to the one we estimated for all countries in table 3. The estimated coefficients are statistically significant at the 1 and 5% level for the Soviet Union and Eastern Europe respectively. The coefficients are not statistically significant for China and Vietnam. We conjecture this is probably due to the fact that the excluded years (before 1945 and after 1980 for China, before 1975 and after 1996 for Vietnam) cover the period during which almost all of the immigration from these countries to the U.S. happened. We are therefore missing virtually all migrations from these countries with our exclusion periods.

The fact that we find similar results with those natural experiments as in our baseline specification in table 3 makes us confident that our exclusion restriction is valid, and that our baseline results are not driven by reverse causality or omitted variables.

4.2.3 Overview of additional robustness checks

Finally, we run a battery of additional robustness checks. The results are presented in the appendix for concision.

Figure 7 confirms our results are not driven by specific foreign origin countries or U.S. destination counties. We estimate our baseline instrument variable specification separately for individual countries, and separately for individual counties . The estimated impact of ancestry on FDI is similar for all foreign countries (top panel of figure 7), and for all U.S. counties (bottom panel of figure 7). Table 11 shows our estimation results separately for the five largest foreign countries (in number of descendants in the U.S.), and the five largest U.S. counties (in total number of foreign ancestry).[16] The estimated impact of ancestry on FDI is similar across these countries and counties.

We remove the principal components from the set of instruments in column 1 of table 3, and our results are essentially unchanged. In column 5 of table 3, we add the most contemporaneous 2010 immigration flows $\left(I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}} \right)$, and find similar results. We explore other specifications in Appendix table 20, removing or not different sets of migrants from the $I^{t}_{o,-r(d)} \frac{I^{t}_{-c(o),d}}{I^{t}_{-c(o)}}$ interaction terms. In panel A, we do not remove any migrants from o to d when computing our push and pull factors: we use $I^{t}_{o} \frac{I^{t}_{d}}{I^{t}}$ interaction terms. In panel B, we only remove migrants from o to d, not other migrants coming from the same continent or settling in the same region: we use $I^{t}_{o,-d} \frac{I^{t}_{-o,d}}{I^{t}_{-o}}$ interaction terms. In panel C, we remove migrants from o to d, and other migrants coming

[16] Appendix table 19 shows the result from separate regressions for all countries. Appendix table 18 shows the results from separate regressions for all sectors.

from the same continent as o: we use $I^t_{o,-d}\frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$ interaction terms. In panel D, we remove migrants from o to d, migrants coming from the same continent as o, as well as migrants settling in counties adjacent to d: we use $I^t_{o,-adj(d)}\frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$ interaction terms. Across all those alternative specifications, our coefficient of interest varies little, and remains significant.

In table 5, we vary the date at which we measure ancestry (2000 or 2010), shifting the dates for our excluded instruments accordingly. We also perform a horserace between ancestry and recent immigration, experimenting with the most recent immigrants (since 2010), and more established immigrants (since 1970). First, our result on the causal impact of ancestry is stable, and robust to changing the precise date at which we measure ancestry. Second, when both ancestry and immigration are used together as RHS variables, we always find that ancestry has a strong and significant impact on FDI, while immigration does not.

Finally, in Appendix table 17, we experiment with various alternative functional forms for our measure of ancestry. We discuss these results in great details in section 6, as they shed light on the mechanism through which ancestry affects FDI. We only note here that our main result, the strong and significant causal impact of ancestry on FDI, is robust to various functional form specifications.

5 The Effect of Ancestry on other Outcomes

5.1 Inward and outward FDI

We first distinguish between inward and outward FDI. To do so, we estimate our standard specification from column 3 of Table 3 separately for inward FDI, where $Y_{o,d}$ is a dummy equal to one if at least one firm in U.S. county d is a foreign affiliate of a parent firm in foreign country o, and for outward FDI, where $Y_{o,d}$ is a dummy equal to one if at least one firm in U.S. county d is the parent of a foreign subsidiary in country o. The results are presented in Figure 6, showing the two coefficient estimates on ancestry and their 95% confidence intervals. Both coefficients are positive and statistically significant. We find a stronger impact of ancestry on outward FDI, an elasticity of 0.2, and on inward FDI, an elasticity of 0.15, although both coefficients are not statistically different from each other.

5.2 The intensive margin of FDI

So far, we have studied the impact of ancestry on the extensive margin of FDI, the probability that at least one firm engages in foreign direct investment. We now turn to the impact of ancestry

on the intensive margin of FDI: conditional on being positive, how large are FDI flows for a given size of the local population with a given foreign ancestry?

The results are presented in Figure 12 and Table 6. Figure 12 presents our results graphically. We show a scatterplot with $\log(ancestry)$ on the horizontal axis, and $\log(FDI\ volume)$ on the vertical axis, for all county-country pairs in our sample. To show results as close as possible to our regression analysis, we first project both variables on the full set of controls and instrumental variables corresponding to our standard specification in column 3 of Table 3, and display the residuals from this projection. We present two panels, corresponding to two alternative measures of the intensive margin of FDI. In panel A, we use the log of the total number firms receiving FDI (subsidiaries) in both the origin and the destination of a given origin-destination pair. In panel B, we use the log of the total number of firms sending FDI (parents) in the origin and the destination. Both panels show a positive relationship between the size of the local community in county d with ancestry from country o and the number of FDI links between county d and country o.

Table 12 presents the results from a formal regression analysis of the impact of ancestry on the intensive margin of FDI. We estimate by instrumental variable various specifications of the following gravity type equation for FDI,

$$\ln FDI_{o,d} = \delta_o + \delta_d + \beta \ln A_{o,d}^{2010} + X'_{o,d}\gamma + \varepsilon_{o,d} \tag{6}$$

where $FDI_{o,d}$ corresponds to various measures of the volume of FDI between foreign country o and U.S. county d, and where we instrument $\ln A_{o,d}^{2010}$ with the same first stage equation (4) as earlier.

Because of our log specification, cases of zero FDI will automatically be dropped from our sample. This creates a selection problem, as counties with non-zero FDI are likely to be systematically different from those with zero FDI. To correct for this potential selection bias, we implement a simple Heckman type correction. Formally, we first estimate an instrumental variable PROBIT regression for the extensive margin of FDI,

$$\rho_{o,d} = \Pr\left(FDI_{o,d} > 0 | observables\right) = \Phi\left(\delta_o + \delta_d + \beta \ln A_{o,d}^{2010} + X'_{o,d}\gamma\right) \tag{7}$$

where $\ln A_{o,d}^{2010}$ is again instrumented as in equation (4). We extract an estimate for $\hat{z}_{o,d} = \Phi^{-1}\left(\hat{\rho}_{o,d}\right)$, the predicted latent variable which determines non-zero FDI. We then include a simple inverse Mills ratio term $\left(\hat{\mu}_{o,d} = \frac{\varphi\left(\hat{z}_{o,d}\right)}{\Phi\left(\hat{z}_{o,d}\right)}\right)$ within our set $X_{o,d}$ of controls in the intensive margin equation (6). This correction for selection, the extensive margin of FDI, is similar to the procedure

in Helpman et al. (2008) for international trade.[17]

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

Across most specifications, and for most measures of the intensive margin of FDI, we find a positive impact of ancestry on the volume of foreign investment. The elasticity of the intensive margin of FDI between country o and county d with respect to the number of residents in county d who claim ancestry from country o, the coefficient β in equation (6), is around 0.3 across most specification. This elasticity is large. A doubling of the number of residents in county d who report ancestry from country o (from the mean, 318, to 636) increases local employment at subsidiaries of foreign firms by 35%. Alternatively, increasing the number of residents in d with ancestry from o by one standard deviation from the mean increases local employment at foreign subsidiaries by 7% of a standard deviation.

With all measures of the volume of FDI, the estimated impact of ancestry is larger with an instrumental variable specification (column 2) compared to the OLS specification (column 1). This finding is similar to our result for the extensive margin of FDI. For all measures of the volume of FDI, the impact of ancestry is larger when origin and destination fixed effects are included (column 2) than when they are not (column 3). Finally, correcting for selection using a Heckman type procedure always leads to a lower estimated impact of ancestry on the volume of

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

FDI (column 4 compared to column 3).[18] In the case of our measures of inward FDI (panels E, F and G), the impact of ancestry becomes insignificant, with slightly negative point estimates. Except for those three cases, correcting for selection does not change our results substantially.

To conclude, foreign ancestry not only affects the extensive margin of FDI, but it also has a sizable impact on the intensive margin of FDI. More descendants of foreign migrants not only increases the likelihood that local firms engage in FDI, but it also increases the number of firms which do so, and the local employment by foreign owned firms.

5.3 Ancestry and immigration

To test our reduced form model of migrations, and to get a quantitative assessment of the importance of both ancestry and the interaction between push and pull factors for contemporaneous migrations, we estimate various specifications of the following equation,

$$I^t_{o,d} = \delta_o + \delta_d + \beta \ln A^{t-1}_{o,d} + bI^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}} + X'_{o,d}\gamma + \varepsilon_{o,d}$$

where $I^t_{o,d}$ is the total number of immigrants from country o who settle in county d over the decade from $t-1$ to t, and $A^{t-1}_{o,d}$ is the total number of county d's resident who claim ancestry from country o at time $t-1$. $\ln A^{t-1}_{o,d}$ is instrumented using the same first stage equation (4) as earlier. We run regressions for two different decades, $t = 2000, 1990$.

The results are presented in table 7. The effect of ancestry on future immigration is significant and large, as is the impact of the interaction between the push factor $\left(I^t_{o,-r(o)}\right)$ and the pull factor $\left(I^t_{-c(o),d}/I^t_{-c(o)}\right)$. These regressions also give us a quantitative measure of the relative importance of the recursive factor $\left(A^{t-1}_{o,d}\right)$ and the interaction between the push and pull factors $\left(I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}\right)$. For the 1980-90 decade, the recursive factor is about twice as large as the interaction between push and pull factors. For the 1990-2000 decade, the ratio increases from 2/1 to 10/1. In other words, both the current economic conditions (the pull factor) and network of existing descendants (ancestry) matter for contemporaneous migrations.

[18] Note that the number origin and destination fixed effects is too large for a PROBIT estimation of the extensive margin of FDI to be computationally feasible. Moreover, Greene et al. (2002) suggest using monte carlo simulations that PROBIT regressions tend to give biased estimates in the presence of a large number of fixed effects. For both reasons, in column 4, we only implement a Heckman type correction for selection in a specification without origin and destination fixed effects.

5.4 International trade

To conclude this section, we compare our methodology to the existing literature on the effects of common ancestry, which has primarily focused on international trade, starting with the seminal contribution of Rauch and Trindade (2002). We now study the impact of ancestry on international trade.

We do not have access to micro data on international trade by U.S. firms that would allow us to measure trade flows at the level of individual U.S. counties. Instead, we use readily available trade data from the Commodity Flow Survey at the level of U.S. states. Our instrument for the composition of ancestry is the same as in equation (4) for FDI, except that all variables are defined at the state level, and not at the county level.

To compare the magnitude of the effect of ancestry on international trade, we focus our attention on the intensive margin of trade, rather than the extensive margin of trade, as most of the literature has done. Our procedure is similar to the one we use for the intensive margin of FDI. We estimate by two-stage least squares the following gravity equation,

$$\ln Trade_{o,d} = \delta_o + \delta_d + \beta \ln A_{o,d}^{2010} + X'_{o,d}\gamma + \varepsilon_{o,d} \tag{8}$$

where $Trade_{o,d}$ stands either for total exports from U.S. state d to foreign country o, or total imports by U.S. state d from foreign country o. The first stage regression for $\ln A_{o,d}^{2010}$ is as in equation 4. As for the intensive margin of FDI, we allow for a Heckman-type correction for the selection bias due to zero trade.

Table 8 presents the results of the estimation of various specifications of (8). Panel A shows the results for the intensive margin of FDI aggregated to the state level for comparison. Panels B and C show the corresponding results for exports and imports. The main finding emerging from the table is that when ancestry is properly instrumented for, and when a full set of country and U.S. state fixed effects is included, we continue to find a robustly positive and significant effect of ancestry on FDI, but we do not find a robust positive impact of ancestry on the patterns of international trade of U.S. states. This can be seen in columns 2 and 3 of Table 8, where the effect of log ancestry on trade is either insignificant or negative.

This finding is in stark contrast with earlier findings in the literature, started by the seminal contribution of Gould (1994) and Rauch and Trindade (2002), which finds the presence of migrants from a given country facilitates both exports and imports from that country.[19] We do not

[19] Rauch and Trindade (2002) find the larger the ethnic Chinese communities in two countries, the more these countries trade. Combes et al. (2005) find the larger migrations between regions within France are, the more trade between these regions. Using a similar dataset as ours, Gould (1994) finds migrations between foreign countries

find any such positive impact of ethnic origin (ancestry) on international trade. A closer look at the data suggests two important features are essential in reaching this negative conclusion: a formal identification for the causal impact of ancestry using instrumental variables, and the inclusion of destination (U.S. state) fixed effects.

In column 1 of table 8, we estimate equation (8) with simple OLS. We do find a strong and statistically significant correlation of ancestry with both imports and exports. This finding is in line with earlier findings in the literature. However it is likely the same unobserved characteristics which facilitate trade between a U.S. state and a foreign country also facilitate migrations from that country to that state. The exclusion restriction $cov\left(\ln A_{o,d}^{2010}, \varepsilon_{o,d} | X_{o,d}\right) = 0$ is unlikely to be satisfied. The results in column 1, as in most of the existing literature, is evidence of a strong correlation between migration and trade, but not of a causal impact of migrations on trade.

In column 4 of Table 8, we estimate equation (8) with instrumental variables, but without controlling for a destination (U.S. state) fixed effect. This finding is similar to Cohen et al. (2015) on trade with Japan, and Parsons and Vezina (2014) on trade with Vietnam. As those papers consider trade with a single foreign country, they cannot, by construction, control for destination (U.S. state) fixed effects. In panels C and D of columns 1 and 4 of table 8, we replicate those results and estimate equation (8) for a single country, Japan or Vietnam. We get a large and statistically significant positive impact of ancestry on both imports and exports, in line with the finding in Cohen et al. (2015) and Parsons and Vezina (2014). This is true whether we use simple OLS as in column 1, or instrument for ancestry as in column 4.[20] However, it is well understood in the empirical trade literature (see for instance Anderson and van Wincoop (2003) or Head and Mayer (2014)) that not controlling for importer and exporter effects may lead to inconsistent estimates. Systematic differences in size, local prices, local efficiency of firms and workers will systematically affect the patterns of trade. Origin and destination fixed effects systematically control for those differences, and are therefore crucial to estimating a gravity equation.

When both a proper instrument is used and origin and destination fixed effects are included, as in columns 2 and 3 of table 8, we find the impact of ancestry on both imports and exports is

and U.S. states are correlated with international trade at the state level, a finding confirmed by Dunlevy (2006). Felbermayr and Toubal (2012) use data on the composition of the stock of migrants and bilateral trade flows among OECD countries and find a strong correlation between migrations and trade. Two papers use an instrumental variable approach: Cohen et al. (2015) use Japanese internment camps during WWII as an instrument for the current distribution of Japanese migrants in the US, and find a strong impact of the presence of Japanese migrants on trade and FDI to Japan; Parsons and Vezina (2014) use the quasi-random allocation of Vietnamese boat people in the US as an instrument for the current distribution of Vietnamese migrants in the US, and find a strong impact on trade with Vietnam.

[20] We cannot implement a Heckman correction for trade by Japan or Vietnam. Both countries import and export to all 51 states, so that there is no variation in the extensive margin of trade to estimate the predicted probability of positive import/export, ρ_{od}.

either insignificant or even negative. We conclude that the previous finding in the literature of a systematically positive impact of co-ethnic networks on trade should be interpreted with some caution. While it is possible that the patterns in our state-level U.S. data are systematically different from those in the international data or, these findings may also be due to the lack of a proper instrument, or due to the use of data on trade with a single foreign country. When either of these conditions are not met (instrumental variable or origin/destination fixed effect), the positive impact of ancestry on trade disappears in our analysis.

We should also note as an aside that correcting for the systematic selection bias of using only positive trade flows with a Heckman type correction does change the estimated impact of ancestry on trade markedly. This can be seen by comparing the results with origin fixed effects in columns 2 and 3. The estimated impact of ancestry on trade goes toward zero when selection bias due to the extensive margin of trade is corrected for. This is in line with recent developments of the empirics international trade, as in Silva and Tenreyro (2006) or Helpman et al. (2008) for instance.

5.5 Quantitative implications

Having estimated the impact of ancestry on a range of different outcomes, we now further illustrate the quantitative implications of our findings using two counterfactual calculations. In particular, we conduct two thought experiments. First, we estimate how investment relations between U.S. counties and China might have evolved if Chinese migrants had not been effectively barred from entering the U.S. between 1882 and 1965. Second, we report how FDI relationships between Los Angeles and the world might have evolved if Los Angeles had had an influx of population in the 1800s similar to that resulting from the San Francisco gold rush.

5.5.1 The effect of Chinese exclusion

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

We first derive a prediction for the time path of aggregate Chinese migration to the United States in this counterfactual scenario. To this end we use our panel data on immigration from different origins over time to run a regression of the form

$$I_{o,t} = \delta_{t,r} + \delta_o + \xi \cdot D_{China,t} + \nu_{o,t},$$

where $D_{China,t}$ is a dummy which is 1 for all Chinese migration between 1880 and 1970 , and $\delta_{t,r}$ and δ_o are time-region and origin fixed effects, respectively. The coefficient of interest, ξ, estimates the average impact of the Chinese Exclusion Act across the years of its existence and census regions on immigration from China. We define the counterfactual time-path of immigration as $\tilde{I}_{o,t} \equiv I_{o,t} - \xi \cdot D_{o,t}^{China}$. We predict the change in $Log\ Ancestry_{o,d}$ using the estimates from our standard first stage regression as

$$dA_{o,d} \equiv \sum_t \hat{\alpha}_t^f \cdot (\tilde{I}_{o,t} - I_{o,t}) \frac{I_{-c(o),d}^t}{I_{-c(o)}^t},$$

where are the estimated first stage coefficients from our standard specification in Column 5 of Table 2. We calculate the counterfactual incidence of FDI relations with China at the county level as $dY_{o,d} \equiv \hat{\beta} \cdot dA_{o,d}$, where $\hat{\beta}$ is the estimated second stage coefficient from column 3 in Table 3.

Our estimates suggests that in the absence of the Chinese Exclusion Act the number of individuals with Chinese ancestry in 2010 would be 5.1 million higher. This increase would have been highly unequally distributed: some states would have virtually no additional inhabitants with Chinese ancestry, while California, New York and Illinois, would have some 15 to 30 thousand more individuals per county with Chinese ancestry. This change in the incidence of individuals with Chinese ancestry translates into large, but heterogenous changes in the incidence of FDI relationships with China. Figure FIXME depicts the expected change in the fraction of counties with FDI relations with China within each US state. The immigration of Chinese would have translated to substantially stronger FDI relationships between China and several of the North-Eastern US states on the extensive margin. For example, we predict that in New York state some 26% of the counties would have measured FDI relations with China instead of 18%, as observed in our data. Counterfactual predictions pertaining to US or foreign immigration policies can be derived using a similar two step procedure, where the first step estimates the counterfactual inflow of migrants to each US census region during each decade, and the second step predicts the implications of this policy on contemporaneous FDI relationships. Note that similar predictions

could be derived looking at alternative policies.

Secondly, we use our parameter estimates to predict how the patterns of FDI relationships between Los Angeles county and the world would look like had Los Angeles developed significantly earlier. In particular, we derive predictions on the intensity of FDI relationships with the world if the number of immigrants in 1880 had been fivefold the actual number of immigrants to Los Angeles. Table 9 presents the results of this thought experiment. [to be completed]

6 Understanding the Effect of Ancestry

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

6.1 First-generation immigrants

Having shown that the historical stock of ancestry predicts subsequent migrations, a natural question is whether the effect of ancestry on FDI requires a sustained inflow of migrants from the same origin. To address this question Table 5 compares the (causally identified) effect of ancestry to that of first-generation immigrants. Column 1 replicates our standard specification for comparison. Column 2 replaces our measure of ancestry in (1) with the log of one plus the number of immigrants from a given origin alive in 2010 (again measured in thousands, using the same functional form as for our measure of ancestry), while again instrumenting as in (4). As expected, we obtain a positive and statistically significant coefficient (the correlation between the two variables is 0.59). However, when we simultaneously include both endogenous variables in the specification, the coefficient on ancestry remains positive and statistically significant at the 1% level while the coefficient on immigrants alive in 2010 is close to zero and insignificant in the OLS specification in column 3 and turns negative (albeit smaller than the positive coefficient on ancestry) in the IV specification in column 4. Column 5 replicates the same result using the number of immigrants alive in 2000. These results suggest that first-generation immigrants have a significantly smaller effect on FDI than their descendants and that the effect of ancestry on FDI fully develops only over long periods of time, consistent with the temporal pattern of reduced-form coefficients shown in Figure 11.

6.2 Heterogenous Effects

We next explore how the effect of ancestry on FDI varies across origins, destinations, and sectors. We begin by dropping the destination fixed effects from (1) and running 112 separate IV regressions of the FDI dummy on ancestry for each origin country that has at least one FDI link with the U.S. Of the 112 coefficients, 74 are positive and statistically significant at the 5% level. The top panel in Figure 7 plots the coefficients on ancestry from these regressions, over the reciprocal of the standard error, yielding a funnel plot, where the size of the circles is proportional to the share of each origin country in the total ancestry of the U.S. All coefficients to the North-East of the red line are statistically significant at the 5% level. To make the plot easier to read it excludes coefficient estimates larger than one (the full set of results is in Appendix Table 19). The plot shows that the estimates for larger origin countries tend to be more precise and clustered in a tight band. The coefficients for the five largest origin countries yield coefficients ranging from 0.174 (s.e.=0.011) for Mexico to 0.265 (s.e.=0.009) for the UK. The plot also suggests that there may be some heterogeneity in size of the effect.

In panel A of Table 12 we probe for patterns in the heterogeneity of the effect across countries by returning to the simple IV specification of column 2 in Table 3 (relating the FDI dummy to ancestry, origin and destination fixed effects and controls for distance and latitude difference) and add interactions of ancestry with measures of distance and various country characteristics. Columns 1-3 show the interaction of ancestry with the geographic distance and measures of genetic, linguistic, and religious distance between the US and foreign countries as defined by Spolaore and Wacziarg (2015). The results show a consistently positive and statistically significant effect on the interaction between ancestry and geographic distance. Once this interaction is accounted for, the interactions with the three other measures of distance are statistically insignificant, suggesting that the effect of ancestry on FDI increases with geographic distance but not other measures of cultural distance. The remaining columns also show a consistently positive effect on the interaction between ancestry and judicial quality as defined by Nunn (2007), suggesting that ancestry has a relatively larger effect on FDI when the origin country has good institutions.

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

country itself is more ethnically diverse.

The bottom panel in Figure 7 shows coefficients and standard errors from 100 separate IV regressions of the FDI dummy on ancestry for the 100 largest counties within the U.S., where again the size of the circles is proportional to the size of the total population of the destination. The coefficient is positive and statistically significant for 99 out of these 100 regressions. The size of the coefficient varies from 0.093 (s.e.=0.045) for Pierce county WA to 0.457 (s.e.=0.060) for Luisiana county Orleans.

Table 13 probes the heterogeneity of this effect in some more detail by again reverting to our simple IV specification column 2 in Table 3 and interacting our measure of ancestry with the share of the population that are of foreign ancestry (column 1) and a measure of the ethnic diversity within the destination county (column 2). The table shows when there are either fewer residents with foreign ancestry (a lower foreign share in column 1), or a more diverse set of ancestries (a higher ethnic diversity in column 2), the effect of ancestry is relatively stronger. However, when both measures of the composition of ancestry are put together, only the interaction with ethnic diversity remains significant. Ancestry has a stronger impact on FDI flows to and from more ethnically diverse U.S. counties, but the overall share of residents with foreign ancestry matters little. These results suggest that U.S. counties with more diverse populations may act as hubs for FDI, where the presence of descendants from a wide variety of origins enhances the effect of ancestry on FDI in each bilateral link.

Finally, we probe the heterogeneity of the effect of ancestry on FDI across sectors by reconstructing our dataset separately for all 20 2-digit NAICS sectors, considering in each case only FDI links sent and received by U.S. firms in that sector. Appendix Table 18 shows the coefficient estimated in separate IV regressions for each sector, where we find a statistically significant positive effect of ancestry on FDI in 18 of the 20 sectors. However, because the number of U.S. firms in some of these sectors is less than 500 these results may be hard to interpret. Table 14 instead aggregates sectors in to five groups with a more comparable number of firms. It shows that the effect in the manufacturing sector is largest in manufacturing (0.175, s.e.=0.029) and smallest in sectors dealing in natural resources (0.036, s.e.=0.010). The overall pattern of results appears consistent with the view that the effect of ancestry on FDI may be larger in sectors that involve more differentiated inputs (Nunn (2007)), but we do not have detailed data on sufficiently many sectors to test this hypothesis formally.

6.3 The effect of diversity on FDI

6.3.1 Spill-overs

In Table 10 we test for the presence of spill-overs within states and between migrants from proximate origins. In column 1 of panel A we use our simple specification from column 2 in Table 3, but add the total number of descendants of ancestry o at the state level. We are able to identify the effect of this spill-over at the state level by aggregating our instruments from (4) to the state level and including them as instruments in the specification, such that both endogenous variables (ancestry at the state and county level, respectively) are identified. The coefficient on our measure of ancestry at the state level is -0.029 (s.e.=0.012), suggesting that there is a negative and significant spill-over from a larger presence of descendants from the same origin in the state on the effect of ancestry on FDI at the county level. In column 2 we instead include (and instrument for) the number residents in the nearest adjacent county with ancestry from the same origin country, where we find a negative, albeit insignificant effect. Column 3 includes the number of descendants from origins within the same continent, where we find a positive but insignificant effect. However, when we include in column 4 a measure for the number of descendants from the closest neighboring country, we find an negative and highly significant effect.

Overall the evidence thus points to the presence of negative spill-overs. The effect of ancestry on the extensive margin on FDI, and to a lesser extent on the intensive margin of FDI, falls with the population of migrants from the same origin in the state as a whole, or with origin form neighboring countries. For example, a large Polish contingent in a given county has a lower effect on the probability of FDI with Poland if the state overall features a large Polish contingent. Similarly, if the destination county itself also hosts a large number of descendants from nearby countries, such as the Czech Republic, the Polish contingent has a smaller marginal effect on FDI. There appears to be some substitutability between migrants that come from geographically proximate countries.

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

6.3.2 Concavity of the effect

Up to this point we have used a concave function linking the probability of FDI to the number of descendants from a given origin country. A concave functional form appears natural as it implies, for example, a lower marginal effect of the 1001st descendant of Irish origin than for

the first. However, an immediate implication of such concavity is that a more ethnically diverse population, combining many smaller communities from different origins, should generate more FDI than one large community of foreign descent. That is, the functional form linking ancestry to FDI is itself informative about the effect of ethnic diversity FDI. To investigate the functional form linking FDI to ancestry we perform several non-parametric and parametric tests.

We start with a flexible non-parametric estimation of the impact of ancestry on FDI,

$$\mathbf{1}\left[FDI_{o,d}>0\right]=\delta_o+\delta_d+\sum_{q=1}^{Q}\beta_q\mathbf{1}\left[A_{o,d}^{2010}\in quantile_q\right]+X_{o,d}'\gamma+\varepsilon_{o,d},$$

where we divide ancestry numbers, $A_{o,d}^{2010}$, into quantiles, including the same covariates as in our simple specification from column 2 in Table 3. The ratio $\beta_q/\beta_{q'}$ measures the relative impact of ancestry in quantile q versus q'. We experiment with different number of quantiles: $Q=3,4,5$.[21] The results are presented in panel A of Appendix table 17. Across all specifications, with terciles, quartiles and quintiles, we systematically find that a higher number of county d residents with ancestry from country o increases the likelihood of FDI flows between o and d, and that the impact of ancestry is concave.

In panel B of Appendix table 17, we offer a formal test to justify the functional form we use throughout the paper, $\ln\left(1+\frac{1}{1000}A_{o,d}^{2010}\right)$. To that end, we perform a non-linear least squares estimation of

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

again including the same covariates as in our simple specification from column 2 in Table 3. We find a point estimate of$\beta=0.157$ and $\kappa=0.001$. This forms the basis for our choice of functional form applied throughout the paper. The functional form using natural logarithms is convenient as it offers a compact way to model a non-linear the impact of ancestry. For small ancestry ($A_{o,d}\ll 1000$), the function $\ln\left(1+A_{o,t}/1000\right)$ is approximately linear in $A_{o,d}$. For large ancestry ($A_{o,d}\gg 1000$), the function $\ln\left(1+A_{o,t}/1000\right)$ behaves approximatively like $\ln A_{o,d}$. So for a large number of residents with foreign ancestry, the coefficient β in equation (1) is simply the elasticity of the extensive margin of FDI with respect to ancestry; for a small number of residents with foreign ancestry, the coefficient β measures the proportional impact of ancestry on the extensive margin of FDI.

In Appendix table 21, we further explore the robustness of our results to alternative functional forms. In column 1, we simply measure ancestry in levels, and find a positive and significant

[21] The cutoffs for the number an residents in county d with ancestry from country o are: $\{0;145;655;+\infty\}$ for terciles, $\{0;108;282;1144;+\infty\}$ for quartiles, and $\{0;92;187;455;1660;+\infty\}$ for quintiles.

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

We conclude from this exploration of the functional form of the impact of ancestry on FDI that, (i) the effect of ancestry is robust across different functional forms, and (ii) the data suggests a strongly non linear impact of ancestry on FDI: changes in the number of residents with foreign ancestry matter more for FDI when there are few foreign descendants than many.

6.4 Interpretation

Again, we point the caveat that our analysis does not inform us directly on the precise channel through which ancestry facilitates FDI: familiarity, contract enforceability, language,... etc. We have shown the positive effect of ancestry on FDI is stronger for more remote countries, stronger for countries with a higher judicial quality, stronger in ethnically diverse U.S. counties and in countries with a higher degree of ethnolinguistic fractionalization, the effect of ancestry is concave, and we have found evidence of a local negative spillover at the U.S. county and foreign country level.

All those evidence are consistent with a model where common ancestry mitigates informational frictions, rather than facilitates contract enforcement. In the presence of informational frictions, common ancestry acts as a conduit for transmitting information. This information channel is all the more important for remote countries (the positive interaction with distance). Information transmissions tends to follow the shortest path, so that a small increment in the number of residents with common ancestry matters more when few residents have information about a country than many (the concave impact of ancestry and the negative spillover effects). On the other hand, common ancestry does not facilitate FDI precisely when in weak judicial environments (the positive interaction with judicial quality), so it does not seem to be a substitute for contract enforcement.

However, those evidence on the mechanism through which ancestry facilitates FDI are only indirect. We leave a thorough study of this mechanism for future research.

7 Conclusion

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

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

First, we are able to causally identify and quantify the effect of ancestry on FDI in a setting with a high degree of external validity while guarding against a wide range of possible confounding factors, including unobserved origin and destination effects. We find that a doubling of a U.S. county's residents with ancestry form a given foreign country relative to the mean increases the probability that at least one local firm engages in FDI with that country by 4.2 percentage points.

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

Third, the effect of ancestry on FDI increases with the geographic distance to the origin country and the quality of its institutions, and decreases with its ethnolinguistic fractionalization. Once these factors are controlled for other measures of genetic, linguistic, and religious distance do not significantly affect the size of the effect of ancestry on FDI.

Finally, we find a range of results that show a positive effect of ethnic diversity on FDI. The most obvious of these is that there are strong indications concavity in the number of descendants of migrants from a given origin, such that a more ethnically diverse population, combining many

smaller communities from different origins, should generate more FDI than one large community of foreign descent. Further to this, we find negative spill-overs both within states and between migrants from geographically proximate countries, such that a larger community of the same ethnic descent in surrounding counties or a larger community of descent from a neighboring country decreases the effect of ancestry on FDI. In addition, the effect of ancestry on FDI itself significantly increases with the diversity of the community of residents with foreign ancestry. All three findings taken together suggest that ethnic diversity may itself be a quantitatively important driver of FDI.

References

Aleksynska, M. and G. Peri (2014). Isolating the network effect of immigrants on trade. *The World Economy 37*(3), 434–45.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic growth 8*(2), 155–194.

Anderson, J. E. and E. van Wincoop (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*(1), 170–192.

Arkolakis, C., A. Costinot, and A. Rodríguez-Clare (2012). New trade models, same old gains? *American Economic Review 102*(1), 94–130.

Bartik, T. J. (1991, April). *Who Benefits from State and Local Economic Development Policies?* Number wbsle in Books from Upjohn Press. W.E. Upjohn Institute for Employment Research.

Burchardi, K. B. and T. A. Hassan (2013). The economic impact of social ties: Evidence from german reunification. *Quarterly Journal of Economics 128*(3), 1219–1271.

Card, D. (1990). The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relations Review 43 (2)*, 245–257.

Card, D. and J. Di Nardo (2000). Do immigrant inflows lead to native outflows? *American Economic Review 90*(2), 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*(3), 693–708.

Chaney, T. (2014). *Oxford Handbook of the Economics of Networks*, Chapter Networks in International Trade.

Cohen, L., U. Gurun, and C. Malloy (2015). Resident networks and firm value. *The Journal of Finance*, to appear.

Combes, P., M. Lafourcade, and T. Mayer (2005). The Trade-Creating Effects of Business and Social Networks: Evidence from France. *Journal of International Economics 66 (1)*, 1–29.

Cortes, P. (2008). The effect of low-skilled immigration on u.s. prices: Evidence from cpi data. *Journal of Political Economy 116*(3), pp. 381–422.

Docquier, F., G. Peri, and I. Ruyssen (2014). The cross-country determinants of potential and actual migration. *International Migration Review 118*(S1), S37–S99.

Dunlevy, J. A. (2006). The influence of corruption and language on the protrade effect of immigrants: Evidence from the american states. *Review of Economics and Statistics 88*(1), 182–186.

Felbermayr, G. J. and F. Toubal (2012). Revisiting the trade-migration nexus: Evidence from new oecd data. *World Development 40*(5), 928–937.

Friedberg, R. (2001). The Impact of Mass Migration on the Israeli Labor Market. *The Quarterly Journal of Economics 116 (4)*, 1373–1408.

Fuchs-Schuendeln, N. and T. A. Hassan (2015, June). Natural experiments in macroeconomics. Working Paper 21228, National Bureau of Economic Research.

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

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*(2), 302–316.

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology 78*, 1360–1380.

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

Guiso, L., P. Sapienza, and L. Zingales (2009). Cultural Biases in Economic Exchange. *The Quarterly Journal of Economics 124*(3), 1095–1131.

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. *Handbook of International Economics 4*.

Head, K. and J. Ries (1998). Immigration and trade creation: Econometric evidence from canada. *Canadian Journal of Economics 31*(1), 47–62.

Head, K. and J. Ries (2008). Fdi as an outcome of the market for corporate control: Theory and evidence. *Journal of International Economics 74*(1), 2–20.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating trade flows: Trading partners and trading volumes. *Quarterly Journal of Economics 123*, 441–487.

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.

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.

Nunn, N. (2007). Relationship-specificity, incomplete contracts, and the pattern of trade. *Quarterly Journal of Economics 122*(2), 569–600.

Parsons, C. and P.-L. Vezina (2014). Migrant networks and trade: The vietnamese boat people as a natural experiment. *mimeo University of Oxford*.

Portes, R. and H. Rey (2005). The Determinants of Cross-Border Equity Flows. *Journal of International Economics 65*(2), 269–296.

Ramondo, N. (2014). A quantitative approach to multinational production. *Journal of International Economics 93*(1), 108–122.

Rauch, J. and V. Trindade (2002, February). Ethnic Chinese Networks In International Trade. *The Review of Economics and Statistics 84*(1), 116–130.

Razin, A., Y. Rubinstein, and E. Sadka (2003). Which countries export fdi, and how much? Technical report, National Bureau of Economic Research.

Silva, J. S. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641–658.

Spolaore, E. and R. Wacziarg (2015). War and relatedness. *The Review of Economics and Statistics*, to appear.

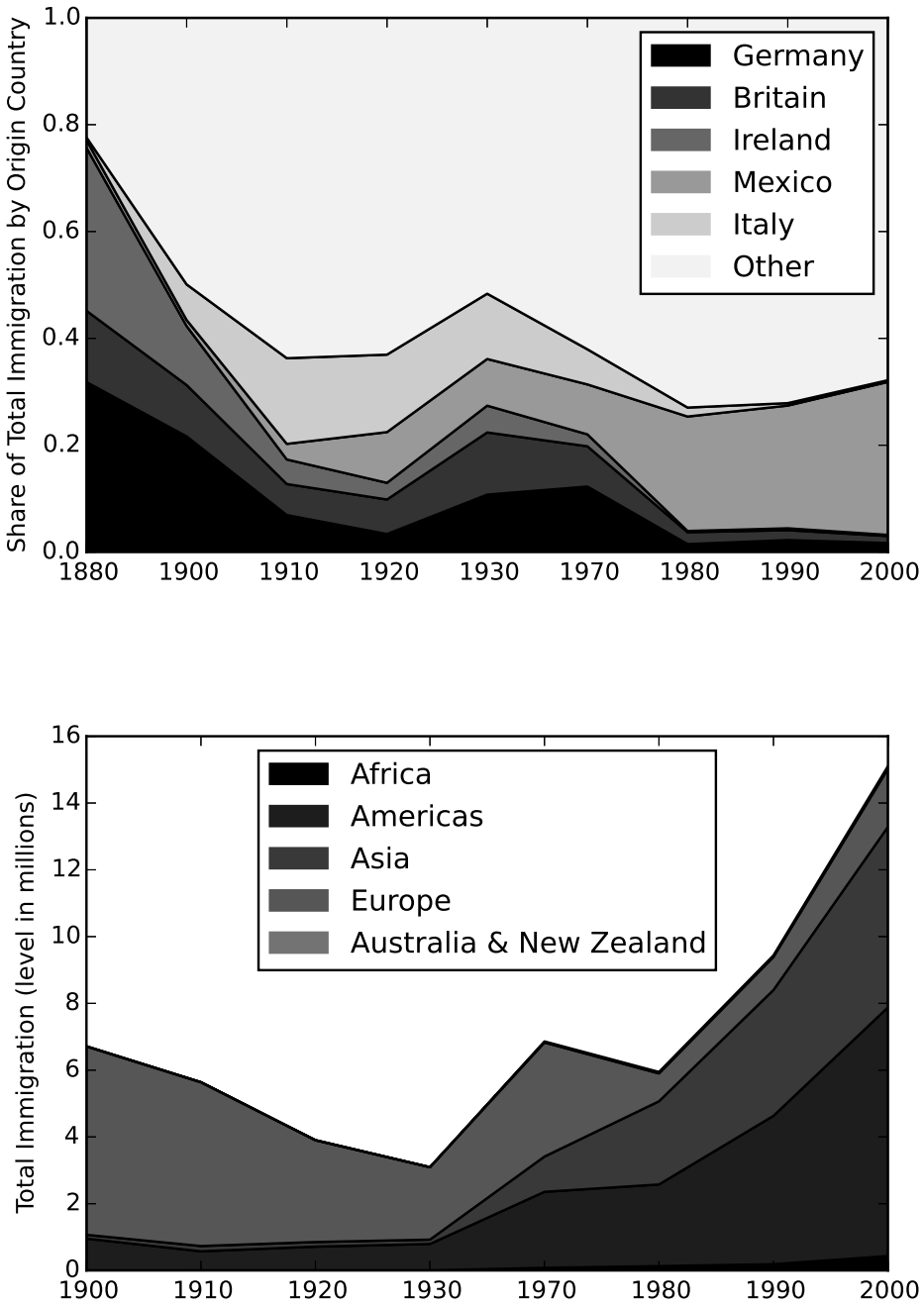


Figure 1: Origins of Immigrants to the U.S.

Notes: The upper panel shows the share of total immigration by origin country, breaking out migrants from the largest senders of migrants to the U.S. overall: Germany, Britain, Ireland, Mexico and Italy. The lower panel shows the the number of migrants (in millions) by continent of origin.

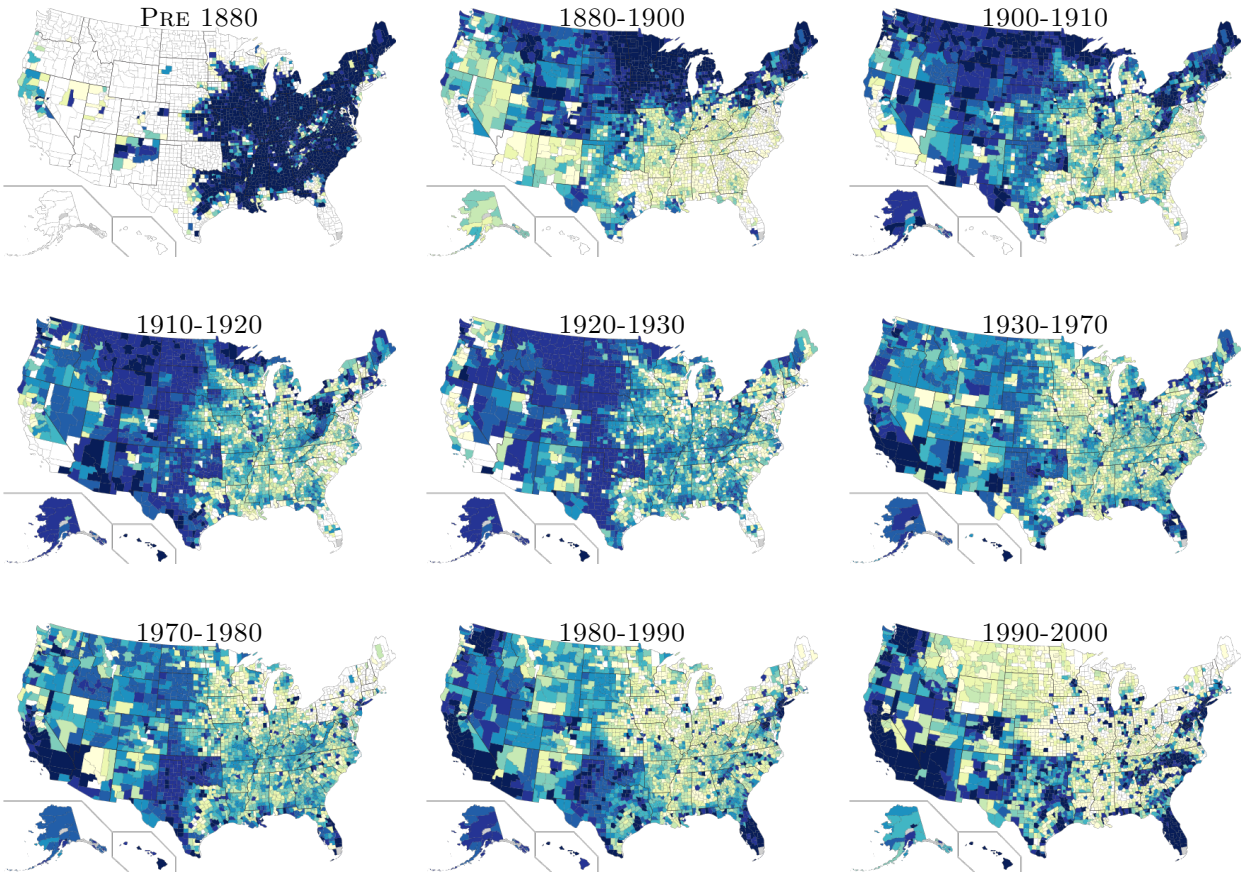


Figure 2: Infolw of Migrants Controlling for Destination & Year FE

Notes: The maps organize the residual immigration flow into the U.S. by census decades. We regress the log number of immigration on destination county c and year t, take the residuals, generate the deciles across all counties and census years, and plot them on the map for each census year. Darker color indicates higher decile.

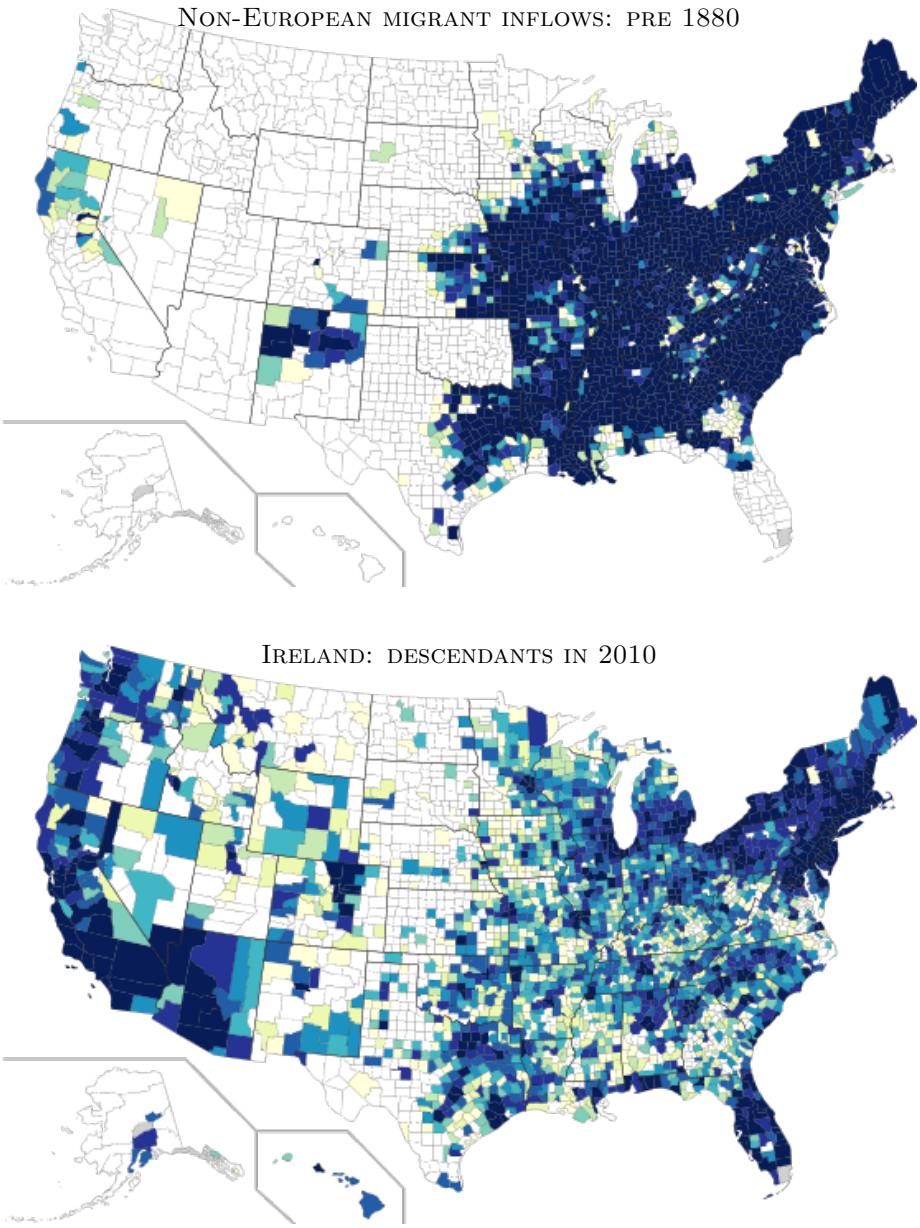


Figure 3: Compare Non-European Immigration Before 1880 and Irish Ancestry in 2010

Notes: The upper part of the map presents the pre-1880, non-European residual immigration flow into the U.S.. The residuals are generated in the same way as in Figure 2. The lower part of the map presents the residual Irish ancestry in 2010. We regress the log number of 2010 ancestry on destination county c and origin country o, take the residuals of Irish, generate the deciles across all counties and plot them on the map. Darker color indicates higher decile.

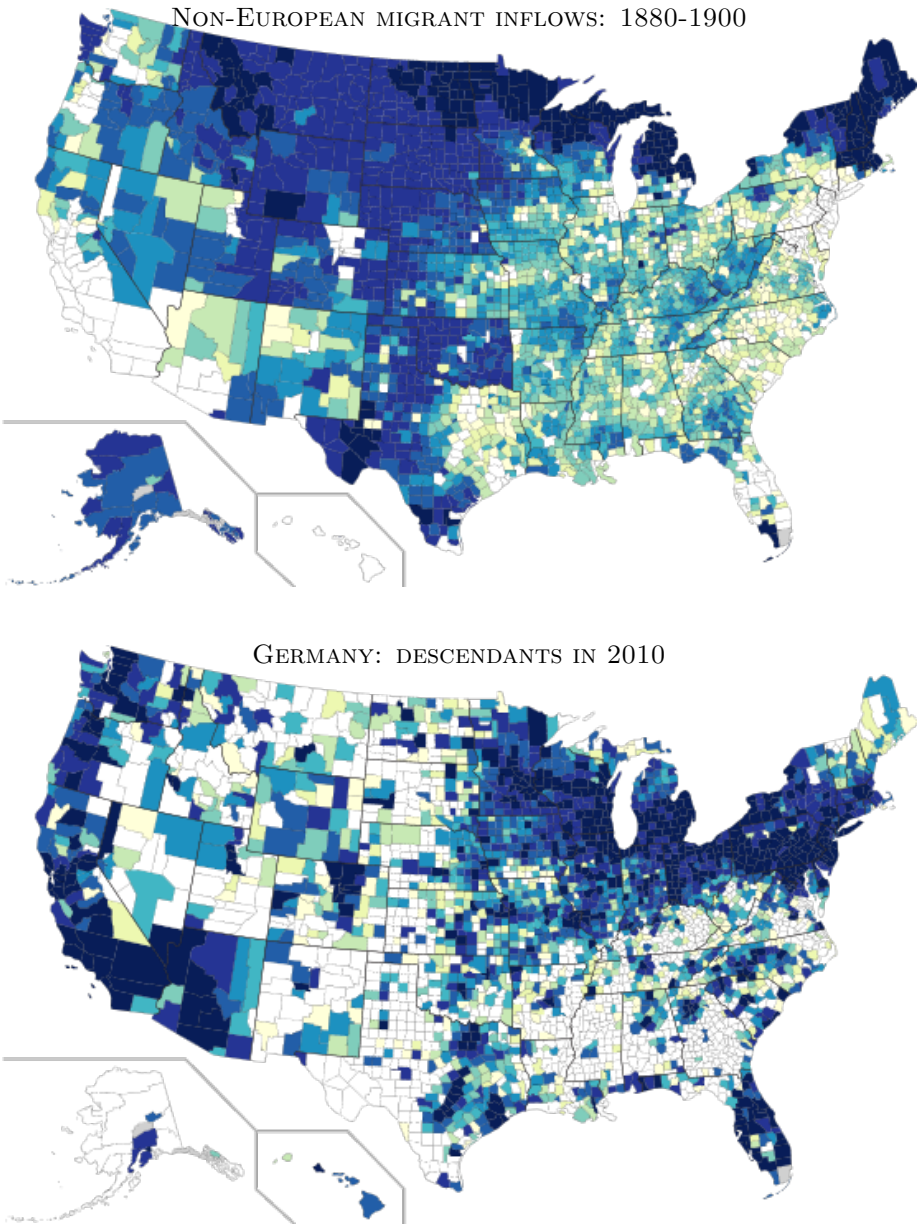


Figure 4: Compare Non-European Immigration in 1880-1990 and German Ancestry in 2010

Notes: The upper part of the map presents the 1990, non-European residual immigration flow into the U.S.. The residuals are generated in the same way as in Figure 2. The lower part of the map presents the residual German ancestry in 2010. We regress the log number of 2010 ancestry on destination county c and origin country o, take the residuals of German, generate the deciles across all counties and plot them on the map. Darker color indicates higher decile.

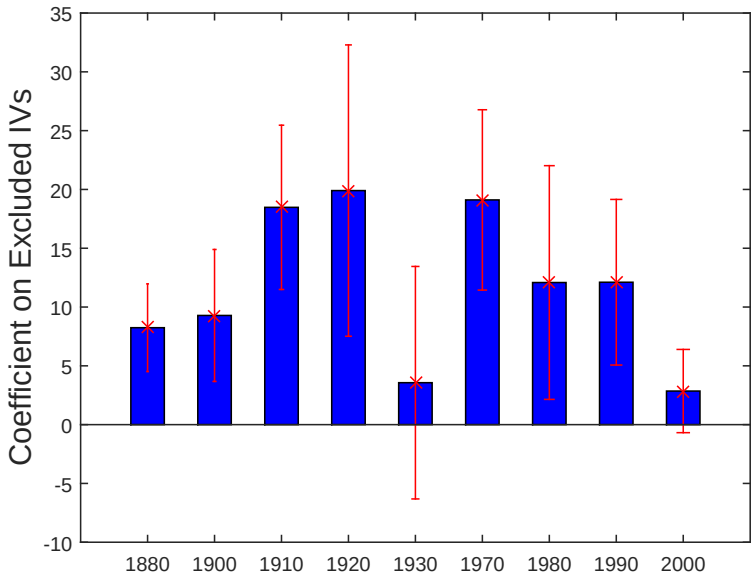


FIGURE 5: THE EFFECT OF IMMIGRATION ON ANCESTRY

Notes: The figure depicts coefficient estimates (and 95% confidence intervals) from Table 2, Column 2 on the excluded instruments $I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$, for $t = 1880, \ldots, 2000$. The dependent variable is the log of 2010 ancestry. The bars in the figure present the respective coefficient estimates on immigration instruments, the red lines give 95% confidence intervals. Robust standard errors are calculated to account for potential heteroscedasticity and clustered at the origin country level.

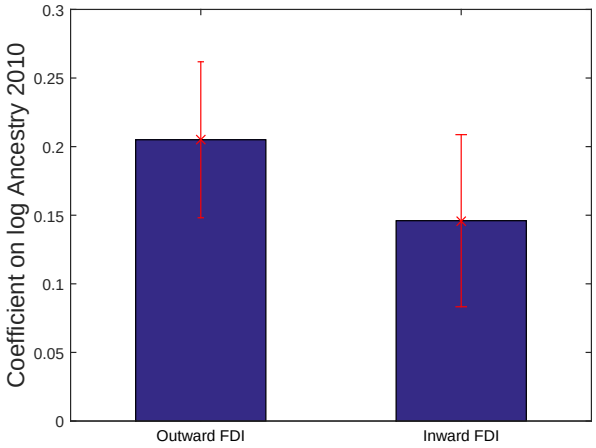


FIGURE 6: COEFFICIENTS ON LOG ANCESTRY (FIRM DATA, EXTENSIVE MARGIN)

Notes: The figure depicts coefficient estimates (and 95% confidence intervals) on the log of 2010 ancestry from extensive margin regressions. The specification is the same as that in Table 3, Column 3. The dependent variable of the left bar is a dummy variable on Outward FDI, which is equal to one if at least one firm in a destination county is the parent of a foreign subsidiary in origin country o. The dependent variable of the right bar is a dummy variable on Inward FDI, which is equal to one if at least one firm in a destination county is a foreign affiliate of a parent firm in origin country. The bars in the figure present the respective coefficient estimates, the red lines give 95% confidence intervals. Standard errors are clustered at the country level to account for potential heteroscedasticity.

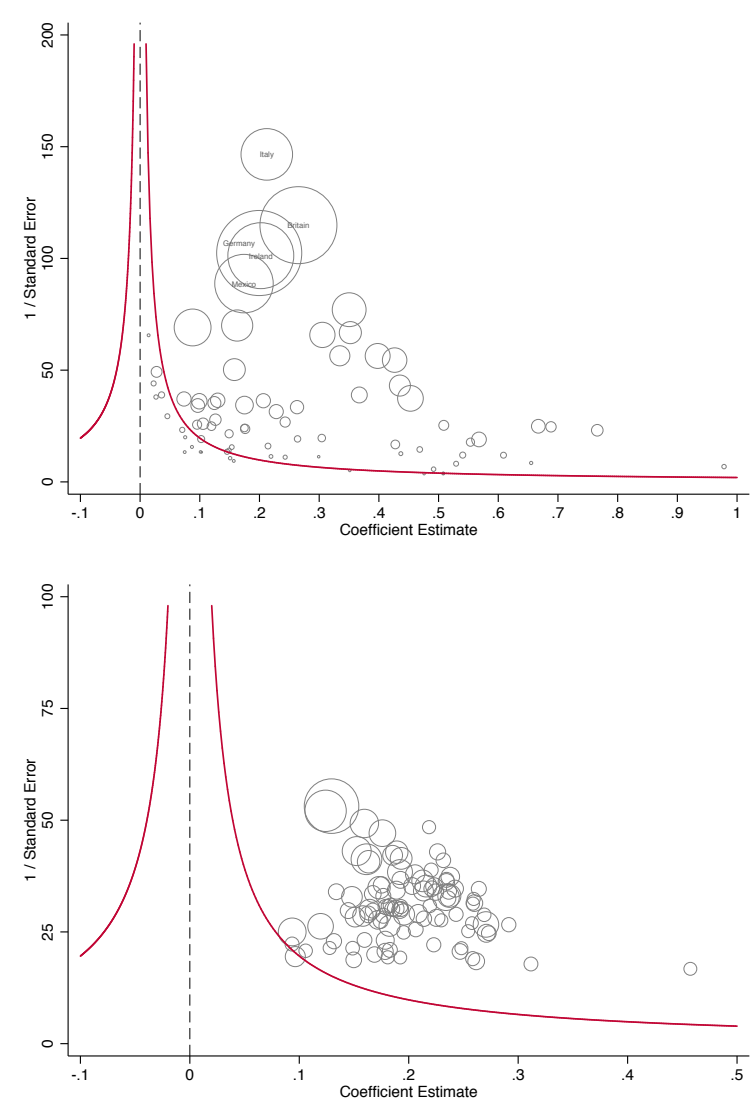


FIGURE 7: FUNNEL PLOTS AT COUNTRY & COUNTY LEVEL

Notes: The figure depicts funnel plots at both country (upper part) and county (lower part) levels. To generate the country-level plot we run IV regression of FDI dummy on the log 2010 ancestry for each origin country. To generate the county-level plot we run IV regression of FDI dummy on the log 2010 ancestry for each destination county. In both parts we use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as excluded instruments, and control for log distance as well as latitude difference. In both parts we plot the estimated coefficients (x axis) against the reciprocal of estimated standard errors on ancestry. The size of the circle is proportional to the size of country ancestry (upper part) and to the size of county population (lower part). The imposed curve is $y = 1.96/x$ for positive x region and $y = -1.96/x$ for negative x region, and circles to the right of the curve indicate statistically significant coefficients.

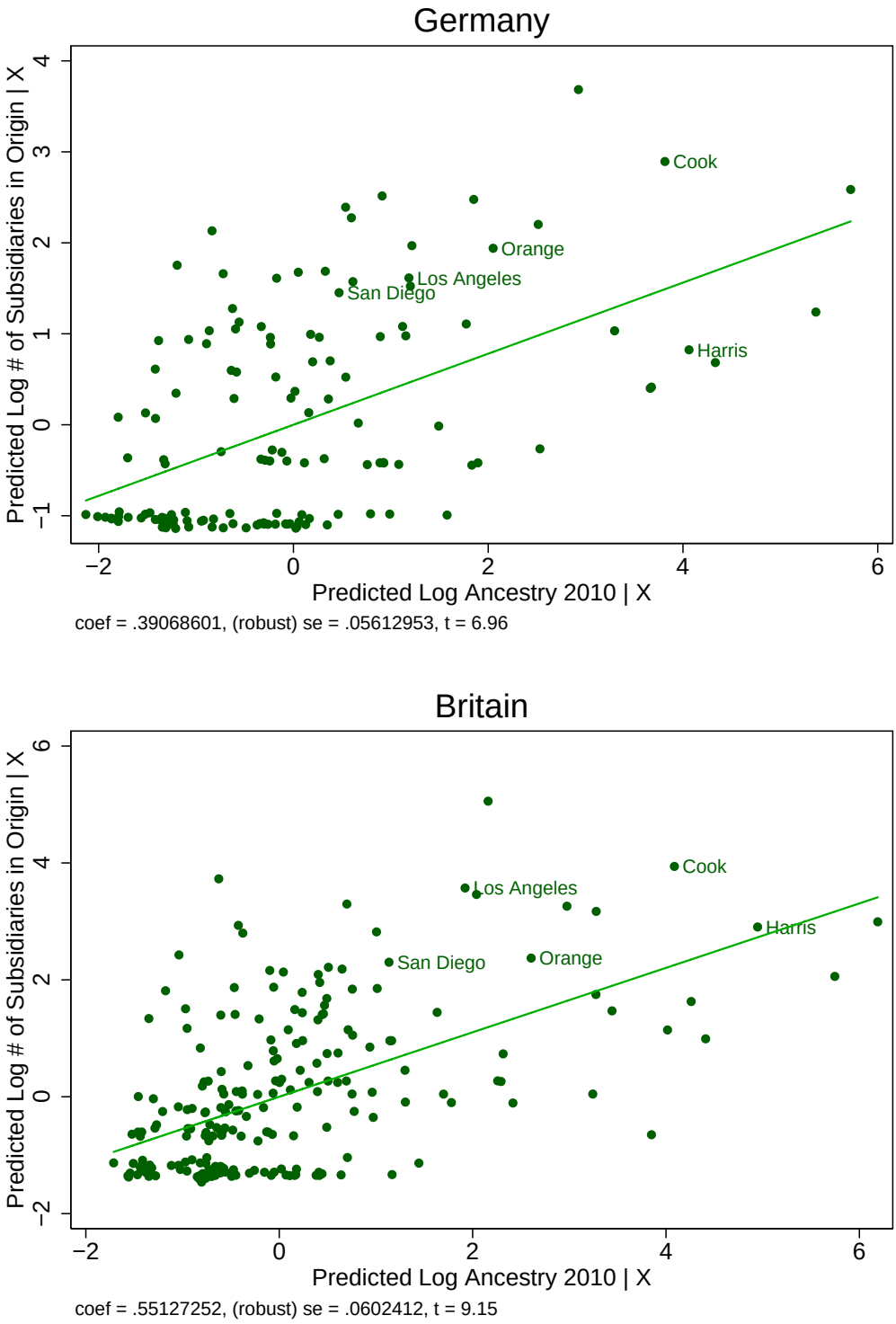
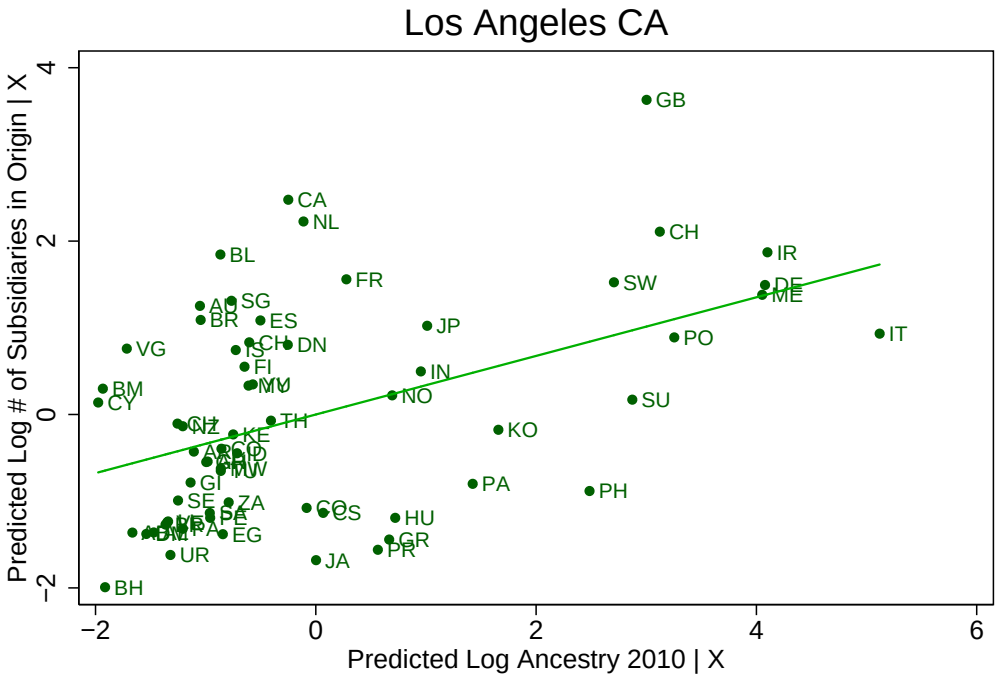


FIGURE 8: SECOND STAGE BY COUNTRY

Notes: The upper part of the figure is a conditional scatterplot of log 2010 German ancestry and log # of subsidiaries in origin country. The lower part is for British ancestry. The corresponding regression uses the same specification as in Table 6, Column 3. The solid line depicts the fitted regression line.



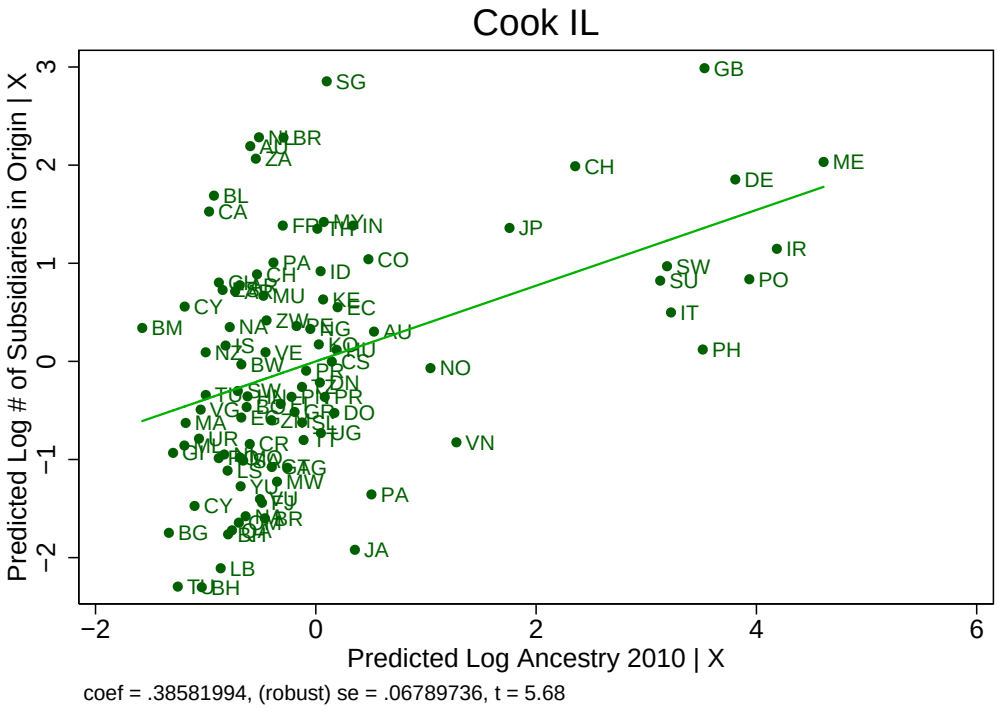


FIGURE 9: SECOND STAGE BY COUNTY

Notes: The upper part of the figure is a conditional scatterplot of log 2010 ancestry in Los Angeles, CA and log # of subsidiaries in origin country. The lower part is for Cook County, IL. The corresponding regression uses the same specification as in Table 6, Column 3. The solid line depicts the fitted regression line.

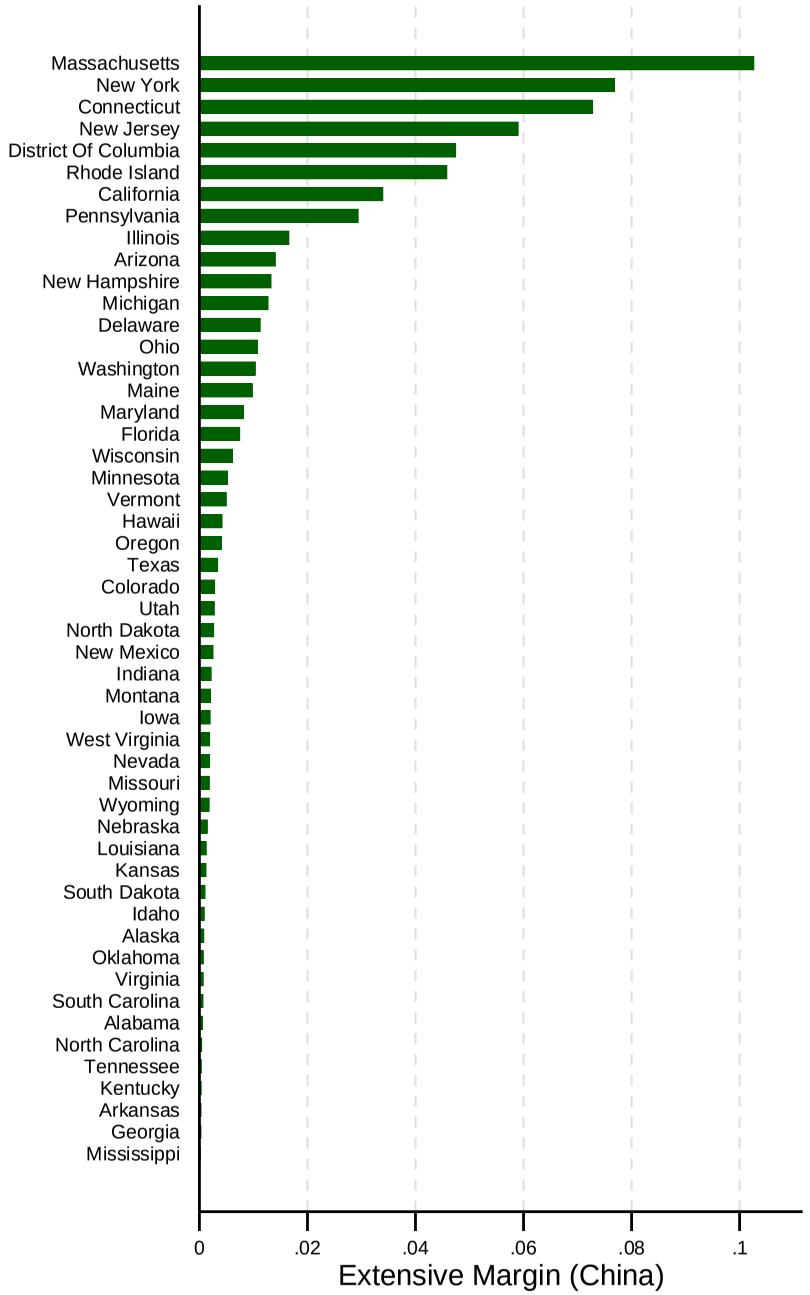


FIGURE 10: CHINA COUNTERFACTUAL EFFECT ON US STATES

Table 1: Summary Statistics

	(1)	(2)	(3)	(4)
	All	*Ancestry > 0*		
		All	*Bottom Quintile*	*Top Quintile*
Panel A: Origin-destination pairs				
FDI Dummy	0.018	0.031	0.003	0.128
	(0.132)	(0.173)	(0.053)	(0.334)
Ancestry 2010 (in thousands)	0.318	0.580	0.000	2.871
	(5.378)	(7.246)	(0.000)	(16.000)
Total Immigrants 2010 (in thousands)	0.029	0.052	0.001	0.226
	(0.352)	(0.474)	(0.013)	(1.018)
Immigrants between 1990-2000 (in thousands)	0.025	0.045	0.001	0.200
	(1.024)	(1.382)	(0.007)	(3.071)
Geographic Distance (km)	9122.393	8397.379	9123.499	7454.962
	(3802.105)	(3763.707)	(4315.761)	(2921.581)
Latitude Difference (degree)	19.440	16.319	19.064	13.512
	(11.312)	(10.902)	(11.397)	(8.632)
# of Total FDI	0.196	0.352	0.028	1.623
	(5.490)	(7.401)	(1.456)	(16.305)
# of Subsidiaries in Origin	0.033	0.060	0.003	0.269
	(1.345)	(1.813)	(0.283)	(3.842)
# of Parents in Destination	0.015	0.027	0.001	0.125
	(0.407)	(0.548)	(0.103)	(1.200)
# of Employees at Subsidiary in Origin (in thousands)	0.039	0.069	0.010	0.319
	(4.941)	(6.661)	(1.298)	(14.749)
# of Subsidiaries in Destination	0.068	0.122	0.011	0.564
	(1.903)	(2.565)	(0.543)	(5.667)
# of Parents in Origin	0.079	0.143	0.012	0.665
	(2.282)	(3.077)	(0.575)	(6.810)
# of Employees at Subsidiary in Destination (in thousands)	1.873	3.398	0.087	16.531
	(86.649)	(116.896)	(5.650)	(260.727)
N	612495	336382	67277	67276
Panel B: Counties				
2010 Share of Population with Foreign Ancestry	0.693	0.694	0.661	0.709
	(0.300)	(0.299)	(0.312)	(0.237)
2010 Diversity of Ancestries	0.785	0.784	0.761	0.831
	(0.073)	(0.073)	(0.068)	(0.076)
2010 Population # (in thousands)	98.389	98.433	6.778	384.363
	(313.143)	(313.238)	(4.873)	(621.566)
2010 Per Capita Income (in thousand dollar)	34.100	34.097	32.708	39.058
	(7.805)	(7.807)	(8.475)	(9.068)
N				
Panel C: Countries				
Genetic Distance	0.103	0.084	0.106	0.064
	(0.053)	(0.041)	(0.050)	(0.036)
N	155	119	18	25
Linguistic Distance	0.950	0.937	0.990	0.922
	(0.110)	(0.121)	(0.010)	(0.114)
N	132	103	8	26
Religious Distance	0.820	0.807	0.923	0.730
	(0.129)	(0.137)	(0.050)	(0.129)
N	131	101	8	25
2010 Per Capita GDP (in thousand dollar)	14.698	18.144	31.903	27.175
	(22.917)	(24.990)	(40.999)	(21.956)
N	165	126	24	27
Judicial Quality	0.503	0.537	0.546	0.681
	(0.208)	(0.214)	(0.224)	(0.197)
N	144	115	15	26
2010 Country Diversity	0.439	0.405	0.433	0.220
	(0.270)	(0.256)	(0.246)	(0.185)
N	163	122	20	27

Notes: The table presents means (and standard deviations). Variables in Panel A refer to our sample of (country,county) pairs used in Table 2, 3, 16, 4, 5, 6, 7, and 10. Variables in Panel B refer to our sample of counties used in Table 13. Variables in Panel C refer to our sample of counties used in Table 12. Column 1 shows data for all observations. Column 2 to 4 show all, bottom quintile and top quintile of observations with positive ancestry. In Panel A, FDI dummy is a dummy varaible which is one if the destination county has either subsidiaries or shareholders in the origin country. In Panel B, population is missing for for 3 counties and per capita income is missing for 8 counties. The ancestry diversity variable is computed as one minus the Herfindahl index of ethnolinguistic group shares in each county. The detail of variables in Panel C will be given in the Data Appendix.

Table 2: First-stage: The effect of historical immigration on ancestry in U.S. counties

	Log Ancestry 2010								*Ancestry 2010*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$	8.3769***	8.2434***	13.5452***	10.3583***	8.2758***	8.2831***	8.7714***		3.0720***
	(1.9147)	(1.9012)	(1.5401)	(1.2960)	(1.1557)	(1.1560)	(1.1126)		(0.3707)
$I^{1900}_{o,-r(d)} \frac{I^{1900}_{-c(o),d}}{I^{1900}_{-c(o)}}$	9.5747***	9.2844***	16.4778***	11.1915***	8.1938***	8.2249***	10.1351***	15.7283***	2.7344*
	(2.8966)	(2.8621)	(3.7077)	(3.7126)	(2.9883)	(2.9910)	(2.8609)	(2.7202)	(1.6462)
$I^{1910}_{o,-r(d)} \frac{I^{1910}_{-c(o),d}}{I^{1910}_{-c(o)}}$	18.3979***	18.4780***	20.4815***	15.2509***	12.1718***	12.1823***	11.4679***	10.6057**	3.8821**
	(3.5849)	(3.5632)	(4.0810)	(4.2301)	(4.3953)	(4.3932)	(4.1029)	(4.2427)	(1.8047)
$I^{1920}_{o,-r(d)} \frac{I^{1920}_{-c(o),d}}{I^{1920}_{-c(o)}}$	19.4012***	19.9029***	26.6142***	23.1033***	23.7992***	23.8072***	19.2352***	29.6357***	9.4739***
	(6.3311)	(6.3202)	(5.6521)	(4.9275)	(2.9739)	(2.9755)	(3.7028)	(4.2132)	(2.2694)
$I^{1930}_{o,-r(d)} \frac{I^{1930}_{-c(o),d}}{I^{1930}_{-c(o)}}$	3.6692	3.5666	4.4446	4.5746	5.3048	5.3817	3.9528	8.0515	9.5596***
	(5.0339)	(5.0432)	(4.9139)	(4.7648)	(3.8929)	(3.9233)	(4.3546)	(4.9773)	(2.4900)
$I^{1970}_{o,-r(d)} \frac{I^{1970}_{-c(o),d}}{I^{1970}_{-c(o)}}$	18.9324***	19.1095***	18.4787***	15.7786***	14.8968***	14.8920***	17.1617***	19.2388***	6.3268***
	(3.9147)	(3.9120)	(3.5246)	(3.4461)	(3.0279)	(3.0339)	(2.5283)	(3.2214)	(0.5087)
$I^{1980}_{o,-r(d)} \frac{I^{1980}_{-c(o),d}}{I^{1980}_{-c(o)}}$	12.1133**	12.0844**	15.5002***	15.0312***	14.4053***	14.4038***	15.5992***	24.6644***	10.2752***
	(5.0645)	(5.0698)	(5.6117)	(5.6083)	(4.9642)	(4.9729)	(3.5548)	(6.6520)	(1.9636)
$I^{1990}_{o,-r(d)} \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$	11.9505***	12.1091***	11.0222***	11.0084***	9.3145**	9.3238**	11.4437***	13.2676**	11.2138***
	(3.5785)	(3.5928)	(3.7020)	(3.6307)	(3.7509)	(3.7620)	(3.5655)	(5.7764)	(3.4601)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$	2.8770	2.8557	3.8857**	4.1176***	4.2494***	4.2594***	2.9149	7.5360***	2.3663*
	(1.8002)	(1.8058)	(1.7177)	(1.5404)	(1.4444)	(1.4451)	(2.4581)	(2.2487)	(1.2365)
$I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$							140.2067***		
							(26.5109)		
R^2	0.58	0.59	0.62	0.71	0.75	0.75	0.77	0.76	0.44
F Stat on excluded IVs	12.674	13.074	2769.129	433.387	91.498	91.127	243.998	85.175	1021.031
p-value on F Stat	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Distance	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Latitude Difference	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Components	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	No	No	Yes	Yes	Yes	Yes	Yes
3rd order poly in dist and lat	No	No	No	No	No	Yes	No	No	No

Notes: All specifications control for origin and destination fixed effects. Standard errors are given in parentheses and clustered at the country level. One, two and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively. Except for column (9) which uses the level of the 2010 ancestry stock as the dependent variable, all the other dependent variables are the log of the 2010 ancestry. All coefficients are multiplied by 100.

Table 3: The effect of Ancestry on FDI

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: IV			*FDI 2014 (Dummy)*			
Log Ancestry 2010	0.243***	0.190***	0.197***	0.197***	0.190***	0.201***
	(0.024)	(0.024)	(0.030)	(0.030)	(0.033)	(0.031)
Log Distance	0.009	0.004	0.026		0.024	-0.024
	(0.010)	(0.009)	(0.031)		(0.029)	(0.027)
Latitude Difference	0.007**	0.006***	0.006**		0.006**	0.004
	(0.003)	(0.002)	(0.003)		(0.003)	(0.004)
N	612495	612495	612495	612495	612495	612300
Panel B: OLS			*FDI 2014 (Dummy)*			
Log Ancestry 2010	0.176***	0.176***	0.155***	0.155***	0.155***	0.171***
	(0.016)	(0.016)	(0.018)	(0.018)	(0.018)	(0.019)
R^2	0.2963	0.2963	0.3633	0.3633	0.3633	0.3932
N	612495	612495	612495	612495	612495	612495
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	No
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from instrumental variable (Panel A) and ordinary least squares (Panel B) regressions at the country-county level. The dependent variable in all panels is the dummy of FDI in 2014. The main variable of interest is *Log Ancestry 2010*. For all columns in Panel A, we include $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ as excluded intruments. 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 country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 4: Diff In Diff Analysis

	Soviet Union	China	Viet-nam	Eastern Europe
Years excluded	1918-1990	1949-1980	1975-1996	1945-1989
	FDI 2014 (Dummy)			
Log Ancestry 2010	0.251***	0.588	0.128	0.109**
	(0.056)	(0.432)	(0.104)	(0.055)
R^2	0.1016	0.3038	0.1967	0.1742
N	3141	3141	3141	18846
F Stat on excluded IVs	5.076	0.404	13.267	37.315

Notes: The table presents coefficient estimates from instrumental variable regressions at the county level. Each column uses data from a subset of origin countries: Soviet Union (Column 1), China (Column 2), Vietnam (Column 3) as well as Albania, Bulgaria, Czechoslovakia, Hungary, Poland and Romania (Column 4). The dependent variable in all columns is the dummy of FDI in 2014. The main variable of interest in all columns is the log of 2010 ancestry, *Log Ancestry 2010*. All columns use as instruments the same set of variables as Column 3 of Table 3, but only the immigration terms between *Closure start* and *Closure end* are excluded instruments; the remaining variables are included as controls. All specifications control for log distance, latitude difference and origin fixed effects. The coefficient estimates on these are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 5: Alternative Ancestry v.s. Immigration (Standard Specification)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	FDI 2014 (Dummy)						
	IV	IV	OLS	IV	IV	IV	IV
Log Ancestry 2010	0.195***		0.146***	0.277***		0.184***	
	(0.011)		(0.019)	(0.031)		(0.015)	
Log Foreign-born 2010		0.257***	0.033	-0.135**			
		(0.023)	(0.044)	(0.056)			
Log Foreign-born 1970					0.523***	0.052	
					(0.045)	(0.053)	
Log Ancestry 2000							0.285***
							(0.029)
Log Foreign-born 2000							-0.149***
							(0.058)
N	612495	612495	612495	612495	612495	612495	612495

Notes: The table presents the coefficient estimates from ordinary least squares (Column 3) and instrumental variable (All the other columns) regressions at the country-county level. The dependent variable in all panels is the dummy of 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 -region and destination-continent fixed effects. The coefficient estimates on these are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively. For Column 4, the Kleibergen-Paap statistic on the excluded instruments is 21.939 with p-value 0.056. For Column 5, the Kleibergen-Paap statistic on the excluded instruments is 21.851 with p-value 0.058.

Table 6: The effect of ancestry on the intensive margin of FDI

	(1)	(2)	(3)	(4)
	OLS	IV	IV	IV
Panel A	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.256***	0.501***	0.350***	0.153***
	(0.051)	(0.047)	(0.022)	(0.030)
R^2	0.7706			
N	10851	10851	10851	10851
Panel B	*Log # of Subsidiaries in destination*			
Log Ancestry 2010	0.282***	0.515***	0.315***	0.249***
	(0.053)	(0.071)	(0.015)	(0.043)
R^2	0.7649			
N	9082	9082	9082	9082
Panel C	*Log # of parents in origin*			
Log Ancestry 2010	0.272***	0.519***	0.325***	0.247***
	(0.056)	(0.086)	(0.016)	(0.041)
R^2	0.7651			
N	9082	9082	9082	9082
Panel D	*Log # of workers employed at subsidiaries in destination*			
Log Ancestry 2010	0.607***	1.419***	0.734***	0.404*
	(0.155)	(0.182)	(0.107)	(0.226)
R^2	0.6931			
N	9082	9082	9082	9082
Panel E	*Log # of Subsidiaries in origin*			
Log Ancestry 2010	0.104**	0.402***	0.244***	-0.051
	(0.042)	(0.045)	(0.024)	(0.044)
R^2	0.7375			
N	4065	4065	4065	4065
Panel F	*Log # of parents in destination*			
Log Ancestry 2010	0.126***	0.484***	0.267***	-0.045
	(0.039)	(0.033)	(0.019)	(0.031)
R^2	0.7614			
N	4065	4065	4065	4065
Panel G	*Log # of workers employed at subsidiaries in origin*			
Log Ancestry 2010	0.219	0.839***	0.450***	-0.179
	(0.161)	(0.081)	(0.058)	(0.175)
R^2	0.6661			
N	4065	4065	4065	4065
Origin × Census Region FE	Yes	Yes	No	No
Destination × Continent FE	Yes	Yes	No	No
Heckman Correction	No	No	No	Yes

Notes: The table presents the coefficient estimates from ordinary least squares (Column 1) and instrumental variable (Column 2 through 4) regressions at the country-county or country-state level. 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 are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 7: The Effect of Ancestry on Immigration

	(1)		(2)	
	Immigration 1990-2000		*Immigration 1980-1990*	
Log Ancestry 1990	9.4796*	9.1846**		
	(5.3527)	(4.5328)		
Log Ancestry 1980			5.3511*	5.4062*
			(3.1992)	(2.9951)
$I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$		0.7409***		
		(0.1264)		
$I^{1990}_{o,-r(d)} \frac{I^{1990}_{-c(o),d}}{I^{1990}_{-c(o)}}$				2.4135*
				(1.2713)
N	612495	612495	612495	612495
F Stat on excluded IVs	11.771	9.353	14.278	12.977

Notes: The table presents the coefficient estimates from ordinary least squares regressions at the country-county level. The dependent variable is the immigration flow from 1990 to 2000 in clolumn 1 and the immigration flow from 1980 to 1990 in column 2. For all columns, we include $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{1980}_{o,-r(d)} \frac{I^{1980}_{-c(o),d}}{I^{1980}_{-c(o)}}$ as excluded instruments. All specifications control for log distance, latitude difference, origin-region and destination-continent fixed effects. The coefficient estimates on these are not reported in the interest of space. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 8: Intensive Margin Of Trade & Investment Data (State Level)

	(1)	(2)	(3)	(4)
	OLS		IV	
Panel A	*Log Total # of FDI relationships*			
Log Ancestry 2010	0.195***	0.210**	0.086*	0.943***
	(0.036)	(0.086)	(0.047)	(0.061)
R^2	0.8490			
N	2384	2384	2199	2384
Panel B	*Aggregate Export*			
Log Ancestry 2010	0.080**	-0.206***	-0.163	1.054***
	(0.034)	(0.053)	(0.111)	(0.089)
R^2	0.8373			
N	7904	7904	4751	7904
Panel C	*Aggregate Import*			
Log Ancestry 2010	0.292***	-0.319***	-0.088	1.350***
	(0.053)	(0.084)	(0.154)	(0.109)
R^2	0.7746			
N	6210	6210	3831	6210
Panel D	*Export To Vietnam*			
Log Ancestry 2010	1.163***			1.204***
	(0.121)			(0.120)
R^2	0.6745			
N	51			51
Panel E	*Export To Japan*			
Log Ancestry 2010	0.869***			1.076***
	(0.199)			(0.125)
R^2	0.4262			
N	51			51
Destination	Yes	Yes	Yes	No
Destination × Continent FE	No	No	No	No
Heckman Correction	No	No	Yes	No

Notes: The table presents coefficient estimates from ordinary least squares (Column 1) and instrumental variable (Columns 2-5) regressions at the country-state level. The main variable of interest is *Log Ancestry 2010*. The dependent variable in Panel B is the aggregate export from each state to each country ,while the dependent variable in Panel C is the aggregate import from each country to each state. The dependent variable in Panel D is the export from each state to Vietnam, while the dependent variable in Panel E is the export from each state to Japan. The dependent variables in all the other panels have the same meaning as in Table 6. For all columns we use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as excluded instruments. All specifications control for log distance, latitude difference and origin fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 9: The Effect of Los Angeles Counterfactuals

	(1)	(2)	(3)	(4)	(5)
Country Name	Ancestry 2010	Predicted Change in Ancestry	FDI #	Predicted % Change in FDI	Predicted Change in Extensive Margin
Germany	365.45	98.09	241	95.95	0.03
Ireland	265.44	92.85	40	89.03	0.25
UK	441.79	40.27	582	31.81	0.04
Norway	43.72	6.93	55	4.87	0.14
Sweden	56.44	6.00	71	4.20	0.05
France	81.06	5.18	278	3.61	0.01
Canada	29.86	3.88	531	2.70	0.00
Switzerland	11.28	3.70	162	2.57	0.02
Czechoslovakia	19.66	3.21	4	2.23	1.26
Netherlands	41.84	2.47	121	1.71	0.01

Notes: The table presents the changes in ancestry, FDI and extensive-margin probability if we increase the share of pre-1880 Los Angeles immigrants to be 5 times as the true size. Only the 10 countries with most change in ancestry are shown in the interest of space. All regressions are reduced-form. We include $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ as instrumental variables, as well as log distance and latitude difference as controls. We used linear regression for ancestry and FDI, probit regression for extensive-margin probability. Column (1) and (3) show true values for each country. Column (2) and (5) show the absolute change in the corresponding variable, i.e. counterfactual predicted value - true predicted value. Column (4) is the precentage change in the corresponding variable, i.e. (counterfactual predicted value - true predicted value) / true predicted value $*100\%$.

Table 10: Spill-Overs Of The Effect (Standard Specification)

	(1)	(2)	(3)	(4)
Panel A: IV		*FDI 2014 (Dummy)*		
Log Ancestry 2010	0.179***	0.180***	0.179***	0.168***
	(0.034)	(0.028)	(0.028)	(0.028)
Log Ancestry 2010, State Level	-0.029**			
	(0.012)			
Log Ancestry 2010 of Nearest County within State		-0.022		
		(0.023)		
Log Ancestry 2010, Continent Level			0.020	
			(0.015)	
Log Ancestry 2010 of Nearest Origin				-0.070***
				(0.022)
N	612495	612495	612495	612495
Panel B: IV		*Log Total # of FDI relationships*		
Log Ancestry 2010	0.215**	0.091	0.166***	0.125***
	(0.099)	(0.074)	(0.048)	(0.040)
Log Ancestry 2010, State Level	-0.179			
	(0.125)			
Log Ancestry 2010 of Nearest County within State		0.042		
		(0.054)		
Log Ancestry 2010, Continent Level			-0.100*	
			(0.059)	
Log Ancestry 2010 of Nearest Origin				-0.046
				(0.064)
N	10851	10851	10851	10851

Notes: The table presents coefficient estimates from extensive-margin (Panel A) and intensive-margin (Panel B) regressions at the country-county level. The dependent variable for Panel A is the dummy of FDI in 2014. The dependent variable for Panel B is the log of total FDI in 2014. For all columns we use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as intrumental variables. All specifications control for log distance, latitude difference, origin-region and destination fixed effects. Standard errors are given in parentheses. Robust standard errors are calculated to account for potential heteroscedasticity. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 11: Extensive Margin Effect On Important Countries/Counties

	FDI 2014 (Dummy)
Panel A: Top 5 Ancestries	*Log Ancestry 2010*
Germany	0.200***
	(0.010)
Britain	0.265***
	(0.009)
Ireland	0.202***
	(0.010)
Mexico	0.174***
	(0.011)
Italy	0.212***
	(0.007)
Panel B: Largest 5 Counties	*Log Ancestry 2010*
Los Angeles,California	0.129***
	(0.019)
Cook,Illinois	0.124***
	(0.019)
Harris,Texas	0.161***
	(0.024)
San Diego,California	0.152***
	(0.023)
Orange,California	0.159***
	(0.020)

Notes: The table presents coefficient estimates from instrumental variable regressions at the country-county level. The dependent variable in both panels is the dummy of FDI in 2014. Panel A presents the coefficient on *Log Ancestry 2010* for each of the top 5 ancestries. Panel B presents the coefficient on *Log Ancestry 2010* for each of the largest 5 counties. We use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as intrumental variables. All specifications control for log distance and latitude difference. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 12: Heterogeneous Effects Across Countries

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A			*FDI 2014 (Dummy)*			
Log Ancestry × Geographic Distance	0.118***	0.117***	0.134**			0.199**
	(0.036)	(0.031)	(0.063)			(0.079)
Log Ancestry × Genetic Distance	-0.812					
	(1.217)					
Log Ancestry × Linguistic Distance		-0.343				
		(0.515)				
Log Ancestry × Religious Distance			-0.710			
			(0.446)			
Log Ancestry × Judicial Quality				0.147*		0.379*
				(0.089)		(0.202)
Log Ancestry × Fractionalization					-0.227**	0.545
					(0.092)	(0.379)
N	486855	414612	411471	452304	508842	446022
Panel B			*Log Total # of FDI relationships*			
Log Ancestry × Geographic Distance	0.445***	0.607***	0.549**			0.848***
	(0.086)	(0.091)	(0.230)			(0.194)
Log Ancestry × Genetic Distance	-6.133					
	(4.391)					
Log Ancestry × Linguistic Distance		-3.562**				
		(1.402)				
Log Ancestry × Religious Distance			-1.542			
			(1.600)			
Log Ancestry × Judicial Quality				1.335***		2.111***
				(0.237)		(0.464)
Log Ancestry × Fractionalization					-1.639***	2.765***
					(0.352)	(0.931)
N	10034	9407	9283	10150	10231	10149

Notes: The table presents coefficient estimates from instrumental variable regressions at the country-county level. The dependent variable for Panel A is the dummy of FDI in 2014. The dependent variable for Panel B is the log of total FDI in 2014. We use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as intrumental variables. 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 country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 13: Heterogeneous Effects Across Counties

	(1)	(2)	(3)
Panel A	*FDI 2014 (Dummy)*		
Log Ancestry × Foreign Share	-0.210**		0.067
	(0.085)		(0.143)
Log Ancestry × Ethnic Diversity		1.373***	1.544***
		(0.241)	(0.429)
N	611910	612495	611910
Panel B	*Log Total # of FDI relationships*		
Log Ancestry × Foreign Share	-0.957**		-0.618
	(0.481)		(0.435)
Log Ancestry × Ethnic Diversity		4.895***	4.637***
		(1.115)	(0.874)
N	10851	10851	10851

Notes: The table presents coefficient estimates from instrumental variable regressions at the country-county level. The dependent variable for Panel A is the dummy of FDI in 2014. The dependent variable for Panel B is the log of total FDI in 2014. We use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as intrumental variables. 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 country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 14: Heterogeneous Effects Across Sectors

	FDI 2014 (Dummy)		
Panel: Sector Group	*Log Ancestry 2010*	N	FDI Dummy
Manufacturing	0.175***	5549	.009
	(0.029)		
Trade	0.172***	3212	.005
	(0.029)		
Information, Finance, Management, and other Services	0.150***	3628	.006
	(0.027)		
Construction, Real Estate, Accomodation, Recreation	0.134***	1637	.003
	(0.024)		
Health, Education, Utilities, and other Public Services	0.046***	689	.001
	(0.022)		
Natural Resources	0.036***	669	.001
	(0.010)		

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from instrumental variable regressions for each of the 5 sector groups at the country-county level. The dependent variable is the dummy of FDI in 2014. We use $I^{1880}_{o,-r(d)}\frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)}\frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as intrumental variables. The third column presents the number of country-county pairs that have FDI link in the corresponding sector. The fourth column presents the mean of FDI Dummy in the corresponding sector group. All specifications control for log distance, latitude difference, origin-region and destination-continent fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Online Appendix

'Migrants, Ancestors, and Investment'

Konrad B. Burchardi

Thomas Chaney

Tarek A. Hassan

A Data Appendix

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

We download the data from the IPUMS website https://usa.ipums.org/usa-action/samples. For the 2010 data, select the 5-year ACS sample. For all the other years, select the largest available sample (for example, there are 1% and 10% sample for 1880 data, and we use the 10% one). For each sample we obtain observations on 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, language.

In these files respondents were asked to report their birthplaces/ancestries, from which we can 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 $A^t_{o,d}$ since 1980. As a general rule, 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, then transform the data into 1990 country-county level using various transition matrices. Details are given in the following sections.

Transition matrices

We need various transition matrices to adjust non-1990 countries to 1990 ones and non-1990 counties/-county groups to 1990 ones, respectively.

- Bpld-to-country: In each year of the US census respondents were asked to report their birthplace. All possible answers (across time) are listed here: https://usa.ipums.org/usa-action/variables/BPL#codes_section. Select "codes" and then "detailed codes". There are roughly 550 different answers to the question of your own birthplace across the censuses from 1850-2012. In each census dataset they are saved in the variable "bpld".Here is a step-by-step tutorial to build those matrices:

  1. Create a transition matrix of zeros, with all possible answers to the 1990 birthplace question as rows and all 1990 countries as columns. In general, the 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 no doubt refers to the area that in 1990 is country c, then we simply change the entry in cell (r,c) from 0 to 1 and moves on to next row; if r does refer to the area that in 1990 is in multiple countries, then we search for the

1990 population of each possible country, and assign probabilities in proportion with the population data. Usually the population information can be found on wikipedia, worldbank database, etc.

- Ancestry-to-country: In the 1980, 1990, 2000 and 2010 census data we have 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. Select "codes" and then "detailed codes". We need to construct transition matrices to map all the answers to the ancestry question into 1990 countries. The procedure is the same as in the bpld-to-country part.

- Group-to-county & puma-to-county: For the years 1970, 1980 and 1990, the census data is at the historic US county group level. A "county group" is an agglomeration of US counties. For the years 2000 and 2010, the census data is at the puma level. A "puma" is also an agglomeration of US counties and is state-dependent. 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 gives a transition matrix from year t county groups to year t counties.

- County-to-county: The list and boundaries of US counties changed over time. In order to merge non-1990 US county-level data with 1990 county-level data, we need to have a transition matrix. There is one transition matrix for each 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 US counties, indexed m, in year 1990. Each cell of the transition matrix takes a value which answers the question: Which fraction of the area of the county c in year t is in 1990 part of county m. Here is a step-by-step tutorial to build those matrices:

  1. Download the year-specific map files.For 1880 us counties, obtain maps from Atlas: http://publications.newberry.org/ahcbp/downloads/united_states.html. Download the 503MB GIS file and find out the 1880 part. For 1900, 1910, 1920 and 1930 counties, obtain maps from IPUMS: https://usa.ipums.org/usa/volii/ICPSR.shtml. For 1970, 1980 and 1990 counties, obtain maps from NHGIS: https://data2.nhgis.org/main.
  2. Project non-1990 maps onto 1990 ones. We use the intersect command in ArcGIS to map year-specific counties onto 1990 counties based on areas. This gives a transition matrix from non-1990 counties to 1990 counties.

Post-1880 flow of immigrants

For each census wave after 1880, we count the number of individuals in each historic US county d with historic foregin country o as birthplace (as identified by birthplace variable "bpld" in the raw data) that immigrated to the US since the last census that contains the immmigration variable (not necessarily 10 years earlier). Then we transform this data

- from non-1990 foreign country ("bpld") level to 1990 foreign country level using bpld-to-country transition matrices.

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

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

- from post-1990 US county to 1990 US county level based on the information from https://www.census.gov/geo/reference/county-changes.html.

Pre-1880 stock of immigrants

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

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

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 ancestry answer ("ancestr1") level to 1990 foreign country level using ancestry-to-country transition matrices.
- from US county group/puma level to US county level using group/puma-to-county transition matrices.
- from non-1990 US county level to 1990 US county level using county-to-county transition matrices.
- from post-1990 US county to 1990 US county level based on the information from https://www.census.gov/geo/reference/county-changes.html.

A.2 Details on the construction of FDI data

We purchase the data from the Bureau van Dijk ORBIS. For each U.S. firm, the raw dataset 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 U.S. and foreign firms. Here are the steps to build the data:

Clean up the postcode information

Postcode is the unique identifier for the county location of U.S firm, so we need to make sure that one county corresponds to one postcode:

- If more than one counties have the same postcode, the one with less population (by Google 2012 population data) are dropped.
- Duplicated postcodes for the same county are made unique by random sampling.
- Missing postcode discovered from main dataset are hard coded. There is only one case: 75427 for Dallas.

Build the parent data

There are 38 variables in the parent data: "Mark" "Company name" "BvD ID number" "Ticker symbol" "Country ISO Code" "City" "Postcode" "Type of entity" "NAICS 2007 Core code (4 digits)" "NAICS, text description" "NACE Rev. 2 Core code (4 digits)" "Core NACE Rev. 2, text description" "Operating revenue (Turnover) th USD 2013" "Number of employees 2013" "Total assets th USD 2013" "Current market capitalisation th USD" "Listed/Delisted/Unlisted" "No of recorded shareholders" "No of recorded subsidiaries" "No of recorded branch locations" "Shareholder - Name" "Shareholder - Ticker symbol" "Book value per share USD" "Shareholder - BvD ID number" "Shareholder - City" "Shareholder - Postal code" "Shareholder - Telephone number" "Shareholder - Type" "Shareholder - NACE Rev. 2, Core code" "Shareholder - NACE Rev. 2, text description" "Shareholder - NAICS 2007, Core code" "Shareholder - NAICS 2007, text description" "Shareholder - Country ISO code" "Shareholder - Direct %" "Shareholder - Total %" "Shareholder - Operating revenue (Turnover) mil USD" "Shareholder - Number of employees" "Shareholder - Total assets mil USD". Here "shareholder" is equivalent to "parent" in our context, so we can use them interchangeably. The data building steps are as follows:

1. Data cleaning-up:
    - Blank rows are filled in within the same company with more than one parents.
    - Numerate string variables: "Listed", "Unlisted" and "Delisted" got a numerical counterparty listed with 1, 0, -1, respectively.
    - Numerical variables in string are adjusted to match Stata format, e.g., n.a., - into ., canceling string , inside numbers.
    - Assign number to Shareholder Direct and Shareholder Total:
        - Mysterious values like "MO", "WO" are recovered from a paper using the same dataset for a simlar topic: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2407845. It says "When the stake of a shareholder is described by the following initials, we replace it with the appropriate number as follows: MO, majority owned, is replaced by "75%"; JO, jointly owned, is replaced by "50%"; NG, negligence, is replaced by "0%"; CQP1,— -, is replaced by "50.01%"; BR, branch, is replaced by "5.01%"; and if the holding of a shareholder is wholly owned (WO), then we delete the firm from the sample, as this firm should not be considered a publicly traded company".
        - Ambiguous numbers are made clear: for values with a ">" before, e.g. $> 25.00 > 30.00 > 50.00 > 75.00$, are replaced by the original number plus 10; for values with a "<" before, e.g. $< 34.00 < 50.00 < 75.00$, are replaced by the original number minus 10; for values with a "±" before, e.g. ± 25.00 ± 75.00, are replaced by the original number.
    - After adjustment, numerical variables in string are destringed.
2. Postcode Matching: both companies and their US parents are matched (foreign parents are ignored in this step), with the postcode data just built. Besides the original string variable postcode, new variables postcode5digit and postcodeextension are generated and labeled "Postal code (5 digit)" and "Postal code (extension)". Similarly shareholders have shareholderpostcodeUS5digit and shareholderpostcodeUSextension (Note the spelling postal code in shareholder variables are unified to postcode).
3. Country Code Matching: both companies and their parents are matched. Each firm has 4 country variables: numerical country code, country name, 2- and 3- digit ISO country code. Then we adjust those modern country codes to 1990 ones based on the information on post-1990 country changes.

Build the subsidiary data

There are 53 variables in the subsidiary data: "Mark" "Company name" "BvD ID number" "Ticker symbol" "Country ISO Code" "City" "Postcode" "Type of entity" "NACE Rev. 2 Core code (4 digits)" "Core NACE Rev. 2, text description" "NAICS 2007 Core code (4 digits)" "NAICS, text description" "Operating revenue (Turnover) th USD 2007" "Total assets th USD 2007" "Number of employees 2007" "Current market capitalisation th USD" "Listed/Delisted/Unlisted" "No of recorded branch locations" "No of recorded subsidiaries" "No of recorded shareholders" "Book value per share USD" "Subsidiary - Name" "Subsidiary - BvD ID number" "Subsidiary - Ticker symbol" "Subsidiary - Country ISO code" "Subsidiary - City" "Subsidiary - Postal code" "Subsidiary - Telephone number" "Subsidiary - Type" "Subsidiary - NACE Rev. 2, Core code" "Subsidiary - NACE Rev. 2, text description" "Subsidiary - NAICS 2007, Core code" "Subsidiary - NAICS 2007, text description" "Subsidiary - Operating revenue (Turnover) mil USD" "Subsidiary - Total assets mil USD" "Subsidiary - Number of employees" "Subsidiary - Level" "Subsidiary - Direct %" "Subsidiary - Total%" "Subsidiary - Status" "Branch - Name" "Branch - BvD ID number" "Branch - Country ISO code" "Branch - City" "Branch - Postcode" "Branch - Telephone number" "Branch - NAICS 2007, Core code" "Branch - NACE Rev. 2, Core code" "Branch - Operating revenue (Turnover) mil USD" "Branch - Total assets mil USD" "Branch - Number employees" "Branch - NACE Rev. 2, text description" "Branch - NAICS 2007, text description". Most cleaning processes are similar to that of the parent data. As follows are the extra cleaning for subsidiary data:

1. Subsidiary status: similarly to shareholder status strings are coded in subsidiarystatus. From an online documentary of the dataset, subsidiarystatus is UO+ if all links found in the path have a percentage over 97.99% or are "UO links" indicated by a source; *UO if all links found in the path have a percentage over 50%, but one or more links are between 50.01 and 97.99%; *UO- if all links have a percentage over 25%, but one or more links are between 25.01% and 50%. To prevent confusion in later analysis, theses string are not numerated to average percentage of their range but integers: UO-=1, UO=2, UO+=3 and missing values "-" to be 0.

2. Merge subsidiaries with branches using variables to denote the difference, that is, rename 13 branch variables to subsidiary* to match the 13 out of the 19 subsidiary variables and reshape the data to move branch rows under subsidiaries rows. Thus, each observation is a subsidiary or branch of the firm. Denoting variables are issubsidiary (=1 if subsidiary and =0 if branch), companyBranchcount (count of branches under each firm), companySubsidiarycount (count of subsidiaries under each firm), and companyRowcount (count of both branches and subsidiaries under each firm).

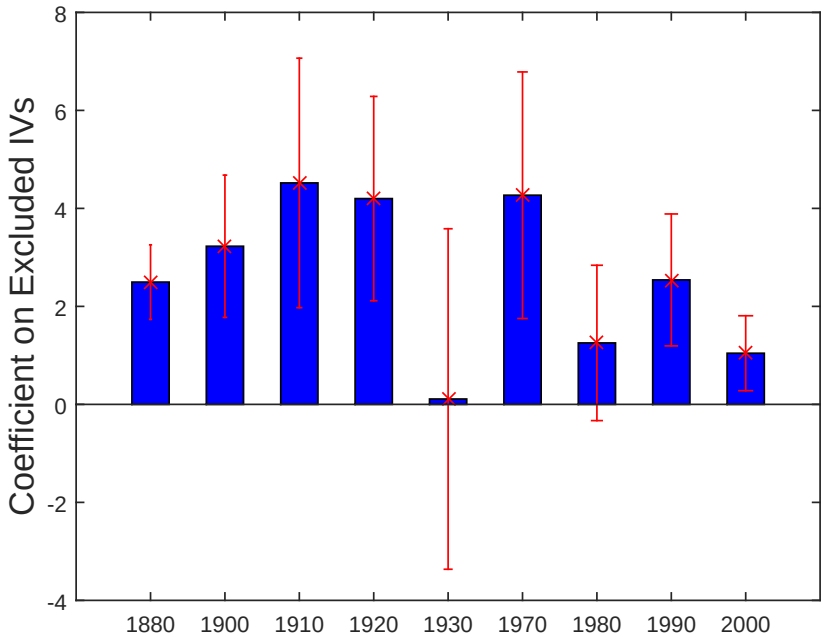


FIGURE 11: THE REDUCED-FORM EFFECT OF IMMIGRATION ON FDI IN 2014

Notes: The figure depicts coefficient estimates (and 95% confidence intervals) from the reduced-form regression corresponding to the specification in column 2 of Table 2. All coefficients are multiplied by 100. The dependent variable is the 2014 FDI dummy. The explanatory variables are the excluded instruments $I^t_{o,-r(d)} \frac{I^t_{-c(o),d}}{I^t_{-c(o)}}$, for $t = 1880, \ldots, 2000$, origin and destination fixed effects, as well as controls for distance and latitude difference. The bars in the figure present the respective coefficient estimates on the interaction terms. The red lines give 95% confidence intervals. Standard errors are clustered at the country level. The R^2 of this regression is 21.8%.

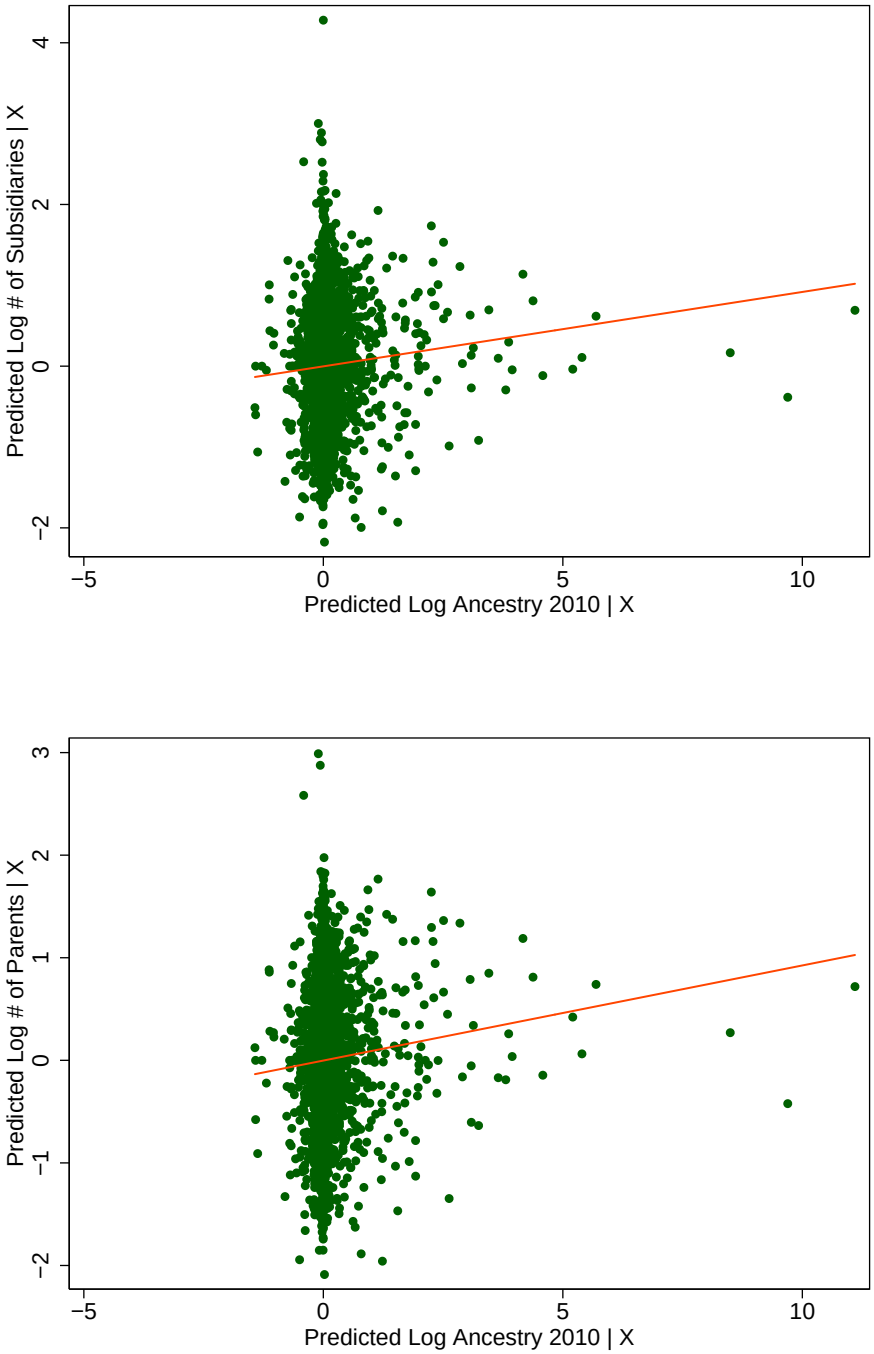


FIGURE 12: SECOND STAGE ADDED VARIABLE PLOT (INTENSIVE MARGIN)

Notes: The figure is a conditional scatterplot of log 2010 ancestry against log (FDI volume). The upper part depicts the log of the total number of domestic subsidiaries of foreign firms plus foreign subsidiaries of domestic firms. The lower part depicts the log of the total number of domestic parents of foreign firms plus foreign parents of domestic firms. The corresponding regression uses the same specification as in Table 6, Column 3. The solid line depicts the fitted regression line. The slope of the upper part line is 0.092 with standard error 0.017. The slope of the lower part line is 0.093 with standard error 0.016.

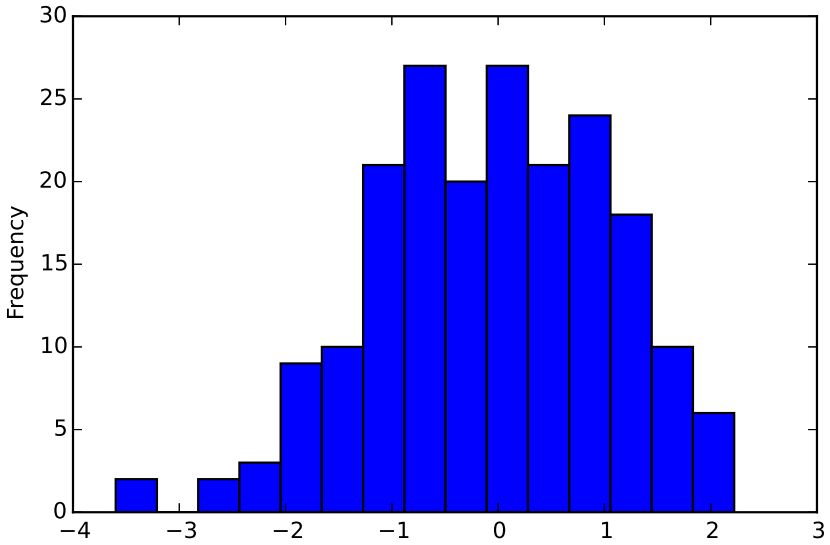


FIGURE 13: THE T-STAT HISTOGRAM FOR STANDARD ERRORS CLUSTERED AT COUNTRY

Notes: The figure presents a placebo experiment as an extension to Table 16 Panel B. Each time we assign each country to some random one on a different continent, run the same specification as in Table 16 Panel B Column 3, and report the t statistic on the estimated coefficient of *Log Ancestry 2010*. We repeat the procedure 200 times and generate the histogram. Using ±1.96 as cut-off values, the false positive rate is 2.5% and the false negative rate is 4.5%.

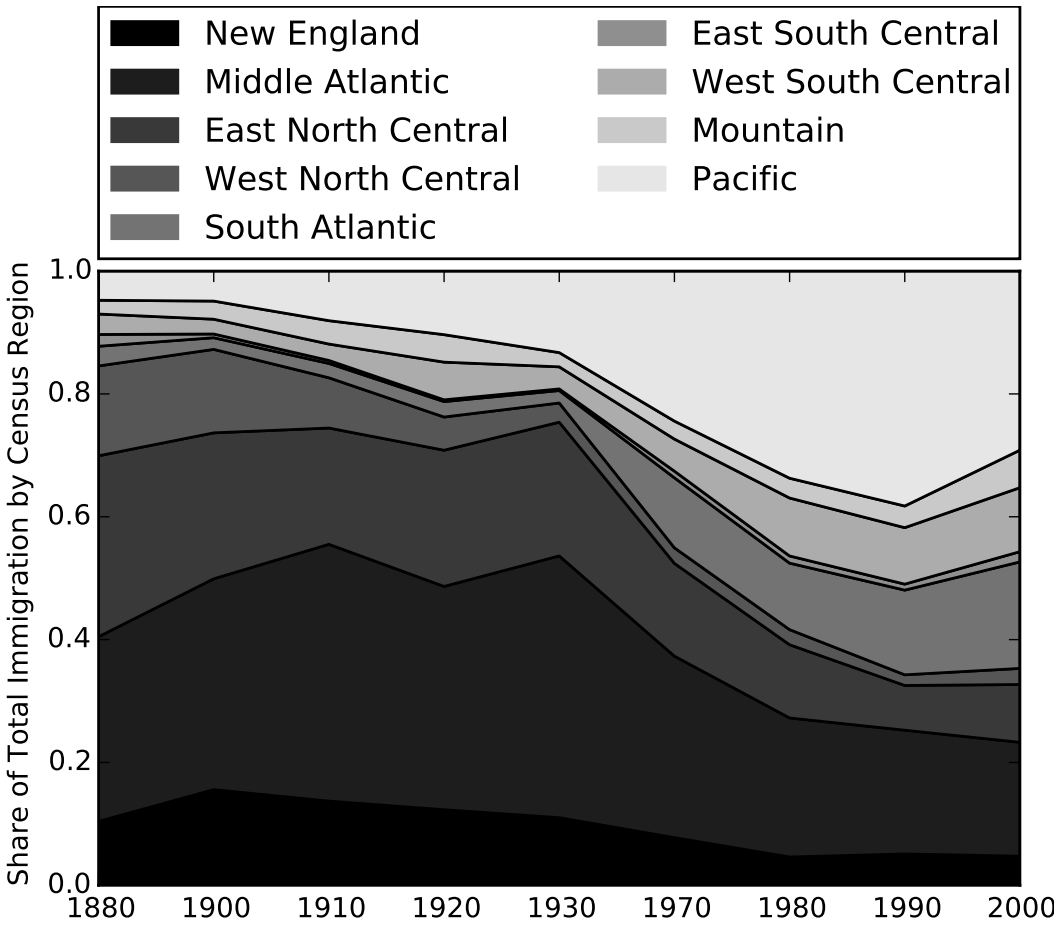


FIGURE 14: IMMIGRATION INTO US CENSUS REGIONS

Notes: Figure 14 shows the time series of total immigration share into nine US census regions as shown in the legend.

TABLE 15: 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

Table 16: Placebo Regression

	(1)	(2)	(3)	(4)	(5)	(6)
	FDI 2014 (Dummy)					
Panel A	*Assign to alphabet neighbor*					
Log Ancestry 2010	-0.012	-0.006	0.014	0.014	0.000	0.017
	(0.021)	(0.015)	(0.031)	(0.031)	(0.032)	(0.035)
N	612495	612495	612495	612495	612495	612300
Panel B	*Assign to alphabet neighbor on a different continent*					
Log Ancestry 2010	-0.031	-0.022	0.012	0.012	-0.005	0.016
	(0.022)	(0.014)	(0.036)	(0.036)	(0.035)	(0.041)
N	612495	612495	612495	612495	612495	612300
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	No
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from two placebo regressions at the country-county level. In Panel A we assign each country to its alphabet neighbor. In Panel B we assign each country to its alphabet neighbor on a different continent. The dependent variable in all panels is the dummy of FDI in 2014. For all columns, we include $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ as excluded instruments. 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 country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 17: Non-Parametric Estimation

	(1)	(2)	(3)
Panel A: OLS	*FDI 2014 (Dummy)*		
Ancestry 2010 Quantile 1	0.003	0.003	0.003
	(0.008)	(0.008)	(0.007)
Ancestry 2010 Quantile 2	0.019	0.012	0.012
	(0.018)	(0.013)	(0.011)
Ancestry 2010 Quantile 3	0.128***	0.046*	0.025
	(0.034)	(0.024)	(0.018)
Ancestry 2010 Quantile 4		0.174***	0.068**
		(0.039)	(0.027)
Ancestry 2010 Quantile 5			0.209***
			(0.042)
N	612495	612495	612495

Panel B	*Nonlinear Least Squares*	
	β	γ
Estimates	0.1574***	0.0011***
	(0.0011)	(0.0000)

Notes: The table presents coefficient estimates from ordinary least squares (Panel A) and nonlinear least squares (Panel B) at the country-county level. The dependent variable in both panels is the dummy of FDI in 2014. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively. All coefficients and standard errors are multiplied by 100. For Panel B, we obtain the optimal β and γ by solving the following nonlinear least square problem:

$$\min \sum_{i=1}^{N} \left(\text{FDI dummy}_i - \alpha - \beta * \log(1 + \gamma * 1000 * \text{ancestry}_i) - \eta' * [\text{distance,latidude}]_i\right)^2.$$

Table 18: Second Stage On FDI Data (Sector Specific Effects)

	FDI Dummy		
	Point Estimate	Standard Deviation	N
	20 Sectors Based on 2007 NAICS code		
Manufacturing	17.520	2.850	5549
Wholesale Trade	15.600	3.053	2513
Professional, Scientific, and Technical Services	13.210	2.712	1925
Information	9.234	2.096	906
Retail Trade	9.012	2.249	846
Transportation and Warehousing	8.941	1.795	620
Administrative and Support and Waste Management and Remediation Services	8.869	2.052	855
Real Estate and Rental and Leasing	8.316	2.260	662
Finance and Insurance	7.399	2.187	1143
Other Services (except Public Administration)	5.664	1.594	301
Management of Companies and Enterprises	5.389	1.526	524
Construction	4.442	1.752	510
Accommodation and Food Services	3.400	1.011	239
Arts, Entertainment, and Recreation	3.034	0.677	131
Mining, Quarrying, and Oil and Gas Extraction	2.928	0.986	528
Health Care and Social Assistance	2.828	1.326	291
Utilities	2.461	1.407	338
Educational Services	0.899	0.551	111
Agriculture, Forestry, Fishing and Hunting	0.701	0.288	149
Public Administration	0.118	0.095	10

Notes: The table presents coefficient estimates on *Log Ancestry 2010* from instrumental variable regressions for each of the 20 sectors at the country-county level. The dependent variable is the dummy of FDI in 2014. We use $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ and principal components as intrumental variables. All specifications control for log distance, latitude difference, origin-region and destination-continent fixed effects. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 19: Second Stage On FDI Data (Country Specific Effects)

	FDI Dummy		
	Point Estimate	Standard Deviation	N
United Arab Emirates	14.971	2.972	60
Kuwait	7.777	2.298	22
Finland	3.919	0.482	180
New Zealand	2.898	0.500	107
Oman	2.815	1.857	6
British Virgin Islands	2.810	0.758	100
Australia	2.236	0.384	369
Malaysia	2.058	0.435	90
South Africa	1.804	0.261	80
Tunisia	1.664	0.481	9
Iceland	1.310	0.268	25
Saudi Arabia	1.257	0.178	29
Puerto Rico	1.118	0.265	26
Bahamas	1.047	0.339	44
Israel	0.978	0.146	137
Switzerland	0.766	0.043	371
BelgiumLuxembourg	0.688	0.041	354
Denmark	0.667	0.040	278
Uruguay	0.655	0.118	21
Thailand	0.609	0.084	68
Japan	0.568	0.053	575
Brazil	0.553	0.056	140
Chile	0.541	0.083	73
Panama	0.529	0.123	44
Austria	0.509	0.040	148
Liberia	0.508	0.264	6
Barbados	0.508	0.267	38
Costa Rica	0.492	0.176	30
Malta	0.476	0.273	11
Turkey	0.469	0.069	48
Norway	0.453	0.027	239
Indonesia	0.437	0.079	29
Senegal	0.436	0.362	2
Canada	0.435	0.023	809
Argentina	0.428	0.060	64
Netherlands	0.426	0.018	398
Sweden	0.398	0.018	323
Republic of Korea	0.367	0.026	155
Spain	0.352	0.015	300
Kenya	0.351	0.197	5
France	0.350	0.013	528
India	0.335	0.018	233
China	0.305	0.015	248
Venezuela (Bolivarian Republic of)	0.304	0.051	32
Belize	0.299	0.089	14
UK	0.265	0.009	664

Egypt	0.264	0.052	23
Colombia	0.263	0.030	45
Samoa	0.243	0.091	5
Peru	0.243	0.037	30
Hungary	0.228	0.032	52
Morocco	0.219	0.088	11
Nigeria	0.214	0.062	18
Italy	0.212	0.007	489
Portugal	0.207	0.028	85
Ireland	0.202	0.010	247
Germany	0.200	0.010	608
Romania	0.176	0.042	23
Czechoslovakia	0.175	0.029	54
Pakistan	0.175	0.041	23
Mexico	0.174	0.011	259
USSR	0.163	0.014	97
Philippines	0.158	0.020	50
Sri Lanka	0.157	0.107	6
Bulgaria	0.154	0.064	11
Ghana	0.151	0.094	6
Lebanon	0.149	0.046	20
Bolivia (Plurinational State of)	0.149	0.071	8
Trinidad and Tobago	0.146	0.074	15
Greece	0.130	0.027	42
Jamaica	0.126	0.036	15
Socialist Yugoslav	0.124	0.028	29
Honduras	0.120	0.040	14
Cameroon	0.118	0.094	2
Guatemala	0.105	0.038	14
Fiji	0.103	0.075	5
Nicaragua	0.102	0.052	7
Algeria	0.101	0.075	3
Viet Nam	0.099	0.028	18
Dominican Republic	0.097	0.029	16
Ecuador	0.096	0.039	15
Poland	0.088	0.014	63
Paraguay	0.087	0.064	4
Democratic People's Republic of Korea	0.081	0.088	1
Jordan	0.076	0.050	7
Sudan	0.075	0.075	1
El Salvador	0.074	0.027	13
Albania	0.071	0.043	3
Bangladesh	0.046	0.034	2
Cambodia	0.036	0.026	3
Haiti	0.028	0.020	2
Ethiopia	0.027	0.026	1
Syrian Arab Republic	0.023	0.023	1
Myanmar	0.014	0.015	1
Afghanistan	0.004	0.004	1

Guyana	0.003	0.003	1
Iraq	0.002	0.002	1
Cuba	-0.000	0.000	1
Libya	-0.020	0.021	1
Mauritania	NA	NA	0
Equatorial Guinea	NA	NA	0
Iran (Islamic Republic of)	NA	NA	0
Somalia	NA	NA	0
Mongolia	NA	NA	0
Greenland	NA	NA	0
Sierra Leone	NA	NA	0
State of Palestine	NA	NA	0
Cape Verde	NA	NA	0
Tonga	NA	NA	0
Lao People's Democratic Republic	NA	NA	0
Nepal	NA	NA	0
Yemen	NA	NA	0

Notes: The table is an extension of Table 11 Panel A, where we only show the results for top 5 ancestries. The fourth column presents the number of counties that have FDI link with the corresponding country.

Table 20: Second Stage On Firm Data (County Level, Con't)

	Extensive Margin					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A	$I_o * I_d/I$					
Log Ancestry 2010	0.209***	0.208***	0.179***	0.179***	0.181***	0.228***
	(0.020)	(0.020)	(0.027)	(0.027)	(0.031)	(0.019)
Log Distance	0.006	0.006	0.022		0.022	-0.005
	(0.010)	(0.010)	(0.028)		(0.029)	(0.020)
Latitude Difference	0.006***	0.006***	0.006**		0.006**	0.010***
	(0.002)	(0.002)	(0.003)		(0.003)	(0.003)
N	612495	612495	612495	612495	612495	612300
Panel B	$I_{o,-d} * I_{-o,d}/I_{-o}$					
Log Ancestry 2010	0.217***	0.207***	0.174***	0.174***	0.179***	0.224***
	(0.021)	(0.020)	(0.028)	(0.028)	(0.031)	(0.019)
Log Distance	0.007	0.006	0.021		0.022	-0.005
	(0.010)	(0.010)	(0.028)		(0.028)	(0.020)
Latitude Difference	0.007***	0.006***	0.006**		0.006**	0.010***
	(0.003)	(0.002)	(0.003)		(0.003)	(0.003)
N	612495	612495	612495	612495	612495	612300
Panel C	$I_{o,-d} * I_{-c(o),d}/I_{-c(o)}$					
Log Ancestry 2010	0.231***	0.225***	0.189***	0.188***	0.186***	0.240***
	(0.023)	(0.022)	(0.027)	(0.027)	(0.031)	(0.021)
Log Distance	0.008	0.007	0.024		0.023	-0.003
	(0.010)	(0.010)	(0.029)		(0.028)	(0.021)
Latitude Difference	0.007**	0.007**	0.006**		0.006**	0.009***
	(0.003)	(0.003)	(0.003)		(0.003)	(0.003)
N	612495	612495	612495	612495	612495	612300
Panel D	$I_{o,-adj(d)} * I_{-c(o),d}/I_{-c(o)}$					
Log Ancestry 2010	0.240***	0.209***	0.200***	0.200***	0.193***	0.248***
	(0.024)	(0.021)	(0.024)	(0.024)	(0.030)	(0.021)
Log Distance	0.007	0.005	0.021		0.019	-0.002
	(0.009)	(0.008)	(0.022)		(0.022)	(0.018)
Latitude Difference	0.007**	0.006***	0.005*		0.005**	0.008***
	(0.003)	(0.002)	(0.003)		(0.003)	(0.003)
N	640764	640764	640764	640764	640764	640560
Principal Components	No	Yes	Yes	Yes	Yes	Yes
Destination × Continent FE	No	No	Yes	Yes	Yes	Yes
Origin × Census Region FE	No	No	Yes	Yes	Yes	No
3rd order poly in dist and lat	No	No	No	Yes	No	No
$I^{2010}_{o,-r(d)} \frac{I^{2010}_{-c(o),d}}{I^{2010}_{-c(o)}}$	No	No	No	No	Yes	No
Origin × State FE	No	No	No	No	No	Yes

Notes: The table presents coefficient estimates from instrumental variable (Panel A) and ordinary least squares (Panel B) regressions at the country-county level. The dependent variable in all panels is the dummy of FDI in 2014. The main variable of interest is *Log Ancestry 2010*. The excluded instruments are indicated by the title in each panel. In Panel D "adj" refers to the adjacent states for the state of county d. 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 country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 21: Alternative Functional Form & Ancestry

	FDI 2014 (Dummy)					
	(1)	(2)	(3)	(4)	(5)	(6)
Ancestry 2010	0.003***					
	(0.001)					
Log Ancestry 2010 (-1 for $-\infty$)		0.197**				
		(0.098)				
(Ancestry 2010)$^{1/3}$			0.203***			
			(0.026)			
Log Ancestry 1980				0.221***		
				(0.037)		
Log Ancestry 1990					0.215***	
					(0.034)	
Log Ancestry 2000						0.202***
						(0.030)
N	612495	612495	612495	612495	612495	612495

Notes: The table presents coefficient estimates from instrumental variable regressions at the country-county level. The dependent variable is the dummy of FDI in 2014. The main variable of interest is the various ancestry variables as 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. The specification is the same as that in Table 3, Column 3. Standard errors are given in parentheses. Standard errors are clustered at the country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 22: Compare Standard Errors

	(1)	(2)	(3)
		FDI Dummy (2014)	
Panel A: Bootstrap	Raw	Country Panel	County Panel
Log 2010 Ancestry	0.1970***	0.1970***	0.1970***
	(0.0104)	(0.0350)	(0.0188)
N	612495	612495	612495
Panel B	Robust	County Cluster	State Cluster
Log 2010 Ancestry	0.1970***	0.1970***	0.1970***
	(0.0101)	(0.0200)	(0.0221)
N	612495	612495	612495
Panel C	Country Cluster	State-Continent Cluster	State-Country Cluster
Log 2010 Ancestry	0.1970***	0.1970***	0.1970***
	(0.0300)	(0.0151)	(0.0133)
N	612495	612495	612495

Notes: The table presents results from experiments of alternative standard errors. Panel A runs a bootstrap experiment with 50 repetitions in each column. Panel A, Column 1 runs bootstrap on the pooled data. Panel A, Column 2 runs bootstrap but keeps the country as panel (so it randomly draws 200 country panels with replacement). Panel A, Column 3 runs bootstrap but keeps the county as panel (so it randomly draws 3000 county panels with replacement). Panel B and C present results from intrumental variable regressions. The specification is the same as that in Table 3, Column 3. Standard errors are given in parentheses. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 23: Composition of Sector Groups

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

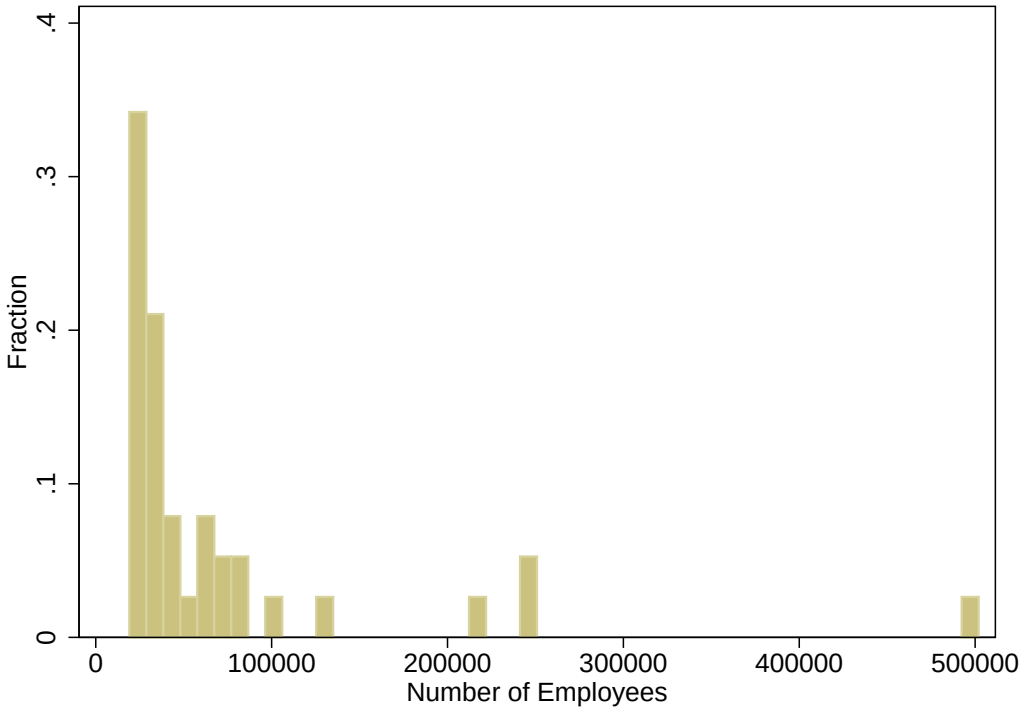
Table 24: Summary Statistics on the Intensive Margin

Origin-destination pairs	(1)	(2)	(3)
Ancestry 2010 (in thousands)	9.796	10.729	15.645
	(36.739)	(39.193)	(55.681)
# of Total FDI	11.057		
	(39.766)		
# of Subsidiaries in Destination		4.572	
		(14.951)	
# of Parents in Origin		5.354	
		(17.975)	
# of Employees at Subsidiary in Destination (in thousands)		11.872	
		(44.522)	
# of Subsidiaries in Origin			5.018
			(15.739)
# of Parents in Destination			2.318
			(4.431)
# of Employees at Subsidiary in Origin (in thousands)			5.812
			(60.380)
N	10851	9082	4065

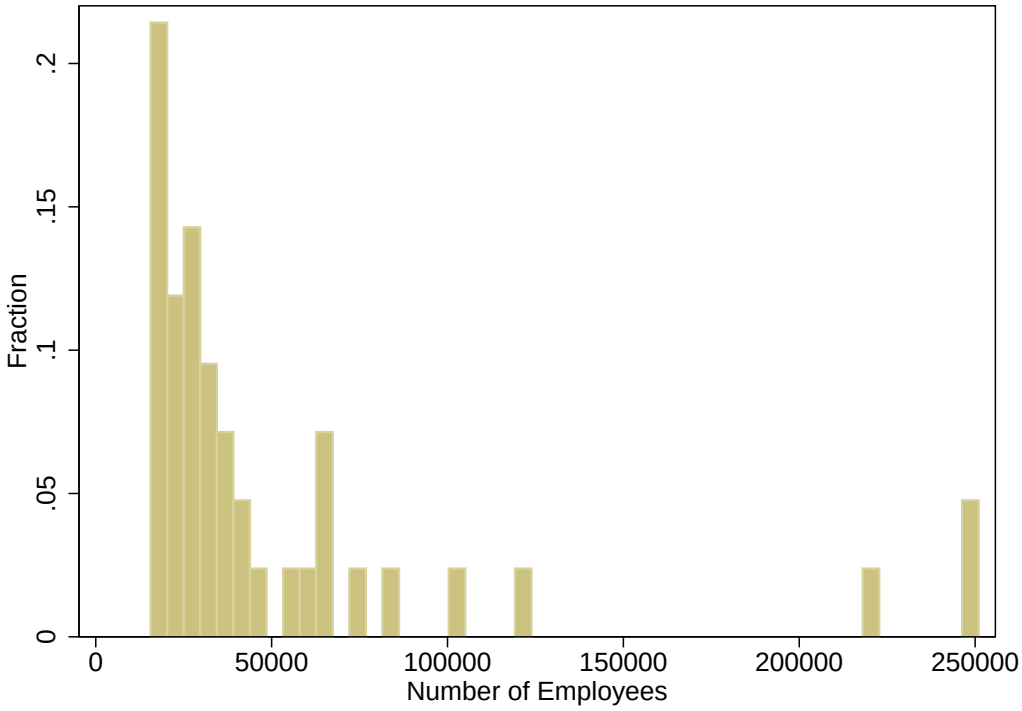
Table 25: Heterogenous Effects on Firm Size

	Below Median Firm Size		Above Median Firm Size	
	(1)	(2)	(3)	(4)
Log Ancestry 2010	0.061***	0.058**	0.124***	0.119***
	(0.022)	(0.027)	(0.017)	(0.021)
Log Distance	0.005**	0.015**	0.008**	0.020
	(0.002)	(0.006)	(0.003)	(0.015)
Latitude Difference	0.002*	0.001*	0.003**	0.003*
	(0.001)	(0.001)	(0.001)	(0.002)
N	612495	612495	612495	612495
Principal Components	No	Yes	No	Yes
Destination × Continent FE	No	Yes	No	Yes
Origin × Census Region FE	No	Yes	No	Yes

Notes: The table presents the result of heterogeneous effects on firm size. Firm size is defined as the number of employees in U.S. firms (subsidiary and parent in destination). The cutoff value between small and big firms is the median employee number, which is 1380 for subsidiaries and 1057 for parents. The main variable of interest is *Log Ancestry 2010*. For all columns we include $I^{1880}_{o,-r(d)} \frac{I^{1880}_{-c(o),d}}{I^{1880}_{-c(o)}}$ through $I^{2000}_{o,-r(d)} \frac{I^{2000}_{-c(o),d}}{I^{2000}_{-c(o)}}$ as excluded intruments. 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 country level. One, two, and three asterisks denote statistical significance at the 10%, 5%, and 1% level, respectively.



Destination Parent



Destination Subsidiary

Figure 15: Firm Size Histogram

Notes: The figure presents the top 10% U.S. firms that have the most number of employees.

Table 26: Separate Final and Intermediate Goods

	(1)	(2)	(3)	(4)	(5)	(6)
		OLS			IV	
Panel A	*FDI Dummy in Final Goods (2014)*					
Log Ancestry 2010	0.094***	0.094***	0.087***	0.170***	0.139***	0.158***
	(0.011)	(0.011)	(0.012)	(0.028)	(0.022)	(0.032)
Log Distance	0.000	0.000	0.007	0.007	0.004	0.024
	(0.006)	(0.006)	(0.014)	(0.007)	(0.006)	(0.023)
Latitude Difference	0.002	0.002	0.005**	0.004*	0.003**	0.004*
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)
N	612495	612495	612495	612495	612495	612495
Panel B	*FDI Dummy in Intermediate Goods (2014)*					
Log Ancestry 2010	0.120***	0.120***	0.111***	0.193***	0.158***	0.172***
	(0.012)	(0.012)	(0.014)	(0.026)	(0.021)	(0.033)
Log Distance	0.001	0.001	0.010	0.007	0.004	0.024
	(0.006)	(0.006)	(0.016)	(0.007)	(0.006)	(0.024)
Latitude Difference	0.004**	0.004**	0.005***	0.006**	0.005***	0.005**
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
N	612495	612495	612495	612495	612495	612495
Principal Components	No	Yes	Yes	No	Yes	Yes
Destination × Continent FE	No	No	Yes	No	No	Yes
Origin × Census Region FE	No	No	Yes	No	No	Yes