The Persistent Effects of Brief Interactions: Evidence from Immigrant Ships.

Diego Battiston*

London School of Economics

2018

Abstract

This paper shows that brief social interactions can have a large impact on economic outcomes when they occur in high-stakes decision contexts. I study this question using a high frequency and detailed geolocalized dataset of matched immigrants-ships from the age of mass migration. Individuals exogenously travelling with (previously unrelated) higher quality shipmates end up being employed in higher quality jobs at destination. Several findings suggest that shipmates provide access and/or information about employment opportunities. Firstly, immigrants' sector of employment and place of residence are affected by those of their shipmates' contacts. Secondly, the baseline effects are stronger for individuals travelling alone and with fewer connections at destination. Thirdly, immigrants are affected more strongly by shipmates who share their language. These findings underline the sizeable effects of even brief social connections, provided that they occur during critical life junctures.

*Department of Economics and CEP, London School of Economics, Houghton Street London WC2A 2AE, United Kingdom. Email: d.e.battiston@lse.ac.uk. I thank to Oriana Bandiera, Jordi Blanes i Vidal, Soledad Giardili, Alan Manning, Guy Michaels, Steve Pischke, Johannes Spinnewijn, Marcos Vera Hernandez, Alwyn Young and participants of the LSE seminars for their useful comments and suggestions. I am greatly indebted to to Matt Nelson and the Minnesota Population Center team for their generous help and advice to access and set the data. Special thanks also to Patricia MacFarlane and the Gjenvick-Gjnvik Archives for sharing genealogical information from their archives.

1 Introduction

It has long been shown that social connections play an important role in shaping economic outcomes. (Jackson, 2011; Topa, 2011; Beaman, 2016; Breza, 2016). Evidence to date has focused on connections established over lengthy periods, or among individuals strongly related in their demographic characteristics. However, many social interactions are circumstantial, brief and with previously unknown individuals. These interactions could also have measurable effects, especially for individuals facing critical moments in their lives. For instance, Bandura (1982) argues that "Some fortuitous encounters touch only lightly, others leave more lasting effects, and still others lead people into new life trajectories.". Chance encounters are also at the heart of theories such as those explaining agglomeration economies (Jacobs, 1969; Glaeser, 1999; Sato & Zenou, 2015). The potential value of brief fortuitous interactions has also been recognized by many organisations, which have implemented reforms to encourage these interactions.[1] Despite their potential, brief interactions have received little empirical attention due to endogeneity and measurement issues.[2]

This paper studies migrants travelling to the US by ship during the first half of the 20th Century. Migrants were placed together in trips lasting no more than a few days. Many faced the need to rapidly learn about potential jobs and final destinations. The dataset follows a large number of individuals who first met while travelling to the US and measures their outcomes many years after arrival. Therefore, this setting provides a unique opportunity to study the value of brief interactions in high-stakes decision contexts.

The dataset links 350,000 male immigrants to their ships of arrival and includes rich geographical information on towns of origin and ports of departure.[3] For every individual, I construct proxies for the quality of his connections upon arrival, exploiting information on the settled immigrants from his town of origin.[4]. More specifically, for each individual,

[1]The following quote by Scott Birnbaum, Vice President of Samsumg Semiconductors is instructive: *"... our data suggest that creating collisions - chance encounters and unplanned interactions between knowledge workers, both inside and outside the organization-, improves performance."* (Waber, et al., 2014).

[2]A body of literature has studied the role of indirect and/or weak (e.g. acquaintances rather than friends) connections. This paper differ from this literature with its focus on the transitory and fortuitous character of the direct interactions between individuals.

[3]Previous studies relying on matched historical data have also used male samples (e.g. Ferrie, 1996; Abramitsky, 2012, 2016). One of the main reasons is that surnames changes were common for females and this makes it difficult to match them across different datasets. In addition to this, female labor force participation is low in this period (Maurer & Potlogea, 2017).

[4]A number of studies have shown the importance of settled immigrants in the assimilation process of new arrived immigrants (Massey et al., 1987, Munshi, 2003; Edin et al., 2003; Lafortune & Tessada, 2012; Beaman, 2015)

I measure two aspects of his potential connections upon arrival: (a) the average earnings (in the US) of previous migrants from his town of origin and (b) the number of previous migrants from his town of origin. Then, I use these variables to proxy the average quality of an individual's previously unknown shipmates.

The empirical strategy relies on the assumption that, conditional on their towns of origin, individuals departing from the same port and in the same week, were plausibly exogenously assigned to ships. This differential assignment creates variation in the characteristics of the (previously unknown) shipmates of an individual. The identification strategy thus compares individuals (exogenously) allocated to travelling in ships that differ in the quality of previously unknown shipmates. A number of balancing tests supports the notion that, conditional on baseline controls, the assignment of passengers to ships was uncorrelated with the characteristics of their previously unknown shipmates. I also provide evidence that the data matching procedure does not induce correlation among shipmates characteristics. In this sense, I perform a number of tests that suggest that, conditional on baseline controls, the probability that a passenger is matched to a census record is uncorrelated with any characteristic of the ship or the individual.

My findings are as follows. Firstly, individuals travelling with higher quality (i.e. better connected) shipmates, end up being employed in higher earnings occupations. This effect is economically significant and persistent in time. For instance, a movement from the lowest to highest quintile in terms of the shipmates' quality is associated with a 4% increase in US labor earnings. This baseline result is robust to: (a) using different measures of occupational earnings, (b) including a large set of additional controls, like, ship-route characteristics, date of arrival and vessel fixed effects, (c) using variation only from individuals boarding at different stops of the same trip and (d) using variation only from repeated trips of the same vessel.

My second set of results suggests that the main mechanism consists of shipmates providing access and/or information about employment opportunities and attractive final destinations. Firstly, I find that the sectors where migrants end up working are affected by the sectors of employment of their shipmates' contacts. Similarly, their final destinations are also affected by the locations of their shipmates' contacts. Secondly, when ships include migrants with different languages, the baseline effects are driven by shipmates speaking the same language. This suggests that some form of verbal communication mediated the effect. Thirdly, the baseline effects are stronger for individuals likely to benefit more from additional connections: (a) individuals travelling by themselves and (b) individuals with poor connections in the US. Overall, my findings provide strong

evidence that migrants benefit from their shipmates' information and/or contacts.[5]

Contribution This paper provides, to the best of my knowledge, the first causal evidence on the economic importance of brief social interactions in high-stakes situations. Equally important is the finding that the effects are largely contingent on individual characteristics. In particular, those travelling alone and with fewer connections at destination are more affected than those with a better network at destination. This suggests the existence of a substitution effect between pre-established interpersonal connections and circumstantial contacts.

Findings from this paper have implications beyond its particular historical setting. First, it is possible that there are many situations where individuals face critical decisions that are irreversible or have long term consequences. Examples include, parental choice of school or students choice of college major. Second, results are consistent with studies showing that labor market entry conditions have persistent effects on job assignment and wages (Oreopoulos, et al., 2006; Oyer, 2006; von Wachter & Bender, 2008). In this paper, I show that short-lasting events that take place just before job search started can affect earnings in the long run. Third, this paper contributes to the economic literature on immigrants assimilation process (Borjas, 1995, 2015, Bleakley & Chin, 2009) by providing evidence that information and conditions upon arrival can determine newcomers future economic success.

Finally, this paper also provides a methodological contribution. It is well known that for large datasets, popular record linkage approaches like Fellegi & Sunter (1969) or Feigenbaum (2016) become unfeasible due computational limitations. I develop a Machine Learning approach to link US immigrant and passenger lists that improves the efficiency of previous methods and can serve as a guide to other researchers matching records across large historical datasets.

Related Literature This paper relates to a number of areas of research. First, a large body of literature has shown the effects of networks and social connections in the context of labor markets (Montgomery, 1991; Marmaros & Sacerdote, 2002; Bayer et al., 2008; Ioannides & Loury, 2004; Bentolilla et al. 2010; Dustmann et al., 2015, Bramoullé et al., 2016; Glitz, 2017).[6] Most of this literature has focused on the importance of job referrals

[5]My dataset is not well suited to disentangling a pure information effect (e.g. shipmates providing information on attractive sectors of employment or final destinations) from a direct access effect (e.g. shipmates providing job referrals or other type of support), and I leave this for future work.

[6]There is also a rich theoretical literature in the area of social networks. Recent reviews can be found in Jackson (2002, 2010), Goyal (2015) and Jackson et.al. (2017).

and job search methods to access better quality jobs.

Related to the role of immigrant networks, a number of articles have measured the importance of connections for newly arrived individuals (Munshi, 2003, 2014; Edin et al., 2003; McKenzie & Rapoport, 2007; Beaman, 2015, Battisti et al., 2017). This paper differs from these studies in that I focus on the role of links created while travelling to destination rather than in the role of pre-existing contacts. This also suggests a link with a growing literature documenting how entry conditions to the labor market can have long-run effects on earnings (Brunner & Kuhn, 2009; Yuji et al., 2010; Oreopoulos et al., 2006). Also, Kramarz & Skans (2014) find that strong social ties (parents) are an important determinant for the first job of young workers and that social ties become more important when information on potential openings are likely to be scarce.

Theoretical models from different fields have assigned an important role to random social interactions. For instance, in the seminal work of Jacobs (1969) random interactions foster innovation and transmission of ideas and in Glaeser (1999), they influence learning of skills.[7] Despite this theoretical work, there are no empirical studies measuring the importance of random encounters in this field. A notable exception is Fitjar & Rodriguez-Pose (2016) who surveyed 542 Norwegian firms engaged in innovation partnerships. They find that 10% of partnerships emerged from random encounters.

A number of previous studies have analyzed the effects of connections established over long periods (e.g. Sacerdote, 2001; Angrist & Lang, 2004). This paper separates from that literature in that the (initial) exposure to social interaction is short, 10 days on average. On the contrary, peer-effects studies typically focus on connections established over long periods.

This paper also relates to the literature on weak ties. Early research, mainly by sociologists (Granovetter, 1973, 1983) found that a significant number of individuals find their jobs through connections such as "friends of friends". This literature emphasizes the role of weak ties in conveying information not prevalent among relatives or close friends. A recent number of studies have analyzed the "strength of weak ties" hypothesis using recent available data (Yakubovich, 2005). Related to immigrant outcomes, Goel & Lang (2016) study the role of weak ties in job search of recent immigrants to Canada and Giulietti et al. (2014) find that the rural-urban decision is largely affected by weak ties. The type of interactions studied in this paper diverge from the concept of weak ties, usually defined as a subset of acquaintances with lower probability to be socially involved with one another.[8]

[7] For a complete review of this literature see Ioannides (2012).

[8] Weak ties are defined in different ways in the literature. For instance, Giulietti et al. (2014), define an immigrant's weak ties as those individuals from his same community who are not his relatives. The

Finally, this paper relates to a body of research that study the process of immigrants' assimilation (Chiswick, 1978; Borjas, 1995, 2015; Bleakley & Chin, 2009). A number of determinants have been explored, including the role of language proficiency, age of arrival, macroeconomic conditions or the performance of settled immigrants. Findings from this paper suggest that the first social connections made by immigrants can affect the later economic success of immigrants.

Plan I describe the historical background and institutional setting in Section 2. I summarize the construction of the matched census-ships dataset in Section 3. The empirical setting and identification strategy is discussed in Section 4. Section 5, presents the main results of the paper and discuss the economic relevance of them. In Section 6, I provide evidence on additional outcomes and heterogeneous effects to establish the social interaction explanation as the preferred interpretation of results. Section 7 concludes.

2 Historical setting

The period 1850-1924 is often referred to as "The Age of Mass Migration". Official statistics indicate that during this period, more than 30 million individuals arrived into the US (Hatton & Williamson, 1998). This was a period of low administrative barriers to immigration that ended after the imposition of the 1924 Immigration Act which sharply reduced immigrant flows (Goldin, 1994).[9]

The vast majority of immigrants arriving after 1892 entered the US through Ellis Island in New York Harbor.[10] During peak years, Ellis Island registered more than 10,000 arrivals per day. Once arrived, immigrants were inspected and authorized to enter the country. The sub-sections below explain the typical stages of the immigration process. This starts when individuals buy their tickets and finishes with the standardized inspection

theoretical model of Sato & Zenou (2015) associate the idea of "random encounters" to weak ties, although they acknowledge the difference with respect to previous studies.

[9]The immigration act of 1892 stated a minimum requirement by banning from entry any person "*unable to take care of himself or herself without becoming a public charge*" (Hutchinson, 1981). In practice this excluded individuals with poor health conditions (including insane) or with criminal records as well as those travelling without enough money to support themselves for few days after arrival. By the end of this period, legislation gradually increased the barriers to immigration (Reisler, 1976; Scruggs, 1988). For instance, the 1917 Literacy Act increased the head tax and introduced a literacy test. The 1921 Emergency Immigration Act introduced a system of quotas mainly directed to reduce immigration from eastern and southern Europe. Another exception was the 1882 Chinese Exclusion Act which banned immigration of Chinese workers. The increase in restrictions was mainly driven by the increase of critical perceptions an attitudes towards immigration (Goldin, 1994).

[10]According to official statistics, more than 75% of total arrivals were through Ellis Island and this percentage increased considerably for European immigrants (Ferenczi-Willcox, 1929).

process at Ellis Island.

Before Departure A typical immigrant would buy his ticket from an agent of the many shipping companies existing at the time.[11] The Passenger Act of 1819 required each vessel arriving from abroad to provide a manifest listing all passengers. Although the information covered by manifests improved over time, after 1904 manifests registered the universe of passengers from any class and nationality (Bandiera, et al., 2016). Given that the cost of any deportation was levied on shipping companies, they faced strong incentives to screen passengers before departing and check that information was accurate. Therefore, individuals were typically required to provide travel documents in advance in order to comply with manifest creation. Additionally, shipping companies carried out their own medical inspection and disinfection before departure.[12] As a result of these requirements, individuals attended the port some days before departing.[13]

The Immigrant Journey Once the medical inspection procedure was completed, passengers were allowed to board the ship for departure. The conditions on the ship were poor for the vast majority, who travelled in steerage class. Rooms usually accommodated large groups and most spaces were shared with other steerage shipmates. Although some individuals traveled with relatives or acquaintances from their home town, a large number of social interactions are likely to have occurred among individuals who had never met before. The duration of the voyage depended on the route and port of departure. By 1910, a trip from Liverpool to New York could take between 6 and 9 days, but departures from Mediterranean ports could take more than two weeks if the route included intermediate stops. Although there was some variation in the duration of the trip, the adoption of the steam engine and other improvements in shipping technology notably reduced the importance of weather conditions (Hopkins, 1910).[14]

Some individuals, specifically those with prepaid tickets and strong connections in the US, had a final destination decided. Indeed, some individuals would have purchased train tickets in advance or relatives would have been waiting in the NY port. However,

[11] Another common arrangement for travelling was prepaid tickets purchased in advance by relatives residing in the US. These tickets required to follow the same steps and procedures than standard tickets.

[12] Passengers usually received a card certifying the medical inspection and additional information like names, ship and manifest page/line. Passengers were instructed to attach the card to their coats and to show it to inspectors upon arrival.

[13] Some ports had facilities for those passengers waiting for departure. In other cases passengers had to pay for their own accommodation.

[14] This contrasts with transatlantic voyages during the late 19th century. For instance, there is a well documented evidence that during the Irish famine migration (1840-1850), weather conditions could delay the departure and the arrival of ships by many weeks (Laxton, 1996).

many passengers travelled with poor information and few contacts on arrival. Lafortune & Tessada (2016) compare the immigrants' answer regarding their intended final destination (if any) with the actual states of residence of recently arrived individuals in the census. They find that only a 45% of answers match with the actual geographical distribution of recent arrivals. Anecdotal evidence suggests that shipmates played an important role in either conveying information on potential destinations and sector of employment or in directly providing job referrals, accommodation and financial support after arrival.[15].

Arrival at Ellis Island When a ship arrived at New York Harbor, immigration officers requested the certified manifests and steerage passengers were conducted to Ellis Island station.[16] Due to the characteristics of inspection facilities, passengers were divided into groups of (approximately) 30 people following their order in the manifest. Passengers who bought tickets together had close manifest numbers. Therefore, families and close acquaintances were typically inspected as part of the same group and queued at the same desk in the Registry Hall. Immigrants had to pass a quick visual medical screening and then immigration clerks in the Registry Hall checked that the inspection cards and the manifest information matched. Finally, passengers answered a series of questions (with the help of official translators) attempted to detect those with criminal records, extreme political affiliations (e.g. anarchists) or likely to became a public charge.[17] Individuals suspected of not meeting the minimum entry standards were separated for further investigation, a procedure that could take several hours or even days. Despite the strict inspection procedure, official statistics reveal that only 2% of passengers were finally deported (US Bureau of the Census, 1975). After inspection, individuals were discharged to enter the US. At this point, many of them faced the decision of where to seek a new life and/or in which sector to apply for a job. The station had money exchange facilities and

[15]For instance, Taylor (2010) provides an example of how destination within US were sensitive to shipmates' suggestions: *"...His mom gave him all the money she had and told him to go to America. He travelled south on foot until he reached Italy, boarded a ship, and landed in New York. People whom he'd met on the ship told him to go to the city of Buffalo because many Polish people lived there..."*. In a second example, Grossman (2009) illustrates that shipmates were also important in providing jobs and accommodation: *"... He took a boat from Cork to New York City. A priest he had met on the ship got him a room to stay in and his job at New York City's Biltmore Hotel..."*. Anecdotal evidence also document a large number of marriages among partners who met during the trip. Indeed, the "Records of the Board of Trade and of successor and related bodies" from the UK, officially registered 133 marriages *while* travelling to the US.

[16]First class and cabin passengers were usually inspected on board and discharged to enter the US without going through the main station.

[17]In practice, the criteria for excluding someone for being *likely to became a public charge*, was circumscribed to passengers with several health conditions or those with not enough money to pay for accommodation and food for a few days after arrival.

many railway agencies from whom they could buy tickets to any destination, including New York City. This paper studies how contacts established during the trip could have influenced decisions at this critical stage.

3 Ships-Census Matched Dataset

In this section I summarize the construction of the dataset and main variables used in the study. Some technical details are relegated to Appendix B where I explain in detail the steps involved in the matching process.

Data Sources The main dataset in this paper combines information from Passenger Lists and historical Censuses. The Passenger Lists contain the universe of 34,000 ship arriving to the New York port during the period 1909-1924.[18] The set of individual variables available in electronic format are: full name, age, gender, race, marital status and last place of permanent residence. I also observe the date of arrival, port of departure and name of the vessel. I compile additional information on ships' characteristics, ports of departure and European cities from multiple online sources.[19] For most of the analysis, I restrict the sample to ships sailing from non-US ports and located at a distance of 3,000 kilometers or more from the port of New York.[20] Individual census information corresponds to the full count of male immigrants from the Integrated Public Use Microdata Series (IPUMS) for years 1920 and 1930 (Ruggles et al 2015). Figure 1 shows the yearly flow of passengers and the immigrant stock in Census for different sub-samples of the population. As discussed in Bandiera et al. (2016), discrepancies between passenger inflows and Census stock are largely driven by return migration and the large drop in immigration inflows after 1914 is due to the WWI.

Matching Census and Ships Data I match passengers' data with census records using first name(s), surname, year of birth and year of immigration. Passengers are

[18]Information from passenger lists is considered accurate and reliable (Weintraub, 2017). The manifests corresponds to the National Archives and Records Administration microfilms series M237 and T715. Similar data has been used in Bandiera et al. (2016) who discuss in detail the accuracy and coverage of passenger lists during the period.

[19]I obtained information available from a number of websites including *www.jewishgen.org, www.stevemorse.org* and *www.theshiplist.com*. I also used information on passenger lists from the series of Family Archives CDs by Gale Research. Patricia MacFarlane provided generous access to the Immigrant Ships Transcribers Guild (ISTG) database which contains digitized passenger manifests and information on immigration during the period of my study.

[20]This excludes all Caribbean, Mexican and Canadian ports which usually account for voyages of short duration. It also excludes a large number of small vessels transporting workers and supplies from and to the Panama Canal zone. Canadian and Mexican citizens are also excluded from the sample.

matched to the closest census year after arrival (i.e. arrivals between 1909 and 1919 are matched to the 1920 census and the remaining to the 1930 census). This dataset allows me to observe the characteristics of immigrants once they are settled in the US, but also the details of the voyage to US, including the characteristics of his shipmates.

The main challenge when matching passenger lists to Census records is the large volume of data.[21] Popular approaches (e.g. Fellegi & Sunter, 1969; Feigenbaum, 2016) can become unfeasible even after following the standard blocking strategy.[22] In Appendix B, I outline a Machine Learning procedure based on Levenshtein Automata that allows me to match records across large datasets. The approach is related to Feigenbaum (2014, 2016) but introduces a number of algorithmic improvements to increase the speed at which the method identifies individuals with similar names and/or surnames.[23] The matched sample consists of 351,289 individuals, 52% of them corresponding to the 1920 census year. The matching rate relative to the Census is around 12%.[24] After excluding individuals sailing from less than 3000 kilometers from New York or missing information on the town of origin or age outside the range 14-65, the sample is reduced to 206,383 individuals.

Geocoding Ports, Routes and Places of Origin I use an algorithm based on the Google Places API to obtain the latitude, longitude and (harmonized) name of departure ports for the universe of ships in the Passenger List data. In total, I identify around 500 different ports, including those located at Caribbean countries, Mexico or Canada. Figure 2 displays the ports identified outside the area excluded from the analysis. Using all the ports declared by passengers (regardless of whether the passenger is matched to the Census or not), I reconstruct the whole route of the ship. Appendix C provides more

[21]Matching based on names and surnames requires calculating string similarity measures, which are computationally demanding. Increasing the sample size exponentially increases the number of string comparisons and this usually becomes unfeasible unless further restrictions are imposed.

[22]Blocking restricts the search of potential matches within a smaller set of records, typically individuals with similar years of birth or arrival. Unfortunately, in my setting blocks are so large that the problem remains.

[23]Intuitively, these modifications reduce the number of repeated calculations required to compare among strings. This is (to the best of my knowledge) the first paper in economics implementing this efficient search approach to match historical data (e.g. Radix Tries Search and Block-Specific Dictionaries). A recent literature in Computer Science have studied the problem of matching large string data (e.g. Baeza-Yates & Gonnet, 1996; Schulz & Mihov, 2002). Unfortunately, there is no existing code or software implementation for these methods and most of them remain as theoretical contributions.

[24]The matching rate is comparable to studies tracking immigrants across census years (Ferrie, 1996; Abramitsky, 2012, 2016). However, as explained in Appendix B the Machine Learning approach requires a human trained random sample of matched individuals. When creating this sample, I use an strict criteria that resulted in a low number of false positive matches. Cross validation exercises reveal that the matching procedure is highly accurate with a false positive rate below the 0.1%. As discussed in a recent paper by Bailey et al. (2017), false positive matches in linked data are more problematic than false negative matches.

details on the geolocalization procedure.

I also geocode information on the "last town of permanent residence" for passengers in the matched sample. The algorithm resembles that used for geocoding ports but it requires some pre-processing steps in order to correct for common typos and abbreviations, towns that disappeared over time and places reported in their original language.[25] The full procedure is described in detail in Appendix C . Overall, I identify around 11,000 different places of origin. Figure 3 displays the location of places identified in the matched sample. Appendix Figure A1 shows the relative frequency of the main ports of departure and countries of origin.

Labor Outcomes Since the 1920 and 1930 censuses did not record information on individual income, I follow previous studies (Abramitsky et al. 2012, 2016; Maurer & Potlogea, 2017) and use the *Occupational Earnings Score* which assigns each individual the percentile rank of his occupation in terms of median earnings in 1950. Naturally, this measure is invariant to wage differences within occupations but it captures whether an individual is employed in a job that pays relatively more. As a robustness check, I use two additional measures. The first one is the *Duncan Socioeconomic Index*, which assign a (subjective) prestige rating to each occupation based on earnings, education and the 1947 National Opinion Research Center Survey (NORC). The second additional measure is the *Nam-Power-Boyd Index* (Nam & Boyd, 2004) which measures the percentage of the labor force employed in occupations with combined levels of education and earnings below the incumbent occupation.[26] Finally, in order to aid the interpretation of the results, I construct a measure of occupational earnings by assigning to each individual the median earnings of his occupation in 1940. Information on sectors of employment and occupations is created and harmonized by IPUMS based on unstructured text questionnaires answers.[27]

Summary Statistics Table 1 presents some summary statistics of the data. Panel A reports aggregated information on the number of individuals, ships and places of origin for different sub-samples and data sources. The first column (full sample) includes individuals

[25]The algorithm generates the following information: latitude and longitude of the place, name identified by the Google Places Api and the south-west/north-east coordinates of the smallest rectangle containing the place. A 20% of the records have missing information on the place of origin and a 15% of the observations are geocoded with a precision above the locality level (e.g. province).

[26]All these variables are created by the Minnesota Population Center and are comparable across individuals and census years (Ruggles et al. 2015).

[27]Although these variables are not directly comparable with more recent industry or occupation classifications (e.g. SIC or NAICS for industries or SOC for occupations), the disaggregation is comparable to 3-digits level and consistent accross census years.

from any origin and age group. The matching rate, defined as the number of matched individuals with respect to the individuals observed in the Censuses, is 12.4%. Matched individuals are observed in approximately 34,000 different ships, departing from 422 ports and proceeding from 10,900 different places of origin.[28] After restricting the sample to individuals in the age group 14-65 with non-missing information on the place of origin and to ships departing from ports at a minimum distance of 3000 km. from New York, approximately 206,000 individuals from 15,000 ships, 170 ports and 8,200 places of origin remain in the sample.

Panel B reports basic statistics on individual and ship characteristics. Ships in the regression sample travelled an average distance of 6,500 kilometers (whole route). This distance would take about 10 days at 15 nautical knots, the average speed for steamers in that period. In the full passenger list data, an average ship transported 173 male passengers in the age group 14-65 (excluding those boarding at less than 3000 km from New York). Ship size is consistent with the findings in Bandiera et al. (2013) for the same period.[29]The average number of passengers per ship observed in the matched sample was about 20. Ships were very diverse in terms of places of origin: an average ship transported individuals from 15 different towns of origin (in the matched sample). A large proportion of passengers were single and travelled without any relative. At destination, most immigrants settled in urban places and 21% were observed living in New York in the next Census after their arrival.

4 Empirical Setting

In this section, I explain the empirical strategy to estimate the effects of brief social interactions, and then justify it with a set of balancing tests. Establishing this causal effect is not an easy task. In addition to considering the exogenous allocation of individuals across ships, I need to consider the possibility that shipmates' characteristics can affect earnings through channels that do not require social interaction. I postpone the discussion of these confounding effects to Section 6, were I provide additional evidence on the social interaction mechanism.

[28]Table 1 indicates that 15% of places of origin are geographical units above the locality level (e.g. province). As a robustness check, in Appendix B I re-estimate the main results excluding these geographical units

[29]Bandiera et al. (2013) find that for the period 1892-1924, the average number of passengers per ship was approximately 500. However, after 1911, the average number of passengers drops below 200 per ship. After accounting for the gender, age and port restrictions in my sample, the average number of passengers is in the same range.

Defining Brief Social Interactions The first step in the analysis requires defining the set of individuals who met for the first time during the voyage. For every individual, I identify this set by *excluding* any shipmate such that 1) shares the same town of origin or 2) has a similar surname, defined as a *Jaro-Winkler* distance below 0.1.[30,31] Along the paper, I will refer to them as the set of *unrelated shipmates*. In Section 5, I perform a set of exercises to rule out the chance that effects are driven by a weak definition of unrelated shipmates.

Connections on Arrival An important variable that I use below is the quality of potential contacts that immigrants had in the US. This is a key variable in the empirical strategy as I will proxy the quality of shipmates based on this dimension. Following a number of influential papers (e.g. Wegge, 1998; Munshi, 2004; McKenzie & Rapoport, 2007, 2010) I define the set of potential contacts at destination, as those individuals who emigrated in the past from the same place of origin. There are two additional reasons to use the community of origin as the relevant unit to define the social network at destination. First, there is a strong consensus among historians on the importance of settled immigrants in triggering chain migration and supporting new arrivals from the same community (Daniels, 2002). Second, during this period the outcomes of newcomers are strongly correlated with the characteristics of settled immigrants from the same community.

To measure the quality of contacts on destination, I focus on two variables:[32]

1) *The average earnings score of settled immigrants from the same town of origin.*

2) *The number of individuals from the same town who emigrated to the US in the past.*[33]

[30]The Jaro-Winkler distance (Winkler, 1999) measures the similarity between two words based on the number and position of common characters.

[31]In addition to these conditions, I use the smallest rectangular area containing the place of origin to exclude any shipmate with area overlapping above 50%. This additional condition assures that no shipmate is considered "unrelated" due to a poor geocoding information (e.g. a shipmate with the same province of origin but without information on the exact town of origin). In Section 5, I show that the main results are robust to more strict conditions (e.g. excluding close towns)

[32]As a robustness check, in Section 5, I re-estimate the main results using alternative definitions of connections on arrival.

[33]The earnings of settled immigrants are calculated only for towns observed in the matched sample as I have no information on earnings of non-matched individuals. The number of emigrants from each town is calculated using the full flow of passengers observed in the passenger lists since 1900. For a given immigrant, either variable is calculated using only individuals who travelled at least one month before him.

The first variable proxies the economic status of potential contacts, based on the notion that wealthier connections can provide information or referrals on better jobs. The second variable proxies the size of the network at destination.[34]

Formally, I define $x_{c(k),t(k)}$ as the earning score for an individual k from town $c(k)$ and who travelled in period $t(k)$. This notation emphasizes the fact that each individual in the data is associated to a unique town of origin and emigration period. The average earnings of potential connections on land for individual j is defined as $X_{c(j),t(j)} = \sum_{r(k)=1}^{t-1} x_{c(k),r(k)}/N_{c(j),t(j)}$ with $N_{c(j),t(j)}$ being the number of individuals from town $c(j)$ who emigrated before period $t(j)$ and are observed in the census.[35] The number of potential contacts upon arrival for individual j, defined as $Z_{c(j),t(j)}$, can be measured as the size of emigration flows from town $c(j)$ to the US *before* period $t(j)$. Note that $Z_{c(j),t(j)}$ is measured using the whole passenger list but $X_{c(j),t(j)}$ and $N_{c(j),t(j)}$ are calculated using the matched sample only. This underlines the complementarity of the two measures. Table 1 Panel B, reports summary statistics about these variables. Earnings of potential contacts are measured in the scale of 0 to 100 and the average in the sample is 49.7. The average number of potential contacts of an individual is 9,300.

Figure 4 illustrates the relevance of previous definitions. Each panel of the figure displays the coefficients of the following regressions between individual outcomes and the quintiles of his potential contacts' characteristics, conditional on ship and predetermined individual characteristics:

$$Y_i = \sum_{q=1}^{5} \beta_q ContactsChar_i^q + \sigma_{s(i)} + \alpha I_i + \epsilon_i \tag{1}$$

where Y_i is an outcome of individual i (measured at the next Census after arrival), $ContactsChar_i^q$ is a dummy for the quintile q of some characteristic of the potential contacts of the individual (e.g. the number of individual's contacts $Z_{c(i),t(i)}$). Each regression controls for ship fixed effects $\sigma_{s(i)}$ and a set of predetermined individual characteristics I_i. Panel A shows the correlation between individual earnings and the average earnings (and number) of settled immigrants from the same town of origin. Panels B to D shows that the location of individuals and the sector of occupation are strongly correlated with those of previous emigrants from the same place. Thus, even if newcomers never interact with settled immigrants, we can think that at the moment of the trip, the previous definitions are predetermined predictors of immigrants' economic success.

[34]Previous studies have measured the migrant network size in different ways. For instance, Munshi (2003) measures it as the share of immigrants from the home community while Beaman (2012) uses the number of individuals from the same country living in a given city.

[35]Note that earnings scores of individuals arrived in different years are usually observed in the same census year.

Identification Strategy In order to identify the effects of brief social interactions, I rely on the assumption that, conditional on their towns of origin, individuals departing from the same port and in the same week, were plausibly exogenously assigned to ships. The plausibility of this assumption is empirically validated later in this section. The intuition behind the identification strategy can be illustrated with the following example: Assume that an individual with residence in Benevento (Italy) has decided to emigrate from the port of Naples (the closest to his town). Naturally, individuals departing in different years or seasons, may face different conditions at departure or arrival. Consequently, shipmates' characteristics can be correlated with unobserved determinants of the individual's earnings at destination. Consider, however, all the ships departing from Naples within a relatively narrow time horizon (e.g. a week). The identification strategy relies on the assumption that the individual assignment is uncorrelated with the characteristics of the unrelated shipmates boarding the same ship.[36]

A number of historical facts support this assumption. First, the selection among passengers of different income took place mainly within ships, as every vessel had different classes and service upgrades. For instance, wealthy individuals usually travelled in first or cabin classes. Second, during a short window of time, the fares for lower class categories (e.g. third class or steerage) were remarkably similar across shipping lines for a given route.[37] The vast majority of immigrants travelled in steerage class. Third, delays due to paperwork or unexpected changes announced by the shipping company were common. Finally, passengers bought their tickets days or weeks in advance, without being able to anticipate the characteristics of their potential shipmates. Naturally, the exogeneity claim must be validated in the data, and in this section I discuss a number of empirical exercises that support this assumption.

A potential concern is that some vessel characteristics (for instance, their external look or capacity) can influence the individual decision, creating some endogenous sorting of passengers. In Section 5, I show that results are robust to the inclusion of a large set of ship characteristics and even of vessel fixed effects. Moreover, as shown below in this section, ship characteristics are strongly balanced with respect to the average shipmates' quality.

The exogenous allocation across ships, creates quasi-experimental variation in the

[36]In Section 5, I explore two alternative identification strategies based on the variation created by repeated voyages of the same vessel and by individuals boarding at different ports during the same trip.

[37]For instance, Hopkings (1910) reports that in 1909, all the steamers covering the Mediterranean service of the Cunard Line, North German Lloyd, White Star Line and Italian Royal Mail Lines had a basic minimum fare of $65 for third class (steerage). Indeed, when including all routes and services, more than 80% of steamers had a basic minimum fare between $55 and $65. This basic fare excluded any additional service or railway transportation.

pool of (unrelated) shipmates of each passenger. This implies that similar individuals can be exposed to a pool of shipmates with different quality of connections on land. An advantage of this strategy follows from the fact that the characteristics of contacts upon arrival are predetermined variables at the moment of the trip, thus not affected by any shock occurring after departure.

Estimating Equation The baseline estimating equation is:

$$Y_i = \beta_1 \bar{X}_i + \beta_2 \bar{Z}_i + \theta_{p(i)} \times \lambda_{w(i)} + \delta_{c(i)} \times \pi_{t(i)} + \epsilon_i \tag{2}$$

where Y_i is a labor market outcome for immigrant i in the US. Consistently with the earlier discussion, I control for the interaction between $\theta_{p(i)}$ (a fixed effect for the port of departure) and $\lambda_{w(i)}$ (the fixed effect for the week of arrival).[38]

The main variables of interest, $\bar{X}_i$ and $\bar{Z}_i$, measure the quality of the connections of i's shipmates. The first variable is the average earnings score of the potential connections on land among i's shipmates. The second measure, is the average number of potential contacts among i's shipmates. As discussed in Section 3, potential connections on land for individual j are defined as the set of emigrants from the same town of origin. Formally, if $u(i,s)$ is the subset of passengers travelling in ship s and unrelated to i, I define $\bar{X}_i = \sum_{j \in u(s,i)} X_{c(j),t(j)}/n_{u(s,i)}$ with $n_{u(s,i)}$ being the number of unrelated shipmates for individual i. Similarly, I define $\bar{Z}_i = \sum_{j \in u(s,i)} Z_{c(j),t(j)}/n_{u(s,i)}$.[39] As defined before in this Section, for a given individual j, $X_{c(j),t(j)}$ is the average earnings in the US among individuals from town $c(j)$ who emigrated before period $t(j)$ and $Z_{c(j),t(j)}$ is the total emigration flow from town $c(j)$ to the US before period $t(j)$.

The baseline specification also controls for the interaction between $\delta_{c(i)}$ (a fixed effect for the town of origin of immigrant i) and $\pi_{t(i)}$ (a fixed effect for the semester of arrival). The inclusion of this interaction serves two purposes. First, it controls for

[38]Note that I do not observe the week of departure, however, conditional on the port of departure, this is similar to control for the week of departure. Moreover, the route of the ship accounts for almost all the variation in voyage duration. In Section 5, I present evidence that results are robust to the inclusion of the route fixed effects.

[39]Some technical aspects involved in the calculation are worth mentioning: (a) Note that both variables are averaged across unrelated shipmates, thus unaffected by their number; (b) As discussed in Section 2, most social interactions are likely to be among passengers boarding at the same port. For this reason I only calculate the average characteristics among this set of unrelated shipmates. In Section 5, I modify this definition and use the characteristics of shipmates from different ports; (c) I only use the characteristics of shipmates in the matched sample. As discussed by Ammermueller & Pischke (2009) and Sojourner (2013), failing to account for the full set of relevant peers, can introduce some attenuation bias in the results. Of course, the identification strategy assumes that the probability that shipmates' are matched is not systematically correlated with unobserved characteristics of the individual, after conditioning for the baseline controls. I address this concern later in this Section.

unobserved time-variant characteristics that could result in individuals from specific towns boarding certain ships with higher probability. This would be the case, for instance, if agencies sold tickets for different ships with varying intensity across regions of the country. Second, given that potential connections on land are defined at the town of origin level, it absorbs any characteristic of individual's own contacts. As discussed in Caeyers & Fafchamps (2017), this strategy eliminates any negative exclusion bias (Guryan et al., 2009) introduced by the fact that i's connections are excluded in the calculation of $\bar{X}_i$ and $\bar{Z}_i$.[40] All regressions cluster standard errors at the week of arrival level. In Appendix Table A2, I show that baseline estimates are robust to alternative clustering choices.

Balancing Tests and Evidence of Exogenous Sorting This subsection discusses a number of tests supporting the identifying assumption outlined before. This is critical to establish a causal interpretation of the effects of shipmates' characteristics on future labor outcomes.

The first test consists of studying the correlation between the predetermined variables of an individual and those of his unrelated shipmates. The exogeneity claim requires that this correlation must be zero after conditioning on the interaction between the port of departure and the week of arrival. Therefore, for every individual in the matched sample, I calculate the average characteristics of his unrelated shipmates. In order to avoid the negative mechanical bias of leave-one-out correlations, I follow Baker et al. (2008) and sample one individual per ship when performing these calculations. Column 1 of Table 2 reports the unconditional correlations and Column 2 conditions on Port of Departure X Week of Arrival.[41] Results indicate that the unconditional correlations are high and significant but all of them become low and insignificant (at 5% level) after controlling for Port of Departure X Week of Arrival.[42]

The second set of tests is given by standard balance regressions. This consists of OLS regressions of a number of predetermined passenger and ship characteristics on the two main variables of interest, $\bar{X}_i$ and $\bar{Z}_i$. The results in Figure 5, where I label each row in the left axis by the dependent variable, plot the estimated 95% confidence intervals of the regression. Panel A plots the confidence intervals for the average earnings

[40] I define $\pi_{t(i)}$ at semester level due to the relatively small size of most towns of origin. For instance, I observe very few week-port cells with more than one individual from the same town boarding different ships. In Section 5, I show that results are robust to controlling for the interaction between town of origin and the month of arrival.

[41] A number of predetermined characteristics in the test vary at the town of origin level, for this reason, I do not control for the town of origin fixed effect, but on a larger geographical level (e.g. provinces in the case of italy). Note however, that this imposes a more demanding condition for balance.

[42] Significance levels are bootstrapped by repeating 500 times the procedure of sampling one individual per ship.

of unrelated shipmates' contacts on land. Similarly, Panel B corresponds to the average number of shipmates' potential connections on land. To illustrate the importance of the baseline controls, I report the estimates with and without the Port of Departure X Week controls.[43] To ease interpretation, all variables in the regressions are standardized.

I find that shipmates' characteristics are (unconditionally) correlated with individual and ship characteristics: the estimates are statistically significant for most dependent variables. The introduction of the baseline controls, however, greatly decreases the estimates which become extremely small in magnitude. For any left hand side variable, the coefficients imply that one standard deviation in either the number or the earnings of unrelated shipmates' contacts on land, has an effect lower than 0.05 standard deviations. Indeed, after controlling for Port X Week, only two of the 32 displayed coefficients are statistically different from zero at the 5% level.[44]

Overall, I interpret the results of this subsection as supporting the exogeneity of the variation of shipmates' characteristics among unrelated individuals departing from the same port during a given week. Consequently with these findings, In Section 5 I provide additional support for the identification assumption, by showing that the results are robust to the inclusion of a large set of additional controls.

Census-Ships Data Matching and Non-Random Sampling A potential concern in the study is that the matching process creates a non-random sample of the ships. A number of additional findings suggest that, conditional on baseline controls, matching is not systematically correlated with individual or ship characteristics.

First, note that the dependent variable in the last row of Figure 5 is the (standardized) share of matched passengers within the ship. Conditional on the Week X Port controls, the correlation is extremely low in magnitude: One standard deviation increase in $\bar{X}_i$ or $\bar{Z}_i$, changes the matching rate in less than 0.02 standard deviations. Figure 6 further explores this idea and estimates the balance equation for quintiles of the shipmates' contacts characteristics.

Second, I estimate the correlation between the ship matching rate and a set of individual predetermined characteristics conditional on similar controls than those in the balance regressions. Figure 7 plots this regression. Estimated coefficients are insignificant

[43]Following the discussion in footnote 41, regressions include fixed effects for large administrative units. Additionally, in order to eliminate any potential downward exclusion bias (Guryan et al., 2009), I control for the earnings and number of passenger's own potential connections. Appendix Figure A3 displays similar balancing tests using the same controls and sample used in the baseline specification (variables defined at town of origin level are then excluded)

[44]Since the right hand side variables can be correlated with each other, Appendix Figure A2 displays the F-statistics of the joint significant test of each regression.

for 12 out of 13 variables and low in magnitude in every case. Along with the balance tests, this evidence suggests that conditional on baseline controls, the matching algorithm does not correlate with individual outcomes. This is not surprising as surname characteristics are the main determinants of the matching rate, and within the Week X Port cell, they are not systematically different.

Finally, I use the full Passenger List data to study whether the probability of being matched correlates with ships characteristics. I regress a dummy variable indicating if the passenger was matched to Census on the full set of Ship fixed effects. Table 3 reports the F-statistic for the joint significance test of Ship fixed effects. Column (1) shows that without further controls, Ship fixed effects have significant predictive power on the matching rate. However, as shown in Column (2), after including the Week X Port fixed controls, Ship fixed effects are jointly insignificant.[45]

These findings also highlight an advantage of the empirical strategy: Even if matching is non-random for the whole sample (e.g. because some nationalities are easier to match), narrowing the variation to the Week X Port of Departure level eliminates any significant difference in matching rates across ships or individuals.

5 Baseline Results

This section describes and interprets the baseline results of the paper. I also show that the effects of travelling with better connected shipmates persisted for years after the arrival. I then discuss a number of robustness tests aimed to provide additional support for the identification assumption. Finally, I discuss the robustness of results to alternative specifications and clustering of standard errors.

Baseline Estimates Table 4 reports estimates of Equation (2) for different measures of earnings and job quality. Column (1) indicates that both dimensions of shipmates' contacts quality have a positive and significant effect on individual earnings score. Exposure to shipmates with connections employed in jobs one percentile higher in the earnings distribution, increases individual earning score in 0.14 points. Similarly, every thousand additional (average) connections among shipmates increases earnings score by 0.05.

[45] A different concern is related to the partial observability of the relevant network structure. Under (conditional) exogenous sorting of individuals across ships, this would result in coefficients attenuated to some extent as discussed in Ammermueller & Pishcke (2009) & Sojourner (2013). In Appendix D, I discuss how the baseline results vary according to the matching rate and the implications for potential attenuation bias. Additionally, I discuss a number of simulations suggesting that the attenuation bias is relatively low in this setting.

Columns (2) to (3) reports the results for the alternative measures of job quality discussed in Section 3. Estimates indicate effects of a similar magnitude.[46] Although these variables are correlated with the earning score, they measure different aspects of job quality. Understanding the size of effects based on Earnings Score is not straightforward as the earning distribution is typically left-skewed. In order to ease the interpretation of my findings, I also report the estimates of Equation (2) when the dependent variable is the logarithm of the earnings derived from the 1940 Census.[47] Findings in Column (4) mean that an upward shift of 10 percentiles along the income distribution of shipmates' connections, increases individual earnings by 2,7%. Every thousand additional (average) connections among unrelated shipmates, increases earnings by 0.7%.[48]

Equation (2) can hide some non-linear relationship between individual earnings and shipmates' connections quality. A potential concern is that results are driven by few ships with outlier characteristics. Figure 8 displays non-parametric evidence that the effects are increasing in the quintiles of the variables of interest. In the case of shipmates' connections earnings, effects are monotonically increasing and statistically significant for quintiles 3 to 5. Travelling in a ship in the highest quintile, increases individual earnings score in 1.8 points with respect to the lowest quintile (an effect of 4% according to the regression with log-earnings in panel B). In the case of the number of connections, the effects are weakly increasing but only significant for the highest quintile. Travelling in a ship among the highest quintile of this variable, increases individual earnings score by 1 point with respect to the lowest quintile (an increase of 2% based on the regression with log-earnings displayed in panel B). It is useful to compare these figures with the estimated correlations between earnings and the characteristics of individual's own connections in the US (Panel A of Figure 4). Although the later is not necessarily causal, it is a useful benchmark for interpreting the magnitude of the effects. Not surprisingly, the effects of shipmates' connections on earnings are lower than the correlation with respect to the own contacts' characteristics. For instance, relative to the lowest quintile, the effect of travelling with shipmates in the highest quintile of contacts' earnings is three to four times lower than the effects of having connections in the highest quintile of earnings.

Appendix Table A3 explores the interaction between the two measures of quality

[46]The Duncan Socioeconomic Index, reflects the social perception of the "prestige" associated to an occupation. The Nam-Power-Boyd index captures differences in the education-earning composition of different occupations. Both variables have the same scale than the earnings score (0 to 100).

[47]The construction of this variable is described in Section 3.

[48]Appendix Table A1 reports the results for two additional variables based on the 1950 Census. The dependent variable in Column (2) replicates the last column in Table 4 but using 1950 Census. Column (3) assign each individual the median earnings of the percentile associated to his occupation according to the earnings distribution in 1950. Results are robust to these alternative earnings measures.

of shipmates' connections. The estimated coefficients correspond to an OLS regression (analogous to Equation (2)) where the explanatory variables are the interactions between two sets of dummies indicating whether the number of shipmates' connections or their average earnings are above/below the median of its distribution. Both measures of connections' quality are relevant. Starting from a situation where shipmates have low-quality connections in terms of both earnings and number, an increase in either dimension has a positive impact on earnings. Table A3 also suggests that the earnings of shipmates' connections is relatively more important than the number of shipmates' connections.

The baseline effects display some heterogeneity at geographical level. Appendix Figure A5 plots the estimates of Equation (2) where the shipmates contacts' earnings variable is interacted with dummies for the country of origin of the individual. The map shows the relative size of the effects for Europe. Among countries with more emigrants in the data, effects are stronger for Ireland, Poland and Greece. Naturally, other factors correlated with the country of origin can drive the heterogeneous effect. For instance, the estimated effect for Italians is significant but below the median for Europe. This could be partially explained by the fact that Italians from distant regions typically spoke different languages. Unsurprisingly, the potential benefits of social interactions might depend on the ability to communicate with those well connected shipmates.

Persistence of the Effects Due to the low number of arrivals between 1914 and 1919, most immigrants in the data are observed many years after arrival (7.5 years on average). This suggests that effects of social interactions with unrelated shipmates is highly persistent. Figure 9 explores this idea in more detail and displays the estimates of the baseline equation where the right hand side variables are interacted with dummies for each year since arrival. Although this disaggregation can confound other characteristics correlated with the time since arrival, the figure suggest that effects are not only driven by recent migration. Moreover, estimated effects are statistically significant even 10 years after arrival.[49]

[49]There are two main confounders for this heterogeneous effect. First, earlier arrivals are older when observed in the Census, and additionally, given the high rate of return migration in this period, likely positively selected. Second, immigrant cohorts can differ in terms of skills and other unobserved determinants of earnings. Whereas the later can't be controlled for, I alleviate the first concern by controlling for the interaction between the right hand side variables and the age of the individual. An additional source of heterogeneity over time is the 1921 Immigration Act, which mainly affected immigration from eastern and southern European countries. Appendix Table A4 shows the effects of shipmates' contacts characteristics interacted with dummies of pre/post 1921 Immigration Act. Results suggest that baseline findings are mainly driven by arrivals before 1921.

Additional Controls In this subsection I show that results are robust to the inclusion of a large number of additional controls. This evidence is important to rule out some potential threats to the validity of the identification strategy. Table 5 summarizes all these findings. Columns (2) and (3) show that estimates are robust to the inclusion of a set of individual characteristics (age, race, marital status, language, and an indicator for the individual travelling with some relative) and a set of characteristics of the ship and the route (e.g. ship capacity, number of passengers, distance travelled, number of stops, share of male passengers, etc.). Robustness to these controls is consistent with the assumption that, conditional on baseline controls, the pool of shipmates is not correlated with individual or ship characteristics. In a more general way, I want to rule out that individuals select into ships due to unobservable characteristics of the ship. This would be the case if for instance, more educated individuals (which potentially correlates with their connections quality) select into ships with higher capacity or higher speed. Such situation would confound the effect of better connected shipmates with individual's different characteristics. Column (6) shows that effects are similar after controlling for vessel fixed effects and this finding is inconsistent with such interpretation.

Note that the baseline specification (Equation (2)) absorbs any shock at the Town of Origin X Semester level. Although this is an already narrow time-space grid, some concerns may arise regarding the relevant time horizon in which local shocks can affect passengers' predetermined characteristics.[50] Column (4) extends the baseline specification to a shorter window of time by controlling for the interaction between fixed effects of the town of origin and the month-year of arrival. Since most towns are relatively small, there are fewer cells with multiple individuals from the same town boarding different ships within the same month. Despite of the lower number of observations, results remain statistically significant with coefficients of similar magnitudes. Column (5) narrows the time horizon to the week level but uses a larger spatial aggregation grid (administrative units above the locality level, e.g. provinces in the case of Italy). In this case, results are similar for the earnings of shipmates' contacts and non-significant for the number of connections on land, although standard errors are also larger due to the introduction of a large number of fixed effects.

As discussed in Section 4, it is possible that ships departing from the same port during the same week, followed a different route. Although the vessel fixed effect controls for most of this variation, some vessels could have covered different routes over time. Column (7) shows that baseline results are robust to the inclusion of fixed effects for each

[50]For instance, it could be the case if a local shock greatly changes the quality of individual's own connections within a semester.

route identified in the data. This rule out that results are driven by some correlation among shipmates' characteristics created by individuals selecting across ships based on the travelled route.[51]

Finally, Columns (8) and (9) aim to control for a narrow set of individual characteristics and labor market conditions upon arrival. Column (8) includes fixed effects for the NYSIIS phonetic coding of surnames (Atack & Bateman, 1992) which accounts for approximately 8000 groups of surnames.[52] Column (9) includes fixed effects for the date of arrival. Despite of a lower number of observations, estimates are robust to the inclusion of the additional controls. These findings have a number of implications. First, surnames embeds some important unobserved characteristics of individuals. Thus, findings are consistent with the claim that conditional on baseline controls, passengers do not select into ships according to individual characteristics that correlate with earnings. Second, surname is the most critical variable when matching between Passenger Lists and Censuses. Some surnames are more difficult to match either because they are too frequent, or because they are more likely to be misspelled when transcribed. Therefore, results in Column (8) are inconsistent with a non-random matching across ships driving the results. Lastly, results in Column (9) rule out that some correlation between shipmates' characteristics and daily conditions upon arrival explains my findings. This would be the case if for instance, the arrival of passengers from certain towns triggered some events like a higher demand for train tickets to some destinations or a lower availability of temporary accommodation in New York City.

Narrowing the Definition of Unrelated Shipmates One potential concern when establishing a causal interpretation of Equation (2) is the possibility that shipmates from different places of origin are already connected before travelling. Although this is an unlikely event for the vast majority of passengers, I restrict in two ways the pool of shipmates assumed to be unrelated. First, I use the fact that travelling together (or buying the ticket from the same agent) typically implied nearby manifest line numbers. Appendix Figure A4 shows an example of this situation where members of the same family follow the same order.[53] In Table 6, I report the estimates of the baseline equation but

[51]This is not surprising given that the ports concentrating most of the departures in this period are usually covered by few routes, and in many cases by a unique route.

[52]Including surname fixed effects is problematic for two reasons. First, the large variety of different surnames would absorb most of the variation at individual level. Second, a non-negligible part of the variation in surnames can be due to transcription errors or typos.

[53]Although families are almost always in the same block within the manifest, this practice was not universally extended for groups travelling together as a small number of shipping companies sorted entries alphabetically by surname.

for every individual, I restrict the set of his unrelated shipmates by imposing a minimum distance in their ID numbers (which follows the same order than manifest line numbers). The first two rows of the table exclude any shipmate with a difference in ID numbers lower than 10 and 15 respectively. The second way in which I restrict this set is by excluding passengers with towns of residence located at less than 100 kilometers from each other. The last row of Table 6 displays the baseline results after imposing both sets of restrictions (Minimum ID number difference and minimum distance). Point estimates are somewhat lower for the earnings of shipmates' contacts (but they remain statistically significant at 1%) and they are similar for the number of shipmates' connections. Note that either restriction can introduce some attenuation bias if true unrelated shipmates are excluded.[54] Moreover, due to language constraints and social preferences, interaction with unrelated individuals can be more likely to occur among those from closer towns.

Alternative Definition of Connections on Arrival As discussed in Section 4, defining potential contacts in the US at the town of origin level is in line with a number of previous studies. However, in the setting of this paper, it is possible to think that narrower definitions of connections are also relevant (for instance, relatives who emigrated in the past). In Appendix Table A5, I re-estimate the baseline specification using two alternative definitions of potential connections upon arrival to the US. First, in Column (2) I consider individuals with similar surname (based on the NYSIIS coding) who previously emigrated from the same province or large administrative unit. Second, I consider past emigrants from the same town of origin who share a similar surname (Column (3)). For small places, the second definition captures to a large extent, relatives who emigrated in the past. In order to ease the comparison across definitions, I standardize all the right hand side variables. Column (1) corresponds to the baseline definition.[55] Alternative definitions result in estimated effects of similar magnitude, and in both cases, higher than the baseline results. Higher estimates can be due to a number of reasons. First, unique surnames are not included in the pool of unrelated individuals when computing earnings of shipmates contacts. Second, given a narrower definition, within ship variation in shipmates' characteristics is also larger. Finally, connections with settled emigrants of similar surname can be the main source of information and support upon arrival, or just better predictors of economic success for immigrants.

[54] See footnote [45]

[55] Narrowing the definitions for potential contacts significantly reduce the number of observations and statistical power since, for instance, very few individuals from same town and with the same surname migrate in the same semester. For this reason, the specification in Table A5 includes fixed effects for the group at which contacts are defined (e.g. Town of Origin X Surname) but interacted with census year instead of semester fixed effects.

Alternative Identification Strategies I explore two additional sources of variation in the characteristics of shipmates. The first strategy exploits the fact that many vessels traveled from the same port repeatedly during the year. Therefore, I only compare passengers travelling in the same vessel within the same semester. Column (1) in Appendix Table A6 estimates the following equation:

$$Y_i = \beta_1 \bar{X}_i + \beta_2 \bar{Z}_i + \psi_{v(i)} \times \theta_{p(i)} \times \lambda_{y(i)} + \delta_{c(i)} \times \pi_{t(i)} + \eta_{r(i)} \times \pi_{t(i)} + \epsilon_i \quad (3)$$

where $\psi_{v(i)}$ is a vessel fixed effect, $\eta_{r(i)}$ is a route fixed effect and the rest of variables are defined identically to Equation (2). Estimates for this specification are displayed in Column (1). Point estimates are highly significant for the case of earnings of shipmates' contacts and the magnitude is approximately 40% lower than the baseline effects. These results provide additional evidence that baseline effects are not driven by passengers sorting across vessels.

The second alternative variation follows from the fact that some ships stopped at different ports before arriving New York. In this case, I exploit the variation in shipmates' characteristics created by passengers from different ports. A potential concern of this specification is that some ports can be very distant from each other reducing the potential interaction between these shipmates. Moreover, in many cases, shipmates boarding at different ports spoke different languages.[56] Additionally, the sample size is largely reduced because either there were no intermediate stops or because only few individuals boarded at a different port. Indeed, I exclude any ship where more than 90% of the passengers boarded in the same port. Column (2) estimates the following equation:[57]

$$Y_i = \beta_1 \bar{X}_i^{sp} + \beta_2 \bar{Z}_i^{sp} + \alpha_1 \bar{X}_i^{dp} + \alpha_2 \bar{Z}_i^{dp} + \psi_{v(i)} + \delta_{c(i)} \times \pi_{t(i)} + \eta_{r(i)} \times \pi_{t(i)} + \epsilon_i \quad (4)$$

where $\bar{X}_i^{sp}$, $\bar{Z}_i^{sp}$, $\bar{X}_i^{sp}$ and $\bar{Z}_i^{sp}$ are defined similarly to Equation (2) but I distinguish between the characteristics of passengers boarding in the port (superscript sp) and that of those boarding at a different port (superscript dp). Estimates from this equation are displayed in Column (2). Point estimates are higher for the characteristics of same-port shipmates and only significant for the earnings of shipmates' contacts. Finally, in Column (3) I only consider individuals who boarded the ship at the first departing port and use

[56] This was true not only for ships stopping at different countries. For instance, italians boarding at different ports typically spoke different languages/dialects and fluent communication among them was very unlikely.

[57] Note that I don't include the interaction between port of departure and the time dimension in order to exploit the variation across ports of the same route. Instead, I control for the interaction between the route and the semester of arrival. This is a less demanding specification compared to the baseline, but unfortunately, statistical power is too low to include Route X Week fixed effects.

the variation created by shipmates boarding at subsequent ports. Remarkably, point estimates for the earnings of shipmates' contacts are similar to those in Column (2).[58]

6 Mechanism: Establishing a Social Interaction Interpretation

The findings in the previous Section, show a causal link between the short run exposure to a pool of better connected individuals and future performance in the labor market. However, this reduced form result is compatible with a number of mechanisms that do not necessary require social interaction among unrelated shipmates. In this Section, I provide evidence supporting the social interaction interpretation as the most plausible one.

I start by showing, that the effect is stronger for passengers with fewer connections and that results are driven, to a larger extent, by shipmates speaking the same language (a natural mediator of social interaction). Then, I show that the sector of employment and place of residence of shipmates' contacts have predictive power on the occupational and residential outcomes of the individual. Finally, as a reassuring exercise, I show that conditional on arriving in the same week and from the same port, the correlation in labor and residential outcomes is stronger among shipmates.

Heterogenous Effects As described in Section 3, this was a period of high-stakes decisions for most immigrants. Consequently, the effects of brief social interaction are expected to be higher for individuals with poor connections and no access to relevant information. Table 7 displays estimations of the baseline regression where each measure of shipmates' connections is interacted with dummies indicating how well connected is the individual himself. Column (1) explores the quality of connections on board, that is, whether the passenger is travelling alone or with relatives.[59] Individuals travelling alone are more benefited by travelling with higher quality shipmates. Column (2), shows that individuals with poor connections on land[60] are more affected by their shipmates' contacts characteristics. Finally, Column (3) shows that effects are stronger for individuals travelling alone *and* with poor connections on land.

[58] This also illustrates that my identification strategy is robust to exclusion bias (see Angrist, 2014).

[59] In order to avoid confounding the effect with the surname prevalence, I include surname NYSIIS code fixed effects. This explains why, consistent with Table 4, the average effect is higher compared to the baseline.

[60] An individual is defined as low connected when the median earnings and the median number of past emigrants from his town are below the median.

A subset of ships in the sample contains shipmates who spoke different mother tongue (using Census definition). As verbal communication is an essential component of social interaction, I expect that the characteristics of shipmates who speak the same language are more relevant.[61] Table 8 displays the estimates of the baseline equation but separating among the characteristics of unrelated shipmates with similar and different language. Although the average effects are lower compared to baseline results, the coefficients associated to shipmates of similar language are always higher compared to those of different language.[62] In the next subsection, I find evidence that shipmates that spoke the same language are also more relevant in explaining the sector of employment and place of residence of immigrants.

Sector of Employment and Residence Place of Shipmates' Connections According to the social interaction hypothesis, shipmates are important in providing information on potential destinations within US. They can also affect labor decisions either by granting access to their networks on arrival or by directly providing job referrals. Consequently, I expect that a number of immigrants migrated toward places where shipmates' contacts concentrates. Similarly, a number of immigrants should have got jobs in sectors where shipmates' contacts were employed with more intensity. I explore this idea with a number of tests.

First, I run three OLS regressions where the dependent variables are dummies indicating whether the individual is employed in primary activities, manufactures, or services.[63] The main explanatory variables are the share of shipmates' contacts employed in primary activities and the share employed in manufactures (services is the ommited category). Regressions also include the set of fixed effects in Equation (2). Table 9 displays the results of this exercise. Notably, results reveal that individuals travelling with shipmates' contacts employed more intensively in some sector, are also more likely to be employed in that sector.

Second, given that New York City was the most popular destination for immigrants, I study to what extent this desicion depended on the place of residence of shipmates' contacts. Figure 10 plots the OLS regression of a dummy variable indicating whether

[61]For instance, Bertrand et al. (2000) use common language to measure links within neighborhoods.

[62]A large number of ships are dropped from the sample because all matched passengers spoke the same language, thus, a number of reasons can explain the lower average effects. First, the remaining ships are larger than the average, with social interactions more difficult to detect or subjected to higher attenuation bias. Second, departures of "multilingual" ships are more concentrated after 1921, where social interactions were less important as shown in Appendix Table A4. Finally, it could be the case that remaining ships covered routes and ports where individuals were less prone to social interaction or less benefited from it.

[63]Based on IPUMS detailed industry classification.

the immigrant remained in New York on the share of shipmates' contacts living in New York (I estimate this non-parametrically for the quintiles of the explanatory variable and controlling for the same set of fixed effects in baseline Equation (2)). The estimated effect is monotonically increasing and significant for the two highest quintiles.

Finally, similar conclusions are obtained using with a more granular definition of sector of employment and place of residence. I show this by estimating the following OLS regression(s):

$$Y_{ij} = \beta S_{ij} + \ \gamma_i \ + \phi_{j(i)} \times \theta_{p(i)} \times \lambda_{w(i)} + \phi_{j(i)} \times \delta_{c(i)} \times \pi_{t(i)} + \epsilon_i \tag{5}$$

where Y_{ij} is an indicator variable that takes one if individual i is employed in sector j (or lives in place j) and zero otherwise, γ_i is an individual fixed effect[64] and $\phi_{j(i)}$ is a sector of empoyment (or place of residence) fixed effect. Consistently with the main identification strategy, $\phi_{j(i)}$ is interacted with the fixed effects in the baseline Equation (2). The main variable of interest is S_{ij}, the share of shipmates' contacts employed in sector j (or living in place j).

Table 10 displays the estimates of Equation (5). In Panel (A), the sector of employment is defined alternatively at one and two digits based on the IPUMS detailed classification. In either case, coefficients are highly significant. An increase of 10 percentage points in the share of shipmates contacts employed in sector j, increases by 0.8% the probability of working in that sector. In Panel B, I use two definitions for the place of residence. Column (1) shows the result for the state of residence and Column (2) for the city of residence.[65] Coefficients have a magnitude comparable to those in Panel A. Finally, Panel (C) displays the estimates of Equation (5) for the Sector of Occupation and the State of Residence with S_{ij} measured separately for shipmates of same and different language. Similar to previous findings, estimates are significantly higher for the characteristics of same-language shipmates.

Correlation in Labor and Residential Outcomes among Shipmates Baseline estimates exploit the variation in predetermined characteristics of shipmates. In this subsection, I follow a different approach by directly looking at labor and residential choices of unrelated shipmates. This exercise is complementary to the previous analysis in two ways. First, it is not affected by measurement issues related to the definition of networks characteristics (e.g. baseline estimates require to measure the earnings of settled immigrants).

[64] Note that each individual enters multiple times in this specification

[65] In the later case I exclude individuals with missing information on the city of residence or living in rural locations.

Second, it can account for social interaction effects, in situations where connections on land are less important for immigrant decisions.

I extend the empirical approach that Bayer et al. (2008) use to identify social interaction effects among neighbors. In this case, I compare the correlation in outcomes among individuals arrived during the same week, conditional on departing from the same port. As already shown in Section 4, predetermined characteristics of shipmates are uncorrelated once we control for the Week X Port interaction. Thus, it is plausible to assume that unobservable determinants of labor and residential outcomes are also uncorrelated. Under this assumption, (conditional) correlation in shipmates' outcomes can be interpreted as the causal effect of travelling together. Naturally, the main limitation of this test is that it does not rule out the presence of common shocks after departure.

More specific, I estimate the following equation using the combination of all possible (non-repeated) pairs of individuals arrived during the same week:

$$Y_{ih} = \beta SameShip + \gamma_i + \gamma_h + \theta_{p(i)} \times \theta_{p(h)} \times \lambda_{w(ih)} + \delta_{d(i)} \times \delta_{d(h)} + \epsilon_{ih} \tag{6}$$

where Y_{ih} is a measure of similarity between the outcomes of (unrelated) individuals i and h. Variables γ_i, γ_h are passenger fixed effects. In order to compare individuals departing from the same port and week, the regression controls for the interaction between $\theta_{p(i)}, \theta_{p(h)}$ (port of departure fixed effects) and $\lambda_{w(ih)}$ (week fixed effect). As suggested above, common shocks experienced during the voyage or upon arrival can create some correlation in individual outcomes even in the absence of social interaction. To alleviate this problem, I control for $\delta_{d(i)} \times \delta_{d(h)}$, the interaction between the fixed effects for the dates of arrival of each passenger in the pair. Although this does not eliminate ship-specific shocks, it controls for any shock affecting passengers arrived during the same day. For instance, some types of jobs could have been advertised only during weekends.

Table 11 displays the estimates of Equation (6) for different outcomes. The dependent variable in Column (1) takes one if the pair of individuals works in the same sector and has the same occupation.[66] Travelling in the same ship, has an effect of 0.15 percentage points which corresponds to a 10% increase in the mean of the dependent variable. In Column (2), the dependent variable measures whether the pair works in the same sector within the same state. In this case, the effects are in the magnitude of 26% over the mean of the dependent variable. Columns (3) an (4) suggest that travelling in the same ship creates some geographical agglomeration. Column (3) shows that travelling in the same ship, is associated with a 3% reduction in the distance between the US residence place of (unrelated) individuals. Column (4) shows that the probability of living in the same city

[66]Occupations and sectors are defined at the most detailed level available in IPUMS created variables.

is 0.2 percentage points higher for unrelated shipmates.

Finally, I estimate the effects of *SameShip* interacted with a number of pair-specific characteristics. Consistent with previous findings in this section, Appendix Table A8 shows that the effects are only driven by pairs of individuals who spoke the same language. Appendix Table A9 explores the idea that pairs of individuals with strong connections upon arrival, should be less affected by brief social interactions. Results are consistent with this interpretation.

7 Conclusions

Although the role of chance encounters with previously unknown people has been long recognized by academics, and more recently by companies and managers, empirical evidence on this subject is largely absent. This paper provides causal evidence that brief social interaction with unknown people has economic relevance, provided they occur during critical life junctures. In particular, I study the effects of interactions among immigrants who met for the first time while travelling to the US during the period 1909-1924. Using a dataset of matched immigrants-ship with detailed geographical information, I have shown that conditional on their town of origin, individuals travelling with (previously unrelated) better connected shipmates, ended up being employed in better quality jobs. I identify this effect using the variation within the same port and week of departure and controlling for the town of origin. A number of tests show that this variation is plausibly exogenous and thus, results are credibly driven by differences in shipmates' characteristics.

A second set of estimations, provides suggestive evidence that the underlying mechanism is related to shipmates providing access and information on potential job opportunities or places of destination within the country. At the same time, heterogeneous baseline results highlight that random social encounters are more important for individuals with lower access to pre-established networks (i.e. contacts with immigrants from the same community and that had settled in the US).

This paper prompts a number of implications beyond the particular setting of the study. First, my results indicate that the benefits of brief social interactions are larger for uninformed individuals or individuals with lower access to stronger forms of networks, like friends or relatives. Second, my results highlight the influence that interactions with unknown people can have in situations where individuals have to make critical decisions and information is scarce. This extends to a large number of settings, for instance, parental choice among schools or students choice of college major. A closely related implication is that economic outcomes among recent waves of refugees to Europe can be affected

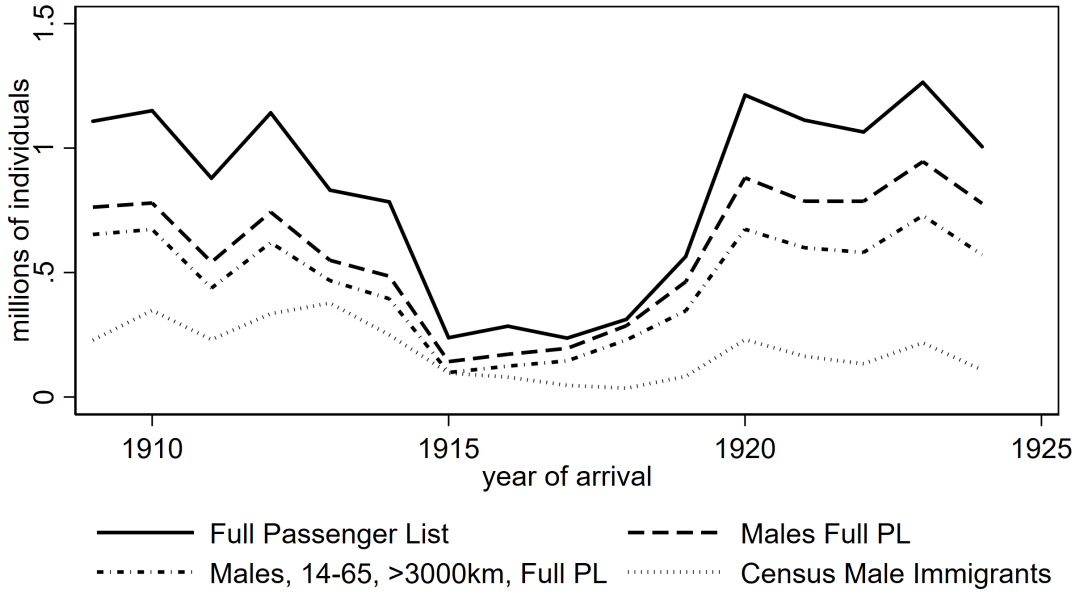
by the characteristics of those who they interact in the days surrounding the voyage (which include boat-mates but also border agents, NGO workers, etc.). More generally, results from this paper illustrates that brief episodes can have long-lasting effects on future earnings.

In recent years, a growing volume of individual level data has become available for researchers. In many cases, information is dispersed across multiple sources and merging across them relies on noisy string variables. Examples vary from historical full count census to recent automatic web generated data. This paper illustrates that incorporating tools from Computer Science can be highly valuable for applied researchers.

Finally, this paper leaves a set of open questions for future research. The extent to which brief social interactions matter in less critical situations can't be answered in the context of this study. Similarly, the setting is not suitable to disentangling between the pure information effect of brief interactions from the direct effect of providing access to better connections or financial support. In this sense, it would be relevant to study settings where individuals meet for a brief period and they never meet again. Finally, despite of recent trends in management practices, the productivity effects of chance encounters within organizations remains largely unexplored.

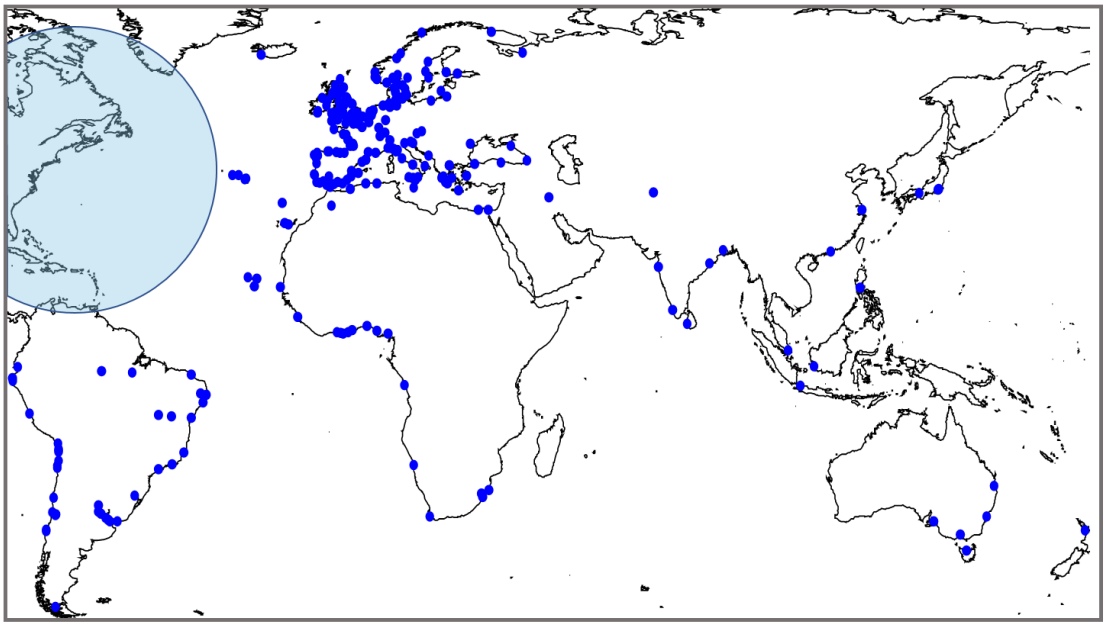
8 FIGURES

Figure 1: Passenger List and Census Data



The graph displays the number of passengers arrived in the period 1909-1924 in the passenger list data and the number of foreign born individuals in the 1920 and 1930 census. For passenger lists, The samples displayed correspond to the full number of individuals departed from any port, the subsample of male individuals and the subsample of males 14-65 years old who departed from ports more than 3000km from New York port. The census figure correspond to male immigrants with country of birth other than Mexico and Canada.

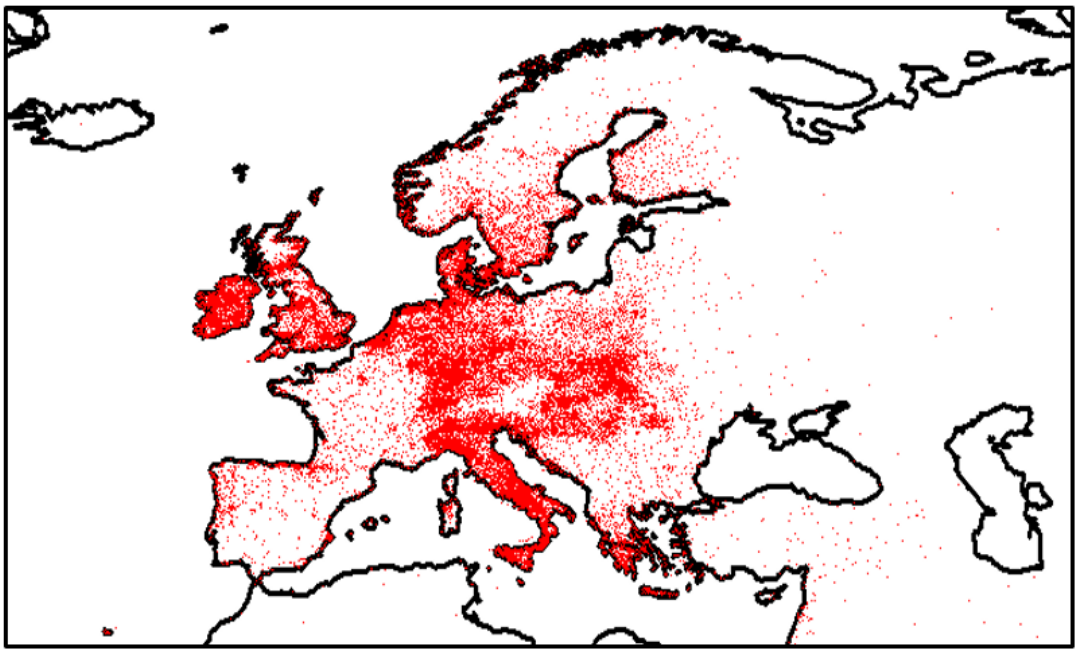
Figure 2: Ports of Departure



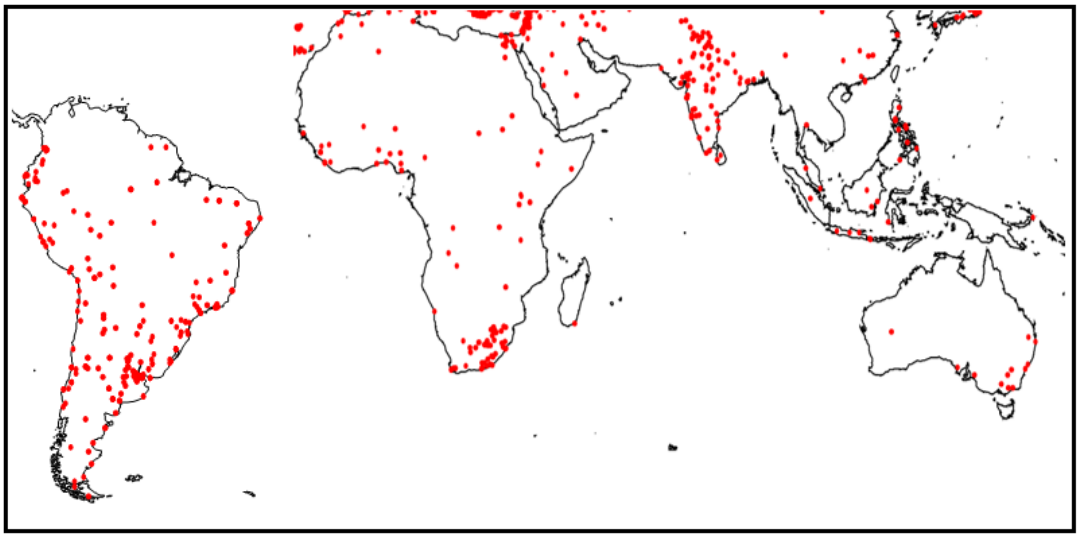
The figure shows all the geolocalized ports of departure for years of arrival 1909-1924. Ports located at less than 10km are displayed as a single port. The shaded area indicates the ports located at a distance below 3000km from New York City.

Figure 3: Places of Origin in Matched Sample

A - Europe

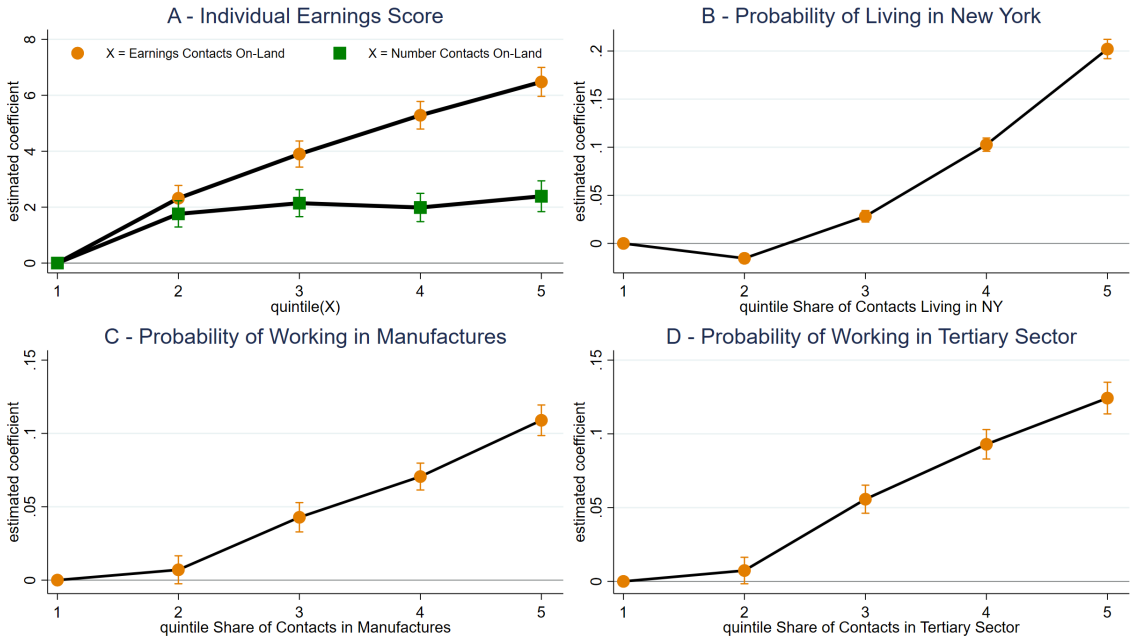


B - Other Regions



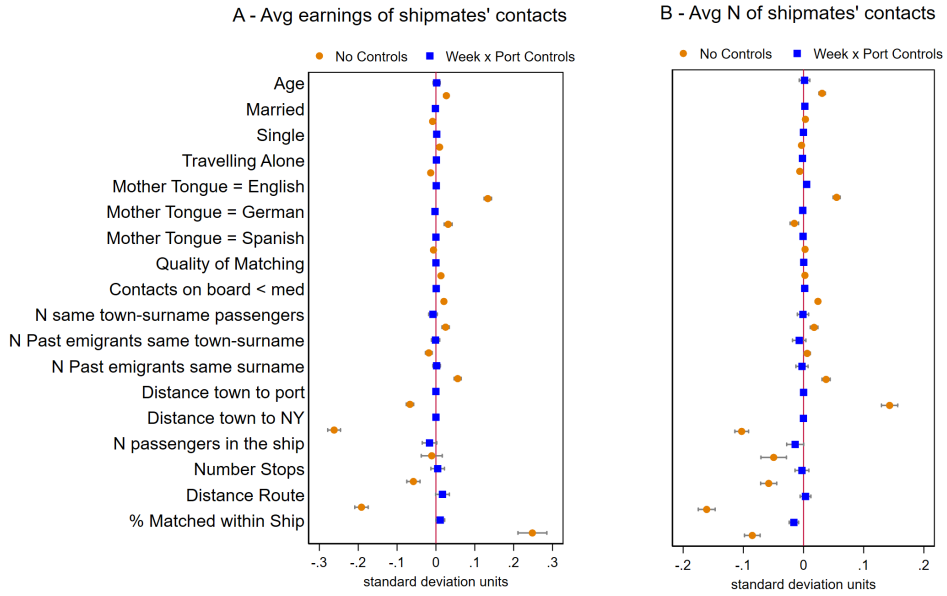
The figure shows all the geolocalized places of origin declared by male passengers in the matched sample for years of arrival 1909-1924. For places identified with precision above the locality level, the map reports the centroid of the administrative unit.

Figure 4: Individual Outcomes and Settled Immigrants from Same Town



Panel A displays the coefficients of OLS regressions of the individual earnings score on the quintiles of the average earnings and the average number of immigrants from the same town of origin. Panel B, displays the coefficients of a regression of a dummy indicating whether the individual lives in New York city as a function of the share of immigrants from the same town of origin settled in New York city. Panels C and D, show the coefficients of a regression for dummies of sector of occupation on the share of immigrants from the same town of origin working in that sector. All regressions control for Ship fixed effects and individual characteristics. Standard errors clustered at the week of arrial level.

Figure 5: Balance of predetermined characteristics

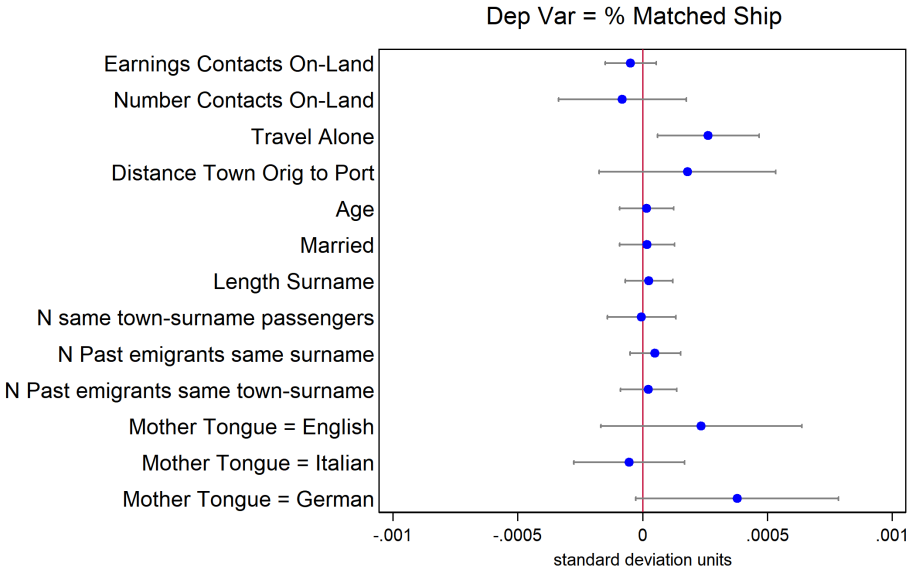


Each panel displays the OLS estimates of regressions of predetermined variables on the average earnigns of (unrelated) shipmates' contacts and on the average number of (unrelated) shipmates' contacts. Each row is a different regression. All regressions control for the characteristics of the individual's contacts. Shipmates' contacts are defined as individuals from the same town who emigrated in the past. Regressions with squared (blue) markers control for fixed effects of the interaction between week of arrival and port of departure and also include fixed effects for large administrative region interacted with port. Standard errors clustered at week of arrival.

Figure 6: Data Matching and Quality of Own Contacts

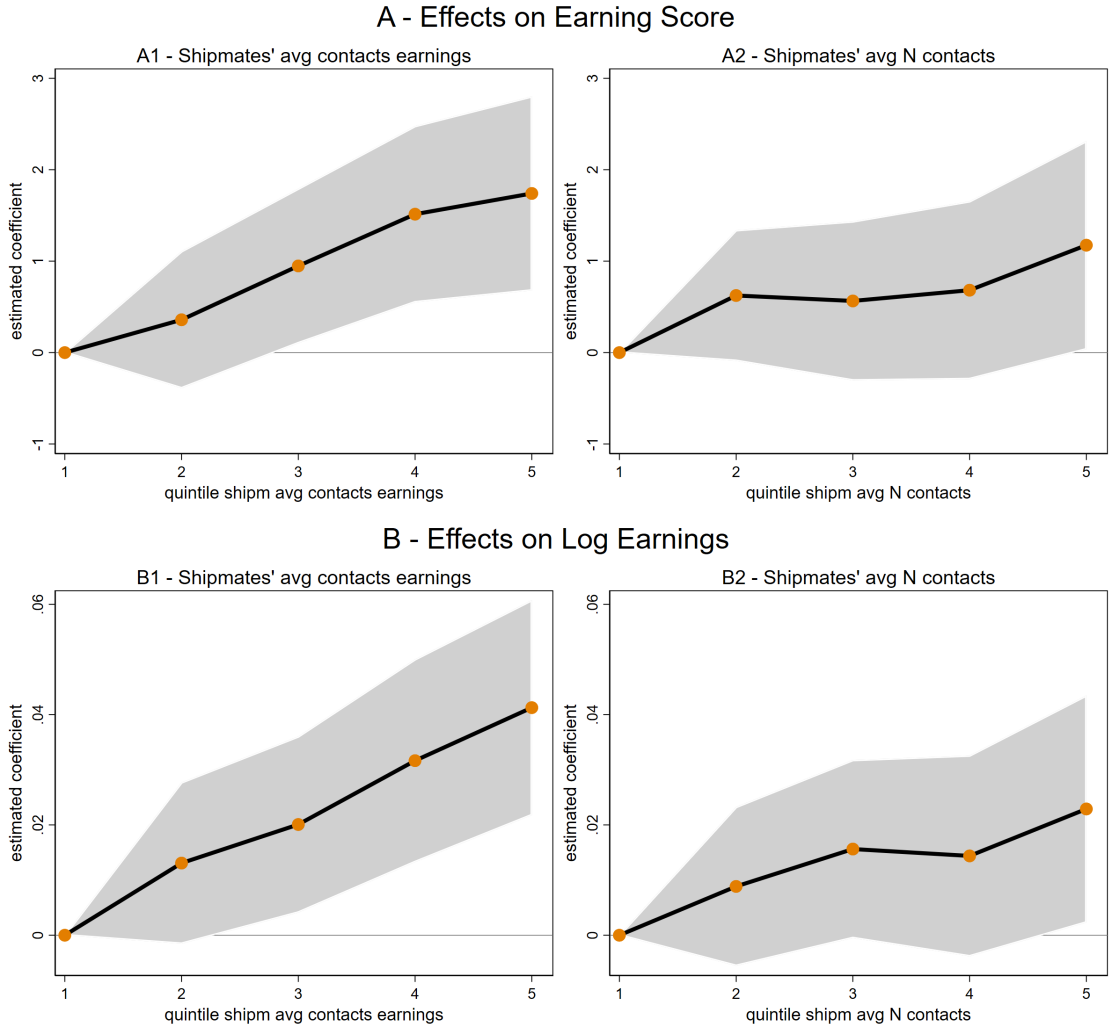
The figure displays the coefficients of an OLS regression where the dependent variable is the % of passengers matched with census within the ship. Circle markers correspond to the quintiles of earnings of individual's contacts on land. Squared markers correspond to the quintiles of the number of individual's contacts on land. Contacts on land are defined as previous emigrants from the same town of origin. Standard errors clustered at week of arrival.

Figure 7: Data Matching and Individual Characteristics



The figure displays the point estimates and confidence intervals of an OLS regression of the percentage of passengers matched within the ship on a set of individual characteristics of the passenger. The regression controls for Port of Departure X Week of Arrival and for the administrative unit of origin. Standard errors clustered at the week of arrival level.

Figure 8: Effect of Shipmates' Connections on Earnings

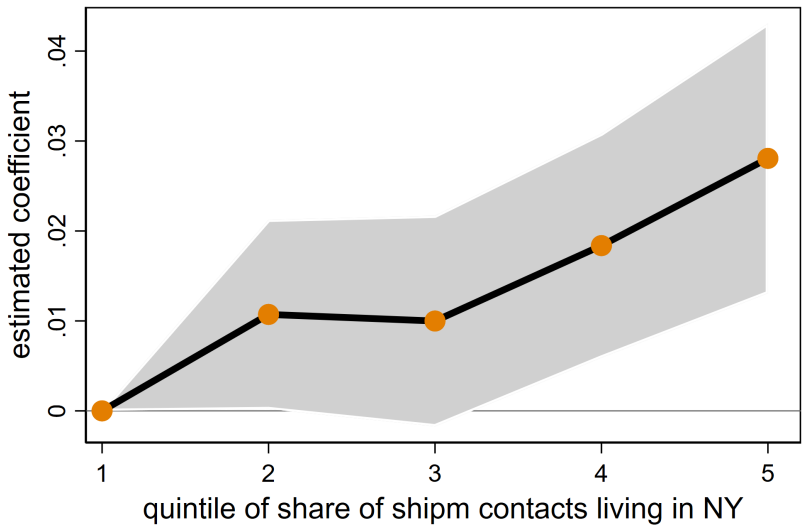


Each panel displays the estimated coefficients of a regression of earnings on the quintiles of shipmates' characteristics. In Panel A, the dependent variable is the occupational earning score and in Panel B the log occupational earnings (1940 based). Sub-panels A1 and B1 shows the effect for the quintiles of the average earnings of (unrelated) shipmates' contacts. Sub-panels A2 and B2 shows the effect of the quintiles of the average number of (unrelated) shipmates' contacts. Shipmates' contacts are defined as individuals from the same town of origin who emigrated in the past. Regressions control for fixed effects of the interaction between the week of arrival and port of departure and the interaction between the town of origin and the semester of arrival. Standard errors clustered at week of arrival.

Figure 9: Effects on Earnings by Time Since Arrival

The figure displays the coefficients of an OLS regression of the individual earnings score on the average characteristics of (unrelated) shipmates contacts interacted with dummies for the number of years since arrival to the country. Regressions control for the Week x Port of Departure, Place of origin X Census Year linear trends and the interaction between the age of arrival and the characteristics of shipmates contacts. Robust standard errors clustered at the week of arrival level.

Figure 10: Probability of staying in NY as a function of shipmates' contacts residing in NY



The figure displays the OLS estimation of the probability of residing in New York city as a function of the quintiles of the share of (unrelated) shipmates'contacts living in New York city. Regressions also controls for quintiles of individuals contacts living in New York city, the number of contacts of the individual and of his (unrelated) shipmates. Regressions include the baseline controls mentioned in the text. Robust standard errors clustered at the week of arrival level.

9 TABLES

Table 1: Descriptive Statistics

Panel A	**Full Sample**	**Reg Sample**[1]		
N Male Individuals Full Passenger List	9,297,026	4,716,934		
N Male Immigrants Census 1920-1930	2,836,404	2,469,503		
N Matched Individuals	351,289	206,383		
N of Ships Matched Sample	34,091	14,910		
N of Vessels Matched Sample	5,138	1,152		
N Ports Matched Sample	422	166		
N Routes Matched Sample	865	454		
N Places of Origin Matched Sample[2]	10,909	8,250		
Panel B	**Avg**	**Std**	**Min**	**Max**
Min Linear Distance Travelled (thousands of km) [3]	6.5	1.2	3	31
Estimated Days Full Voyage at 15 Knots Speed	9.7	1.9	4.6	46.5
Distance Town to Port of Departure	526.6	913.1	0	19214
Passengers per Ship in Passenger List [4]	173	303.2	1	3749
Passengers per Ship in Matched Sample [4]	20.1	23.2	1	262
Past Emigration from Same Place (thousands)	9.3	22.7	0	168
Earnings of Past Emigrant from Same Place	49.7	11.6	.6	100
Avg N of Potential Contacts of Shipmates (thousands) [5]	6.2	9.6	0	168
Avg Earnings of Potential Contacts of Shipmates [5]	49.8	6.4	3.1	100
N of Different Places of Origin in the Ship	14.9	17	1	178
Age at arrival	23	10.4	0	68
Married at arrival	.29	.45	0	1
Share Travelling Alone [6]	.74	.44	0	1
Share Living in Urban Places at Destination	.82	.38	0	1
Share Individuals Staying in New York City	.21	.4	0	1
Average N of Ships in Week X Port	2.8	1.8	0	15

[1] The regression sample includes individuals 14-65 years old. For the case of Passenger List information, it only includes ships departing from ports more than 3000km away from New York port and without missing information on the place of origin. [2] Places of origin with at least two matched individuals during one semester in the regression sample. [3] The Minimum Linear Distance of the voyage is estimated as the sum of the straigth distance between subsequential ports identified within the route, sorted by their proximity to New York port. [4] Only individuals in the regression sample. [5] Potential contacts are defined as past emigration from the same town or place of origin. [6] Individuals travelling alone are defined as those without any other passenger in the ship with same place of origin and surname.

Table 2: Correlation of Characteristics within Ship

	(1) Unconditional Correlation	**(2) Conditional on Week x Port**
Age	0.079***	-0.007
Married	0.088***	0.003
Single	0.094***	0.003
Travelling alone	0.075***	0.001
Mother tongue = English	0.642***	0.004
Mother tongue = German	0.488***	0.002
Mother tongue = Spanish	0.462***	-0.011
Quality of machting	0.110***	0.008
N same town passengers	0.079***	0.031*
N same town-surname passengers	0.085***	0.006
N Past emigrants same town-surname	0.008	-0.004
N Past emigrants same surname	0.114***	-0.006
N Past emigrants same town	-0.014***	-0.015
Avg earnings of land contacts	0.176***	-0.010
Distance town to port	0.165***	-0.004
Distance town to NY	0.701***	0.000

The table displays unbiased estimates of the correlation between individual and average shipmates' characteristics, excluding those residing in the same place or with similar surname. Unbiased estimations are obtained by sampling one random passenger per ship (see Bayer et al.(2008)). Column (2) controls for Week of arrival X Port of Departure and Adm Region X Port. Sample of 14-65 males not residing in the US before departure. Bootstrapped significance levels.

Table 3: Probability of Matching Passenger List - Census

	Dep Var = Passenger Matched with Census	
	(1) No Controls	(2) Week X Port
F-Stat Joint Significance of Ship FE	5.92	0.60
p-value	0.00	1.00
N Individuals	5008017	4996193

The table reports the joint significance F-statistic for the Ship Fixed Effects, in a regression where the dependent variable is a dummy for whether the passenger is matched in the census. The sample is the full passenger list for non-american citizens in the age group 14-65.

Table 4: Effect of Shipmates' Connections on Earnings and Job Quality

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Average Contacts Earnings	0.14***	0.08***	0.11***	0.27***
	(0.03)	(0.02)	(0.03)	(0.06)
Number of Contacts	0.05**	0.04**	0.05***	0.07**
	(0.02)	(0.02)	(0.02)	(0.04)
Mean DepVar	50.89	23.24	44.36	881.88
N individuals	97395	97818	97395	96484
R2	.338	.359	.368	.384
F excl	12.2	9.9	11.4	12.9

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. † Coefficients multiplied by 100 in this column. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. In all regressions, the first reported explanatory variable is the average earnings score of the potential contacts of (unrelated) shipmates. The second explanatory variable is the average number of potential contacts of (unrelated) shipmates. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table 5: Additional Controls

	Depvar = Earnings Score								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Avg. Earnings	0.14***	0.14***	0.14***	0.09**	0.11**	0.16***	0.15***	0.18***	0.21***
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)	(0.05)
N of contacts	0.05**	0.04**	0.04**	0.07**	0.02	0.04*	0.04*	0.07**	0.07**
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)	(0.03)	(0.03)
N individuals	97395	95115	95115	67765	74775	95096	95069	78127	77952
R2	.338	.342	.342	.41	.394	.346	.348	.467	.488
Baseline Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓
Individual Controls	-	✓	✓	✓	✓	✓	✓	✓	✓
Ship-Trip Controls	-	-	✓	✓	✓	✓	✓	✓	✓
Town Origin X Month Arriv	-	-	-	✓	-	-	-	-	-
Week X Port X Admin Reg	-	-	-	-	✓	-	-	-	-
Vessel FE	-	-	-	-	-	✓	✓	✓	✓
Route FE	-	-	-	-	-	-	✓	✓	✓
Surname Nysiis FE	-	-	-	-	-	-	-	✓	✓
Date of Arrival FE	-	-	-	-	-	-	-	-	✓

Each column displays estimates of OLS regressions of the earnings score on the the average earnings of (unrelated) shipmates potential contacts and their average number of contacts (in thousands). The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. Potential contacts are defined as past emigrants from the same town of origin. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. Baseline controls include fixed effects for Week of Arrival X Port of Departure and Place of Origin X Semester. Individual controls are age, marital status, indicators for individuals travelling with relatives, white ethnicity and english native tongue. Ship-Route controls are the number of passengers in the ship, max capacity of the ship, number of stops, total distance, number of trips made by the ship, average number of stops of the ship, days since the last trip observed in the sample, share of married passengers, share of male passengers and number of US resident passengers. Surnames fixed effects are based on the NYSIIS codification system. Column (4), include interactions between the week of arrival, port of departure and the administrative region of the place of origin. Robust standard errors clustered at week of arrival level.

Table 6: Investigating Potential Contacts Before Travelling

	(1) Shipmates contacts avg earnings	(2) Shipmates avg Number of contacts
Baseline	0.14***	0.05**
	(0.03)	(0.02)
$\|ID_i - ID_j\| > 10$	0.12***	0.04**
	(0.03)	(0.02)
$\|ID_i - ID_j\| > 15$	0.12***	0.04**
	(0.03)	(0.02)
$Dist(Town_i, Town_j) > 100km$	0.09***	0.04***
	(0.03)	(0.02)

Each row in the table reports the coefficients of a different OLS regression of the earnings score on the shipmates contacts characteristics. Each column variable is a different explanatory variable of the same regression. The second and third row exclude any shipmate j with ID number difference below 10 and 15 respectively. The last row excludes any shipmate j with ID number difference below 15 and with town of origin located at less than 100km of individual's town of origin. All regressions control for the characteristics of individual own contacts. All regressions include the baseline controls described in the text and fixed effects fo rthe interaction between port of departure and week, and the interaction between port of departure, administrative area of residence and year-semester. Robust standard errors clustered at week of arrival level.

Table 7: Estimated Effects by Individual's Connections On-Board and On-Land

	Depvar = Earnings Score		
	(1)	(2)	(3)
Definition of Low Connections:	No Contacts On Board (Same Town or Surname)	Quality of Potential Contacts on Land	No Contacts On Board + Quality of Land Contacts
Shipmates Contacts Earnings x Low Connections	0.24***	0.19***	0.34***
	(0.06)	(0.06)	(0.07)
Shipmates Contacts Earnings x High Connections	0.15***	0.11***	0.12***
	(0.04)	(0.03)	(0.03)
Shipmates N of Contacts x Low Connections	0.07*	0.10**	0.13**
	(0.04)	(0.05)	(0.05)
Shipmates N of Contacts x High Connections	0.07**	0.05**	0.05**
	0.03	0.02	0.02

Each column shows the coefficients of a different OLS regression of the earnings score on the average earnings of contacts and on the number of contacts of (unrelated) shipmates' interacted with a dummy variable indicating the quality of connections of the individual. In Column (1), an individual is defined as low connected if he is travelling without any person of same surname from the same place of origin. In Column (2) an individual is defined as low connected if the number of persons from the same place in the ship is below the median and if the average earnings of past emigrants from same place is below the median. Column 3 defines an individual as low connected if the number of emigrants and the average earnings of past emigrants from the same palce of origin is below the median and if there is no other passenger from the same place of origin in the ship. Surname similarity is defined based on nysiis phonetic coding. All regressions include fixed effects of Week X Port of Departure, Place of Origin X Semester of Arrival, indicators for the route and a dummy variable indicating if the individual is high or low connected according to the definition in the column. Column (1) includes fixed effects for each nysiis surname category. Standard errors clustered at the week of arrival level.

Table 8: Effects by Language of Shipmates

	(1) Earnings Score	(2) Log Earns†
Average Earnings of **Similar Language** Shipmates' Contacts	0.05**	0.14***
	(0.02)	(0.04)
Average Earnings of **Different Language** Shipmates' Contacts	0.03	0.06
	(0.03)	(0.06)
Average Number of **Similar Language** Shipmates' Contacts	0.01	0.03
	(0.01)	(0.02)
Average Number of **Different Language** Shipmates' Contacts	-0.01	-0.02
	(0.01)	(0.02)

Each column of the Table displays estimates of an OLS regression of a measure of individual earnings on the average characteristics of (unrelated) shipmates contacts. The main explanatory variables are calculated separately for shipmates who spoke similar and different mother tongue. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Column (2) the dependent variable is the (log) median earnings of the occupation in 1940.The sample includes all (male 14-65) matched passengers in the period 1909-1924 with at least one shipmate speaking a different mother tongue. Mother tongue definition is constructed based on IPUMS categories. † Coefficients multiplied by 100 in this column. Regressions also control for baseline controls as defined in the text. The number of observations in the regressions is 62,890. Standard errors clustered at the week of arrival level.

Table 9: Effects on Sector of Employment

Shipmates Characteristics	(1) Primary Sector	(2) Manufactures Sector	(3) Services Sector
Share of Contacts in Primary Sector	0.08**	-0.010	-0.07*
	(0.03)	(0.04)	(0.04)
Share of Contacts in Manufactures	0.01	0.07*	-0.08**
	(0.03)	(0.04)	(0.04)
Mean DepVar	0.3	0.4	0.4
N individuals	83459	83459	83459

This table displays estimates of OLS regressions of the sector of employment of the individual on the share of (unrelated) shipmates contacts employed in each sector. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable a dummy indicating whether the individual is employed in agriculture and other primary activites. In Column (2) the dependent variable a dummy indicating whether the individual is employed in the manufacturing sector. In Column (3) the dependent variable a dummy indicating whether the individual is employed in services or public sector. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions control for the share of individual contacts in each sector, the number of contacts of the individual, the average number of contacts of his shipmates, the average earnings of the shipmates contacts, the average earnings of his own contacts and indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table 10: Shipmates Effects on Sectors of Occupation and Place of Residence

Panel A: Sector of Occupation of Individual

	(1) Sector 1 digit	(2) Sector 2 digits
Share of Shipmates Contacts Working in the Same Sector	0.079***	0.076***
	(0.014)	(0.008)

Panel B: Place of Residence of Individual

	(1) State of Residence	(2) City of Residence
Share of Shipmates Contacts Living in Destination Place	0.084***	0.073***
	(0.009)	(0.010)

Panel C: By Language of Shipmates

Share Contacts Working/Living in Same Sector/State:	(1) Sector of Occup (1d)	(2) State of Residence
Shipmates of **Similar Language**	0.078***	0.086***
	(0.012)	(0.007)
Shipmates of **Different Language**	0.015	0.016**
	(0.012)	(0.008)

Panel A displays the coefficients of an OLS regression of $Y_{ij}(t)$, a dummy that takes one if individual i works in sector j, on X_{ij}^{SM}, the share of (unrelated) shipmates contacts working in sector j. Regressions include individual fixed effects, fixed effects of the interaction between sector of occupation, week and port of departure and fixed effects of the interaction between sector of occupation, administrative region of origin and year of arrival. Regressions also control for the share of individual contacts working in sector j. In Column (1) sector of occupation is defined at 1 digit and in Column (2) at 2 digits, in both cases based on the 3 digits classification created by IPUMS. Panel B displays the coefficients of an OLS regression of $Y_{ic}(t)$, a dummy that takes one if individual i lives in place c, on X_{ic}^{SM}, the share of (unrelated) shipmates contacts residing in place c. Regressions include individual fixed effects, fixed effects of the interaction between the place of residence, week and port of departure and fixed effects of the interaction between place of residence, administrative region of origin and year of arrival. Regressions also control for the share of individual contacts living in place c. In Column (1) the place of residence is defined as the state of residence. In Column (2) the place of residence is based on 85 cities with the highest share of individuals from the sample, excluding those residing in non-classified cities or small rural areas. In Panel C, the share of shipmates contacts working in different sectors or living in different states are calculated separately for shipmates with similar and different mother tongue. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. The number of observations are 712,440 for Panel A Column(1), 5,303,720 for Panel A Column(2), 7,563,689 for Panel B Column(1), 8,971,750 for Panel B Column(2), 464,445 for Panel C Column (1) and 5,110,210 for Panel C Column (2). Standard errors clustered at the week of arrival level.

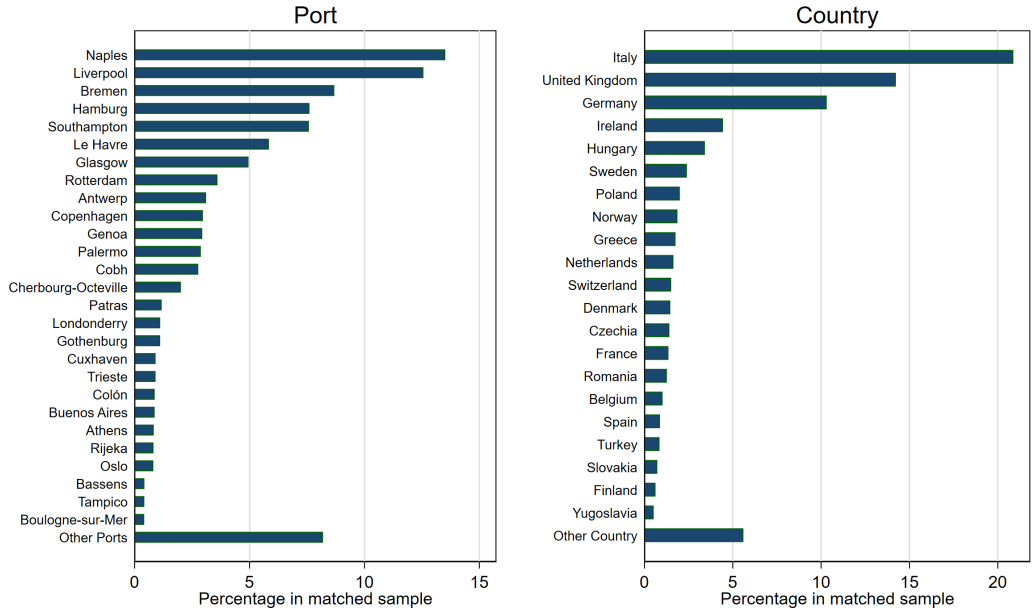
Table 11: Correlation in Labor and Spatial Outcomes of Shipmates

	(1) Work Same Ind x Occ	(2) Work Same Ind x State	(3) Log Dist Residence	(4) Live Same City
Same Ship	0.15***	0.09***	-3.15***	0.20**
	(0.04)	(0.03)	(0.60)	(0.09)
% Effect Over the Mean	9.4	11.09	-3.15	2.18
N individuals	134974	134974	193551	137602
N observations	18556160	18556160	37425892	19775703

All coefficients are multiplied by 100. Table displays the OLS regressions of individual-pair level outcomes on a dummy variable indicating whether the pair travelled in the same ship. The sample consists of all matched male passengers arrived during the period 1909-1921 grouped into non repeated pairs of individuals who arrived during the same week. The sample only include pairs of individuals with different surname (defined as Jaro-Winkler distance above 0.1) and from different places of origin. In Column (1) the dependent variable is a dummy for whether the pair works in the same occupation and industry. In Column (2) the dependent variable is a dummy for whether the pair works in the same industry and lives in the same state . In Column (3) the dependent variable is a measure of the log distance between the county of residence of each individual in the pair. In Column (4) the dependent variable indicates whether the pair lives in the same city. Regressions include indicators for each individual in the pair, fixed effects of Week of Arrival X Port Departure(i) X Port Departure(j) and fixed effect of Date Arrival(i) X Date Arrival(j) where i and j index individuals in the pair. Standard Errors clustered at week of arrival level.

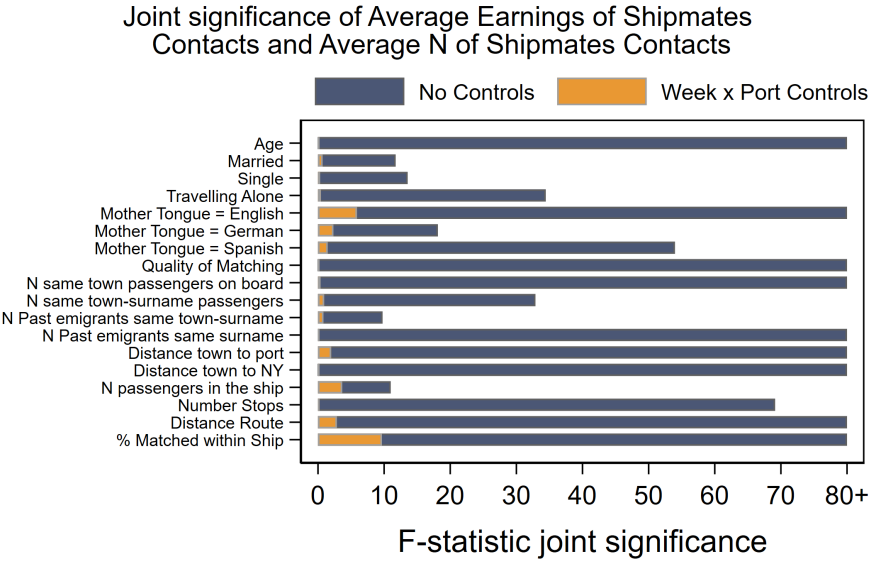
APPENDIX A: Additional Tables and Figures

Figure A1: Main Ports of Departure and Countries of Origin



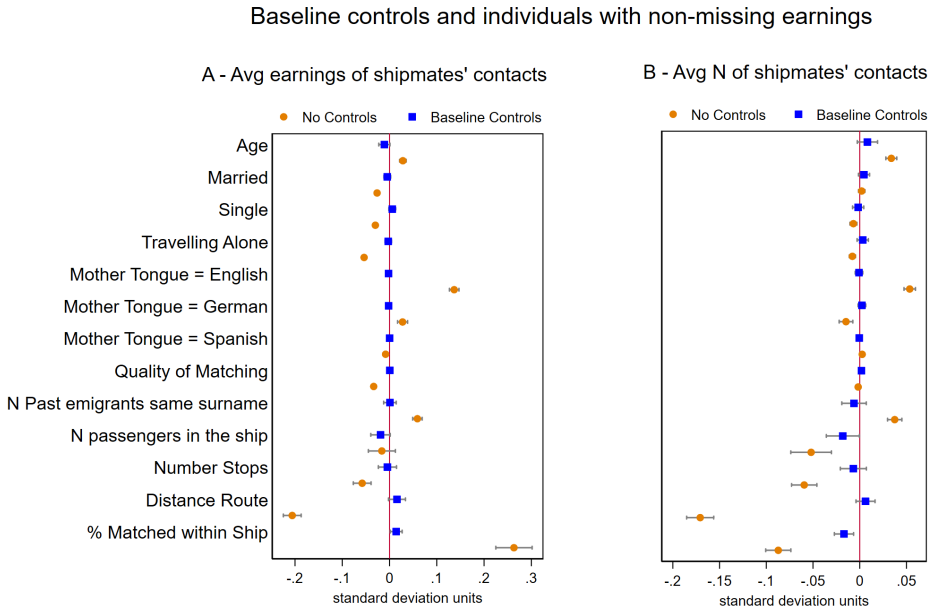
Data include all matched passengers during in 1909-1924 who are not US citizens. Data exclude arrivals from ports located at less than 3000km from the port of New York.

Figure A2: Balance Regressions



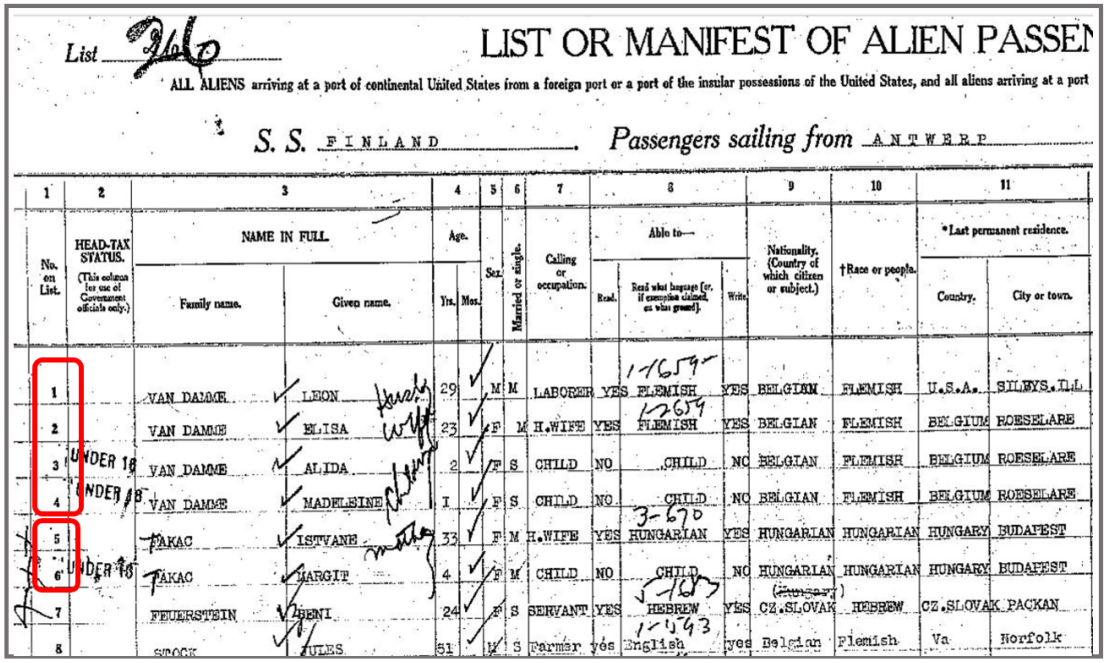
Each panel displays the joint F statistic of the OLS estimates of regressions of predetermined variables on the average earnigns of (unrelated) shipmates' contacts and on the average number of (unrelated) shipmates' contacts. Each row is a different regression All regressions control for the characteristics of the individual's contacts. Shipmates' contacts are defined as individuals from the same town who emigrated in the past. The regression with additional controls include fixed effects for the interaction between week of arrival, port of departure and fixed effects for large administrative region interacted with port. Standard errors clustered at week of arrival level.

Figure A3: Balance of predetermined characteristics



Each panel displays the OLS estimates of regressions of predetermined variables on the average earnigns of (unrelated) shipmates' contacts and the average number of (unrelated) shipmates' contacts. Each row is a different regression. All regressions control for the characteristics of the individual's contacts. Shipmates' contacts are defined as individuals from the same town who emigrated in the past. In panel B, regressions control for fixed effects of the interaction between week of arrival and port of departure and also include fixed effects for the interaction between town of origin and semester of arrival. Standard errors clustered at week of arrival.

Figure A4: Passengers Travelling Together



List 7160

LIST OR MANIFEST OF ALIEN PASSEN

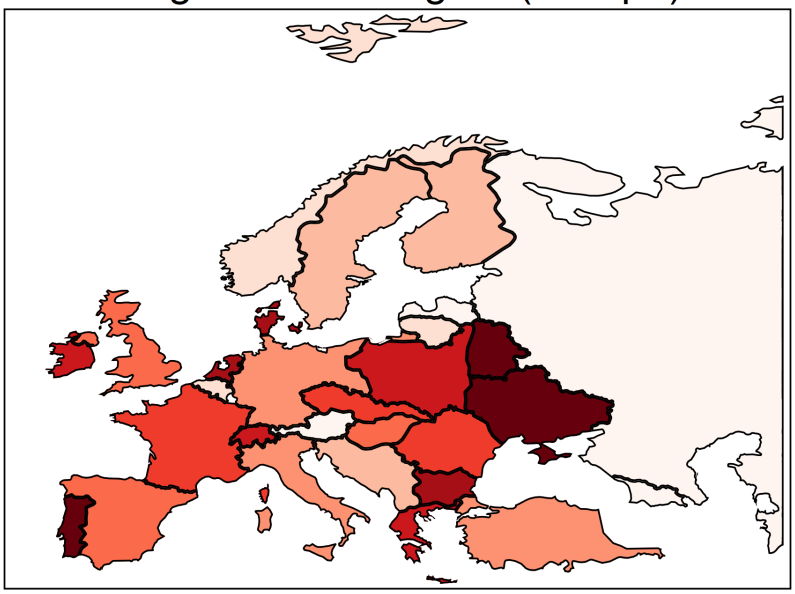
ALL ALIENS arriving at a port of continental United States from a foreign port or a port of the insular possessions of the United States, and all aliens arriving at a port

S. S. FINLAND. Passengers sailing from ANTWERP

1	2	3		4		5	6	7	8			9	10	11	
No. on List.	HEAD-TAX STATUS. (This column for use of Government officials only.)	NAME IN FULL. Family name.	Given name.	Age. Yrs.	Mos.	Sex.	Married or single.	Calling or occupation.	Able to— Read.	Read what language (or, if exemption claimed, on what ground).	Write.	Nationality. (Country of which citizen or subject.)	†Race or people.	*Last permanent residence. Country.	City or town.
1		VAN DAMME	LEON	29		M	M	LABORER	YES	FLEMISH 17659	YES	BELGIAN	FLEMISH	U.S.A.	SILVYS.ILL
2		VAN DAMME	ELISA	23		F	M	H.WIFE	YES	FLEMISH 17659	YES	BELGIAN	FLEMISH	BELGIUM	ROESELARE
3	UNDER 16	VAN DAMME	ALIDA	2		F	S	CHILD	NO	CHILD	NO	BELGIAN	FLEMISH	BELGIUM	ROESELARE
4	UNDER 16	VAN DAMME	MADELEINE	1		F	S	CHILD	NO	CHILD 3-670	NO	BELGIAN	FLEMISH	BELGIUM	ROESELARE
5		FAKAC	ISTVANE	33		F	M	H.WIFE	YES	HUNGARIAN	YES	HUNGARIAN	HUNGARIAN	HUNGARY	BUDAPEST
6	UNDER 16	FAKAC	MARGIT	4		F	M	CHILD	NO	CHILD	NO	HUNGARIAN	HUNGARIAN	HUNGARY	BUDAPEST
7		FEUERSTEIN	BENI	24		F	S	SERVANT	YES	HEBREW 17543	YES	CZ.SLOVAK	HEBREW	CZ.SLOVAK	PACKAN
8		STOCK	JULES	51		M	S	Farmer	yes	English	yes	Belgian	Flemish	Va	Norfolk

Source: The Statue of Liberty - Ellis Island Foundation

Figure A5: Heterogeneous Effects by Country of Origin of Passengers (Europe)



The figure displays the OLS estimation of the effects of shimates' contacts earnings interacted with dummies for countries of origin of the individual. Regressions control for dummies of country of origin, week of arrival and port of departure interactions fixed effect and town of origin fixed effects interacted with the year of arrival. The plotted coefficients are normalized in the scale zero to one. Effects are significant at 10% for the following list of countries/regions: Belarus, Czechoslovakia, Denmark, Germany, Greece, Hungary, Ireland, Italy, Netherlands Poland, Portugal, Switzerland, UK, Ukraine.

Table A1: Alternative Measures of Earnings Based on 1950 Census

Shipmates Characteristics	(1) Log Earnings Occupation 1940	(2) Log Earnings Occupation 1950	(3) Log Earnings Percentile 1950
Average Contacts Earnings	0.27***	0.25***	0.28***
	(0.06)	(0.06)	(0.08)
Number of Contacts	0.07**	0.07*	0.08
	(0.04)	(0.04)	(0.05)
Mean DepVar	881.88	2219.54	2235.79
N individuals	96484	97395	97395

This table displays estimates of OLS regressions of different measures of earnings on the average characteristics of shipmates contacts. All coefficients are multiplied by 100. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the (log) median earnings of the occupation in 1940. Column (2) is similar to Column (1) but using 1950 1% census sample. In column (3) the dependent variable is the (log) median earnings of the percentile ranking of the occupation in 1950. In all regressions, the first reported explanatory variable is the average earnings score of the potential contacts of (unrelated) shipmates. The second explanatory variable is the average number of potential contacts of (unrelated) shipmates. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table A2: Alternative Clustering of Standard Errors

	Clustering Level				
	(1) Week of Arrival	(2) Month of Arrival	(3) Ship x Port Depart	(4) Region of Origin	(5) Multicluster Week-Ship
Avg. Earnings	0.14***	0.14***	0.14***	0.14***	0.14***
	(0.031)	(0.037)	(0.028)	(0.032)	(0.031)
N of contacts	0.05**	0.05**	0.05**	0.05**	0.05**
	(0.020)	(0.020)	(0.019)	(0.019)	(0.020)
N individuals	97391	97391	97391	97391	97391

Each column shows the coefficients of a different regression for alternative levels of clustering in standard errors. The dependent variable is the earnings score. The reported explanatory variables are the average earnings of (unrelated) shipmates potential contacts and their average number of contacts (in thousands). Potential contacts are defined as past emigrants from the same town of origin. All regressions control for the characteristics of individual own contacts. Baseline controls include interaction for week of arrival and port of departure and the interaction between administrative region of origin and port of departure.

Table A3: Effects by Interactions Between Variables of Interest

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
(Contacts' Earnings High) x (N Contacts High)	1.73***	1.23***	1.74***	3.91***
	(0.48)	(0.36)	(0.43)	(0.90)
(Contacts' Earnings High) x (N Contacts Low)	1.42***	0.93***	1.26***	2.51***
	(0.47)	(0.35)	(0.43)	(0.89)
(Contacts' Earnings Low) x (N Contacts High)	0.70	0.64**	0.70*	1.63**
	(0.43)	(0.30)	(0.38)	(0.79)
N individuals	97395	97818	97395	96484
R2	.338	.358	.367	.383

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. Each column shows the coefficients for the interaction between two set of dummies. The first set of dummies indicates whether the shipmates connections earnings are above or below the median of its distribution and the second set of dummies indicates whether the shipmates number of connections is above or below the median of its distribution. The omitted category is shipmates below the median of contacts earnings and contacts number. † Coefficients multiplied by 100 in this column. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. Standard errors are clustered at the Week of Arrival level.

Table A4: Effects Before and After the 1921 Emergency Quota Act

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Avg Contacts Earnings x Pre-Quota	0.15***	0.09***	0.13***	0.30***
	(0.03)	(0.02)	(0.03)	(0.07)
Avg Contacts Earnings x Post-Quota	0.07	0.03	0.03	0.12
	(0.07)	(0.05)	(0.06)	(0.12)
Number of Contacts x Pre-Quota	0.05**	0.04**	0.05**	0.08*
	(0.02)	(0.02)	(0.02)	(0.04)
Number of Contacts x Post-Quota	0.04	0.04	0.05	0.04
	(0.04)	(0.03)	(0.03)	(0.07)
N individuals	97391	97814	97391	96480

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. In all regressions, the reported explanatory variables are the average earnings score of the potential contacts of (unrelated) shipmates and the average number of them, in both cases interacted with a dummy indicating whether the individual emigrated before or after the introduction of the 1921 Emergency Quota Act. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. † Coefficients multiplied by 100 in this column. Standard errors are clustered at the Week of Arrival level.

Table A5: Alternative Definitions of Contacts on Land

Shipmates Characteristics	(1) Baseline Definition (Same Town)	(2) Same Admin Region and Similar Surname	(3) Same Town and Similar Surname
Avg Contacts Earnings	1.25***	2.77***	2.78***
	(0.14)	(0.32)	(0.47)
Number of Contacts	0.38**	0.72**	0.87*
	(0.15)	(0.35)	(0.47)
N observations	130684	35552	17297

Each column shows the coefficients of a different regression of the individual earning score on the average earnings and number of potential contacts of (unrelated) shipmates'. Explanatory variables are standardized with zero mean and standard deviation one in every regression. Each column corresponds to a different definition of potential contacts residing in the US. In Column (1), potential contacts are defined as past emigrants from the same town of origin. In Column (2), potential contacts are defined as past emigration from same administrative area of origin and with similar surname. In Column (3), potential contacts are defined as past emigrants from the same town of origin and with similar surname. Surname similarity is based on nysiis phonetic coding. All regressions control for fixed effects for Week of Arrival X Port of Departure and fixed effects for the group at which potential contacts are defined interacted with census year. Standard errors clustered at the week of arrival level.

Table A6: Alternative Identification Strategies

	(1)	(2)	(3)
		Different Stops of Same Ship	
	Repeated Trips of Same Vessel	Any Port in the Route	Only First Port of Departure
Boarding at the Same Port:			
Average Earnings of	0.08***	0.11***	-
Shipmates' Contacts	(0.02)	(0.04)	-
Average Number of	0.03*	0.02	-
Shipmates' Contacts	(0.02)	(0.02)	-
Boarding at a Different Port:			
Average Earnings of	-	0.07**	0.07*
Shipmates' Contacts	-	(0.03)	(0.04)
Average Number of	-	-0.01	0.01
Shipmates' Contacts	-	(0.02)	(0.02)
N observations	93305	48463	23791
Vessel × Port × Year Arriv	✓	-	-
Place of Origin × Semester	✓	✓	✓
Route × Semester	✓	✓	✓
Vessel FE	-	✓	✓

This table displays estimates of OLS regressions of the occupational earnings score on the average characteristics of shipmates contacts. Each column is a different regression. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. All regressions control for the characteristics of own contacts, the number of passengers and the days elapsed since the previous trip of the vessel. Column (1) only includes vessels with at least two trips during the year. The characteristics of the shipmates in rows are the average earnings of contacts in land and the average number of contacts in land, calculated separately for shipmates boarding the ship at the same port and at different ports of the same route. Columns (2) and (3) exclude any ship with more than 90% of total passage boarding in the first port. Column (3) only includes passengers boarding in the first port of the route. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. Standard errors are clustered at the Week of Arrival level.

Table A7: Subsample of Places of Origin Geolocalized with High Precision

Shipmates Characteristics	(1) Earnings Score	(2) Duncan Index	(3) NPB Index	(4) Log Earns Occ1940†
Average Contacts Earnings	0.25***	0.12***	0.20***	0.46***
	(0.05)	(0.04)	(0.04)	(0.14)
Number of Contacts	0.04	0.05**	0.05**	0.08
	(0.03)	(0.02)	(0.03)	(0.08)
Mean DepVar	52.04	24.09	45.53	714.03
N individuals	70925	71257	70925	70295

This table displays estimates of OLS regressions of different measures of earnings and job quality on the average characteristics of shipmates contacts. The sample includes all male passengers arrived in the period 1909-1924 matched with census years 1920-1930 and with non-missing information on occupation. The sample is restricted to those individuals for whom the town of origin is geocoded with locality or sublocality precision level. In Column (1) the dependent variable is the occupational earnings score created by IPUMS. In Columns (2) and (3), the dependent variable is the Duncan Socioeconomic Index and the Nam-Power-Boyd Index. In Column (4) the dependent variable is the (log) median earnings of the occupation in 1940. In all regressions, the first reported explanatory variable is the average earnings score of the potential contacts of (unrelated) shipmates. The second explanatory variable is the average number of potential contacts of (unrelated) shipmates. Potential contacts are defined as past emigration from the same town of origin. For every individual, the pool of unrelated shipmates excludes any passenger with same town of origin or with similar surname. Surnames are defined as similar when the Jaro-Winkler distance is below 0.1. All regressions include indicators for Week of Arrival X Port of Departure and Place of Origin X Semester. † Coefficients multiplied by 100 in this column. Standard errors are clustered at the Week of Arrival level.

Table A8: Correlation in Outcomes by Spoken Language

	(1) Work Same Ind x Occ	(2) Work Same Ind x State	(3) Log Dist Residence	(4) Live Same City
SameShip x Same Lang	0.20***	0.19***	-6.63***	0.52***
	(0.05)	(0.04)	(0.77)	(0.10)
SameShip x Diff Lang	0.03	-0.05	2.13***	-0.36***
	(0.05)	(0.04)	(0.70)	(0.12)
N individuals	134974	134974	193551	137602
N observations	18556160	18556160	37425892	19775703

All coefficients are multiplied by 100. Table displays the OLS regressions of individual-pair level outcomes on a dummy variable indicating whether the pair travelled in the same ship, interacted with a dummy indicating if the pair speaks the same mother tongue (based on census categories). The sample consists of all matched male passengers arrived during the period 1909-1924, grouped into non repeated pairs of individuals who arrived during the same week. The sample only include pairs of individuals with different surname (defined as Jaro-Winkler distance above 0.1) and from different places of origin. In Column (1) the dependent variable is a dummy for whether the pair works in the same occupation and industry. In Column (2) the dependent variable is a dummy for whether the pair works in the same industry and lives in the same state. In Column (3) the dependent variable is a measure of the log distance between the county of residence of each individual in the pair. In Column (4) the dependent variable indicates whether the pair lives in the same city. Regressions include indicators for each individual in the pair, fixed effects of Week of Arrival X Port Departure(i) X Port Departure(j) where i and j index individuals in the pair and fixed effects for Date Arrival (i) X Date Arrival (j). Standard Errors clustered at week of arrival level.

Table A9: Correlation in Outcomes by Contacts On-Land

	(1) Work Same Ind x Occ	(2) Work Same Ind x State	(3) Log Dist Residence	(4) Live Same City
SameShip x (HighCont-HighCont)	0.12***	0.06**	-2.34***	0.16*
	(0.04)	(0.03)	(0.59)	(0.10)
SameShip x (HighCont-LowCont)	0.16***	0.13***	-3.82***	0.24**
	(0.06)	(0.04)	(0.79)	(0.11)
SameShip x (LowCont-LowCont)	0.29***	0.21***	-7.09***	0.43***
	(0.07)	(0.06)	(1.02)	(0.13)
N individuals	134974	134974	193551	137602
N observations	18556160	18556160	37425892	19775703

All coefficients are multiplied by 100. Table displays the OLS regressions of individual-pair level outcomes on a dummy variable indicating whether the pair travelled in the same ship, interacted with a set of dummies indicating if both individuals have a high number of contacts on land, only one individual has high contacts on land or both have high number of potential contacts on land. Contacts on land are defined as the number of past emigrants from the same town of origin. The sample consists of all matched male passengers arrived during the period 1909-1921, grouped into non repeated pairs of individuals who arrived during the same week. The sample only includes pairs of individuals with different surname (defined as Jaro-Winkler distance above 0.1) and from different places of origin. In Column (1) the dependent variable is a dummy for whether the pair works in the same occupation and industry. In Column (2) the dependent variable is a dummy for whether the pair works in the same industry and lives in the same state. In Column (3) the dependent variable is a measure of the log distance between the county of residence of each individual in the pair. In Column (4) the dependent variable indicates whether the pair lives in the same city. Regressions include indicators for each individual in the pair and fixed effects of Week of Arrival X Port Departure(i) X Port Departure(j) where i and j index individuals in the pair. Standard Errors clustered at week of arrival level.

APPENDIX B: Matching Passenger Lists and Census using Machine Learning

In this section I provide further details on the matching procedure used to merge Passenger Lists with Census Data. I start by describing the potential problem faced by researchers dealing with large historical records. Then, I explain the steps involved in the matching algorithm and the techniques used to increase its speed.

The Dimensionality Problem An important challenge when matching across large datasets follows from the need of relying on fuzzy and noisy variables like names and surnames. Economists have used a number of approaches to address this problem, for instance, Fellegi & Sunter (1969), Christien & Churches (2005), Goeken (2011) and more recent Feigenbaum, (2016). However, in many cases, these approaches become unfeasible when data is large.[67] Not surprising, many studies relying on historical data have tried to overcome this problem by either using small random sub-samples or by imposing restrictive assumptions during the matching process.

Although recent advances in computer science have improved the search and matching techniques (see for instance, Schulz & Mihov, 2002), they remain unfamiliar and probably inaccessible to most applied Economists. The lack of easy implementations and the high entry costs to this literature has contributed to their low adoption. In this Appendix, I address the problem of matching across large historical datasets by improving on existing Machine Learning approaches (Feigenmabum, 2016). I introduce some simple modifications, popular among Computer Scientists, which significantly increase the speed and reduce the computational requirements of the matching process.

Two problems contribute to make matching unfeasible. First, the number of calculations required to compare records increases exponentially with the sample size. Intuitively, if there are N individuals in each dataset, the matching process involves comparing the name similarity of each pair of individuals which result in N^N calculations. Second, measuring similarity between string variables, involves computationally intensive algorithms. For instance, the most extended measure to compare two strings is the Levenshtein Distance (Levenshtein, 1966). It is defined as the minimum number of character insertions, deletions or substitutions required to transform the first string into the second one. Some statistical packages include commands to calculate Levenshtein distances but they are typically slow due to the complexity of the algorithm (usually based on Wagner & Fischer,

[67] I tried replicating the approaches described by Christien & Churches (2005) and Feigenbaum (2016) using a 20% random sample of the data. Both procedures resulted unfeasible for a desktop PC with intel-i7 processor and 24GB ram.

1974).

Blocking In some cases, researchers alleviate the first problem by narrowing the subset of potential matches *before* comparing names. In my setting, this *blocking* strategy, consists in defining for every individual i in the passenger list, a set of Census individuals such that: 1) They arrived to the US during the same year than i and 2) The distance in reported year of birth with respect to i is below 2. Then, for each passenger, I search for census individuals with similar names and surnames, *only* within the relevant block. In some cases, blocking solves the dimensionality problem and matching performs reasonably well.

Unfortunately, in many cases like in my setting, blocks are too large and the number of pair comparisons remain unfeasible. Some restrictive assumptions (like blocking on phonetic coding, or on the first two characters of the surname) are not recommended, particularly when dealing with non-English surnames, as they significantly reduce the accuracy of the matching.[68]

Matching Procedure The whole procedure follows a number of steps described below. Some steps are similar to those in Feigenbaum (2016), but some modifications are introduced to increase the feasibility and accuracy of the method. For efficiency reasons, the direction of the match is performed from the Passenger List to the Census data.

1. **Preliminar Cleaning:** I start by using a dictionary of US places (states, cities and acronyms), to detect passengers that are either US citizens, or have residence in the US. These individuals are excluded from the matching. Then, I use a dictionary of names acronyms and abbreviations (e.g. Jno. for John) and replace them in Passenger Lists and Census.[69].

2. **Unmatchable Cases:** I drop multiple observations with same name, surname, year of arrival and year of birth. These individuals cannot be distinguished from each other in the Census data, and therefore matching them is not possible.

3. **Set of Candidates:** For every passenger arriving during year y_a with year of birth y_b, find a set of "potential matches" in the census with year of immigration y_a and year of birth $y_b \pm 2$ and with a Levenshtein distance in given name and surname

[68] An important advantage of the algorithm used in this paper is that the Levenshtein distance, although computationally more intensive than the Jaro-Winkler distance, captures to a larger extent different sources of string differences, (e.g. not only typos but also phonetic transcriptions, etc.)

[69] This dictionary is constructed based on information from genealogy sites

below a threshold d.[70] This is the key step in the procedure and usually unfeasible if performed without any additional restriction. I explain later in this Section, two modifications that allow to identify candidates with similar names and surnames significantly faster compared to existing algorithms available in some statistical packages. Lastly, I drop any passenger for whom the set of candidates includes multiple census individuals that match exactly in name, surname and year of birth.

4. **Human Trained Sample:** The previous step defines a set of potential matches for each passenger. I randomly sample 2000 sets, and for each one, I decide whether there exists a candidate who is a "true match" for the reference passenger. As noted by Feigenbaum (2016), human criteria to detect true matches is highly reliable and accurate compared to automatized heuristic procedures. In a recent paper Bailey et al.(2017) find that supervised procedures, based on human trained samples, result in higher matching quality compared to other methods like Ferrie (1996). The training step is performed using all information available to the researcher, this includes the distance in names, surnames and year of birth but also additional information on the whole set of candidates and even the whole sample. Note that it is possible that no candidate is declared a true match. This would happen in two situations. First, if no candidate looks similar enough to the reference passenger (e.g. surnames are too different to be considered a typo or phonetic translation). Second, because more than one passenger looks similar to the reference passenger. When deciding whether a candidate is a true match, I also consider the number of candidates in the block, how similar is the second best candidate, how popular is the name or surname, and any type of information that can be relevant. In this step, the researcher sets the level of accuracy of the match as the following steps are aimed to "imitate" the heuristic behavior of the researcher.[71]

5. **Prediction of True Matches:** Based on the human trained sample, I use a Machine Learning approach to predict the true matches for the whole sample. Feigenbaum (2016) proposes a double-threshold probit procedure and Goeken et al. (2011) describe a Support Vector Machine approach[72]. In my case, I use a Random Forest Classifier (Breiman, 2001) due to its well known out-of-sample prediction properties. Additionally, the inclusion of a large set of variables describing the whole set

[70] Census data can be affected by rounding bias in the year of birth. For this reason, I also include the closest round year of birth.

[71] Other linking approaches that use human trained samples are Goeken et al. (2011) and Cristien & Churches (2005).

[72] This is similar to the procedure used by IPUMS to create census linked samples.

of candidates combined with the ability of the method to detect highly non-linear patterns, notably reduces the number of multiple predictions (i.e. two candidates are matched with the same passenger).[73] Indeed, cross-validation exercises reveals that the method results in a negligible number of false positives matches.[74] Bailey et al. (2017) shows that the bias introduced by false positive links are more harmful than the biased resulting in smaller matched samples and suggest that the quality of inference can be improved by increasing the precision of match (at the cost of reducing the number of matches). Table B1 at the end of this Section describes the main variables used as inputs in the Random Forest Classifier.

6. **Refining Predictions:** The fact that each Census candidate can belong to the set of potential candidates of multiple passengers implies that for a small number of cases, the same Census individual is matched with two different passengers. In those cases, I use the matching probability of the Random Forest model to assign as a true match the pair with highest probability. Then I run the Random Forest Classifier again excluding from the set of Census candidates those already matched to a passenger.[75]

As mentioned above, Step 3 is unfeasible even after blocking on year of birth and year of arrival. Some improvement in the algorithm that searchs among similar names and surnames is required to make any progress. The modifications I propose are the following: **1)** Reduce the number of comparisons by using indexed dictionaries of names and surnames specific for every block. **2)** Use a Levehnstein automata approach for searching among "similar names". A Levehnstein automata is a function that identify all the words within a list that are below a certain string distance. The automata significantly reduces the speed of calculations by transforming the dictionaries of names and surnames into a data structure called "radix trie" which decomposes words into a tree of common suffixes. Intuitively, the speed gain comes from the fact that when two words are detected to be above a certain string distance, every word sharing the same "branch" of the second word, will be at least at the same distance, and many searches are skipped.

[73]For the few cases where multiple matches are predicted, I only consider the highest probability match. Alternatively, the difference in the matching probability between the best and the second best matching can be considered, but I find no significant differences in my case.

[74]The Scikit Phyton package includes an straightforward implementation of the Random Forest Classifier

[75]All the results in the paper are robust to dropping individuals who were originally matched to multiple passengers.

Indexed Dictionaries A simple way of increasing search speed is by eliminating repeated calculations. This is achieved by creating a set of dictionaries for names and surnames, specific to every year of immigration and year of birth block. For instance, the target dictionary of surnames for a passenger arrived in 1911 with year of birth 1891, will contain the set of (non-repeated) surnames in Census data corresponding to all individuals arrived in 1911 with years of birth 1889 to 1893. Each surname is associated to a numerical id number. Similarly, names and surnames in the Passenger List are stored in dictionaries specific to the year of immigration and year of birth. Denote $W_S^P(y_b, y_a)$ to the dictionary of surnames constructed with individuals in the passenger list arrived in year y_a and born in year y_b. In a similar way, denote $W_S^C(y_b, y_a)$ to the dictionary of surnames based on census individuals arrived in year y_a and born in year $y_b \pm 2$. Instead of comparing among individuals, dictionary search is reduced to find for every entry in $W_S^P(y_b, y_a)$, a set of entries in $W_S^C(y_b, y_a)$ below a maximum Levenshtein distance defined by the researcher.

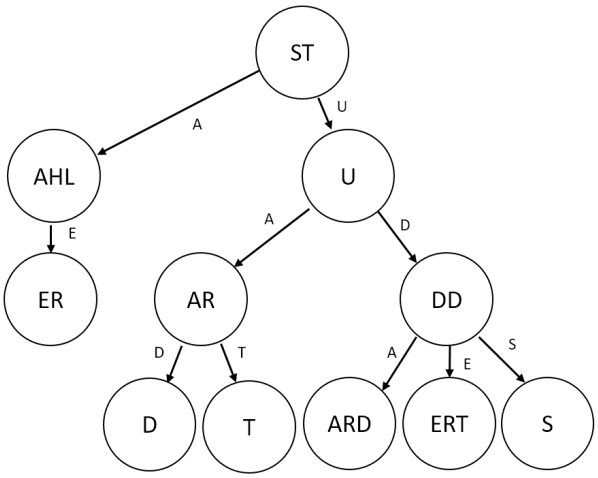
Levenshtein Automata and Radix Tries Search across dictionaries is more efficient than comparing individuals records, however, calculating string distance measures is slow. If dictionaries are too large, search remains unfeasible. I start by decomposing each dictionary into a Radix Trie, a structure that store words as a combination of suffixes (nodes) and paths connecting them.[76] Figure B1 below, shows an example of it for a dictionary of 8 surnames. Note that each word is associated to a parent branch and child nodes can emerge after a word terminates.

After transforming dictionaries into Radix Tries, I program a Levenshtein Automata that searches within the Trie and that for each entry in $W_S^P(y_b, y_a)$, retrieves a set of "similar" surnames from $W_S^P(y_b, y_a)$ (similarly for given names dictionaries). This Levenshtein Automata is thousands of times faster than any sequential word comparison. The reason is the lower number of required computations. Intuitively, as words are organized into branches, once the Automata detects a word not satisfying the similarity criteria, it stops searching into subsequent nodes. Remaining words in the branch, won't satisfy the criteria as well.[77]

[76] Radix tries are a common way in Computer Science to storage large volumes of string data. Beyond the search speed increase, they are also useful to storage information in a sequential way.

[77] http://personal.lse.ac.uk/BATTISTO/LevAutom.py is a simple Python implementation of a Levenshtein Automata based on Radix Tries. The code can be directly implemented using Stata datasets and export results to Stata format. The program is a simplified version of the program used in this paper.

Figure B1: Radix Trie



Dictionary: {*stahl, stahler, stud, stuart, stuard, studdard, studdert, studs*}

In order to further increase the speed, I add two additional elements. First, search is adaptative: for short words I start with a lower tolerance (maximum distance of 2) and only increase this threshold if few similar words are found. For longer words, the Automata starts to search with a tolerance of 3. The reason is that setting a high tolerance bound for short words is inefficient as it would retrieve most of the target words of similar length. Second, I store results as numerical matrices, where each cell contains the id number that indexes the word and the first column correspond to the Passenger List dictionary entries.[78]

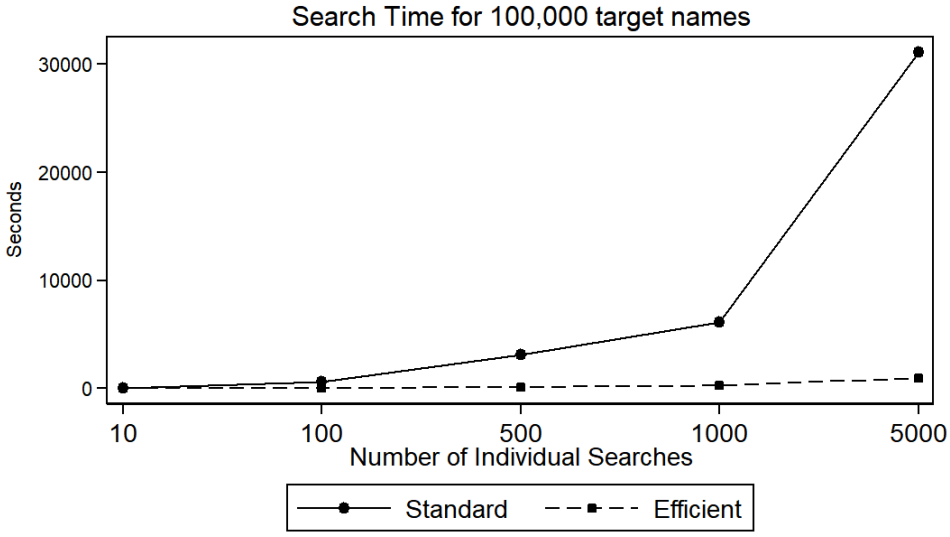
The final step to find the "set of potential candidates", is as follows: for every individual in the Passenger List, find the Census individuals with given names *and* surnames identified in the numerical matrices mentioned above. Since this step entirely relies on numerical variables, the process is fast even for large volumes of data. Figure B2 below illustrates the efficiency gain of the improved algorithm. The Figure compares the time required to find potential candidates for different number of individual records using a target database with 100,000 individuals.[79] The standard method uses the stata command

[78]I further restrict the number of "similar words" to the closest 300 entries identified by the Automata. This number is non-biding for the vast majority of names, but restricts the matrix dimension for a small number of short names that match with any word of similar length. The criteria to sort entries is based on the Jaro-Winkler distance (a variation of the Levenshtein distance that accounts for the length of the string and the relative position of the unmatched characters, (Lynch & Winkler, 1994)). This is convenient because it has a denser scale compared with Levenshtein distance. Furthermore, Feigenbaum (2016) uses a Jaro-Winkler treshold of 0.2 to restrict the pool of potential matches

[79]This size corresponds to the average size of a Year of Immigration X Year of Birth block, although the number of searches is substantially lower than the one performed to construct the dataset.The calculations

strdist to calculate Levenshtein distances and sequentially searches for candidates with names and surnames at a maximum distance of 3. The efficient algorithm incorporates Radix Tries Search and Dictionaries as explained in the text. The difference is significant, for instance, the standard algorithm takes more than 8 hours to perform 5,000 candidates searches while the improved algorithm does the same job in 16 minutes.

Figure B2: Comparing Search Algorithms



The Figure displays the search time in seconds of a standard search procedure and an improved version incorporating Radix Tries Search and Dictionaries as explained in the text. The improved algortihm used in this figure is a simplified version of the algorithm used to create the main dataset in the paper as it does not incorporate further improvements like efficient memory allocation (using numerical codes for strings storage) or alternative search methods for composed names. The target dictionary contains 100,000 individuals and the horizontal axis corresponds to different number of searches.

were performed with an i7-7th generation Intel processor and 24 GB of ram memory.

Table B1: Variables used for Random Forest Matching

Pair specific variables	Jaro Winkler Distance in first names
	Jaro Winkler Distance in surnames
	Jaro Winkler Distance of names and surnames combined
	Any match in the first name (relevant when multiple first names)
	First names match in Soudex code
	Surnames match in Soudex code
	Difference in age
	Round year of birth in Census
	Round year of birth in Passenger List
	Exact first name-surname match
	Exact first name-surname-yearbirth match
	First letter of first name matchs
	First letter of surname matchs
	Last letter of first name matches
	Last letter of surname matches
	Midle name initial matches (when multiple names)
	First name case Census(e.g. multiple names, middle initial, etc.)
	First name case Passenger List (e.g. multiple names, middle initial, etc.)
Block and aggregated variables	Number of potential candidates (and square)
	N of first name matches within block of candidates
	N of surname matches within block of candidates
	Average first name (Jaro Winkler) distance to all candidates in block
	Average surname (Jaro Winkler) distance to all candidates in block
	Jaro Winkler distance in first name to next candidate in block
	Jaro Winkler distance in surname to next candidate in block
	N of exact name-surname matches within block of candidates
	Frequency of surname in Census
	Frequency of first name in Census
	Frequency of surname in Passenger List
	Frequency of first name in Passenger List
	Frequency of first name-surname combination
	N of individuals in census year of birth cell

Note: The table does not list interactions between the variables included in the model.

APPENDIX C: Geocoding geographical information

This section describes the algorithm used to geocode the geographical units used in the main analysis.

Places of Origin The data contains information on the "last town of permanent residence". I first identify those individuals reported as US residents and exclude them from the matching process. For the matched sample, I pre-process the data by correcting for common typos and abbreviations in city or country names (e.g. Liverpool abbreviated as lpool). Then, I run a geocoding algorithm that uses the Google Places Api to identify the following information: Latitude and Longitude of the place, Name identified by Google Places and the Shouth-West/North-East coordinates of the smallest rectangle that contains the place. This rectangle is used in the main analysis to further restrict the set of shipmates assumed to be unrelated before the voyage.

The algorithm runs in several steps. It first starts by running an automatized search of the place of origin reported in the Passenger List (after cleaning). I only keep the cases where Google Places retrieves a unique place and it refer to a locality (city, village, etc.). For the remaining cases, I use a dictionary of country abbreviations and acronyms to split the sample by country of origin. Then, I search with Google Places using biasing parameters corresponding to the country. In a second step, I set the language parameter consistently with the country[80]. Finally, I manually search for the remaining cases where more than one observation is observed in the data. In many cases, the manual process consists in homogenizing names spelled with typos and re-running the Google Places search. In other cases, it consists in checking genealogy sites, and simple Google search for towns' name changes or translations.

In a number of cases (around 18% of the sample), the exact town can't be identified either because the individual report a broader administrative unit (e.g. the Italian province or region instead of the town), or only a larger administrative unit transcription is recognized by the algorithm, or the exact town does not exist anymore.[81] These cases are codified under the larger administrative region and the corresponding rectangle accounts for this. Finally, a number of observations can only be associated to disappeared historical regions (e.g. Kingdom of Galicia in the actual border between Poland and Ukraine). For these cases, I manually assign a coded name and the rectangle that covers the area of the historical region.

[80]This is useful for some eastern European cities, transcribed in their native language

[81]The advantage of using towns of origin instead of regions or provinces, is that fewer towns changed their names during the 20th Century.

For towns identified with high precision, I use a reverse geocoding algorithm to find the larger administrative region containing it. Broadly, this corresponds to the Google Place Api *administrative_area_level_3* category. Of course, due to changes in political divisions during the 20th century, measurement error can be significant for this codification.

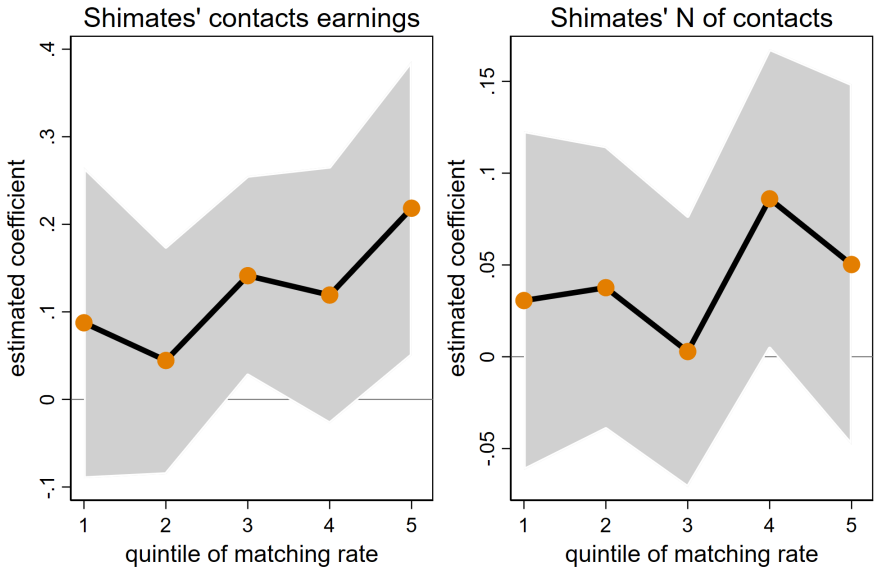
Ports and Routes Following similar steps than those used to geocode the town of origin, I obtain the latitude and longitude of every port of departure in the whole sample (including not matched observations). When two ports belong to the same city or they are located at less than 10 kilometers, I group them into the same unit (e.g. Liverpool and Birkenhead). Some observations include not only the port of departure but a whole list of ports covered during the voyage. In those cases, I only consider the first port reported by the individual as the departure port. Notably, the procedure geolocalizes the port of departure for more than 99% of the passenger records in the period 1909-1924.

Using all the ports of departure in the ship identified at the passenger level, I reconstruct the whole route of the vessel and calculate the total distance of the trip. I assume that stops are sorted by their distance to New York port and the travel distance is calculated as the sum of the minimum linear distance connecting the stops. In the case of ships that stop at Caribbean ports, when constructing Route Fixed Effects, I group them into the same category. There are three reasons for this grouping. First, the distance between Caribbean ports is small and total distance, other trip characteristics, and Caribbean ports' conditions are quite similar if we ignore this variation. Second, routes are identified based on the port of embarkation of all the passengers within the ship. Given that relatively few passengers board the ship in these small ports, differences in the estimated route can be due to measurement error. Finally, the main analysis do not use individuals departing from these ports.

APPENDIX D: Baseline Effects by Ship's Matching Rate and Potential Attenuation Bias

Figure D1 reports the baseline effects by quintiles of the ship's matching rate.[82] Effects are weakly increasing in the matching rate for both measures of shipmates' connections. This suggests that some attenuation bias could be expected due to the partial observability of the set of unrelated shipmates'. However, the fact that effects are not uniquely driven by the highest quintile also indicates that attenuation bias is not extremely large. This is not surprising given that many passengers within the ship shared either the same town or the same region of origin. Thus, since matching is orthogonal to individual characteristics (conditional on baseline controls), the sampling variation is lower relative to a case where shipmates' characteristics vary at individual level (i.e. within towns of origin). For instance, in the extreme case where individuals are matched proportionally to the share of their towns of origin within ship, there is no attenuation bias if the matched sample is large enough to include at least one individual per town of origin in the ship.

Figure D1: Estimated effects on individual's earning score by ship's matching rate



The figure displays the OLS estimation of the effects of shimates' contacts earnings interacted with dummies for the quintiles of the percentage of individuals matched within the ship on individual's earnings score. Quintiles are calculated conditional on Port of Departure. Regressions control for quintiles of the ship matching rate and other controls included in the baseline specification Regressions exclude ships with less than 10 individuals matched. Robust standard errors clustered at the week of arrival level.

In order to explore this idea more explicitly, I perform a series of exercises based

[82] To avoid some confounding effects, the quintiles are calculated conditional on the Port of Departure and the Year of Arrival.

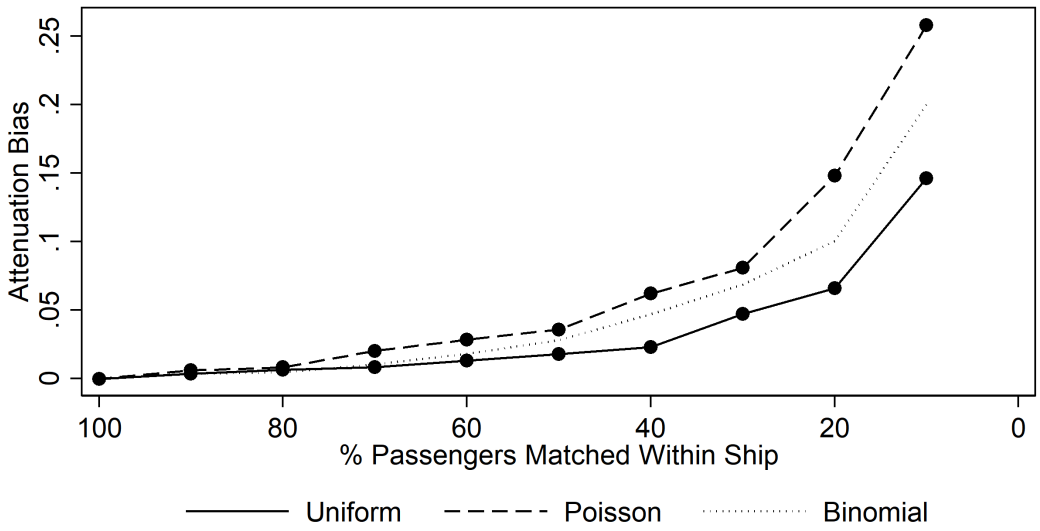
on simulated data. Using a distribution of ships and passengers that replicates the one observed in the full Passenger List, I generate the individual earnings as $Y_{i(c)} = \alpha + \bar{X}_{i(-c)} + X_{i(c)} + \epsilon_i$, where $X_{i(c)}$ is a simulated town of origin-specific component and $\bar{X}_{i(-c)}$ is the average town of origin component across all the unrelated shipmates' contacts (in other words, it replicates the construction of the average earnings of unrelated shipmates' contacts as used in the previous sections.) The term ϵ is an idiosyncratic individual component. The variance of $X_{i(c)}$ and ϵ_i are calibrated based on the distribution of their analogues observed in the matched sample data. Then, I create a random sample of passengers for each ship and recalculate the variable $\bar{X}_{i(-c)}$ using only the sampled passengers (this simulates the fact that only a subset of passengers are matched in the actual data). Finally, I calculate the attenuation bias for different sampling percentages using OLS estimations of the earnings equation.

The main difficulty when estimating the distribution of towns of origin within the ship (before sampling), is that this variable is harmonized only for the matched sample. The distribution of towns plays an important role in the attenuation bias as all the shipmates' characteristics ultimately depend on their town of origin. Hence, I simulate the distribution of passengers among towns of origin using three alternative assumptions: Uniform, Poisson and Binomial distributions. The parameters of each distribution is calibrated to replicate the number of average towns per ship in the matched sample data. [83]

Figure D2 shows the results of the simulations discussed above. The exercise reveals that for all distributional assumptions, the attenuation bias is relatively low. For instance, even for matching rates of 10%, the attenuation bias varies from 15% to 25%. The low attenuation bias is mainly driven by the fact that the number of different towns within the ship is not extremely large. Although these simulations rely on a number of arbitrary assumptions (e.g. homogenous matching rate across ships), the findings from this section suggest that baseline results shouldn't be seriously downward biased due the partial observability of the pool of shipmates.

[83] More specific, I start by simulating the distribution of towns with a low value for the distributional parameter. The distributional parameter is expressed as a percentage of the size of the ship (e.g. a uniform distribution with parameter of 0.5 implies that, on average, there is a town every two passengers within the ship). Then, I create a random sub-sample of passengers for each ship where the sampling rate is equal to the average matching rate in the data (approximately 12%). Finally, I calculate the average number of different towns per ship in the simulated random sub-sample. If this value is below the average number of different towns per ship in the matched data, I increase the distributional parameter and repeat the process until these values match.

Figure D2: Simulated Attenuation Bias



The graph displays the estimated attenuation bias using a simulated dataset with a distribution of ships and passengers similar to the one observed in the full Passenger Lists in the period 1909-1924. The distribution of passengers across different towns of origins is simulated assuming alternatively a uniform, poisson or exponential distribution. The parameter of each distribution is set set proportionally to the ship's size and calibrated to match the average number of towns in the matched sample. The attenuation bias is based on the OLS estimation of a linear m odel where the dependant variable is the individual income and the explanatory variable is the average income of shipmates' town of origin. Simulated income is calibrated to have a similar variance than the observed in the data. The horizontal axe measures the simulated share of passengers matched within the ship.

References

Abramitzky, R., Boustan, L.P and Eriksson, K. (2014), "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration" *Journal of Political Economy* 122, June: 467506.

Abramitzky, R., Boustan, L.P and Eriksson, K. (2012), "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review*, 102 (5): 1832-56.

Aho, Alfred V. and Corasick, Margaret J. (1975), "Efficient String Matching: An Aid to Bibliographic Search". *Communications of the Association for Computing Machinery*. 18 (6): 333340. doi:10.1145 /360825. 360855.

Ammermueller, S. and Pischke, J. S. (2009), "Peer Effects in European Primary Schools: Evidence from PIRLS," *Journal of Labor Economics* 27, July 2009, 315-348

Atack, J. and Bateman, F. (1992), "Matchmaker, Matchmaker, Make Me a Match: A General Personal ComputerBased Matching Program for Historical Research" *Historical Methods* 25, 2: 5365.

Athey, S. (2016), "Machine Learning and Causal Inference for Policy Evaluation". KDD '15 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Pages 5-6

Baeza-Yates, R. A. and Gonnet, G. H. (1996) Fast text searching for regular expressions or automaton searching on tries. *Journal of the Association for Computing Machinery*43, 6, November, 1996, 915-936.

Bailey, M., Cole, C. Henderson, M. and Massey, C.(2017), "AHow Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth" *NBER Working Paper* No. 24019, November 2017.

Bandiera, O., Rasul, I. and Viarengo, M. (2016), "The Making of Modern America: Migratory Flows in the Age of Mass Migration", *Journal of Development Economics*, Vol. 102, May 2013, pp 23-47.

Battisti, M., Peri, G. Romiti, A. (2016), "Dynamic Effects of Co-Ethnic Networks on Immigrants' Economic Success" , *Mimeo*, October 2016.

Bayer, P., Ross, S. L. and Topa, G. (2008), "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy*, 116(6), 11501196.

Beaman, Lori (2016) " Social networks and the labor market." In: Y. Bramoull, A. Galeotti B. W. Rogers (Hrsg.), *The Oxford Handbook of the economics of networks*, (Oxford Handbooks), Oxford: Oxford University Press, S. 649-671.

Bentolilla, S., Michelacci, C. and Suarez, J. (2010), "Social Contacts and Occupational Choice," *Economica*, 77(305), 2045.

Bertrand, M. Luttmer, E. F. P., Mullainathan, S. (2000), "Network Effects and Welfare Cultures", *Quarterly Journal of Economics*, 115(3), pp. 1019-1055, August 2000.

Bleakley, H. and Chin, A. (2010), "Age at Arrival, English Proficiency, and Social Assimilation among U.S. Immigrants" *American Economic Journal: Applied Economics* 2 (1): 16592.

Borjas, G. J. (2000), "Ethnic Enclaves and Assimilation," *Swedish Economic Policy Review*, 7, 89122.

Bramoullé Y., Galeotti A. and Rogers B. (eds.) (2016), "The Oxford Handbook of the Economics of Networks" *Oxford University Press*.

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45(1), 5-32, 2001

Brunner, B. and Kuhn A. (2009), "To Shape the Future: How Labor Market Entry Conditions Affect Individuals Long-Run Wage Profiles," *IZA Discussion Paper Series*, No. 4601.

Caeyers, B. and Fafchamps, M. (2016), "Exclusion Bias in the Estimation of Peer Effects," *NBER Working Papers* 22565, National Bureau of Economic Research, Inc.

Christien P. and Churches T. (2005), "Febrl - Freely extensible biomedical record linkage.", *Manual*, release 0.3, edition 2005.

Daniels, R. (2002), "Coming to America: A History of Immigration and Ethnicity in American Life", Second Edition, *Harper Collins*, isbn=9780060505776.

Dustmann, C., Glitz, A., Schonberg, U. and Brucker, H. (2015), "Referral-based Job Search Networks," *forthcoming The Review of Economic Studies*.

Edin, P.A., Fredriksson, P. and Aslund, O.(2003), "Ethnic Enclaves and the Economic Success of Immigrants: Evidence from a Natural Experiment", *Quarterly Journal of Economics*, vol 118, s329-357

Feigenbaum, J. J. (2016), "Automated Census Record Linking: A Machine Learning Approach", *Working Paper*.

Fitjar, R. D. and Rodrguez-Pose, A., (2016) "Nothing is in the Air" . *CEPR Discussion Paper* No. DP11067.

Genda, Yuji, Ayako Kondo and Souichi Ohta (2010), "Long-Term Effects of a Recession at Labor Market Entry in Japan and the United States," *Journal of Human Resources*, Vol. 45, No. 1, pp. 157-196.

Giulietti, C., Wahba, J. and Zenou, Y. (2014), "Strong versus Weak Ties in Migration," *IZA Discussion Papers* No. 8089.

Glaeser, E. L. (1999), "Learning in cities," *Journal of Urban Economics* 46, 254-277

Glitz, A. (2017), "Coworker networks in the labour market", *Labour Economics*, 44, issue C, p. 218-230.

Goeken, R., Huynh, L, Lenius, T. and Vick, R. (2011), "New Methods of Census Record Linking" *Historical Methods* 44:7-14.

Goel, D. and Lang, K. (2016), "Social Ties and the Job Search of Recent Immigrants," *IZA Discussion Papers* 9942, Institute for the Study of Labor (IZA).

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

Granovetter, M. S. (1973), "The strength of weak ties," *American Journal of Sociology*, 78, 13601380.

Granovetter, M. S. (1983), "The strength of weak ties: A network theory revisited," *Sociological Theory*, 1, 201233.

Gusfield, D. (1997), "Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology." *Cambridge University Press.*

Ho, T. K. (1995), "Random Decision Forests". *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 1416 August 1995. pp. 278282

Hopkins, A. A. (1910), "The Scientific American Handbook of Travel" *Munn and Co.*, New York, NY.

Hutchinson, E.P. (1981) "Legislative History of American Immigration Policy", 1798-1965, Philadelphia: *University of Pennsylvania Press*, 410.

Ioannides, Y. (2012), "From neighborhoods to nations : the economics of social interaction" Princeton, N.J. : *Princeton University Press.*

Ioannides, Y. M. and Loury, L. D.(2004), "Job Information Networks, Neighborhood Effects and Inequality," *Journal of Economic Literature*, 42 (4), 1056.1093.

Jackson, M. (2011) "An Overview of Social Networks and Economic Applications" in the *The Handbook of Social Economics*, edited by J. Benhabib, A. Bisin and M.O. , North Holland Press 2011

Jacobs, J. (1969) The Economy of Cities. New York: *Random House*

Kramarz, F. and Skans, O. N. (2014) "When Strong Ties are Strong: Networks and Youth Labour Market Entry", *The Review of Economic Studies*, Volume 81, Issue 3, 1 July 2014, Pages 11641200.

Kugler, A. (2003), "Employee Referrals and Efficiency Wages," *Labour Economics* 10, 531-556.

Laxton, E. (1996) The famine ships: The Irish exodus to America, 1846-51. *London: Bloomsbury.*

Lynch, M. P. and Winkler, W. E. (1994), "Improved String Comparator," Technical Report, Statistical Research Division, Washington, DC: *U.S. Bureau of the Census.*

Marmaros, D. and Sacerdote, B. (2002), "Peer and Social Networks in Job Search," *European Economic Review* 46, 870-879.

McKenzie, D. and H. Rapoport (2007), "Network effects and the dynamics of migration and inequality: Theory and evidence from Mexico," *Journal of Development Economics* 84, 1-24.

McKenzie, D. and H. Rapoport (2010), "Self-selection patterns in MexicoUS migration: The role of migration networks," *Review of Economics and Statistics* 92, 811-821.

Maurer, S. E. and Potlogea, A. V. (2017), "Male-biased Demand Shocks and Womens Labor Force Participation: Evidence from Large Oil Field Discoveries," *Working Paper Series of the Department of Economics, University of Konstanz* 2017-08, Department of Economics, University of Konstanz.

Montgomery, (1991), "Social Networks and Labor-Market Outcomes Toward an Economic Analysis" *American Economic Review*, vol 81, No 5, 1408-1418

Munshi, K. (2003), "Networks in the Modern Economy: Mexican Migrants in the US Labor Market", *Quarterly Journal of Economics*, 549-599

Nam, C. B. and Boyd, M. (2004), "Occupational Status in 2000: Over a Century of Census-based Measurement," *Population Research and Policy Review* 23, 2004: 327-358.

Oreopoulos, P., von Wachter, T. and Heisz A. (2006), "The Short- and Long-Term Career Effects of Graduating in a Recession: Hysteresis and Heterogeneity in the Market for College Graduates," *NBER Working Paper*, No. 12159.

Patel, K. and Vella, F. (2013), "Immigrant Networks and their implications for Occupational Choice and Wages," *The Review of Economics and Statistics*, 95(4).

Sato, Y. and Zenou, Y. (2015), "How Urbanization Affect Employment and Social Interactions" *European Economic Review* 75:131155.

Schulz, K. U. and Mihov, S. (2002), "Fast String Correction with Levenshtein-Automata". *International Journal of Document Analysis and Recognition.* 5 (1): 6785.

Sedgewick, R. and Wayne, K. (2001) Algorithms, 4th Edition, Addison-Wesley, 2011.

Sojourner, A. (2013), "Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR" *Economic Journal.* 123(569): 574-605

Taylor, N. F. (2010), "Grandma, Tell Us a Story", *Xlibris Corporation LLC*, 9781450071819

Topa, G. (2001), "Social Interactions, Local Spillovers and Unemployment," *The Review of Economic Studies*, 68, 261.295.

Topa, G.(2011) "Labor Markets and Referrals", in the *The Handbook of Social Economics*, edited by J. Benhabib, A. Bisin and M.O. , North Holland Press, 2011

US, Bureau of the Census (1975) "Historical Statistics of the United States, Colonial Times to 1970". *U.S. Bureau of the Census.* Pt.1

Waber, B. Magnolfi, J. and Lindsay, G. (2014) "Workspaces That Move People" , *Harvard Business Review*, October 2014 Issue (https://hbr.org/2014/10/workspaces-that-move-people).

Wagner, R. A. and Fischer, M. J. (1974), "The string-to-string correction problem. *Journal of the Association for Computing Machinery*", 21, 168173.

Wegge, S. A. (1998), "Chain Migration and Information Networks: Evidence from Nineteenth-Century Hesse-Cassel" *Journal of Economic History* 58 (4): 957-86

Yakubovich, V. (2005), "Weak ties, information, and influence: How workers find jobs in a local Russian labor market," *American Sociological Review*, 70, 408421.