Like an Ink Blot on Paper

Testing the Diffusion Hypothesis of Mass Migration, Italy 1876–1920

Yannay Spitzer
yannay.spitzer@huji.ac.il

The Hebrew University of Jerusalem and CEPR

Ariell Zimran
ariell.zimran@vanderbilt.edu

Vanderbilt University and NBER

December 2019

Click Here for Most Recent Version

Abstract

Why were the poorer countries of the European periphery latecomers to the Age of Mass Migration? We test the *diffusion hypothesis*, which explains the *delayed migration puzzle* by arguing that mass migration from peripheral countries was delayed by a lack of exposure to migration networks, and that the geographic expansion of these networks in a process of spatial diffusion was the main factor that eventually triggered mass migration. Focusing on post-unification Italy, we construct a comprehensive commune- and district-level panel of annual emigration data over four decades. We develop a new set of stylized facts on Italian emigration that are consistent with four main predictions of the diffusion hypothesis. Among these, we show that Italian mass migration to North America began in a few separate *epicenters* and expanded from there in an orderly pattern of spatial expansion over time. Using a new instrumental variables strategy, we show that this pattern was the product of a mechanism in which emigration from one commune was affected by emigration from its neighbors. These findings strongly support the diffusion hypothesis, and call for a revision of our understanding of one of the most important features of the Age of Mass Migration.

Acknowledgements For helpful comments, we thank Ran Abramitzky, William Collins, Joseph Ferrie, Michela Giorcelli, Joel Mokyr; seminar participants at the Hebrew University of Jerusalem, Northwestern University, Stanford University, the University of Michigan, the University of California Davis, and Vanderbilt University; and conference participants at the 2019 Galatina Summer Meetings. Christine Chang, Elizabeth Nelson, Sarah Robinson, Claire Whittaker, and Danielle Williamson provided excellent research assistance. This project was supported by a Vanderbilt University Research Scholar Grant. Work on this paper was completed while Ariell Zimran was a W. Glenn Campbell and Rita Ricardo-Campbell National Fellow and the William C. Bark National Fellow at the Hoover Institution, Stanford University; funding from the Hoover Institution is gratefully acknowledged. This material is based upon work supported by the National Science Foundation under Grant No. SES-1425598.

1 Introduction

During the *Age of Mass Migration* (1840–1914), nearly 30 million Europeans migrated to the US (Abramitzky and Boustan 2017; Hatton and Williamson 1998).[1] Although there is a consensus among scholars that economic incentives played a central role in creating this movement, some important patterns of the Age of Mass Migration are difficult to reconcile with the canonical push-pull model of migration (Sjaastad 1962; Todaro 1969). One of the most important is what we refer to as the *delayed migration puzzle*. Despite having the highest real wages in Europe, western European countries, such as Britain and Germany, were the early leaders in transatlantic migration, beginning in the 1840s or earlier (Cohn 2009). Countries in the southern and eastern European periphery, such as Italy, Austria-Hungary, and the Russian Empire, although far poorer, lagged behind for several decades (Hatton and Williamson 1994). It was not until the 1890s that the periphery suddenly surged to dominance of US immigration, taking the lead in both the absolute number of migrants and in emigration rates (Barde, Carter, and Sutch 2006; Hatton and Williamson 1998).[2] What made Europeans from the poorest countries latecomers to mass migration? If real income gaps were the prime movers of transatlantic migration, why did millions of potential migrants from the poorest countries forego for decades the opportunity to earn higher wages abroad, before suddenly starting to exploit this "technology"?

This paper tests the *diffusion hypothesis* as an explanation for the delayed migration puzzle. Initially proposed by social historians (Baines 1995; Gould 1980; Lowell 1987; Moya 1998), this hypothesis highlights the importance of networks of chain migration in determining the timing of mass migration. According to this hypothesis, in places without existing chain migration networks, most individuals were initially unable to migrate, regardless of the incentive to do so. Networks of chain migration spread over time from places already experiencing migration to their neighbors based on short-distance social links or information flows. This resulted in a spatial process, in which "the contagion of emigration spread over the map much like an ink blot on paper" (Moya 1998, p. 113).[3] The diffusion hypothesis's explanation for the delayed migration puzzle is thus that the main factor that delayed migration from the European periphery was the lack of exposure to social networks of chain migration and the slow pace at which the networks diffused over space. Areas with a high propensity to produce migration due to their low living standards did not do so until their

[1] During the same period, nearly 55 million Europeans migrated to the New World (Hatton and Williamson 1998).

[2] Formally, this implies that the correlation between origin-country real wages and emigration was not negative (as the push-pull model predicts) until the 1890s. Cross-sectional real wage differences were also poor predictors of emigration rates within countries (Baines 1995, ch. 4).

[3] Gould (1980, p. 283) used a similar analogy: "One might describe this process as one of 'diffusion,' at least in the mechanical sense in which a drop of ink on a small piece of blotting paper gradually 'diffuses' over the whole area."

pent-up demand for migration was released by the arrival of chain migration networks.

The diffusion hypothesis does not explain *why* the European periphery was relatively unlinked to migration networks before the 1890s. Its main tenets are that networks *were* missing in those decades; networks diffused slowly and spatially; the potential for mass migration existed before migration networks arrived; the arrival of migration networks was the main factor unleashing this potential; and there need not have been an internal cause for the rapid switch from no migration to mass migration, other than the arrival of migration networks.

Due to limited data and methodological constraints, existing evidence on the diffusion hypothesis is sparse. Whether the spatial diffusion of networks actually explains the timing of migration, and indeed whether a spatial diffusion process operated in the European periphery at all during the Age of Mass Migration, remain open questions. The diffusion hypothesis is not accepted by economic historians as the main explanation for the delayed migration puzzle. Instead, the dominant view is that it was caused by delayed economic and demographic modernization in the European periphery (Hatton and Williamson 1998); we refer to this explanation as the *internalist view*. In this paper, we provide the first systematic test of the validity of the diffusion hypothesis based on an explicit theoretical framework, a novel dataset of sufficient detail, and an empirical strategy that examines various testable predictions. We find a wide array of evidence indicating that a spatial diffusion process operated in Italy over the period 1876–1920. This finding provides the strongest support to date of the diffusion hypothesis as an explanation for the delayed migration puzzle.

Our investigation of the diffusion hypothesis focuses on Italian emigration during the period 1876–1920, which Gould's (1980) seminal paper cites as a paradigmatic case of migration driven by the diffusion of chain migration networks. We can observe this movement from its early stages in the 1870s through its surge in the 1900s and peak in the 1910s.[4] Italy constitutes an ideal laboratory in which to test for the existence of a diffusion process. It was sufficiently large and had enough internal variation to enable us to observe diffusion processes evolving gradually within its borders. Moreover, Italy was among the largest migrant source countries, with more than 4 million migrants to the US and nearly 2.5 million migrants to South America during the Age of Mass Migration.[5] Italian emigration statistics document this movement in fine detail, and form the basis for a novel panel dataset of emigration rates at the district and commune level,

[4]Mass migration to the US was effectively ended by World War I and by the imposition of the Literacy Test in 1917 and various quotas in 1921 and 1924. We also have data on the other two main Italian migratory flows—to South America and to other European countries—and use these data where comparing the flows to different destinations is useful.

[5]Between 1890 and 1914 Italians were 23.1 percent of all alien passengers entering the US. Only the migration from the Austro-Hungarian Empire (Central Europe other than Germany and Poland in the data of Barde, Carter, and Sutch 2006) was larger, at 23.3 percent of all arrivals (Barde, Carter, and Sutch 2006). Given limited migration from other sources to South America, Italy was responsible for more transatlantic migrants than any other source country.

covering more than 8,300 communes and 40 years. The scope and detail of these data enable us, for the first time, to rigorously characterize the spatial evolution of the Italian emigration and to test the diffusion hypothesis at the time and place in which it is suspected to have operated.

To illustrate the basic evidence motivating this paper, Figure 1 presents the average annual rates of emigration to North America at the district level for eight half decades between 1876 and 1914.[6] These maps support the view that the evolution of Italian emigration was governed by a process of spatial diffusion. In particular, they show a pattern in which emigration to North America began in a small number of separate *epicenters* and then spread from these epicenters to the rest of the country in an orderly spatial expansion over a period of about three decades. By World War I, it appears that most southern Italian districts had reached relatively stable levels of migration that were unrelated to initial leadership. Similar patterns are evident for migration to South America and Europe.[7] These patterns do not appear to be driven by some alternative underlying trend in economic incentives, as the development of migration to the different major destinations (i.e., North America, South America, and Europe) followed different spatial and temporal patterns.

To formalize the diffusion hypothesis, we propose a theoretical framework that combines the traditional push-pull framework with an underlying model of diffusion over a spatial network.[8] It treats migration as a technology that becomes available to individuals once a person to whom they are linked has migrated. Once they gain the option of emigration, individuals can then choose whether to move or not based on typical push and pull factors, such as income differences. Personal links within communes enable the intra-communal diffusion of the migration technology, and links with persons in nearby communes account for the inter-communal diffusion. *Pioneers*, who migrate without begin linked to a prior migrant, are allowed by this framework; as long as they are sufficiently rare and social connections at long distances are sufficiently uncommon, our framework can explain a spatial diffusion of emigration and a persistent delay in emigration from emigration-prone but unlinked regions.

Our theoretical framework makes four testable predictions. First, the country should exhibit β- and σ-convergence in emigration rates across communes and districts over time. Second, local emigration rates should follow an S-pattern over time, initially low, then rapidly increasing as networks arrive, and finally reaching a state of saturation in which they plateau around the level determined by push and pull factors. Third, mass emigration should be observed initially in only a small number of places and should expand

[6]We explain the construction of these data in section 4.
[7]See Appendix Figures A.1 and A.2.
[8]For simplicity, we rely on an informal description, but we develop a formal model in Online Appendix B.

spatially from there. This implies a prediction that we call the *frontier effect*—the rate of emigration and the probability that a commune or district entered mass migration in a given period should be inversely related to its distance from the nearest place that experienced mass migration in the previous period. Finally, the correlation between the destination choices of migrants from any two places in the country of origin should diminish with respect to the distance between them, because of spatial correlation in exposure to specific networks.

Our empirical analysis proceeds in two steps. In the first, we find that the four testable predictions of our framework are borne out in the data. In support of the convergence prediction, we find that the coefficient of variation in commune-level emigration rates to North America was more than halved between 1890 and 1914—evidence of σ-convergence. We also find that this was driven by laggards catching up with leaders in a pattern of β-convergence—communes in the bottom quartile of pre-1900 emigration rates to North America experienced a 50-times greater increase in average annual emigration rates after 1900 than communes in the top quartile of pre-1900 emigration rates. We also find clear evidence that local time series of emigration followed an S-shape. In support of the frontier effect, we find that a one-standard deviation increase in distance from the frontier of mass migration was associated with an 11 percentage point decline in the probability of achieving mass migration in the next period. Finally, in support of the correlated destinations prediction, we find that a one-standard deviation increase in distance between two provinces was associated with an increase in the dissimilarity index of their destinations of 0.104, or just under half of a standard deviation.

In the second step, we test whether, as the diffusion hypothesis predicts, the spatial expansion of migration and the frontier effect were the product of a *spatial contagion* mechanism in which a commune's emigration rate was causally affected by recent emigration from nearby communes. We find a positive correlation between a commune's emigration rate and a spatially weighted average of neighbors' lagged emigration rates, even when controlling for commune and year fixed effects, the lag of own emigration, and a variety of measures of development. This positive correlation is reassuring, but clearly the primary identification challenge is that it might simply capture the spatial diffusion of, or spatially correlated shocks to, fundamentals that cause emigration. To address this concern, we develop an instrumental variables approach. Intuitively, we compare two identical communes equidistant from an epicenter of emigration. The only difference between them is that the neighboring population around the first commune is on average closer to the epicenter relative to the second. Since emigration expanded geographically from these epicenters over time, the first commune will have an earlier contact with neighboring migrants than the second. The diffusion hypothesis predicts that the first commune will experience a greater increase in emigration than the second, and our empirical analysis tests whether this was the case. Formally, we implement this approach by constructing a measure of exposure to neighbors' emigration that is determined only by their distance to emigration epicenters and using this as an instrument for actual emigration exposure while controlling directly for a commune's own distance to such epicenters. Under the assumption that the spatial distribution of the neighboring population around a commune was independent of the commune's own emigration, except insofar as it affected exposure to chain migration networks,[9] this approach can identify a causal contagion effect of neighbors' migration on a commune's migration. When implementing this approach, we find a positive and statistically significant relationship between a commune's lagged emigration exposure and its emigration. In the benchmark results, the estimated lagged-neighbors'-emigration elasticity of own emigration is on the order of 0.4–0.5. Balancing tests confirm that the instrument is not correlated with other observable characteristics that could affect migration, alleviating the concern that spatial contagion reflected any other diffusion process that was not based on personal links with migrants from neighboring communes.

The accumulated evidence from the two steps of our analysis leads to the conclusions that there was a spatial diffusion of emigration and that this diffusion was likely transmitted through social contacts. Moreover, we find evidence that the diffusion processes were specific to the destination. This leaves the hypothesized social chain migration networks as the only plausible explanation for the pattern of spatial diffusion. These findings alone are insufficient to enable us to conclude than an absence of widespread networks of chain migration was the main explanation for the delayed mass migration from Italy.[10] But they do constitute the most comprehensive evidence yet of the existence of a diffusion process operating during the Age of Mass Migration (or in any other context), thus supporting the validity of the diffusion hypothesis.

Most directly, this paper contributes to the literature seeking to explain the delayed migration puzzle (e.g., Faini and Venturini 1994; Gould 1980; Hatton and Williamson 1994, 1998). Although the role of networks in shaping migration flows is widely recognized by economists and other social scientists (e.g., Fernández-Huertas Moraga 2013; Hatton and Williamson 1998; Massey et al. 1993; McKenzie and Rapoport 2010; Munshi 2003; Nelson 1959; Spitzer and Zimran 2018; Wegge 1998), the dominant view by economic historians of the Age of Mass Migration does not accept the diffusion of chain migration networks as an explanation for the delayed migration puzzle (Faini and Venturini 1994; Hatton and Williamson 1998). Instead, the onset of mass migration from an area depended, according to the internalist view, on its entry

[9] In fact, the required assumption is even weaker than this—it is only that any effect of a commune's neighboring population distribution on its own emigration be time invariant. The main danger is that other factors driving migration might also be transmitted along social lines.

[10] This conclusion would require structural estimation of the model in Online Appendix B.

into economic and demographic modernization, which released poverty traps, loosened connections to the land, and increased demographic pressures (Hatton and Williamson 1998). These processes occurred earlier in the leader countries, implying that the delayed mass migration from the European periphery was a result of delayed development. Such explanations based on the internal characteristics of immigrants' places of origin can rationalize the main feature of the delayed migration puzzle—that the correlation between real wages and emigration was at first positive and only later became negative. But they leave a number of important features of the Age of Mass Migration unexplained, as we detail below. As a result, a satisfactory explanation for the delayed migration puzzle that is backed by strong quantitative evidence has been lacking.

More generally, the literature on the economics of the Age of Mass Migration has expanded in recent years as new data and methods have been brought to bear on fundamental questions of this phenomenon (Abramitzky and Boustan 2017; Hatton and Ward 2019). But since Hatton and Williamson's (1998) espousal of the internalist view, there has been little advance in understanding the timing of mass migration and its geographic evolution.[11] In this sense, this paper seeks to fill an important gap in the recent literature and to revive the debate regarding the determinants of mass migration.[12] This paper, by revisiting and systematically testing the diffusion hypothesis with clear theoretically derived predictions, a multi-faceted empirical approach, and data at a level fine enough and a scope large enough to observe diffusion in action, lays the groundwork for an important revision of our understanding of the causes of mass migration from the European periphery.

This paper also makes a theoretical contribution to the economics of immigration. This contribution is strengthened by the fact that the open-border policy that the US maintained during the Age of Mass Migration makes it possible to study the economics of migration in the absence of restrictive policy and the other limitations of modern data (Abramitzky and Boustan 2017). In particular, the delayed migration puzzle is simply one historically important example in which individuals did not migrate despite standing to benefit tremendously from doing so.[13] It is well understood that liquidity constraints pose a significant impediment to migration from developing economies, and often the conclusion is that extreme poverty must be alleviated before mass migration is generated.[14] The lesson that we draw from the Italian migration

[11]A notable exception is Spitzer (2016), who showed evidence suggestive of a diffusion process playing a key role in the evolution of Jewish migration from Russia over the half century prior to WWI.

[12]Our paper thus complements that of Karadja and Prawitz (2019), who study the causes and consequences of emigration from Sweden during the Age of Mass Migration.

[13]For instance, Bryan, Chowdhury, and Mobarak (2014) document a surprisingly low seasonal migration in Bangladesh and Wozniak (2010) shows that non-college-educated individuals are less likely to respond to potential wage gains through migration than are college-educated individuals.

[14]McKenzie and Rapoport (2007) find an inverse-U relation between household wealth and emigration. Gray, Narciso, and Tortorici (2019) show that income shocks coming from changes in commodity prices also led to increases in emigration from Italy, presumably due to relaxation of liquidity constraints.

is that it is possible that the friends and relatives effect may be so strong that it can, by itself, switch a region from little or no migration to extremely high rates of migration within a short period of time and independently of any structural changes, such as poverty alleviation, urbanization, or sectoral shifts.

A better understanding of the factors that caused the delay of mass migration from the southern and eastern European periphery is also important because of the tremendous impact of the Age of Mass Migration on US economic, cultural, and political development. The eventual arrival of the "new immigrants" from the European periphery marked a turning point in American immigration history. It is impossible to predict what would have happened if the migration of Italians and others from the periphery had begun as early as the 1840s or 1850s. But it is clear that the American economy and the American nation would have developed very differently in either case. The diffusion hypothesis thus provides an answer to a question with fundamental implications to our understanding of American history—why did millions of immigrants arrive when they did, and why did millions of would-be Americans never come to the United States?

2 The Delayed Migration Puzzle

Prior to 1870, virtually all European immigrants to the US came from Germany, Ireland, or Britain. As can be seen in Figure 2, the composition of American immigration quickly changed in the 1890s, such that by 1900 almost 80 percent of immigrants came from the European periphery, in particular Italy, Russia, and Austria-Hungary.[15] These patterns conflict with the traditional push-pull paradigm (Sjaastad 1962; Todaro 1969), which predicts a negative relationship between living standards and the propensity to emigrate. As Figure 3 shows, the European cross-country correlation between real wages and emigration rates was positive during the 1870s and 1880s (panels a and b). It only became negative in the 1890s (panels c and d), as emigration rates in the poorer periphery increased.

There is a consensus among historians that this sudden change was not the result of changes in the cost of transatlantic travel. Fares of transatlantic passage were affordable and stable from the 1860s onwards.[16]

[15]These are shares of immigration from Europe. We focus on immigration from Europe because it was by far the largest in the nineteenth and twentieth centuries. Asian immigration had been effectively banned by the Chinese Exclusion Act of 1882 and by the "Gentlemen's Agreement" of 1907. There was also a large Scandinavian immigration beginning in the 1870s, but it was never one of the dominant flows of arrivals to the US.

[16]During the middle of the nineteenth century there was a sharp reduction in transatlantic steerage fares that is argued to have increased migration (e.g., Killick 2014). However, from the 1860s onwards, prices were generally sufficiently low that immigrants could repay them after a few weeks of labor in the US; indeed, there is evidence that fares were stable or even slightly rising in this period (Feys 2007; Hatton and Williamson 1998, pp. 14–15; Keeling 2012, Appendix 7). To be sure, as pointed out by Keeling (2013), there was a steady reduction in journey time and improvements in safety and in conditions aboard the ships, which amounted to an effective reduction in costs. Moreover, conditions were better in ships embarking from British ports and worse in Mediterranean ports, as indicated by the share of second class and steerage closed-berth out of the ship capacity (Keeling 2013, Figure 5.2). But clearly, there was no differential trend in fares between old and new source countries that can explain the large and rapid swing in the sources of transatlantic migrants.

Similarly, restrictions on the immigration of able-bodied Europeans and their dependents were minimal in the US until the imposition of the literacy restriction in 1917. Thus, although "the poorest had the most to gain by a move to higher living standards" (Hatton and Williamson 1994, p. 56), and although the US was open to their immigration, potential immigrants from the poorer periphery did not exploit the opportunity of migrating to the US for decades. Why did the masses of the European periphery abstain from overseas migration for so long? And why did they suddenly take the lead in both absolute numbers and rates of emigration in the 1890s? If the delay was because of a lack of incentives for migration, why were they lacking, and why did these incentives quickly arrive in the 1890s? If the incentive to migrate was always present, what prevented potential migrants from exploiting this opportunity, and why was this constraint suddenly removed in the 1890s?

While recent research has made considerable advances in studying the experiences of immigrants in the US, little progress has been made in understanding the causes of emigration.[17] Indeed, despite the importance of the delayed migration in the economic history of the Age of Mass Migration, it has not been studied in the recent resurgence of research into the Age of Mass Migration (surveyed by Abramitzky and Boustan 2017 and Hatton and Ward 2019). Instead, the canonical economic explanation for the delayed migration puzzle dates to an earlier wave of studies of the Age of Mass Migration. Faini and Venturini (1994) argue that the delayed emigration from Italy specifically was driven by the presence and eventual loosening of poverty constraints. Hatton and Williamson (1998), on the other hand, find no evidence of a role of poverty constraints in determining emigration rates at either the country-decade level for all of Europe, or at the province level in Italy for 1902 and 1912, and point to a different mechanism linking emigration to development. They find that the determinants of migration during the Age of Mass Migration were real wage gaps between the origin and destination, the lagged rate of natural increase, and the level of employment in agriculture (a measure of industrialization). Based on these findings and the fact that the response of the "Latin" countries to these migration incentives were similar to the responses of other countries, Hatton and Williamson (1998, pp. 45–46) conclude that "mass emigration in Europe had to await the forces of industrialization at home and a glut in the mobile age cohort driven by a demographic transition that industrialization produced." In sum, the dominant view in the economic history literature holds that the delayed emigration of the southern and eastern European periphery can be explained by these areas' persistent underdevelopment, and that the onset of mass migration at the end of the nineteenth century was caused by some internal process of economic and demographic modernization. We refer to this as the

[17] Gray, Narciso, and Tortorici (2019), Karadja and Prawitz (2019), Spitzer (2016), and Spitzer, Tortorici, and Zimran (2019) are notable exceptions.

internalist view of the delayed migration puzzle.

A number of important patterns in the Age of Mass Migration challenge the internalist view. For example, Ireland's early leadership in migration far predated its industrialization (Mokyr and Ó Gráda 1982), and there is no evidence of demographic pressure in Ireland when its emigration rate first began to surge before the Great Famine (Cohn 2009; Mokyr 1983). Jewish migration from the Pale of Settlement in the Russian Empire also challenges the internalist view. This migration began in a few poverty-stricken provinces in the northwest; only later did it spread to the nearby centers of Polish industrialization and the farther southern provinces, in which living standards were higher (Spitzer 2016). In general, the evidence based on within-country correlations does not show any systematic negative or positive relation between economic conditions and emigration in the Age of Mass Migration (Baines 1995, pp. 27–28).

The Italian migration also exhibits several patterns that do not fit with the internalist view. At the national level, Ardeni and Gentili (2014) show that the explanatory factors studied by Faini and Venturini (1994) and Hatton and Williamson (1998) have little to no power in explaining the time series of Italian emigration at the national level when the analysis is adjusted to address the flow of Italian migration to multiple destinations. Moretti (1999) also argues that Hatton and Williamson's (1998) model fares poorly in explaining the timing and destination distribution of Italian migration.

The internalist view also fares poorly in explaining the surge in emigration from southern Italy around 1900 (panel c of Figure 4). Although this was indeed the period in which Italy's industrialization began in earnest (Toniolo 2013), industrial development occurred almost entirely in the northwestern "industrial triangle" (Ciccarelli and Fenoaltea 2013). Indeed, conditions in southern Italy seem to have worsened or at best stagnated rather than improving in this period (Federico, Nuvolari, and Vasta 2019; Iuzzolino, Pellegrini, and Viesti 2013).[18] Demographic pressures identified by Faini and Venturini (1994) and Hatton and Williamson (1998) also have some difficulty in explaining this surge. Del Panta (1997, p. 10) explains that it was the north rather than the south of Italy that experienced improvements in infant mortality during the late nineteenth century, and Vecchi (2011, Table S6) finds that the increase in life expectancy at birth was considerably larger in the north than in the south.[19] Thus, it is difficult to link the surge in emigration from the south to any sense of economic modernization or demographic pressure.

[18]Ciccarelli and Fenoaltea (2013) do point out that the leaders in industrial and general economic *growth* immediately after unification in 1871 were in the south. But southern industry appears to have been largely artisanal or based on traditional extractive industries, such as in the sulphur-mining areas of southern Sicily, rather than the industrial and especially factory growth that made up the economic modernization of the Industrial Revolution. Moreover, they show that after 1881 southern industry was in relative decline. For more details on regional growth in post-unification Italy, see Daniele and Malanima (2007) and Felice (2005).

[19]Livi Bacci and Breschi (1990) point out that fertility declines in Italy did not come until after the period that we consider.

The main contender to the internalist view is the diffusion hypothesis. This explanation for the timing of mass migration from the periphery was originally proposed by Gould (1980, p. 302), who argued that "the great upsurge in emigration from South and East Europe was caused . . . by the progress of the diffusion of emigration in potentially emigration-prone areas, the achievement of 'saturation' levels in which occurred almost simultaneously in the ten or fifteen years before the outbreak of the War." Gould (1980) views Italy as the canonical case of the diffusion of migrant networks, and Lowell (1987) and Moya (1998) apply the framework to explain emigration patterns from Scandinavia and from Spain, respectively. Baines (1995, p. 30) similarly sees the spread of emigration within countries as evidence of diffusion of information, rather than a reflection of trends in underlying incentives for migration.

The evidence provided by these studies to support the diffusion hypothesis has tended to be informal. Gould (1980), for instance, provides evidence of convergence and argues that migration rates followed S-shaped local time series. Hatton and Williamson (1998), on the other hand, provide one of the first attempts to formally test the diffusion hypothesis. Focusing on Italy, they test specifically for persistence in emigration rates of provinces over time, and for a relationship between literacy (as a factor that could facilitate the spread of information) and emigration rates. While they do find evidence of persistence,[20] they do not find role for literacy in determining emigration. They ultimately conclude that diffusion "offers few empirical predictions and says nothing about why emigration rates eventually declined" (p. 15) and that "while such forces [as diffusion] mattered, there is little evidence that persistence or literacy dominated [Italian] provincial emigration rates with anything like the force often assigned to them in the qualitative literature" (p. 121).

Two factors, however, have limited these previous tests of the diffusion hypothesis. First, prior research has lacked a complete theoretical framework from which to derive testable predictions of the diffusion hypothesis. Second and more importantly, in most previous studies, the data are insufficiently detailed to test for the presence of the crucial inter-communal expansion of migration and for a mechanism of spatial contagion by which emigration from one commune causes migration from other nearby communes. For instance, Gould's (1980) and Hatton and Williamson's (1998) studies of diffusion in Italy relied on data at the province level or coarser, at which inter-communal diffusion cannot be effectively observed. The strongest evidence that could thus be gleaned was in regards to convergence and persistence, leaving open the question of whether diffusion can explain patterns of migration during the Age of Mass Migration.

The strongest existing evidence on the diffusion hypothesis is provided by Spitzer (2016), who shows that patterns of mass emigration from the Pale of Settlement were dominated by β-convergence between followers

[20]Gomellini and Ó Gráda (2013) find an even larger estimate of the role of persistence in determining the quantity of emigration.

and leaders of migration. He also documents a spatial expansion of the region experiencing mass emigration from an initial epicenter. Our analysis goes beyond that of Spitzer (2016), making advances in theory, data, and empirics, to strengthen the evidentiary basis of the diffusion hypothesis.

3 Theoretical Framework

In this section, we develop the theoretical framework that guides our analysis, which is based on a formal model presented in Online Appendix B. We also enumerate the testable predictions deriving from this framework, which form the basis of our empirical analysis. As in the traditional push-pull framework (Sjaastad 1962; Todaro 1969), individuals' incentives for migration in our framework are determined by push-pull factors, such as real wage gaps between the origin and the destination, as well as other factors espoused by the internalist view, such as industrialization, urbanization, and demographic pressures. We make two main departures from the push-pull model. First, we assume that being linked is a *necessary* condition for migration for all potential migrants except pioneers. Second, we include an explicit structure specifying how such links are generated.

3.1 Basic Setup

There are three types of individuals in our model.[21] Individuals begin as *susceptible* (unlinked). These individuals are not able to migrate regardless of the incentive to do so. Eventually, these individuals may switch to being *linked*. This switch can occur in two ways: an individual switches from susceptible to linked when one of his contacts migrates; alternatively, the switch can occur spontaneously, in which case this individual can be a pioneer. Linked individuals make a choice of whether or not to migrate based on the usual push and pull factors. If they migrate, they become a *migrated* individual and their susceptible contacts become linked. For simplicity, the following discussion will focus on the case of a single destination. When there are multiple destinations, the progress of individuals from susceptible to linked to migrated is separate for each destination and individuals linked to more than one destination decide whether to migrate to one of them or to remain in the origin.

Figure 5 shows a hypothetical chain of events that illustrates the main concepts of our framework. There are three communes, A, B, and C. The first individual to migrate was a_1 from commune A. Individual a_1 was a pioneer, meaning that he switched spontaneously from susceptible to linked without contact with a

[21] The classification is based on that of infectious disease models (e.g., Bernoulli 1776; Kermack and McKendrick 1927). Such models have been applied in economics by Burnside, Eichenbaum, and Rebelo (2016), among others.

prior migrant, and then migrated. He was connected to two other residents of commune A, a_2 and a_3, and his migration converted them from susceptible to linked.[22] This is a case of *intra-communal* diffusion of the migration technology. Eventually, a_2 and a_3 also decided to migrate, converting four more susceptibles in commune A to being linked. As the process proceeds, commune A is likely to quickly become *saturated*, in the sense that all individuals would become either linked or will have already migrated, and there would be no more susceptibles. At this point, commune A's migration is determined solely by push and pull factors and not by the rate at which the migration technology diffuses. Commune A is an *epicenter*, since migration was already common there in the initial period.

The migration of individual a_3 also linked b_1, an out-of-town contact in neighboring commune B. This is a case of *inter-communal* diffusion of the migration technology, which caused a *contagion* of migration from commune A to commune B. If individual b_1 were eventually to migrate, commune B would likely advance towards saturation and transmit the migration technology to its neighboring communes, and so on. Commune C, on the other hand, was further from A, and before receiving the migration technology through an inter-communal linkage, one of its residents, individual c_1, a pioneer, spontaneously gained the option to migrate. If he migrates he is likely to start a new chain of migration spreading from commune C.

This model does not specify how precisely the migration of one individual enables the migration of those to whom he is linked. We envision two possible mechanisms. First, it could be that migration depends on material support by friends or relatives. Prior migrants may have provided direct financial aid, as well as guidance and social support that enabled smooth absorption in the country of destination—a first address, a first job, furniture, advice, etc. Second, regardless of the actual support provided by friends and relatives, these connections provide an example that overseas migration can be undertaken successfully, much more so than any other form of media, such as newspapers.[23] This seems to have been an important feature of the quick process by which Antonio Squadrito brought migration to his Sicilian hometown, Gualtieri-Sicamino, as described in more detail in Online Appendix C. When he migrated to the US in 1898, his neighbors "all prophesied disaster and misfortune 'in that strange wild land'" (Brandenburg 1904, p. 43), and "every one, even his own family, had been dubious and pessimistic about the venture. Since then [by 1903] more than one tenth of the population has followed him, and any remaining pessimism was restrained, and those who were too poor to go, too old or too well situated to take new chances, vented openly expressions of

[22] Arrows in Figure 5 indicate the direction of the diffusion of the migration "technology."

[23] One way in which Hatton and Williamson (1998) operationalize the diffusion hypothesis is to search for a relationship between literacy rates and migration, based on the assumption that literacy was necessary for the diffusion of information through newspapers and letters. But given the very high rates of return migration from the US to Italy (Bandiera, Rasul, and Viarengo 2013; Hatton and Williamson 1998), it is possible that some of the most useful information was transmitted by visiting or returning migrants, as in the case of Antonio Squadrito's return to his home town.

envy" (Brandenburg 1904, p. 109).[24] These two mechanisms are consistent with the spirit of our framework and they are indistinguishable in our data. We are agnostic as to which one of them carried more weight. Consistent with either of these mechanisms, Spitzer and Zimran (2018, Table A.1) show that 94.9 percent of Italian passengers arriving in the US after 1907 reported that they were meeting someone already in the country, and 41.0 percent reported that they were meeting an immediate family member.[25]

3.2 Comparison to the Canonical Model

As a point of reference, we take Hatton and Williamson's (1998, chs. 3–4) specifications, referring to them as the canonical framework. Our framework and the explicit model underlying it (Online Appendix B) are similar to the traditional way that international migration has been modeled in that the fundamental incentives are the same—migrants are driven by the push-pull factors discussed above, most prominently real wage gaps between origin and destination. Our conceptualization of the friends and relatives effect is different, however. In the canonical framework, this effect is accounted for by controlling for some combination of the stock or lag of migration from the same country or smaller unit of origin, which can be interpreted as a migration cost shifter. In our framework, the friends and relatives effect is construed as a technology; when an individual has no linkage, the cost of migration is infinitely high; when an individual is linked, the cost immediately diminishes, and additional links to more migrants do not help to reduce it further. An Italian who has no friend or relative in the US will not be able to migrate. However, one who knows a single person there can easily move, and as a first order approximation, having more than one acquaintance does not change his incentives.[26]

While the primitives of the friends and relatives effect are thus different, when aggregated at the local level some of the predictions are similar, such as the simple conclusion that, at the local level, current migration is increasing in past migration. But there are also important differences. One is that the diffusion hypothesis predicts an eventual state of saturation, in which the rates of migration stabilize, whereas the canonical model predicts that the rates of migration will rise monotonically.[27]

[24] Gould (1980, pp. 292–293) made a similar point: "lapse of time was required for the idea of emigration to become implanted in the minds of the members of a community, for its net benefits to be firmly demonstrated by the experience of those who had gone first, and for those pioneers to set in train that process of encouragement and financial help which forms the links constituting what we call 'chain migration.'"

[25] In Appendix Figure A.3, we use the data from Spitzer and Zimran (2018) and data from the Battery Conservancy (2009) to show that these connections were overwhelmingly to individuals living in the same or nearby communes, with the distance of connections greater for individuals meeting friends than for individuals family members.

[26] To be clear, our framework can include past migration as a cost shifter without affecting the qualitative conclusions.

[27] Both the canonical framework and our model predict that improvements in the standards of living will eventually reduce migration. According to Hatton and Williamson (1998), this was the cause of the decline in migration often observed in the Age of Mass Migration.

A more significant difference between the two frameworks is in explaining how migration first begins in a particular place. According to the canonical model, there is no spatial contagion across regions. Conditional on local characteristics, the timing of the onset of mass migration is independent of whether or not migration was already present in a region. An area that has not experienced migration but in which individuals are otherwise strongly incentivized to migrate, will spontaneously start to produce migration within a short period of time (e.g., a few years). If a certain area abstains from migration for a long period of time (e.g., a decade or more), this model implies that it must be because its residents lack an incentive to migrate (Hatton and Williamson 1998, p. 39).

Our model allows dynamics of this type. If pioneers are common, then individuals will quickly gain the option to migrate regardless of their and nearby areas' experience of past migration. However, unlike in the canonical model, a persistent lack of migration from a place with high migration incentives can also be explained by a lack of connections to prior migrants in the region. If pioneers and social connections over longer distances are sufficiently rare, then this "pent-up" demand for migration will be unleashed only when the inter-communal diffusion of the migration technology reaches this area. The resulting surge in migration would occur not because the underlying incentives for migration had become stronger but because the intra-communal diffusion would have quickly spread the migration "technology" among the local population. In sum, the canonical model is nested within our framework. The difference is that our framework allows for spatial contagion, and eventually for situations in which migration is delayed for a long period only due to geographic distance from migration networks.

3.3 Testable Predictions

As discussed above, our theoretical framework can capture the diffusion hypothesis's explanation for the delayed migration puzzle if pioneers are sufficiently rare and social connections over longer distances are sufficiently uncommon. Under these conditions, our framework makes the following four testable predictions.

Prediction 1 (Convergence)**.** The overall cross-commune variation in the rates of migration caused by underlying variation in push-pull factors is initially augmented by the variation in the shares of linked individuals. As a growing number of communes reach saturation, the latter source of variation is gradually eliminated, such that the variation in migration levels off around a lower rate, reflecting only variations in push and pull factors. This is one of the main patterns that Gould (1980) identified as clear evidence of diffusion. To be clear, the diffusion hypothesis does not predict that all communes will converge to a similar rate of emigration. Indeed, the prediction is that a significant amount of variation will remain even when all communes are exposed to emigration, based on the variation of push and pull factors across them.[28] This leveling is manifested by a pattern of σ-convergence—the cross-commune standard deviation and coefficient of variation of migration rates will decline steadily, and when the entire country reaches saturation, it will stabilize. Second, this process generates a pattern of β-convergence. Communes that are latecomers to migration due to an initial absence of linkage to prior migrants experience rapidly rising migration rates shortly after linkage, whereas communes that are already saturated have higher rates but little or no growth. With convergence in migration rates coming from laggards catching up, the β-convergence prediction is of a strong negative relationship between current migration rates and future growth in migration.

Prediction 2 (S-Shaped Local Time Series)**.** Before any individual in a commune is linked, the commune's migration rate will be very low.[29] Once a sufficient number of individuals in a commune become linked (either through an external linkage or spontaneously as a pioneer) and migrate, the intra-communal diffusion of migration that this creates will generate a rapid increase in the migration rate as individuals become linked, migrate, and link their connections. Eventually, everyone in a commune will be linked (saturation), and the rate of migration will stabilize around the rate determined by push and pull factors. When combined, these three phases will create an S-shaped local time series of migration rates, and the successive entry of communes into these S-curves should generate the convergence of Prediction 1. This, too, is a prediction made by Gould (1980), and is also a common prediction of the of the technology-adoption literature (e.g., Bass 1969; Comin, Dmitriev, and Rossi-Hansberg 2012; Jovanovic and Lach 1989).

Prediction 3 (Frontier Effect)**.** Starting from the early sources of migration (the epicenters), emigration from one commune will link individuals in neighboring communes, spreading migration to them. Ultimately, this process will result in successive communes beginning to experience migration in a spatial order. A corollary of this is that communes not already experiencing mass migration will only begin to do so if they are close enough to areas in which the migration technology had already arrived. We define the *frontier* at any given period to be the boundary of an area that has or has had emigration above a certain threshold level. The *frontier effect* prediction is that the probability of a commune to enter mass emigration in the next period is inversely related to its distance from the current period's frontier.

This prediction is stronger than simply a spatial expansion of the region experiencing mass migration. Instead, the framework predicts that emigration from one commune, by creating more linked individuals in

[28] Gould (1980, p. 314) points out that "The process of diffusion . . . did not guarantee that pioneer migration would be followed by a mass movement increasing in some predetermined mathematical progression. If the conditions were not propitious: if the income gain was insufficiently large, for example, or the conditions of the migrant community unacceptable in some other way, the pioneer movement would prove still-born."

[29] Formally, the prediction is that it will have no migration at all.

nearby communes, will *cause* an increase in migration from its neighboring communes in a pattern of *spatial contagion*, as one commune "infects" its neighbors.

Prediction 4 (Correlated Destinations)**.** Two neighboring communes will typically share the same networks due to their proximity. Therefore, they should have similar migration options and a similar distribution of destinations. On the other hand, two distant communes are likely to be part of different networks, potentially leading to different destinations. In general, because emigration networks are destination-specific and the probability that two migrants belong to the same network diminishes as the distance between them grows, the dissimilarity in the distribution of migration destinations of two communes should increase with distance.

4 Data

4.1 Sources and Construction

Our main data source is the *Statistica della Emigrazione Italiana per l'Estero*. This series of volumes was published approximately every two years from the 1870s to the 1920s by the Italian *Direzione Generale della Statistica*. We digitized three data panels from this source. The first is a panel of annual emigration counts spanning the period 1884–1920 at the level of the commune (*comune*), of which there were more than 8,300 in Italy.[30] The second is a series of annual emigration counts at the district level (*circondario* or *distretto*), of which there were 284, which enables us to extend our temporal coverage to begin in 1876 rather than in 1884.[31] The last is an annual panel (1877–1920) of emigration counts for 25 consistently defined destinations (usually countries) at the level of the province (*provincia*), of which there were 69 in Italy. We focus in most of our analysis on three main groups of countries: North America, South America, and Europe, which together comprised 96.8 percent of all Italian emigration during the period 1876–1920. These data are based on jurisdictional boundaries at the time that the data were reported. These boundaries experienced changes during our study period, as well as in the century since. Online Appendix D describes how we dealt with these issues to arrive at a series of data at consistently-defined geographic levels with known geographic relationships with one another.[32]

This source has previously been used to study Italian emigration (e.g., Ardeni and Gentili 2014; Faini

[30] Ultimately we have data for 8,352 communes.

[31] The precise number of districts varied over the study period; but we were able to identify 284 consistently-defined units with emigration data.

[32] We are missing data for 1879 for the district and province-by-destination data and for 1888 for the province-by-destination data. In 1916 and 1917, there was virtually no transatlantic migration because of World War I, and consequently there was no volume published for these years.

and Venturini 1994; Hatton and Williamson 1998; Moretti 1999), but only using the province-level data. To our knowledge, we are the first to digitize and analyze the migration data at the finer levels of geographic disaggregation. Their fineness makes ours perhaps the most detailed data in terms of geographic disaggregation and temporal coverage available on a migration flow as large, as geographically varied in origin and destination, and as historically important as that from Italy in the late nineteenth and early twentieth centuries.[33]

The emigration counts reported in the *Statistica della Emigrazione Italiana per l'Estero* are based on the issuance of passports to emigrants by the Italian government. Although it provides the most comprehensive data available on Italian migration, there are some known issues with this source, such as inaccurate reporting by the mayors (*sindaci*) of the Italian communes.[34] One issue that Hatton and Williamson (1998, p. 98) cite as being of special concern is the change in the Italian passport law such that passports were compulsory and free from 1901 onward, and costly and not required, though still helpful, prior to 1901. While our analyses will use year or half-decade fixed effects wherever possible, which should control for this change, it is still potentially concerning that there is a surge in emigration, in particular to the US, between 1900 and 1901 (Figure 4) that may be driven by data issues rather than by true changes.[35] But US arrival data (Barde, Carter, and Sutch 2006) show growth in Italian arrivals from 1900–1901 that closely matches the increase in our data, reducing this concern.[36] A related issue is that the commune-level data for 1884–1903 aggregate some communes with low but non-zero emigration rates in a single report for each district. Communes included in this aggregation will appear to have an emigration rate of zero in these years. Although this restriction is not ideal, the low emigration rates that this excludes are unlikely to drive our results. Moreover, this concern does not affect the district-level data.

Another concern raised by Foerster (1919, ch. 2) and Hatton and Williamson (1998, ch. 6) is that the emigration data do not always enable a distinction between temporary and permanent migration.[37] Unlike Hatton and Williamson (1998), however, we do not view this as a deficiency. A major feature of the mass migration from Italy to the United States was the temporary nature of many of these moves (Bandiera, Rasul, and Viarengo 2013). Our goal is to explain the large movement of labor, whether permanent or not,

[33]Karadja and Prawitz (2019) and Lowell (1987) use highly detailed data on emigration from Sweden (Karadja and Prawitz 2019; Lowell 1987)—a country with less than one-sixth of Italy's population.

[34]See the discussion of the accuracy of the Italian emigration data by Foerster (1919, pp. 10–22).

[35]Foerster (1919, p. 21) claims that the Italian official statistics were less precise than the American immigration data, and that around 1901 there was a change from under- to over-enumeration of Italian emigrants.

[36]This is shown in Appendix Figure A.4. Although there is little difference around 1901, larger differences emerge later in the study period.

[37]We are able to make this distinction for some earlier years; in these cases we use the total of temporary and permanent migration.

which the data enable us to do.

We compute emigration rates using 1901 population (reported in the same source as the emigration data) as the denominator.[38] Because the smallest geographic unit for which destination data are available is the province, we impute destination-specific emigration rates for each commune and district, based on the province-year-specific weights of destinations, as

$$e_{idpt} = e_{ipt} \times s_{dpt}, \tag{1}$$

where e_{idpt} is the imputed rate of emigration in year t from commune or district i in province p to destination d, e_{ipt} is the observed emigration rate in year t from commune or district i in province p to any destination, and s_{jpt} is the observed share of all emigrants from province p in year t who migrated to destination d.

We also digitized data at the district level on age distributions, literacy rates, and occupational distributions from the 1881 Italian census for use as controls in our analysis. Specifically, we collect information on the fraction of the male labor force that was employed in agriculture or in industry, the fraction of the population of a district that was younger than 15 years old (a measure of demographic pressure), and the fraction of males age 15 or above who were literate. To our knowledge, these district-level data are the most geographically disaggregated data available and have not been previously used.[39] We focus on 1881 because this census provides the most recent information available prior to the start of the study period in which we have commune-level data. Analogous data are available for 1901 and 1911, but we do not use them because they may be "bad controls"—endogenous to the emigration that we study. We also collect commune-level information on rail access from Ciccarelli and Groote (2017), which we use to construct a measure of a commune's distance from the nearest rail line in 1881. Finally, we use data from the Shuttle Radar Topography Mission to determine the average elevation of each commune.

[38] We use 1901 population because it is the population reported in the 1904–1905 volume of the emigration statistics, which is the first volume in which the emigration of all communes is reported regardless of magnitude. It thus enables us to secure population measures that are most comparable in the unit of analysis to those of our emigration data. We have also repeated the main results with an alternate measure of 1901 population, provided by ISTAT, which is based on legal rather than actual population (which is the measure that we collect from the emigration statistics) and with 1881 population, which has the advantage of predating any of our commune-level emigration data. Our results are almost identical, in part because most of our regressions include unit fixed effects that absorb the time-invariant population measure that we use as the denominator. Due to this similarity, we report results only with the preferred population measure coming from the emigration statistics.

[39] Prior studies of Italian migration and development focus on province-level data (e.g., Federico, Nuvolari, and Vasta 2019; Hatton and Williamson 1998).

4.2 Summary Statistics

Table 1 summarizes our emigration data. Column (1) presents data at the district level for all district-years. Columns (2)–(6) present data at the commune level, with column (2) presenting data for all commune-years and columns (3)–(6) limiting the data to various time periods. The first four rows of this Table describe data on average emigration rates, showing the increase in aggregate emigration rates over time, in particular for migration to North America and Europe. Little increase over time is evident in the average migration rate to South America, suggesting that we observe the South American migration when it was already well developed (see also Figure 4). The rest of this Table focuses on extensive margins of migration—whether the emigration rate reached a minimal threshold that we consider to constitute mass migration, and whether there was any emigration.[40] The extensive margins show similar patterns to the emigration rates. There is a substantial increase around 1900 in the fraction of commune-years reaching this threshold to North America and Europe, with no similar increase in migration to South America. By the last period, over 40 percent of commune-years exhibit mass migration to North America, and similarly over 40 percent exhibit mass migration to Europe. Only about one-fifth of commune-years produced mass migration to South America. There is a surge, however, in the extensive margin around 1901, in part due to the diminishing tendency to aggregate migration figures for low-migration communes. By the end of the study period, nearly 97 percent of commune-years show some emigration.

These patterns can also be seen in Figure 4, which show a similar pattern at the aggregate level using the province-destination data described above. Panel (b) shows that the large increase in aggregate emigration around 1900, as shown in panel (a), was largely out of the south, which was, as shown in Panel (c), the primary source of emigration to North America. Panel (d) describes the migration to South America, which drew in much more equal proportions from the various Italian regions, and its trend appears to have already reached saturation by the mid-1880s. As might be expected based on geographic proximity, Italian migration to Europe was, as shown in Panel (e), primarily a Northern phenomenon.

Table 2 presents summary statistics for time-invariant variables. In the average district, about half of male labor was employed in agriculture, and just over a fifth in industry, with greater industrial employment in the north. Slightly over half of adult males were literate, with greater literacy in the north. Finally, about a third of the typical district was under age 15. The average commune had an elevation of about 400 meters and was about 10 kilometers from the nearest rail line, with a greater distance from rail in the south.

Table 2 also presents average distance to epicenters of emigration. We define an epicenter of emigration

[40] As mentioned above, this must be treated with caution prior to 1904.

to North America as a district that had an average annual emigration rate of at least 0.001 to North America over the period 1876–1883 (i.e., after the district data begin but before the commune data became available), and which did not have a neighboring district with a greater average annual emigration rate to North American in this period. This criterion defines six epicenter districts, marked in Figure 6: Sala Consilina in Salerno, Isernia in Campobasso, Corleone in Palermo, Chiavari in Genova, Albenga in Genova, and Napoli in Napoli.[41] For South America, we devise a similar definition, though the fact that we begin to observe the South America-bound migration when it was already well developed requires that the precise definition be different. Specifically, we define these epicenters, marked in Appendix Figure A.5, as those with an average annual emigration rate of at least 0.005 over the period 1876–1883 with no adjacent district with a greater average rate. The epicenters are Lagonegro in Potenza, Chiavari in Genoa, Asiago in Vicenza, and Gemona in Udine. We do not define epicenters similarly for Europe; instead, we exploit the fact that the Europe-bound emigration was greatest in the districts sharing a land border with neighboring European countries and use the distance to this border as the measure of distance to the epicenter. For migration to North America and South America, we compute the distance from each commune to the nearest capital commune of an epicenter district. The average commune was about 150 kilometers from an epicenter of North America-bound migration, 170 kilometers from an epicenter of South America-bound migration, and 240 kilometers from the European land border.

5 Patterns

We begin by checking whether each of the four testable predictions laid out in section 3 are present in the data. We focus our analysis on North America-bound migration. As shown above, this is the major transatlantic movement that we observe from its early development through its saturation. We do, however, refer to the migrations to Europe and South America when can illustrate a point. In many cases, we focus our analysis on the half decade as our preferred temporal unit. This approach has the advantage of smoothing out year-to-year fluctuations in annual data and reducing dimensionality to ease visualization while not aggregating to the point that longer-term patterns are obscured.

[41] It is beyond the scope of this paper to determine why these were the initial epicenters of mass migration. But it is notable that all except Isernia are either major ports or adjacent to them, suggesting a role for trade in generating migration.

5.1 Convergence

Gould (1980) and Hatton and Williamson (1998) cite as a key piece of evidence in favor of the diffusion hypothesis the decline over time in the cross sectional variance of emigration rates across Italian regions. Figure 7 examines this at finer levels of geographic disaggregation—the district and the commune—using data on emigration to North America. Our preferred measure of this variance is the coefficient of variation because it is both normalized by scale and can account for cases of zero migration.[42] Using either the district or the commune as the unit of analysis, there is clear evidence of σ-convergence over time, with the coefficient of variation decreasing almost monotonically from around 4 in 1880 to just over 1 in 1910.[43]

This σ-convergence is the product of new areas entering migration while areas already experiencing migration achieved saturation and stabilizing emigration rates. This can be seen in Figure 8, which tests for β-convergence. Specifically, we compute the average annual emigration rate for each district and commune for the periods before and after 1900. We then plot on the horizontal axis the average annual emigration rate prior to 1900. The vertical axis is the relative change around 1900, which is the ratio of average annual emigration rates between the two periods. As the β-convergence prediction implies, there is a strong negative and nearly monotonic relationship between these variables. Areas with the lowest initial emigration rates experienced the greatest growth after 1900, reflecting their entrance into migration. In particular, the average commune in the top quartile of pre-1900 migration experienced approximately a quadrupling of average emigration rates between the pre- and post-1900 periods. Communes in the bottom quartile of pre-1900 migration, on the other hand, experienced a more than 200-fold increase in average emigration rates over the same period.

Importantly, whereas a large fraction of places experienced tenfold or even hundredfold increases in emigration rates, not a single district experienced a reduction. This is a clear evidence that the β-convergence is not a simple case of mean-reversion or of churning of leading and lagging districts due to random shocks. Instead, the average rate of migration in the period prior to 1900 is the effective lower bound for the rates of migration after 1900. This is strongly consistent with the notion of saturation—districts plateaued when they reached their full emigration potential, and tended not to descend from the maximum level they had

[42] An alternative is to use the standard deviation of the natural log of emigration rates, which normalizes by scale but which cannot account for zero migration cases. The results using this measure (which omits cases of zero migration) are qualitatively identical.

[43] Note that due to the tendency to not report the specific migration counts of communes producing little migration in the early years, one might expect to see an exaggeration of the downward trend in cross-commune variation. However, the district totals were almost aways fully reported, and the trends of the decline in variance based on either commune or district data is very similar. This suggests, first, that the problem of omission of low migration measures is not detrimental; and second, that the decline in variance is really occurring across larger units, and not within them. This is consistent with the idea that the arriving tide of migration lifts all boats in the same area.

reached.[44]

A natural concern is that the correlation is spurious (the pre-1900 emigration rate enters positively into the x-axis and negatively into the y-axis); but the presence of a negative correlation between the change and the emigration rate in the *second* period, called the "Falsification Correlation" in the notes to the figures, implies that the pattern is driven by convergence, and is not spurious. If the plotted correlation were spurious, then this "Falsification Correlation" would be positive and of the same absolute magnitude.

Table 3 provides a more detailed test of the β-convergence prediction. Using data at the commune-half decade level, we estimate a regression of the form

$$\log\left(\frac{e_{it}}{e_{it-1}}\right) = \alpha_t + \beta\log(e_{it-1}) + \varepsilon_{it}, \tag{2}$$

where α_t are half-decade fixed effects and e_{it} is the emigration rate to North America in half decade t. This repeats the analysis of Figure 8 for six half-decade periods rather than for two extended ones. Column (1) performs this estimation with no added controls, finding that there was indeed a negative and statistically significant relationship, indicating faster growth by the laggards.[45] Column (2) of Table 3 repeats the same estimation with the addition of a variety of control variables—half decade-specific functions of latitude, longitude, and the other variables described in Table 2. Results are qualitatively similar.

In columns (3) and (4), we take a direct approach to addressing the possibility that the correlation between the change in migration and its lag is purely mechanical. Specifically, we use half decade-specific functions of distance from the nearest epicenter of migration to North America. We discuss this strategy in detail below, where it forms the main part of the identification strategy in our test of the spatial contagion mechanism in section 6. Briefly, it exploits the fact that, as documented above, emigration spread out from the initial epicenters, regardless of the underlying reasons for this spread. This removes the problem of a mechanical negative correlation between the dependent and independent variables. Using this instrumental variables approach reveals a strong first stage, as evidenced by the high first-stage F statistics, and shows that the β-convergence results are robust to instrumentation.

[44] On the other hand, this is incompatible with the notion of an inverse-U curve trend of migration (Hatton and Williamson 1998), though it may be that the downward-sloping part of an inverse-U curve would have occurred after our study period had World War I and American policy not interrupted mass migration from Italy.

[45] Again, there is an obvious concern that this correlation may be spurious, driven by the appearance of lagged emigration on both the right- and left-hand sides of equation (2). To address it, we first include in each column under "Falsification" the coefficient from repeating the estimation of equation (2) with $\log(e_{it})$ as the regressor. In column (1), the fact that this coefficient is positive indicates that there is some merit to this concern; but the fact that we can clearly rule out that the absolute value of the falsification coefficient is of the same magnitude as the coefficient of interest suggests that even if the coefficient is biased in this way, the convergence is not spurious.

5.2 S-Shaped Local Time Series

The convergence patterns documented in section 5.1 can arise from a variety of different time patterns of local migration. Prediction 2 is that this convergence should arise from S-shaped local time series. Moreover, these S-curves should be roughly ordered by distance from the epicenters, with closer areas surging earlier.

Figure 9 plots the average annual emigration rate by septile of distance from epicenters of emigration to North America, restricting distance to under 250 kilometers for communes, or under 300 kilometers for districts. Whether the commune or the district is used as the unit of observation, S-shapes are clearly evident in the time series. The rates of emigration in the earlier periods were low, and lower in areas further from the epicenters. When the rising portion of the S was encountered, the closer areas experienced only small increases while those at a greater distance experienced greater increases. Saturation in most places was then achieved after 1905, with a smaller gap in the emigration rates relative to distance. By the end of the study period, the advantage of the leader areas in migration was considerably reduced.[46]

5.3 Frontier Effect

The basic descriptive evidence supporting this prediction is presented in the maps of Figure 1, described in section 1, at the district level, and in Figure 10 at the commune level.[47] At both levels of geographic disaggregation, there is strong visual evidence of a spatial diffusion of areas of high emigration rates. Both Figures divide communes or districts according to quintiles of the distribution emigration to North America in the period 1911–1914, and the gradual expansion of the areas of high migration from the epicenter areas to nearby areas over time is clear. In Sicily, for instance, it appears that migration expanded from Palermo in the northwest, reaching Messina in the east first and with Siracusa in the southeast trailing furthest behind. Equally evident is the expansion of migration from both Campania and Messina into Calabria, or outward from Liguria and Toscana further north.

To present these patterns in a more compact and formal way, Figure 11 plots local polynomial regressions of average annual emigration rates to North America against distance from the epicenters, separately for each half decade. Epicenter districts are excluded (explaining the lack of a peak at 0). Both at the district and the commune level, places closer to the epicenters are emigration leaders throughout the study period,

[46] Analogous curves for the other major destinations exhibit different patterns, as shown in Appendix Figure A.6 for South America and A.7 for Europe. There are no S-shaped trends for South America. This reflects the fact that, as shown above, we begin to observe the South American migration when it was already nearly mature—that is, after the surging portion of the S. For Europe, S-curves are evident, but distance from the epicenter is still a strong predictor of emigration rates at saturation, likely because proximity to the epicenter played the additional role of reducing migration costs.

[47] Appendix Figures A.8 and A.9 present analogous commune-level figures for migration to South America and Europe, respectively.

with considerable declines from about 50 to 100 kilometers from the epicenters. There is also a gradual incorporation of more distant areas into emigration, as is particularly evident for the district data in Panel 11(a)—the left extreme of the curve rises gradually over time, incorporating slightly more distant areas, before surging after 1900. Similar patterns are evident in Appendix Figure A.10, which focuses on the extensive margin by relating the probability of achieving an average annual emigration rate of at least 5 per thousand to the distance from these epicenters.

In Figure 12, we test whether this pattern is statistically significant and robust to controlling for local characteristics. Panel (a) presents the coefficients from a regression of the log emigration rate on the log distance from epicenter interacted with half-decade indicators and half decade fixed effects. The coefficients in the early part of the study period are strongly negative, indicating that being further from the epicenters of North America-bound migration was associated with less migration. In 1891–1895, the elasticity of emigration rates with respect to distance from the epicenter is -1.2. There is then a nearly monotonic decline in the magnitude of these coefficients, consistent with a spread of migration away from these epicenters and a gradual removal of an obstacle to migration over time. In Panel (b) we repeat the same analysis controlling for half decade-specific functions of latitude, longitude, and the other variables described in Table 2. The results are largely unchanged.

Figure 13 performs the same exercise as in Figure 11, but instead of distance from the epicenters, the x-axis in this Figure is lagged distance from what we define to be the *frontier of mass migration*—the nearest district that had ever achieved a rate of emigration to North America of at least 5 per thousand by a particular half decade. This Figure is restricted to districts that had not yet achieved this rate of emigration in any previous half decade. Thus, unlike in Figure 11, the composition of the sample and the location of a given commune or district along the x-axis changes between half-decades. Whereas Figure 11 shows the overall expansion from the epicenters, Figure 13 focuses on the moving frontier. This Figure provides compelling evidence of the frontier effect. There is clear evidence of a decline in emigration rates with lagged distance from the frontier when pooling the observations from all time periods. These patterns are also evident for each individual time period separately. Importantly, even the surges in migration in the 1901–1905 and 1881–1885 half decades primarily incorporated communes and districts within about 100 kilometers of the frontier. Moreover, the general absence of high emigration rates at distances beyond 200 kilometers from the frontier suggests that proximity to the expanding frontier of mass migration is in fact *necessary* for entrance into mass migration.[48]

[48]Figure A.11 presents analogous evidence for the extensive margin.

5.4 Correlated Destinations

The data on emigration by destination at the province level enable us to test the correlated destinations prediction.[49] Specifically, for every province pair ij for every half decade t in our data, we compute a dissimilarity index between the emigration share vectors as

$$V_{ijt} = \frac{1}{2}\sum_{k=1}^{25} |m_{ikt} - m_{jkt}| \, ,$$

where k indexes the 25 consistently defined destination countriess in the province-destination data and m_{ikt} is the fraction of all emigrants from province i who traveled to destination k in half decade t.[50] The dissimilarity index has the convenient feature of being interpretable as the fraction of emigration from province i that would have to be rerouted to match the destination distribution of province j (or vice versa). We then run a regression of the form

$$V_{ijt} = \alpha_t + \beta \log(d_{ij}) + \varepsilon_{ijt},$$

where α_t are half-decade fixed effects and d_{ij} is the distance between the capitals of provinces i and j.

We present the results of this estimation in Table 4. Columns (1) and (2) use all destinations and all years, with and without control variables, respectively. The controls are absolute differences between provinces in their 1881 agricultural employment share, industrial employment share, literacy rate, and fraction under age 15. Whether or not controls are included, there is a strong positive association between the distance between provinces and the dissimilarity index of their emigration shares, indicating that provinces that were closer together were more similar in their distribution of destinations. The coefficient in column (2) can be interpreted as showing that a 10 percent increase in distance between provinces is associated with a 1.2 percentage point increase in the fraction of migrants who would have to be rerouted for the destination distributions of the provinces to match, relative to a mean dissimilarity index in the sample of 0.60. Column (3) repeats the same analysis, but computes the dissimilarity index including only migration to major destinations.[51] Column (4) includes all other destinations. Results are similar, though the relationship with distance is smaller for the minor destinations. Finally, columns (5) and (6) return to the use of emigration to all destinations, but divide the sample at 1900. Similar results are reached for both periods, though somewhat surprisingly, the coefficient is stronger for the post-1900 period.[52]

[49] Additional suggestive evidence is provided in Online Appendix E.

[50] Appendix Figure A.12 shows an example, mapping the dissimilarity index of each province relative to the province of Palermo over the whole study period.

[51] These are described in the notes to Table 4.

[52] This is surprising because one would expect that as networks diffuse, the options for migration would become more similar

5.5 Discussion

The exercises in this section provide strong new descriptive evidence that the testable predictions of the diffusion hypothesis held in Italy between 1876 and 1920. Some of this evidence, such as the results on convergence in section 5.1, mirror analyses of Gould (1980) and Hatton and Williamson (1998), but at a finer level of geographic disaggregation. Others are entirely novel. The most important novel result is the strong evidence for the existence of spatial expansion and the frontier effect. In this regard, our highly detailed data enable us to provide a test of the crucial inter-communal diffusion prediction of the diffusion hypothesis, whereas prior studies have focused on intra-location diffusion (e.g., persistence). This prediction is particularly important because it enables us to distinguish between the diffusion hypothesis and the internalist view. Whereas Hatton and Williamson (1998, p. 99) point out that evidence on convergence and S-shaped time series "might be explained in other ways" besides diffusion (a critique equally applicable to correlated destinations), spatial expansion of migration is more difficult to rationalize in other ways.

To be sure, such spatial expansion could be driven by the spatial diffusion of some other factor, such as industrialization, that causes migration. For example, it may be that industrialization or demographic pressures were the real cause of migration, and that the diffusion of migration simply reflected the spatial diffusion of these factors. Alternatively, it could be that emigration raises the standards of living by removing local population pressures or through remittances. Emigration in one locality might help to remove liquidity constraints in neighboring provinces, such that the combination of two-way causation between emigration and income with the diffusion of positive income shocks would create a spatial diffusion of migration. Such diffusion of another factor appears to be an unlikely cause for diffusion of migration, particularly given the lack of evidence of industrial or demographic development in southern Italy at the time of the surge of migration, as discussed in section 2. Nevertheless, we treat this as a possible alternative explanation that we tackle in section 6.

across the country, making distance less important as a cause for dissimilarity. Appendix Figure A.13 shows a bin scatter plot of dissimilarity indices against distance for three select half decades representing the beginning, middle, and end of the study period. The positive correlation documented in Table 4 is clearly apparent. Indeed, it appears to be present throughout the range of distance, indicating that it is not simply the outcome of regional differences. Unlike in columns (5) and (6) of Table 4, there is also evidence of convergence, with the shift downward of the dissimilarity index over time, which is consistent with the scenario of simultaneous spread of all networks throughout the country, making destination distributions more similar over time, conditional on distance.

6 The Spatial Contagion Mechanism

Our goal in this section is to strengthen the evidence indicating that the spatial diffusion documented in section 5.3 was the product of social connections transmitting the migration technology. One important concern is that the spatial diffusion of migration was in fact a reflection of the spatial diffusion of some other factor driving migration. At the outset, this alternative hypothesis appears difficult to maintain. First, as discussed in section 2, the broad patterns of the geographic evolution of the Italian migration do not seem to have followed the lines of the modernization of the Italian economy. Second, this idea must be reconciled with the fact that the three main movements that we observe, to North America, South America, and Europe, followed very different paths. It is hard to imagine how any local factor causing emigration will be so pronouncedly specific to a destination. Nonetheless, this is a possibility that must be formally considered.

More specifically, we seek to show that the patterns of spatial expansion and the frontier effect are the product of a mechanism in which one commune's emigration was causally affected by the emigration of its neighbors—what we call the *spatial contagion mechanism.* We begin by developing a regression strategy that operationalizes the empirical testing of this mechanism and show that the descriptive patterns that this estimation reveals in the data are consistent with such a causal effect. We then introduce an instrumental variables strategy to arrive at results that can plausibly be given a causal interpretation. The instrumental variables results show that the geographic expansion of mass migration followed social channels and therefore that social connections were likely to be the key factors determining the spread of mass migration.

6.1 Empirical Strategy

As in section 5, we focus on mass migration to North America. Our baseline estimation equation is

$$\log(e_{it}) = \alpha_i + \alpha_t + \beta \log(e_{\neg it-1}) + \gamma \log(e_{it-1}) + \mathbf{x}_i' \delta_t + \varepsilon_{it}. \qquad (3)$$

In equation (3), e_{it} is the rate of emigration from commune i in period t, constructed as in equation (1); α_i and α_t are commune and period fixed effects; e_{it-1} is a one-period lag of e_{it}; $\mathbf{x}_i$ is a vector of commune-level controls; and δ_t is a period-specific vector of coefficients. The main variable of interest in equation (3) is $e_{\neg it-1}$, which we refer to as *lagged emigration exposure.* This is a lag of a measure of a commune's exposure to emigration that is determined by the emigration rates of all other communes to that destination, with greater weight exerted by nearer and more populous communes. Specifically, we define this object as

$$e_{\neg it} = \frac{\sum_{j\neq i} e_{jt} N_j d_{ij}^{\theta}}{\sum_{j\neq i} N_j d_{ij}^{\theta}}, \tag{4}$$

where N_j is the population of commune j, d_{ij} is the distance between communes i and j, and θ is the rate at which the influence of other communes' j emigration rates on that of commune i decays over distance. A value of $\theta < 0$ implies that more distant communes have a smaller impact than do nearer communes. We use $\theta = -2.28$ for data at the commune-year level and $\theta = -2.38$ for data at the commune-half decade level.[53] We explore robustness of our results to different values of θ in Online Appendix F.

The denominator of equation (4) is equivalent to Donaldson and Hornbeck's (2016) market access statistic, a measure of effective proximity to population. The numerator is analogous to market access but for emigrants rather than population. The measure $e_{\neg it}$ is thus a population- and distance-weighted average emigration rate in the neighborhood of commune i. The scale of $e_{\neg it}$ is that of an emigration rate, and it can be thought of as the probability that any out-of-commune contact that a person may have migrated in the previous period.[54] The coefficient β in equation (3) can be interpreted as an elasticity of emigration with respect to emigration exposure.

The specification in equation (3) is derived from that used by Comin, Dmitriev, and Rossi-Hansberg (2012) to study the spatial diffusion of technology. One of the main departures that we make from their method is the inclusion of the lag of own emigration. We do this to control for the intra-communal network of previous migrants. We also differ from the approach of Comin, Dmitriev, and Rossi-Hansberg (2012) in that we focus on the impact of *lagged* emigration exposure on a commune's emigration rate, rather than on a contemporary impact. This is motivated by the time it takes for any migrant to be able to support others who would follow him, and it also helps mitigate a potential reflection problem (Manski 1993). Throughout our analysis, we cluster standard errors at the level of the province.[55]

[53] In principle, we could estimate equation (3) by OLS, but the parameter θ enters non-linearly through the measure of lagged exposure. Estimating the complete set of parameters non-linearly in multiple specifications, including θ, is not feasible due to the computational burden of this estimation. We approach this challenge by following the procedure implemented by Donaldson and Hornbeck (2016) and Zimran (2019). Specifically, we estimate a simplified version of equation (3), excluding $\mathbf{x}_i$, by non-linear least squares, and use the implied value of θ throughout the remaining analysis. The resulting values of θ imply that a commune twice as far away exerts approximately a fifth the influence on the emigration exposure measure.

[54] The measure $e_{\neg it}$ satisfies two desirable conditions. The first is that it is robust to splitting communes. If we focused only on proximity weighted emigration rates of communes, splitting a single commune into two with the same emigration rate would increase the weight of that commune in the computation of equation (4). By weighting by a commune's population, we avoid this concern. The second is that it is robust to increases in population. If we focused only on the number of emigrants, then doubling the population near a commune (with an accompanying doubling of the emigration rate) would double an individual's emigration exposure. To reflect the limited number of connections that a person can have, our measure is robust to population density, which would have no impact on the proximity-weighted emigration rate. The implied assumption is that the number of links that any individual has outside of his commune is fixed and independent of the population density in the neighborhood.

[55] We have also repeated our main results with Conley standard errors and a distance cutoff of 200 kilometers. Results in

6.2 OLS Results

Table 5 presents the results of estimating equation (3) by OLS for migration to North America with the unit of observation at the commune-year. In column (1), we present results excluding commune fixed effects and including only province and period fixed effects. In the absence of commune fixed effects, these results can tell us little more than that there is a correlation between the emigration rates of nearby places, but they are useful in establishing some basic patterns. The coefficient on lagged own emigration indicates substantial persistence in emigration rates over time within a commune, with estimated elasticity of 0.480. Such persistence is an important prediction of the diffusion hypothesis, and is found in most empirical studies of migration. But neighbors' lagged migration is almost equally important, with an elasticity of 0.337. Column (2) adds the basic controls in levels, and some of the coefficients on the control variables indicate the importance of the push factors emphasized by the internalist view. For example, the share of employment in industry, the literacy rate, and the share of the district population under 15 are positively and significantly correlated with emigration, as expected. The share of employment in agriculture and elevation are also positively correlated with emigration. In fact, elevation, which in post-unification Italy was typically associated with an inefficient rural economy, appears to be a first-order determinant of emigration. Column (3) adds period-specific controls (i.e., interactions of these controls with period fixed effects) and the results are qualitatively unchanged.

Column (4) estimates a basic form of equation (3) with the commune fixed effects but without any controls. The commune fixed effects help to ensure that the apparent contagion is not merely the product of persistently high emigration rates from some areas. The inclusion of commune fixed effects naturally diminishes the correlation between current and lagged own emigration, as the commune fixed effects capture time-invariant factors driving persistence in emigration rates. On the other hand, it increases the estimated elasticity with respect to lagged neighbors' migration to 0.422. This is a strong indication that this elasticity does not simply reflect persistent similarity between nearby places. Column (5) adds the controls consisting of initial commune- and district-level characteristics interacted with year fixed effects, with no meaningful impact on the results. Column (6) also adds province-year fixed effects.[56] The magnitude of the estimated elasticity is decreased considerably by this inclusion, but the basic patterns remain.

Column (7) adds lagged measures of emigration exposure for both South America and Europe. If the positive correlation between emigration exposure and emigration were driven by the spatial contagion of

this case are very similar.

[56]Because the migration-by-destination data vary at the province-year level, this specification reduces the concern that our results are driven by variation in migration at province borders.

underlying push factors or by the effect of emigration on the local economy, rather than through the friends and relatives effect, then we would expect to observe a positive relationship between emigration and emigration exposure, regardless of the destination. Consistent with the predictions of the diffusion hypothesis, only emigration exposure to North America enters significantly; the coefficients on exposure to emigration to other destinations are reassuringly very close to zero and statistically insignificant.[57] Finally, column (8) tests another prediction of the diffusion hypothesis—that emigration exposure should be less important when network links are more saturated. In particular, we interact emigration exposure with own lagged emigration, and find that the interaction is negative and statistically significant, as expected. That is, the greater is the own lagged emigration of the commune, the less the impact of emigration exposure on its emigration.[58]

In Appendix Table A.1, we repeat the same regressions as in Table 5, this time with the unit of analysis being the commune-half decade rather than the commune-year. The results are almost identical, with one exception. When adding commune fixed effects, the coefficient of lagged own emigration diminishes by half in the annual regressions, and by an order of magnitude in the half-decade regressions. This difference in the impact of including commune fixed effects is to be expected. Lagged emigration captures serial correlation in emigration within communes, conditional on the commune-fixed effects. Naturally, serial correlation will be weaker across longer periods. However, the same does not occur to the lagged migration exposure, which remains as strong as in the annual regressions. This suggests that, unlike the own lag of migration, lagged exposure does not capture a simple process of spatial correlation that decays over time, but an actual persistent effect.

In both Tables 5 and A.1, after including for commune-fixed effects, the elasticity with respect to lagged neighbors' emigration is stronger than that of lagged own migration (and as discussed above, much more so in the half-decade regressions in Appendix Table A.1). While this may simply be because the factors generating persistence in migration rates are time invariant and captured by the commune fixed effects, it may also seem suspicious, suggesting that the combined effects of inter-communal connections is greater than of that of intra-communal connections. There are two possible reasons that could make this difference spurious. The first is that lagged emigration exposure, as an average of the data for many communes, suffers

[57] In fact, the point estimates are negative, suggesting substitution between destinations. This evidence is limited by the fact that the emigration-by-destination data are observed only at the province level. This result also does not rule out a scenario in which general push factors diffuse and drive migration, while the friends and relatives effects simply steers the migrants to the same destination but does not, in and of itself, increase their numbers. While theoretically plausible, this mechanism is unlikely to be true given the lack of evidence of a diffusion of push factors in southern Italy at the time of the increase of emigration (see section 2). It is also a concern that is directly addressed in our IV estimation below.

[58] In Appendix Figure A.14, we show the results of including various lags of own emigration and emigration exposure. In both cases, the first lag is most impactful.

from less measurement error than the data based on only a single commune, and that this creates stronger attenuation bias for the own lagged emigration than for lagged emigration exposure. The more concerning explanation is the possible presence of Nickell (1981) bias in this dynamic panel setting, which would tend to asymptotically bias the coefficient on lagged own emigration downward and the coefficient on lagged emigration exposure—our coefficient of interest—upward.[59] The average of 18 observations per commune in the commune-year data is relatively large, reducing this concern; but we cannot completely rule out this bias. We can, however, bound the potential impact of this bias on our results. The results of this exercise, presented in Online Appendix G, reassure us that our conclusions are not the spurious product of this bias.

6.3 Instrumental Variables Strategy

Whether transmitted by social connections or not, we have documented in section 5.3 that the area of high migration expanded geographically from the epicenters. This meant that in a certain period, some communes would experience a surge in emigration as a result of their proximity to the ever-expanding frontier of mass migration. Our instrumental variables strategy asks what happened to a commune's emigration when other communes near it experienced a surge in their migration due to this expansion.[60]

Figure 14 illustrates the intuition of such a strategy. Communes A and B are equidistant from the epicenter E and identical in every way except for the distribution of the neighboring population. Commune A has a greater share of its neighboring population (the black dots) closer to the epicenter relative to commune B. When the mass migration frontier reaches communes C and D (the population between epicenter E and communes A and B) and migration from communes C and D (but not from communes F and G) surges, residents of commune A are more likely to be linked to a migrant than are residents of commune B. Under the diffusion hypothesis, commune A is then likely to experience a greater migration surge in the next period than is commune B. Our instrumental variables strategy asks whether this is in fact true. The key assumption underlying this strategy is that the distribution of neighbors' population relative to epicenters has no impact on emigration except through its effect on exposure to the expanding frontier of mass migration.[61] For instance, if industrial development spread spatially from epicenters over time *and* if

[59]The estimates are unbiased but inconsistent under Nickell (1981) bias when the number of communes but not time periods increases.

[60]In principle, an alternative approach might be to look for specific shocks that induce emigration and to determine the impact of these shocks on neighboring communes. However, such shocks are likely to have considerable spillover effects over space which would interfere with our ability to disentangle the impact of the shock and the impact of diffusion. Moreover, one such shock, the 1908 Messina-Reggio Calabria earthquake, also had a much smaller impact on migration than one might expect (Spitzer, Tortorici, and Zimran 2019). Another approach might consist of a network-based pull factor shock (Boustan 2010), but difficulties in making reliable links between passenger records and the US census limits our ability to implement this approach.

[61]With year or half-decade fixed effects and the interaction of distance from epicenters with time fixed effects, the identification

its spread depended on population density, then this would challenge our interpretation of the results. We return to this concern below, though we find it a priori implausible in this context, as discussed above.

We implement this intuition by constructing a measure of lagged emigration exposure that is a function only of neighbors' distance from emigration epicenters. In particular, our estimation proceeds as follows. First, we estimate

$$\log(e_{jt}) = a_j + a_t + \eta_t \log(z_j) + \mathbf{x}_j'\pi_t + u_{jt}, \tag{5}$$

where a_j and a_t are commune and period fixed effects, z_j is distance from the nearest epicenter of migration to North America, $\mathbf{x}_j$ is a vector of controls, and η_t and π_t are period-specific vectors of coefficients. The coefficient of interest is η_t, which captures the correlation of distance from the epicenter with emigration, separately for each period. After estimating equation (5), we compute fitted $\tilde{e}_{jt}$ as

$$\tilde{e}_{jt} = \exp\left[\hat{\eta}_t \log(z_j)\right]. \tag{6}$$

That is, we construct the fitted values using only distance from epicenters, and excluding the fixed effects and controls.[62] We then use $\tilde{e}_{jt}$ to compute

$$\tilde{e}_{\neg it} = \frac{\sum_{j\neq i} \tilde{e}_{jt} N_j d_{ij}^{\theta}}{\sum_{j\neq i} N_j d_{ij}^{\theta}}.^{63} \tag{7}$$

This is an analog of the measure of emigration exposure in equation (4), but because $\tilde{e}_{jt}$ is a function only of commune j's distance from epicenters and the period,[64] $\tilde{e}_{\neg it}$ is only a function of this distance (and naturally also of the population and distance weights). Equation (7) implements the intuition described above in that, conditional on distance to commune i, it gives greater weight to communes with greater population and because emigration was greater in communes closer to the epicenter, $\tilde{e}_{jt}$ is larger in these communes. Finally, we estimate

$$\log(e_{it}) = \alpha_i + \alpha_t + \beta \log(e_{\neg it-1}) + \gamma \log(e_{it-1}) + \rho_t \log(z_i) + \mathbf{x}_i'\delta_t + \varepsilon_{it} \tag{8}$$

assumption is that there is no such relationship that varies over time.

[62] In principle, we could compute equation (6) using fitted values from equation (5) with fixed effects and controls, but we wish to ensure that any correlations uncovered through this process are purely the products of the instrument and not due to, for instance, correlated fixed effects that are not sufficiently controlled for in the non-linear estimation below with linear fixed effects. Including the fixed effects and the controls in the predicted values of equation (6) strengthens our results.

[63] In computing the fitted values, we must assign $\tilde{e}_{jt} = 0$ for all j for which $e_{jt} = 0$ because they must be omitted in estimating equation (5).

[64] It is a function of period because the estimation of $\eta_t \log(z_j)$ is accomplished by interacting $\log(z_j)$ with period fixed effects.

by instrumental variables using $\log(\tilde{e}_{\neg it-1})$ as an instrument for $\log(e_{\neg it-1})$.[65] Equation (8) is an analog of equation (3), but adds a control for commune i's own distance from the nearest epicenter (z_i). In essence, this approach uses neighbors' distance from epicenters as an instrument for neighbors' emigration. The control for own distance from epicenter ensures that the analysis compares only communes equidistant from epicenters as in the example of Figure 14.

It is important to note that, although our instrumental variables strategy takes as given and utilizes the spatial expansion of mass migration, it does not presuppose what it is meant to test—that the spatial expansion is the result of a causal effect of neighbors' migration.[66] Whatever the cause of the negative correlation between emigration and distance from the epicenter, the instrument satisfies the exclusion restriction so long as the distribution of the population around a commune is random conditional on its distance from the epicenter. Thus, while our approach takes the expansion of mass migration as given, it tests whether this expansion proceeded along spatial lines.

6.4 Instrumental Variables Results

Before moving to the instrumental variables regressions, Table 6 presents the results of balancing tests that indirectly assess the validity of the exclusion restriction. Specifically, we test whether the distribution of population around a commune, captured by the population- and distance-weighted neighbors' log distance from epicenters, is correlated with a commune's own characteristics, conditional on a commune's own log distance from an epicenter. In particular, the measure of neighbors' weighted log distance from the epicenter is

$$\frac{\sum_{j\neq i} \log(z_j) N_j d_{ij}^{\theta}}{\sum_{j\neq i} N_j d_{ij}^{\theta}},$$

which is an analog of equation (4) with emigration rates replaced with log distance from an epicenter. We then regress a variety of 1881 commune characteristics that may have affected migration on this population-weighted measure of neighbors' log distance, controlling for own log distance and for province fixed effects. Reassuringly, there is little evidence of any relationship between these measures and that of the population distribution. Only elevation has a statistically significant relationship with neighbors' distance, with higher elevation areas tending to have neighbors further from the epicenters. This correlation is not directly concerning, as we control for elevation in our analysis, as above, and the other measures are uncorrelated

[65] It is not necessary to adjust standard errors for the use of this generated instrument (Wooldridge 2002, pp. 116–117).

[66] Indeed, it need not have been the case that the spatial expansion of migration documented in section 5.3 proceeded along social lines, as the diffusion hypothesis predicts. In the example of Figure 14, it could be the case that communes A and B, despite their different "social" exposure to neighbors' migration, experienced no different response to this increase in emigration exposure.

with distance. But in any event, such a relationship would likely work against our results, as these higher elevation communes tended to have higher saturation emigration rates.[67]

The results of the main instrumental variables estimation are described in Table 7. Column (1) presents results using commune-year data and no controls. Panel A shows the results of estimating equation (8). The results of this estimation are very similar to those of column (4) of Table 5, which performs the analogous estimation without the added control for own distance from epicenter. This is the OLS version of the equation that will be estimated by instrumental variables.

Panels B and C develop the instrument. In Panel B, we estimate equation (5) and present the F-statistic for the test of joint significance of the interaction of log distance from epicenters with year fixed effects. This is not the typical Staiger and Stock (1997) test of instrument validity (because this is not the first stage of the instrumental variables regression). But the fact that the F-statistic is relatively large at 17.76 suggests that distance from the epicenter can explain a considerable fraction of the variation in emigration rates. We present only this F-statistic because there are too many coefficients to present compactly. Figure 12 shows that indeed the pattern of these coefficients fits with our expectations (using data at the commune-half decade level), with distance from epicenter having a smaller impact on emigration over time as the mass migration frontier expands. Panel C presents the results of the first stage of the instrumental variables estimation, in which the dependent variable is the log of lagged emigration exposure $e_{\neg it-1}$ and the main regressor of interest is the log of lagged instrumented exposure $\tilde{e}_{\neg it-1}$, which is constructed from the estimates of Panel B as in equation (7). The strongly statistically significant coefficient on lagged instrumented exposure indicates that the constructed instrument satisfies the relevance requirement.

Finally, Panel D presents the results of the instrumental variables estimation of equation (8). Column (1), which uses commune-year data without controls, shows that the positive elasticity between lagged emigration exposure and a commune's emigration is robust to instrumentation, though its magnitude is somewhat reduced and it is only marginally statistically significant. Column (2) adds controls, and column (3) also adds province-year fixed effects. In both of these columns, the instrument has a strong first stage (Panel C) and the estimated effect of lagged emigration exposure on emigration is positive, statistically significant, and of approximately the same magnitude as the OLS results of Table 5, though the stage 0 F-statistic is somewhat small in column (3). Columns (4)–(6) repeat the same estimation with commune-half decade data, finding similar results. As in the OLS regressions, moving from commune-year to commune-half decade data reduces the magnitude of the coefficient on the own lag of migration while increasing that of

[67] Online Appendix H presents an additional balancing test.

the lagged emigration exposure. The robustness of the positive and statistically significant coefficient of lagged emigration exposure through the instrumentation indicates that the spatial expansion of emigration did indeed proceed along social lines. While it still leaves open the possibility that some other factor diffused across space and caused migration, it is difficult to explain why the diffusion of such a variable should have been transmitted by social contacts. As a result of these exercises, we argue that the transmission of the migration technology through space-dependent social contacts is the only plausible explanation for the patterns that emerge.

7 Conclusion

The question of why migration from the European periphery was delayed is one of the fundamental puzzles of the Age of Mass Migration. In this paper, we test an influential explanation for this puzzle that attributes the delay to an absence of chain migration networks, focusing on emigration from post-unification Italy. Using a newly constructed dataset on emigration rates at the commune, district, and province-destination level over the period 1876–1920, we provide two new fundamental pieces of evidence supporting the diffusion hypothesis. First, the four testable predictions of the diffusion hypothesis are confirmed by the data. The most novel result in this regard is that we have documented for the first time the spatial expansion of migration to North America, beginning in the epicenters identified above, and expanding at a rate of less than 100 kilometers per half decade. Second, we find strong evidence that the spatial expansion of migration reflected diffusion along social channels, as envisioned by the diffusion hypothesis, rather than some other mechanism.

Although our results provide strong support for diffusion as a factor shaping Italian emigration, thus forming the strongest empirical basis to date for the diffusion hypothesis, they still fall short of refuting the internalist view or proving that the delay in Italian mass migration was in fact due to the lack of connections to prior migrants. Doing so requires taking a structural approach, which would enable the evaluation of counterfactual scenarios for the evolution of Italian emigration, a task that we leave for future research.[68] Nonetheless, the accumulation of evidence that a spatial diffusion process operated, passed through social lines, and generated time trends and convergence patterns consistent with the diffusion hypothesis is strong evidence that the forces envisioned by the diffusion hypothesis shaped the patterns of Italian mass migration.

Besides the relevance of these results to our understanding of the Age of Mass Migration specifically and

[68]In practice, this requires estimating the formal model in Online Appendix B and using the fundamental parameters to simulate scenarios that are based on different distributions of initial exposure to migration.

the economics of migration generally, this analysis provides important insights for contemporary international migration. The overwhelming majority of immigrants to the US currently come from middle-income countries in Central and South America, or East and South Asia. The world's poorest region, sub-Saharan Africa, accounts for only about one in 25 immigrants (Radford 2019), surprisingly little in light of the continent's endemic poverty, political instability, and large refugee populations. Some of the potential of African migration has already been made apparent by the recent European Migrant Crisis, as tens of thousands of migrants from sub-Saharan Africa sought refuge on Europe's southern shores. Following the populist backlash in Europe and the imposition of measures to restrict trans-Mediterranean migration, the numbers have declined from their heights in 2015–2016; but there is little doubt that the potential for mass migration from sub-Saharan Africa still exists, and that millions more would consider that option if it were available. Though small, there are signs that the number of Africans seeking shelter and a future in the US is rising quickly (Feinberg and Petrie 2019; Solomon 2019). The role of networks appears crucial in this migration (Solomon 2019). Most of the refugees have direct family or social connections with previous migrants already in the US, who have supported and often triggered their journeys, and many have indicated that they in turn will support the migration of their friends and relatives still in Africa: "As more Africans learn from relatives and friends who have made the trip that crossing Latin America to the United States is tough but not impossible, more are making the journey, and in turn are helping others follow in their footsteps . . . Those who reach the United States often send advice back home, helping make the journey easier for others" (Solomon 2019)

Why have sub-Saharan countries been so heavily under-represented thus far in US immigration, despite having some of the most powerful push factors for migration? Is the recent increase in numbers an indication of an impending surge in African migration, or rather a temporary reaction to the European change in policy? More broadly, what does it mean when a region such as sub-Saharan Africa, that seems to have all of the incentives to produce mass migration to the US and other Western countries, has failed to do so for decades? Was it because extreme poverty is, in itself, a factor that suppresses emigration? Or has a process of diffusion of migration begun in sub-Saharan African that will generate a surge in the future? We believe that the answers to these questions can be informed by the history of US immigration, and that the delay and eventual beginning of migration from the "new" source countries that we study in this paper is arguably the best precedent available for the surprisingly late surge in migration from a region whose residents stand to benefit tremendously from it.[69]

[69] We recognize that the lack of open borders in the modern case is an important difference from the Age of Mass Migration. Nonetheless, the recent increase in African migration to the US suggests that even under these restrictions, the potential level

In particular, if the diffusion hypothesis is correct, it may mean that poor developing countries that currently produce little migration are not abstaining from it due to economic underdevelopment or liquidity traps, as the standard push-pull model of migration predicts, but primarily due to the lack of links to previous immigrants. Once those links are formed, migration might rapidly accelerate. Importantly, the standard push-pull model of migration predicts that such a surge would typically be the product of a process of transformative economic change. The diffusion hypothesis on the other hand predicts that such swings in migration rates might occur even in a stable economy without any discernible economic trigger. The diffusion hypothesis also has implications for the evolution of already-mature migration flows. It predicts that what appears to be a continuously rising trend in migration from a source country as a whole is in fact a series of small regions taking off and quickly reaching a steady high rate of migration (S-shaped local migration trends). Thus, the diffusion hypothesis predicts that a steady rising trend of migration from a given country may stabilize when the country reaches saturation (as we believe happened in Italy during the first decade of the twentieth century), again without any economic trigger for a change in the trend. These insights are particularly important in the context of the rising migration of African refugees to the United States. Viewing this migration through the lens of diffusion can help to better understand its ongoing development and possible future course.

of migration has not been achieved.

References

Abramitzky, Ran and Leah Platt Boustan (2017). "Immigration in American Economic History." *Journal of Economic Literature* 55:4, pp. 1311–1345.

Ardeni, Pier Giorgio and Andrea Gentili (2014). "Revisiting Italian Emigration before the Great War: A Test of the Standard Economic Model." *European Review of Economic History* 18, pp. 452–471.

Baines, Dudley (1995). *Emigration from Europe 1815–1930.* Cambridge: Cambridge University Press.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo (2013). "The Making of Modern America: Migratory Flows in the Age of Mass Migration." *Journal of Development Economics* 102, pp. 23–47.

Barde, Robert, Susan B. Carter, and Richard Sutch (2006). "Table Ad106–120: Immigrants, by country of last residence—Europe, 1820–1997." In *Historical Statistics of the United States.* Susan B. Carter, Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright (ed.). Cambridge: Cambridge University Press, pp. 1.560–1.563.

Bass, Frank M. (1969). "A New Product Growth for Model Consumer Durables." *Management Science* 15:5, pp. 215–227.

Battery Conservancy (2009). "Castle Garden Database." Retrieved from `http://www.castlegarden.org`.

Bernoulli, Daniel (1776). "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir." In *Mémoires de mathematique et de physique.* Academie Royale des Sciences.

Boustan, Leah Platt (2010). "Was Postwar Suburbanization 'White Flight'? Evidence from the Black Migration." *Quarterly Journal of Economics* 125:1, pp. 417–443.

Brandenburg, Broughton (1904). *Imported Americans: The Story of the Experiences of a Disguised American and His Wife Studying the Immigration Question.* New York: Frederick A. Stokes Company.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak (2014). "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh." *Econometrica* 82:5, pp. 1671–1748.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo (2016). "Understanding Booms and Busts in Housing Markets." *Journal of Political Economy* 124:4, pp. 1088–1147.

Ciccarelli, Carlo and Stefano Fenoaltea (2013). "Through the Magnifying Glass: Provincial Aspects of Industrial Growth in Post-Unification Italy." *Economic History Review* 66:1, pp. 57–85.

Ciccarelli, Carlo and Peter Groote (2017). *Italian Railway Development, 1839–1913* [machine-readable database]. ArcGIS Online.

Cohn, Raymond L. (2009). *Mass Migration Under Sail: European Immigration to the Antebellum United States.* New York: Cambridge University Press.

Comin, Diego A., Mikhail Dmitriev, and Esteban Rossi-Hansberg (2012). "The Spatial Diffusion of Technology." NBER Working Paper 18534.

Daniele, Vittorio and Paolo Malanima (2007). "Il prodotto delle regioni e il divario Nord-Sud in Italia (1861–2004)." *Rivista di Politica Economica* 97:3–4, pp. 267–315.

Del Panta, Lorenzo (1997). "Infant and Child Mortality in Italy, Eighteenth to Twentieth Century: Long-term Trends and Territorial Differences." In *Infant and Child Mortality in the Past.* Alain Bideau, Bertrand Desjardins, and Héctor Pérez Brignoli (ed.). New York: Oxford University Press. Chap. 1, pp. 7–21.

Donaldson, Dave and Richard Hornbeck (2016). "Railroads and American Economic Growth: A 'Market Access' Approach." *Quarterly Journal of Economics* 131:2, pp. 799–858.

Faini, Riccardo and Alessandra Venturini (1994). "Italian Emigration in the Pre-War Period." In *Migration and the International Labor Market 1850–1939.* Timothy J. Hatton and Jeffrey G. Williamson (ed.). London: Routledge, pp. 72–90.

Federico, Giovanni, Alessandro Nuvolari, and Michelangelo Vasta (2019). "The Origins of the Italian Regional Divide: Evidence from Real Wages, 1861–1913." *Journal of Economic History* 79:1, pp. 63–98.

Feinberg, Robbie and Bonnie Petrie (2019). "African Migrants Are Becoming a New Face of the US Border Crisis." URL: `https://www.npr.org/2019/06/20/733682502/african-migrants-are-becoming-a-new-face-of-the-u-s-border-crisis`.

Felice, Emanuele (2005). "Il valore aggiunto regionale. Una stime per il 1891 e per il 1911 e alcune elaborazioni di lungo periodo (1891–1971)." *Rivista di Storia Economica* 21, pp. 273–314.

Ferenczi, Imre and Walter F. Wilcox (1929). *International Migrations.* New York: National Bureau of Economic Research.

Fernández-Huertas Moraga, Jesús (2013). "Understanding Different Migrant Selection Patterns in Rural and Urban Mexico." *Journal of Development Economics* 103, pp. 182–201.

Feys, Torsten (2007). "Where All Passenger Liners Meet: New York as a Nodal Point for the Transatlantic Migrant Trade, 1885–1895." *International Journal of Maritime History* 19:2, pp. 245–272.

Foerster, Robert F. (1919). *The Italian Emigration of Our Times*. New York: Russell & Russell.

Gomellini, Matteo and Cormac Ó Gráda (2013). "Migrations." In *The Oxford Handbook of the Italian Economy Since Unification*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 10, pp. 271–302.

Gould, John D. (1980). "European Inter-Continental Emigration: The Role of 'Diffusion' and 'Feedback'." *Journal of European Economic History* 9:2, pp. 267–315.

Gray, Rowena, Gaia Narciso, and Gaspare Tortorici (2019). "Globalization, Agricultural Markets and Mass Migration: Italy, 1881–1912." *Explorations in Economic History* 74, 101276.

Hatton, Timothy J. and Zachary Ward (2019). "International Migration in the Atlantic Economy 1850–1940." In *Handbook of Cliometrics*. Claude Diebolt and Michael Haupert (ed.). Springer.

Hatton, Timothy J. and Jeffrey G. Williamson (1994). "Latecomers to Mass Emigration." In *Migration and the International Labor Market 1850–1939*. Timothy J. Hatton and Jeffrey G. Williamson (ed.). London: Routledge, pp. 55–71.

——— (1998). *The Age of Mass Migration: Causes and Economic Impact*. New York: Oxford University Press.

Iuzzolino, Giovanni, Guido Pellegrini, and Gianfranco Viesti (2013). "Regional Convergence." In *The Oxford Handbook of the Italian Economy Since Unification*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 20, pp. 571–598.

Jovanovic, Boyan and Saul Lach (1989). "Entry, Exit, and Diffusion with Learning by Doing." *American Economic Review* 79:4, pp. 690–699.

Karadja, Mounir and Erik Prawitz (2019). "Exit, Voice and Political Change: Evidence from Swedish Mass Migration to the United States." *Journal of Political Economy* 127:4, pp. 1864–1925.

Keeling, Drew (2012). *The Business of Transatlantic Migration Between Europe and the United States, 1900–1914*. Zurich: Chronos Verlag.

——— (2013). "The Improvement of Travel Conditions for Migrants Crossing the North Atlantic 1900-1914." In *Points of Passage: Jewish Transmigrants from Eastern Europe in Germany, Britain and Scandinavia*. Tobias Brinkmann (ed.). New York: Berghahn Books.

Kermack, William O. and Anderson G. McKendrick (1927). "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London, Series A, Containing Papers of a Mathematical and Physical Character* 115:772, pp. 700–721.

Killick, John (2014). "Transatlantic Steerage Fares, British and Irish Migration, and Return Migration, 1815–60." *Economic History Review* 67:1, pp. 170–191.

Livi Bacci, Massimo and Marco Breschi (1990). "Italian Fertility: An Historical Account." *Journal of Family History* 15:4, pp. 385–408.

Lowell, Briant L. (1987). *Scandinavian Exodus: Demography and Social Development of 19th Century Rural Communities*. Boulder: Westview.

Manski, Charles F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60:3, pp. 531–542.

Massey, Douglas S., Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J. Edward Taylor (1993). "Theories of International Migration: A Review and Appraisal." *Population and Development Review* 19:3, pp. 431–466.

McKenzie, David and Hillel Rapoport (2007). "Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico." *Journal of Development Economics* 84, pp. 1–24.

——— (2010). "Self-Selection Patterns in Mexico-US Migration: The Role of Migration Networks." *Review of Economics and Statistics* 92:4, pp. 811–821.

Mokyr, Joel (1983). *Why Ireland Starved: An Analytical and Quantitative Study of Irish Poverty, 1800–1851*. Boston: George Allen and Unwin.

Mokyr, Joel and Cormac Ó Gráda (1982). "Emigration and Poverty in Prefamine Ireland." *Explorations in Economic History* 19, pp. 360–384.

Moretti, Enrico (1999). "Social Networks and Migrations: Italy 1876–1913." *International Migration Review* 33:3, pp. 640–657.

Moya, Jose C. (1998). *Cousins and Strangers: Spanish Immigrants in Buenos Aires, 1850–1930*. Berkeley: University of California Press.

Munshi, Kaivan (2003). "Networks in the Modern Economy: Mexican Migrants in the US Labor Market." *Quarterly Journal of Economics* 118:2, pp. 549–599.

Nelson, Phillip (1959). "Migration, Real Income, and Information." *Journal of Regional Science* 1:2, pp. 43–74.

Nickell, Stephen (1981). "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49:6, pp. 1417–1426.

Radford, Jynnah (2019). "Key Findings About US Immigrants." URL: `https://www.pewresearch.org/fact-tank/2019/06/17/key-findings-about-u-s-immigrants/`.

Sjaastad, Larry A. (1962). "The Costs and Returns of Human Migration." *Journal of Political Economy* 70:5, pp. 80–93.

Solomon, Daina Beth (2019). "US Dream Pulls African Migrants in Record Numbers Across Latin America." URL: `https://www.reuters.com/article/us-usa-immigration-africa/u-s-dream-pulls-african-migrants-in-record-numbers-across-latin-america-idUSKCN1U01A4`.

Spitzer, Yannay (2016). "Pogroms, Networks, and Migration: The Jewish Migration from the Russian Empire to the United States 1881–1914." Mimeo., Hebrew University of Jerusalem.

Spitzer, Yannay, Gaspare Tortorici, and Ariell Zimran (2019). "International Migration Responses to the 1908 Messina-Reggio Calabria Earthquake." Mimeo., Vanderbilt University.

Spitzer, Yannay and Ariell Zimran (2018). "Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925." *Journal of Development Economics* 134, pp. 226–247.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression With Weak Instruments." *Econometrica* 65:3, pp. 557–586.

Todaro, Michael P. (1969). "A Model of Labor Migration and Urban Unemployment in Less Developed Countries." *American Economic Review* 59:1, pp. 138–148.

Toniolo, Gianni (2013). "An Overview of Italy's Economic Growth." In *The Oxford Handbook of the Italian Economy since Unification*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 1, pp. 3–36.

Vecchi, Giovanni (2011). *In ricchezza e in povertà. Il benessere degli italiani dall'Unità ad oggi*. Bologna: Il Mulino.

Wegge, Simone A. (1998). "Chain Migration and Information Networks: Evidence from Nineteenth-Century Hesse-Cassel." *Journal of Economic History* 58:4, pp. 957–986.

Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wozniak, Abigail (2010). "Are College Graduates More Responsive to Distant Labor Market Opportunities?" *Journal of Human Resources* 45:4, pp. 944–970.

Zimran, Ariell (2019). "Transportation and Health in the Antebellum United States 1820–1847." NBER Working Paper 24943.

Tables

Table 1: Summary statistics for time-varying variables

	District	Commune				
	(1)	(2)	(3)	(4)	(5)	(6)
Variable	All	All	1884–1890	1891–1900	1901–1910	1911–1914
Emigration Rate	0.016	0.016	0.010	0.011	0.025	0.027
	(0.023)	(0.028)	(0.026)	(0.026)	(0.030)	(0.029)
Emigration Rate to North America	0.004	0.005	0.002	0.002	0.008	0.009
	(0.007)	(0.011)	(0.007)	(0.005)	(0.014)	(0.013)
Emigration Rate to South America	0.003	0.003	0.004	0.004	0.004	0.003
	(0.005)	(0.007)	(0.009)	(0.009)	(0.007)	(0.005)
Emigration Rate to Europe	0.010	0.009	0.005	0.005	0.012	0.015
	(0.022)	(0.022)	(0.021)	(0.022)	(0.023)	(0.024)
Emigration Rate at least 1 percent	0.430	0.413	0.227	0.262	0.633	0.718
	(0.495)	(0.492)	(0.419)	(0.440)	(0.482)	(0.450)
Emigration Rate to NA at least 0.5 percent	0.187	0.226	0.082	0.100	0.352	0.436
	(0.390)	(0.418)	(0.275)	(0.301)	(0.478)	(0.496)
Emigration Rate to SA at least 0.5 percent	0.174	0.202	0.209	0.202	0.267	0.211
	(0.379)	(0.401)	(0.407)	(0.401)	(0.442)	(0.408)
Emigration Rate to Europe at least 0.5 percent	0.329	0.283	0.132	0.134	0.422	0.498
	(0.470)	(0.450)	(0.339)	(0.341)	(0.494)	(0.500)
Any Emigration	0.955	0.653	0.400	0.451	0.859	0.966
	(0.206)	(0.476)	(0.490)	(0.498)	(0.348)	(0.182)
Any Emigration to NA	0.896	0.672	0.397	0.436	0.858	0.966
	(0.305)	(0.469)	(0.489)	(0.496)	(0.349)	(0.182)
Any Emigration to SA	0.943	0.680	0.415	0.452	0.859	0.966
	(0.233)	(0.466)	(0.493)	(0.498)	(0.348)	(0.182)
Any Emigration to Europe	0.954	0.671	0.405	0.427	0.858	0.966
	(0.210)	(0.470)	(0.491)	(0.495)	(0.349)	(0.182)
Observations	10,697	255,488	31,600	79,773	80,200	32,080
Units	284	8,020	7,900	8,020	8,020	8,020

Notes: Observations are at the district-year level in column (1) and at the commune-year level in columns (2)–(6). Standard deviations in parentheses.

Table 2: Summary statistics for non-time-varying variables

	District				Commune			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variable	All	North	Center	South	All	North	Center	South
District Share of Male Labor in Agriculture	0.547	0.555	0.547	0.534	0.545	0.548	0.543	0.539
	(0.109)	(0.101)	(0.135)	(0.109)	(0.119)	(0.127)	(0.126)	(0.100)
District Share of Male Labor in Industry	0.216	0.237	0.191	0.192	0.231	0.263	0.197	0.191
	(0.070)	(0.074)	(0.058)	(0.056)	(0.096)	(0.110)	(0.062)	(0.053)
District Adult Male Literacy Rate	0.468	0.582	0.453	0.279	0.525	0.678	0.473	0.284
	(0.183)	(0.149)	(0.127)	(0.055)	(0.211)	(0.134)	(0.140)	(0.053)
District Population Fraction Under Age 15	0.328	0.334	0.309	0.327	0.326	0.333	0.305	0.323
	(0.024)	(0.024)	(0.026)	(0.018)	(0.023)	(0.022)	(0.023)	(0.015)
Distance to Railroad (1881, km)					9.850	8.701	7.999	12.731
					(12.479)	(12.065)	(9.171)	(13.974)
Mean Elevation (m)	400.889	370.491	416.997	446.065	444.885	462.970	391.517	438.578
	(377.425)	(465.328)	(203.671)	(242.861)	(424.677)	(502.836)	(263.309)	(323.220)
Distance to North America Epicenter (km)	169.622	200.852	146.185	126.381	151.466	161.740	137.325	140.227
	(98.116)	(83.547)	(75.843)	(111.151)	(93.303)	(68.300)	(78.547)	(128.689)
Distance to South America Epicenter (km)	155.776	94.647	208.582	237.222	170.972	124.155	218.179	230.346
	(110.473)	(49.144)	(106.815)	(123.756)	(105.574)	(48.721)	(107.466)	(133.677)
Distance to European Border (km)	242.833	43.356	217.446	599.689	240.766	45.587	235.779	583.824
	(274.776)	(41.963)	(102.302)	(190.421)	(263.405)	(36.662)	(124.165)	(171.721)
Observations	284	154	41	89	8,349	4,528	1,228	2,593

Notes: Observations are at the district level in columns (1)–(4) and at the commune level in columns (5)–(8). Standard deviations in parentheses. Observation numbers are the minimum with observations for all variables. Distance from railroad is 0 for any commune with a rail line passing through it in 1881, and the distance from the nearest commune border to the rail line for all other communes. Distance to the epicenters of North America- and South America-bound emigration are from the commune centroid to the centroid of the epicenter district's capital city. Distance to the European border is from the commune centroid. Averages for the occupation, literacy, and age data in columns (5)–(8) are averages of the data for the commune's district.

Table 3: β-convergence

	(1)	(2)	(3)	(4)
Variables	OLS	OLS	IV	IV
Lagged Own Emigration	-0.228^a	-0.385^a	-0.196^a	-0.393^a
	(0.017)	(0.018)	(0.038)	(0.044)
Observations	36,199	35,899	36,127	35,827
R-squared	0.520	0.628	0.162	0.350
Controls	No	No	No	Yes
1st Stage F			16.923	23.865
Falsification	0.101	0.221		
	(0.025)	(0.023)		

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Standard errors clustered at the province level. Unit of observation is a commune-half decade. Dependent variable is the change in the log of the emigration rate to North America. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. All specifications include half-decade fixed effects.

Table 4: Destination dissimilarity and distance between provinces

Variables	(1) All	(2) All	(3) Major	(4) Minor	(5) Pre-1900	(6) Post-1900
log(Distance)	0.134^a	0.121^a	0.114^a	0.072^a	0.090^a	0.145^a
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	21,114	21,114	21,114	21,114	9,384	11,730
R-squared	0.242	0.250	0.195	0.089	0.164	0.328
Controls	No	Yes	Yes	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Dependent variable is the dissimilarity index in the emigration destination distribution of the two provinces making up a province pair in a given half decade. Unit of observation is a province pair-half decade. Standard error clustered by province pair. All regressions include half-decade fixed effects. R-squared is computed after absorbing half-decade fixed effects. Major destinations are US, Canada, France, Argentina, Uruguay, Switzerland, Austria-Hungary, Germany, and Brazil. Controls are absolute differences in agricultural and industrial employment shares, literacy rates, and fraction under age 15.

Table 5: Spatial contagion of US-bound migration, OLS, annual

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Own Emigration	0.480^a	0.469^a	0.457^a	0.266^a	0.250^a	0.220^a	0.249^a	0.155^a
	(0.013)	(0.013)	(0.012)	(0.016)	(0.013)	(0.011)	(0.012)	(0.044)
Lagged Emigration Exposure	0.337^a	0.291^a	0.315^a	0.422^a	0.405^a	0.254^a	0.433^a	0.315^a
	(0.024)	(0.025)	(0.027)	(0.044)	(0.040)	(0.023)	(0.055)	(0.060)
Latitude		0.092^c						
		(0.054)						
Longitude		−0.040						
		(0.040)						
log(Elevation)		0.068^a						
		(0.013)						
District Ag. Employment Share		0.620^b						
		(0.252)						
District Ind. Employment Share		0.956^a						
		(0.340)						
District Literacy Rate		0.104						
		(0.200)						
District Under 15 Fraction		2.840^a						
		(0.849)						
log(Rail Distance + 1)		0.012^a						
		(0.002)						
Lagged Emigration Exposure (SA)							−0.001	
							(0.040)	
Lagged Emigration Exposure (EU)							−0.059	
							(0.051)	
Interaction								-0.014^b
								(0.006)
Observations	145,126	144,113	144,113	145,081	144,068	144,059	144,068	144,068
R-squared	0.371	0.382	0.445	0.189	0.286	0.091	0.286	0.287
Province FE	Yes	Yes	Yes	No	No	No	No	No
Commune FE	No	No	No	Yes	Yes	Yes	Yes	Yes
Province-Year FE	No	No	No	No	No	Yes	No	No
Controls	No	No	Yes	No	Yes	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Standard errors clustered at the province level. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-year. Emigration rate regressors included as logs. Controls include year-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Interaction is the product of lagged own emigration rate and lagged emigration exposure to North America. R-squared is computed after absorbing fixed effects.

Table 6: Instrument balance

Variables	(1) Ag Share	(2) Ind Share	(3) Young Frac	(4) Lit Rate	(5) Rail Dist	(6) Elevation
Neighbors' Log Distance	0.057	−0.016	−0.007	0.013	1.326	1.160^{b}
	(0.053)	(0.035)	(0.007)	(0.029)	(0.962)	(0.502)
Observations	8,335	8,335	8,335	8,335	8,335	8,278
R-squared	0.420	0.603	0.513	0.930	0.113	0.547

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Standard errors clustered at the district level. Dependent variable listed in column header. Ag Share is 1881 share of district employment in agriculture. Ind Share is 1881 share of district employment in industry. Young Frac is 1881 share of district population under age 15. Lit Rate is the 1881 adult male literacy rate at the district level. Rail Dist is 1881 log distance of the commune to a rail line. Elevation is the log of elevation of the commune. Unit of observation is a commune. All specifications include province fixed effects and condition on a commune's own log distance to epicenter.

Table 7: Spatial contagion regressions, IV

Variables	(1) 1 Yr	(2) 1 Yr	(3) 1 Yr	(4) 5 Yr	(5) 5 Yr	(6) 5 Yr
Panel A: OLS						
Lagged Own Emigration	0.272^{a}	0.249^{a}	0.220^{a}	0.058^{a}	0.053^{a}	0.046^{a}
	(0.017)	(0.012)	(0.011)	(0.015)	(0.015)	(0.014)
Lagged Emigration Exposure	0.316^{a}	0.333^{a}	0.247^{a}	0.372^{a}	0.382^{a}	0.327^{a}
	(0.033)	(0.038)	(0.023)	(0.067)	(0.050)	(0.042)
Panel B: Stage 0						
F-statistic of Instrument	17.757	10.714	9.273	16.272	22.682	3.591
Panel C: First Stage						
Lagged Own Emigration	0.224^{a}	0.174^{a}	0.069^{a}	0.226^{a}	0.187^{a}	0.087^{a}
	(0.015)	(0.009)	(0.003)	(0.014)	(0.012)	(0.005)
Lagged Instrumented Exposure	0.930^{a}	0.915^{a}	1.000^{a}	0.921^{a}	0.828^{a}	0.932^{a}
	(0.091)	(0.055)	(0.046)	(0.109)	(0.087)	(0.072)
Panel D: IV						
Lagged Own Emigration	0.294^{a}	0.220^{a}	0.206^{a}	0.019	−0.034	0.021
	(0.038)	(0.016)	(0.010)	(0.045)	(0.029)	(0.016)
Lagged Emigration Exposure	0.225^{c}	0.480^{a}	0.407^{a}	0.522^{a}	0.795^{a}	0.562^{a}
	(0.128)	(0.075)	(0.057)	(0.170)	(0.146)	(0.118)
Controls	No	Yes	Yes	No	Yes	Yes
Province-Year FE	No	No	Yes	No	No	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: All regressions include commune and year fixed effects and year-specific functions of log distance from epicenter. Controls are year- or half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Standard errors clustered at the province level.

Figures

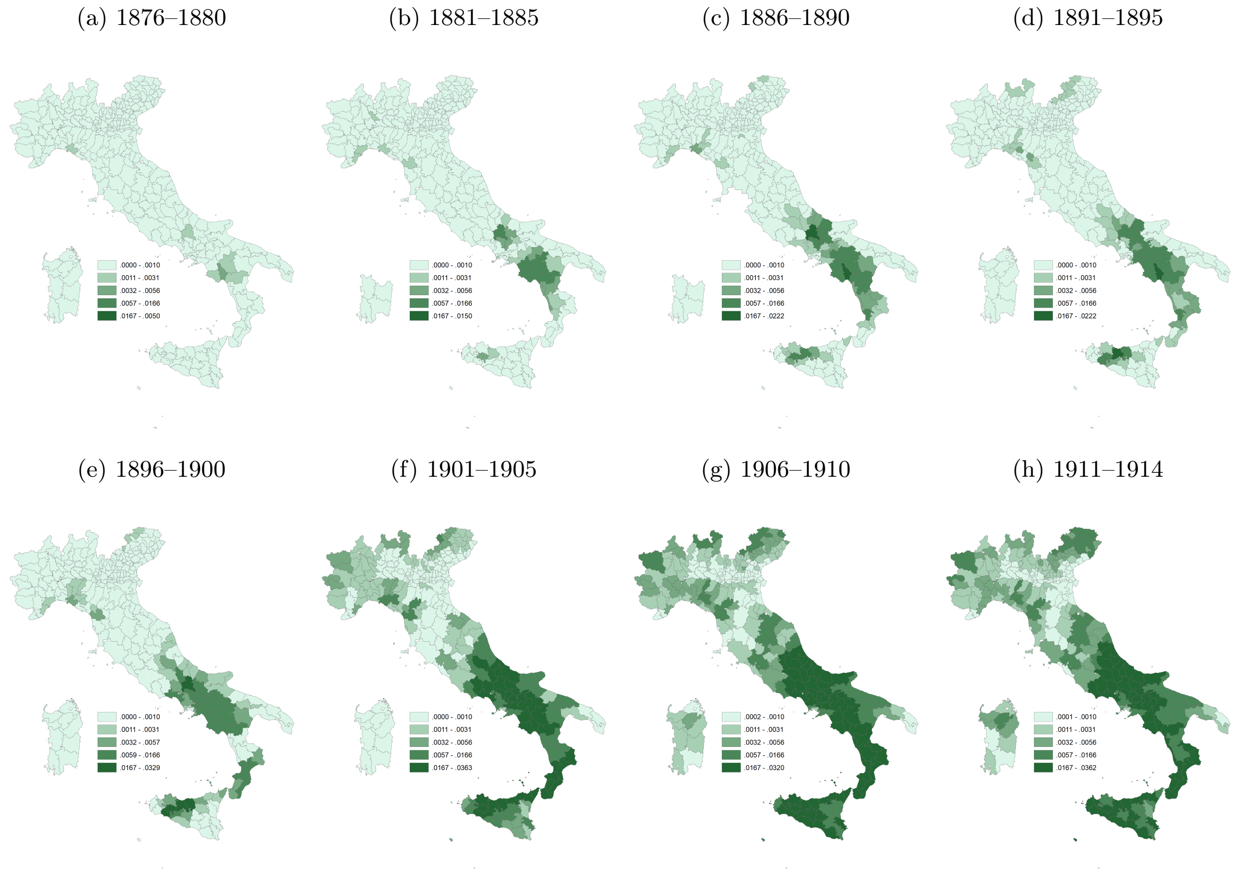


Figure 1: District-level emigration rates to North America

Note: Each panel presents a district's average annual emigration rate to North America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914.

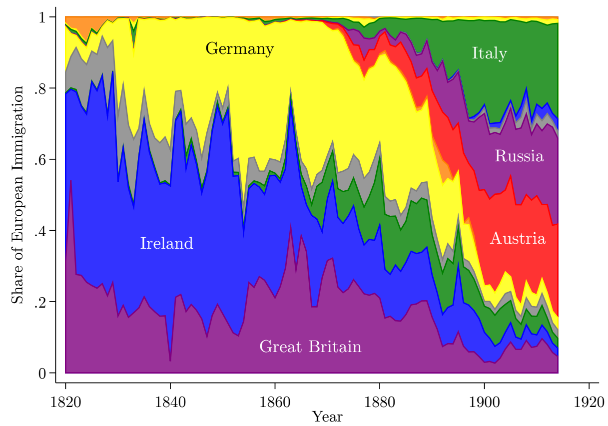


Figure 2: Distribution of origin countries for US immigration from Europe

Source: Barde, Carter, and Sutch (2006)

Note: The "Austria" data are from Barde, Carter, and Sutch's (2006) data for "Other Central Europe," which cover Central Europe other than Germany and Poland.

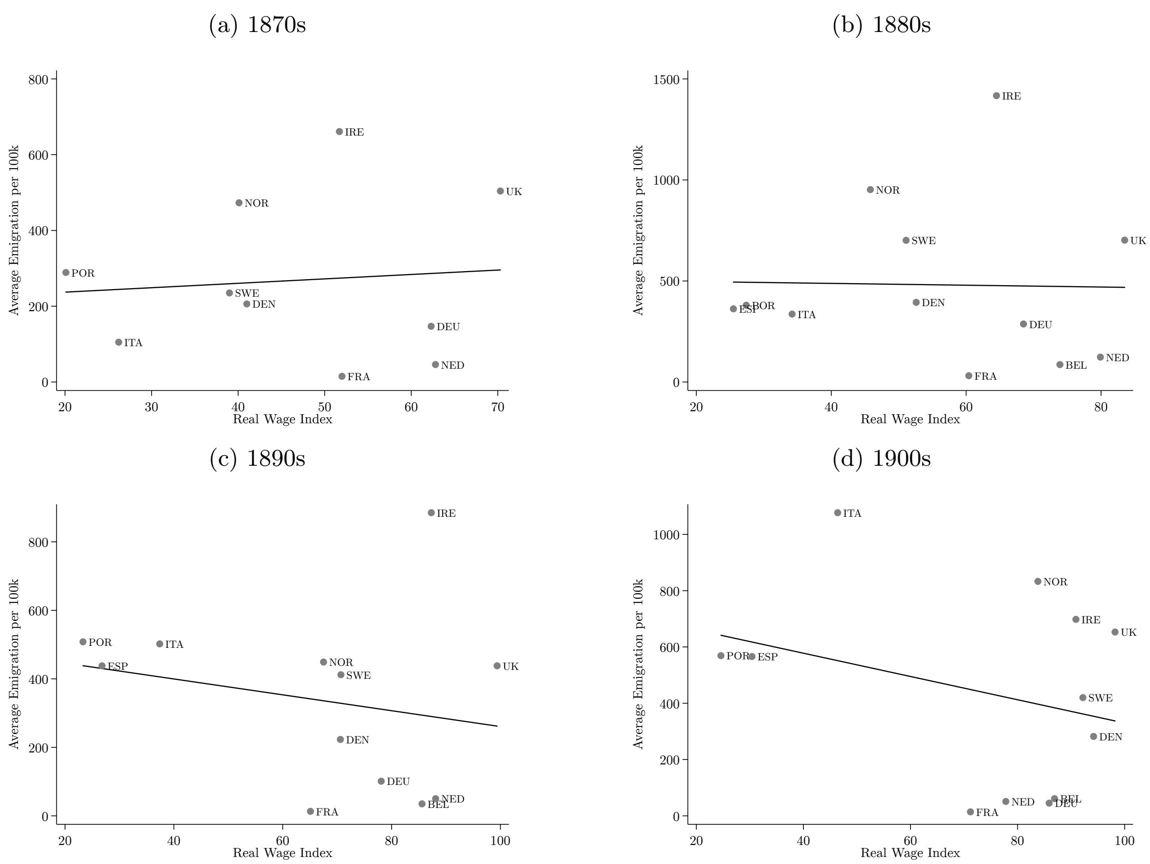


Figure 3: Emigration and real wages

Source: Emigration data are from Ferenczi and Wilcox (1929). Wage data are from Hatton and Williamson (1998).

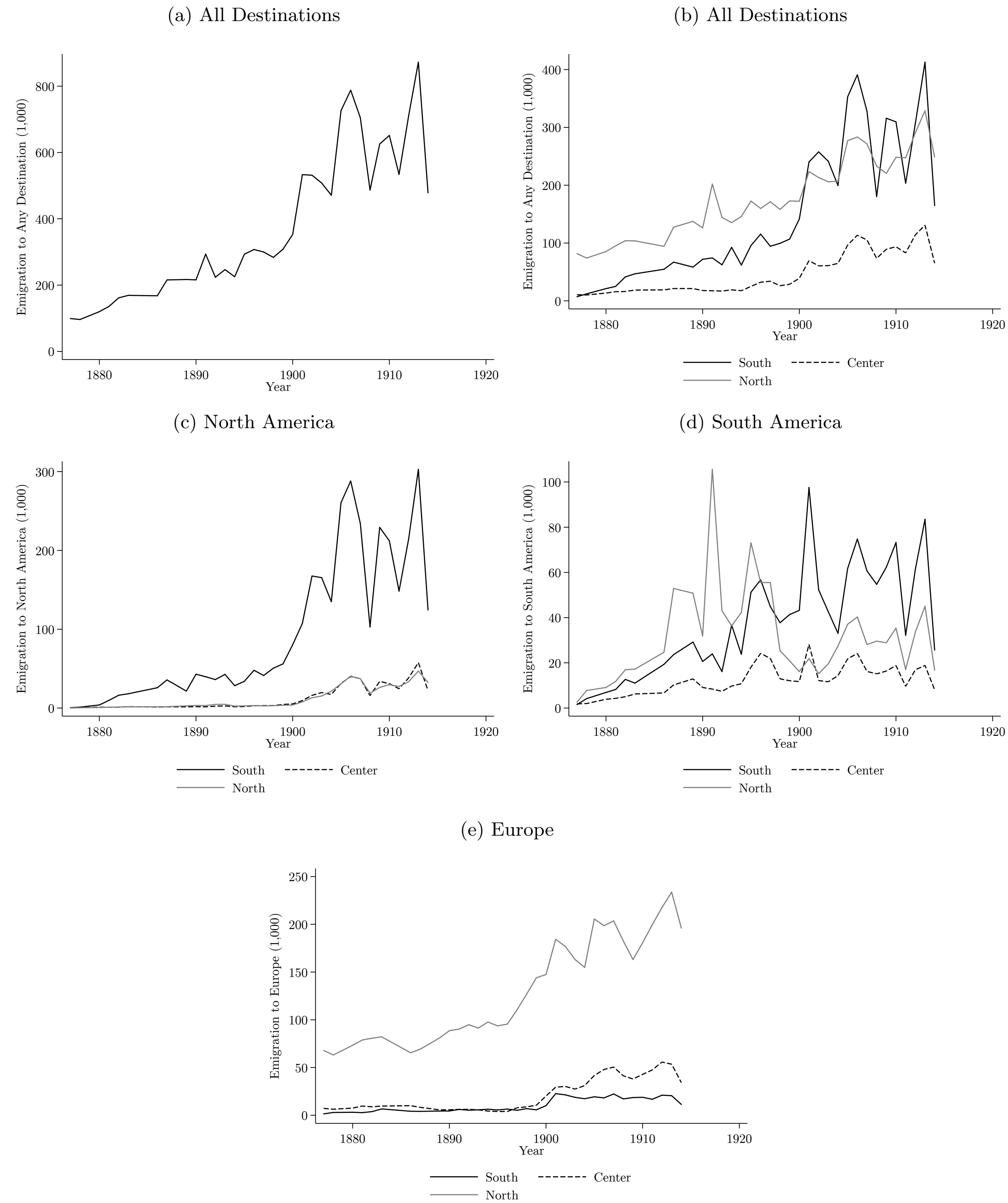


Figure 4: Aggregate emigration by origin and destination

Note: These figures are based on our province-by-destination data. South includes the regions of Abruzzo, Campania, Puglia, Basilicata, Calabria, Sicilia, and Sardinia. Center includes the regions of Liguria, Toscana, Marche, Umbria, and Latium. North includes the regions of Piemonte, Lombardia, Veneto, and Emilia Romagna.

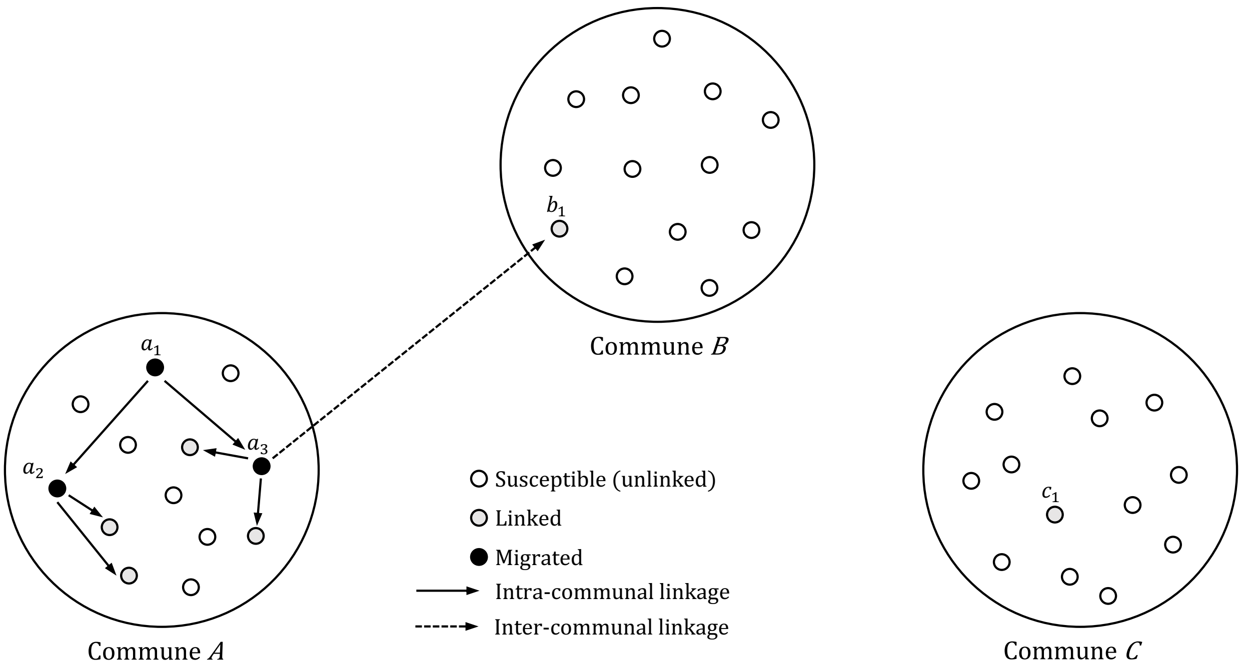


Figure 5: Hypothetical example

Figure 6: Epicenters of mass migration to North America

Note: Epicenters are the districts marked in black. They are defined as described in text.

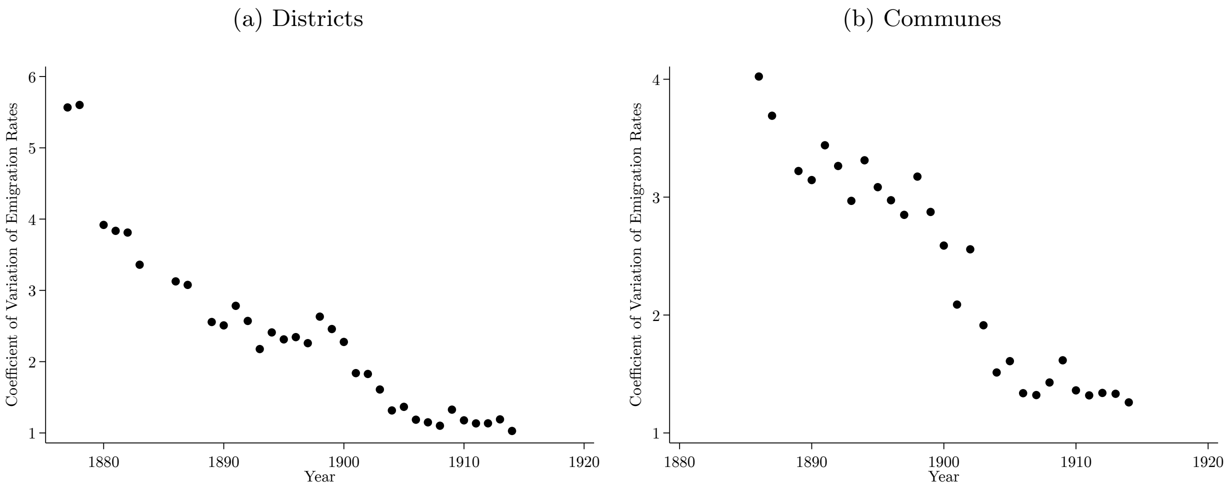


Figure 7: σ-convergence in emigration rates to North America

Note: Each point represents the coefficient of variation in emigration rates to North America in a particular year.

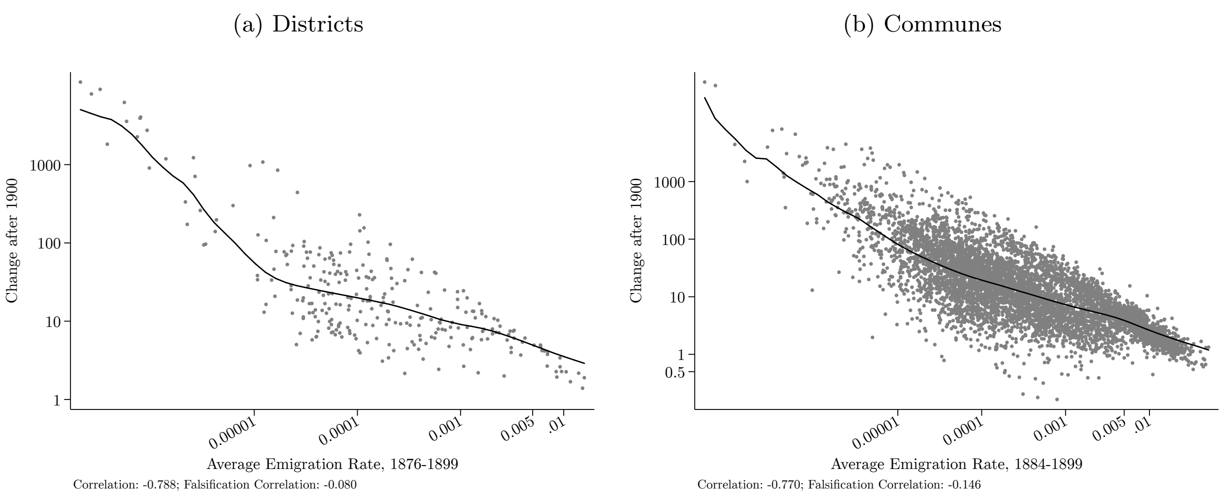


Figure 8: β-convergence in emigration rates to North America

Note: Each point represents a commune or district. The x-axis is the average annual emigration rate for the commune or district for 1884–1899 on a log scale. The y-axis is the ratio of the average emigration rate before and after 1900, also on a log scale.

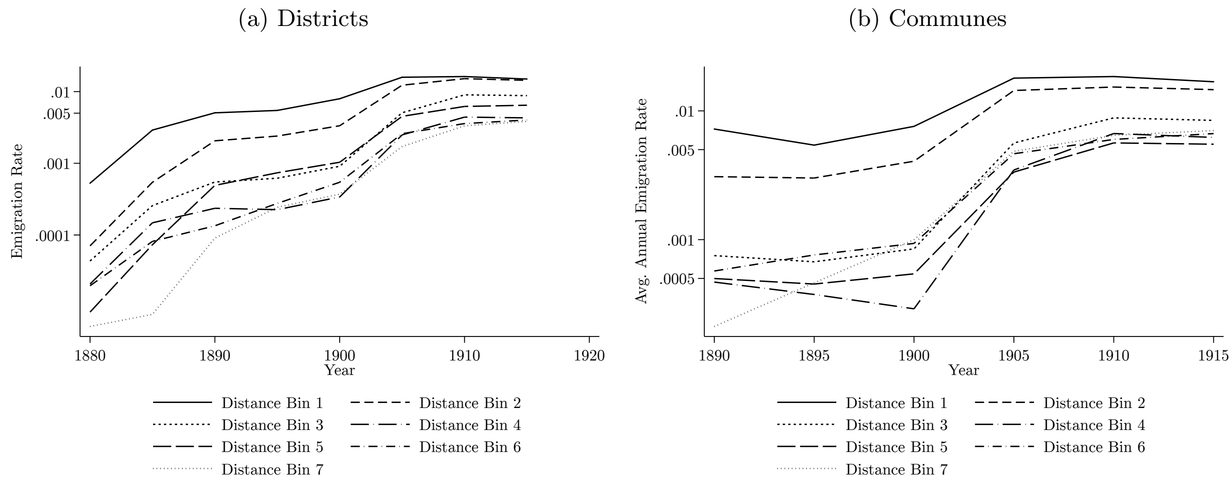


Figure 9: Time series of migration to North America by distance from epicenter

Note: Distance bins are based on septiles of distance to epicenters. Panel (a) excludes districts more than 300 kilometers from an epicenter. Panel (b) excludes communes more than 250 kilometers from an epicenter.

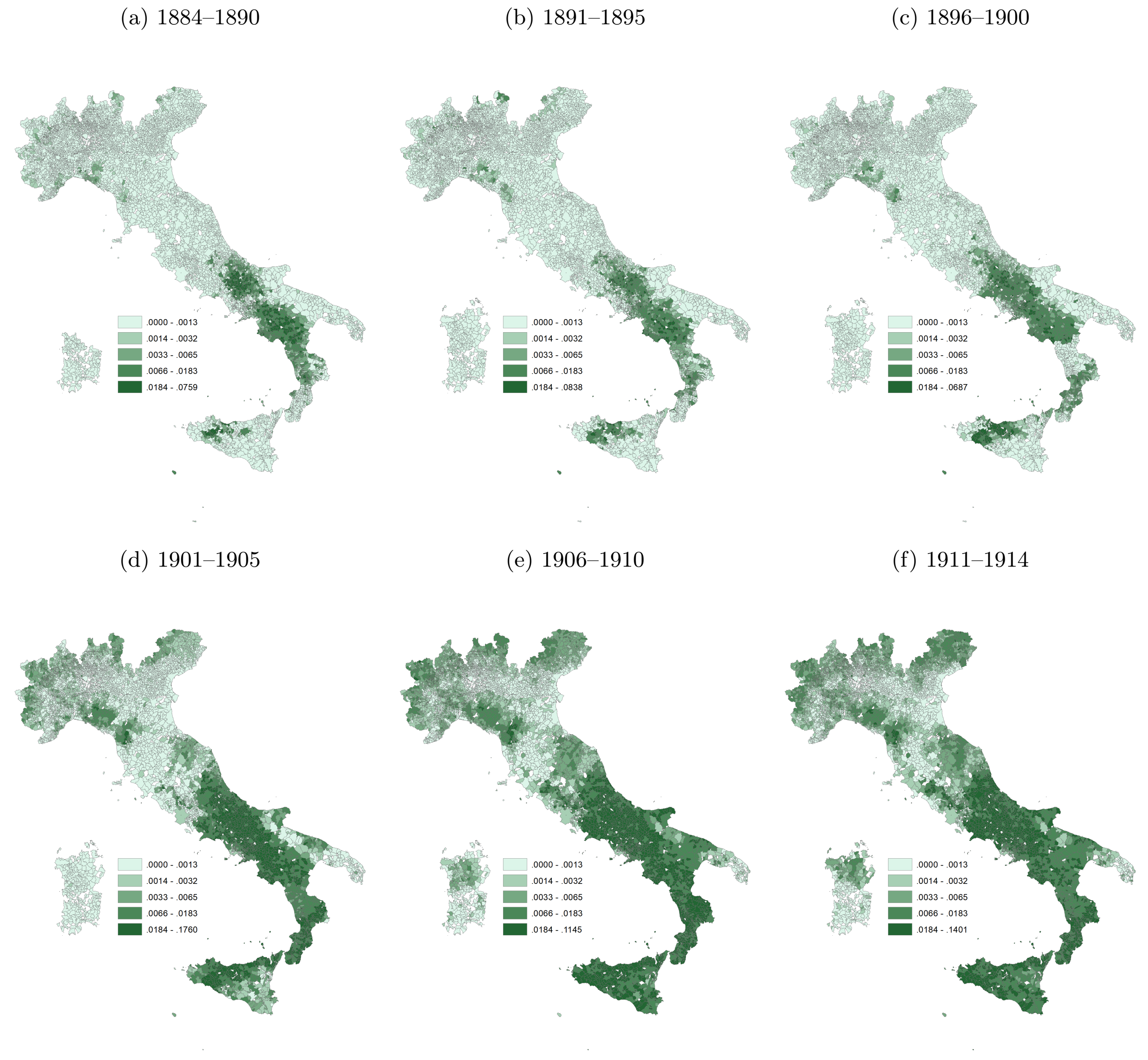


Figure 10: Commune-level emigration rates to North America

Note: Each panel presents a commune's average annual emigration rate to North America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914.

(a) Districts (b) Communes

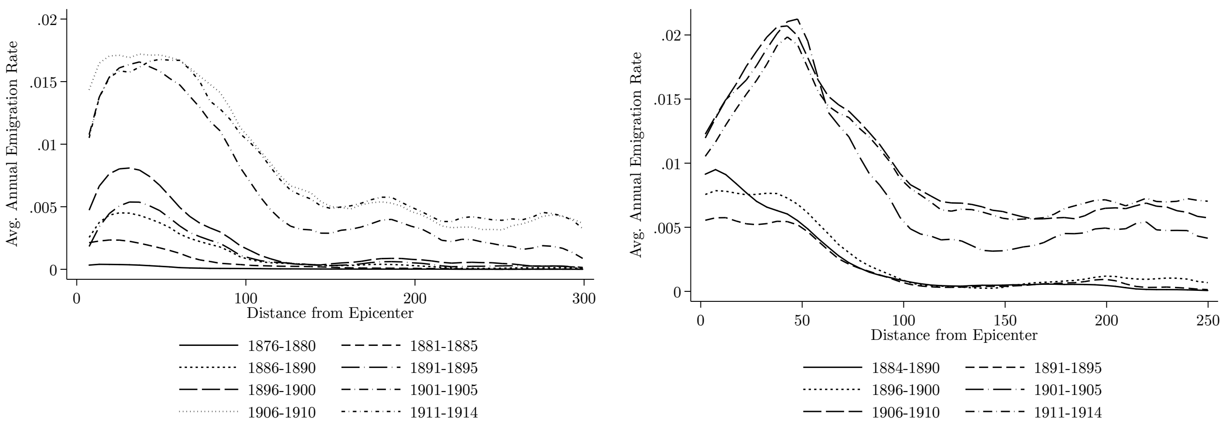


Figure 11: Emigration rates to North America by distance to epicenter (km)

Note: These figures plot local polynomial regressions of the average annual emigration rate for each half decade against distance from the epicenters of emigration to North America. Epicenter districts are excluded.

(a) No Controls (b) Controls

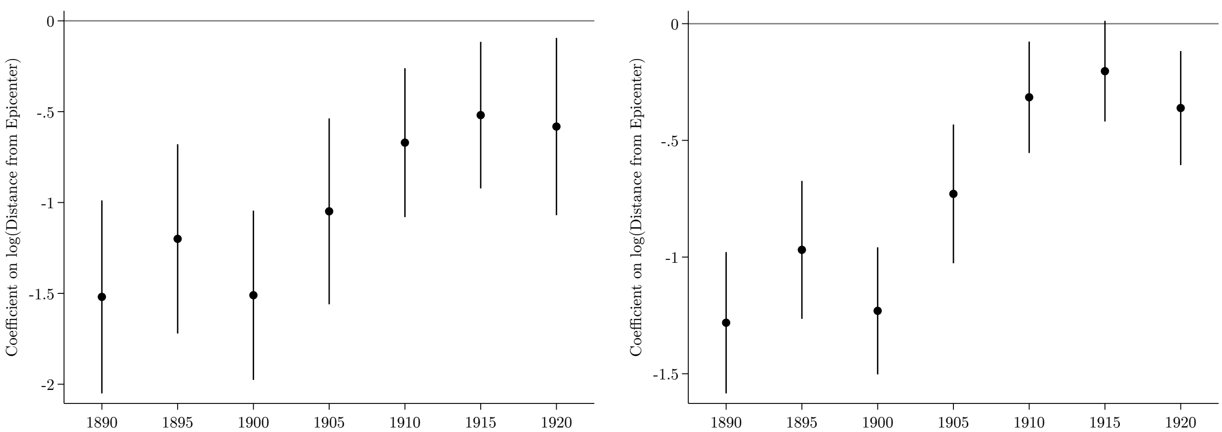


Figure 12: Emigration to North America by distance from epicenters of emigration to North America

Note: These are coefficients from regressions of log average annual emigration rates at the commune-half decade level to North America for each half decade on half-decade fixed effects and log distance to the epicenters of emigration to North America. Half decades are defined by their last year (e.g., 1895 indicates the period 1891–1895).

(a) Districts (b) Communes

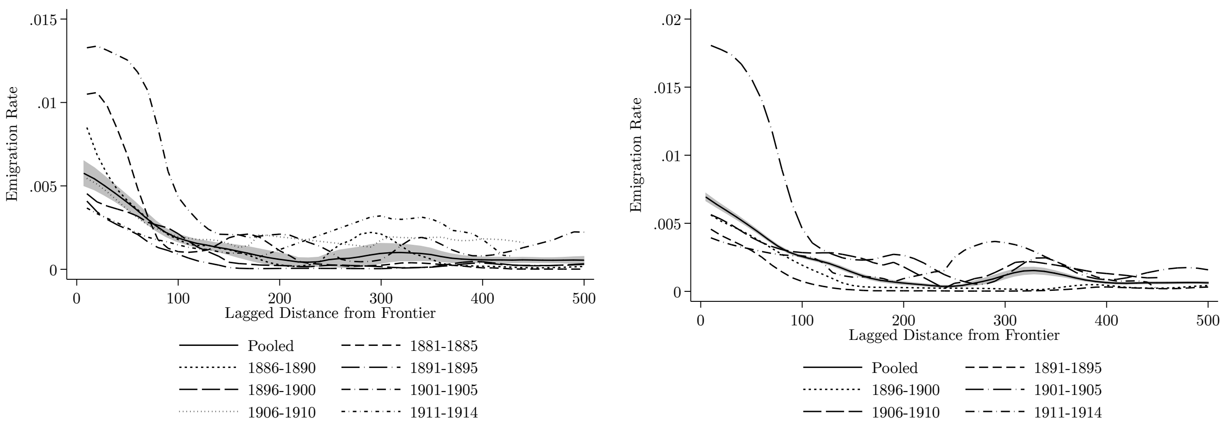


Figure 13: Emigration rates to North America by distance to the mass migration frontier (km)

Note: These figures present local polynomial regressions of average annual migration rates for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 0.005 by the previous half decade, limiting the sample to communes in districts that had not yet achieved this threshold.

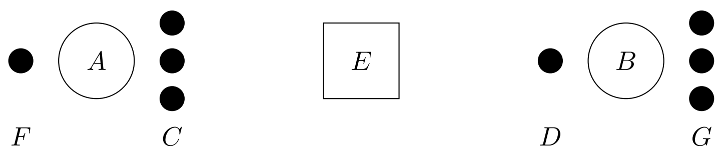


Figure 14: Intuition of our instrumental variables strategy

Note: Communes A and B are equidistant from the migration epicenter E, but commune A has a higher share of neighbors (dark circles) closer than it to the epicenter (in commune C), whereas commune B has a higher share of neighbors farther than it from the epicenter (in commune G) and a smaller share of neighbors closer than it to the epicenter (in commune D).

A Additional Tables and Figures (For Online Publication)

Table A.1: Spatial contagion of US-bound migration, OLS, half decade

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Own Emigration	0.391^a	0.376^a	0.377^a	0.047^a	0.049^a	0.045^a	0.049^a	-0.202^a
	(0.016)	(0.016)	(0.016)	(0.015)	(0.015)	(0.014)	(0.015)	(0.064)
Lagged Emigration Exposure	0.342^a	0.272^a	0.317^a	0.479^a	0.477^a	0.341^a	0.503^a	0.265^a
	(0.057)	(0.057)	(0.042)	(0.089)	(0.043)	(0.042)	(0.062)	(0.079)
Latitude		0.105						
		(0.084)						
Longitude		−0.093						
		(0.066)						
log(Elevation)		0.090^a						
		(0.021)						
District Ag. Employment Share		1.107^a						
		(0.372)						
District Ind. Employment Share		1.668^a						
		(0.497)						
District Literacy Rate		0.355						
		(0.286)						
District Under 15 Fraction		4.990^a						
		(1.139)						
log(Rail Distance + 1)		0.010^a						
		(0.003)						
Lagged Emigration Exposure (SA)							−0.049	
							(0.075)	
Lagged Emigration Exposure (EU)							0.007	
							(0.051)	
Interaction								-0.034^a
								(0.008)
Observations	36,199	35,899	35,899	36,115	35,815	35,812	35,815	35,815
R-squared	0.329	0.353	0.424	0.097	0.226	0.054	0.227	0.234
Province FE	Yes	Yes	Yes	No	No	No	No	No
Commune FE	No	No	No	Yes	Yes	Yes	Yes	Yes
Province-Half Decade FE	No	No	No	No	No	Yes	No	No
Controls	No	No	Yes	No	Yes	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Standard errors clustered at the province level. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Emigration rate regressors included as logs. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Interaction is the product of lagged own emigration rate and lagged emigration exposure to North America. R-squared is computed after absorbing fixed effects.

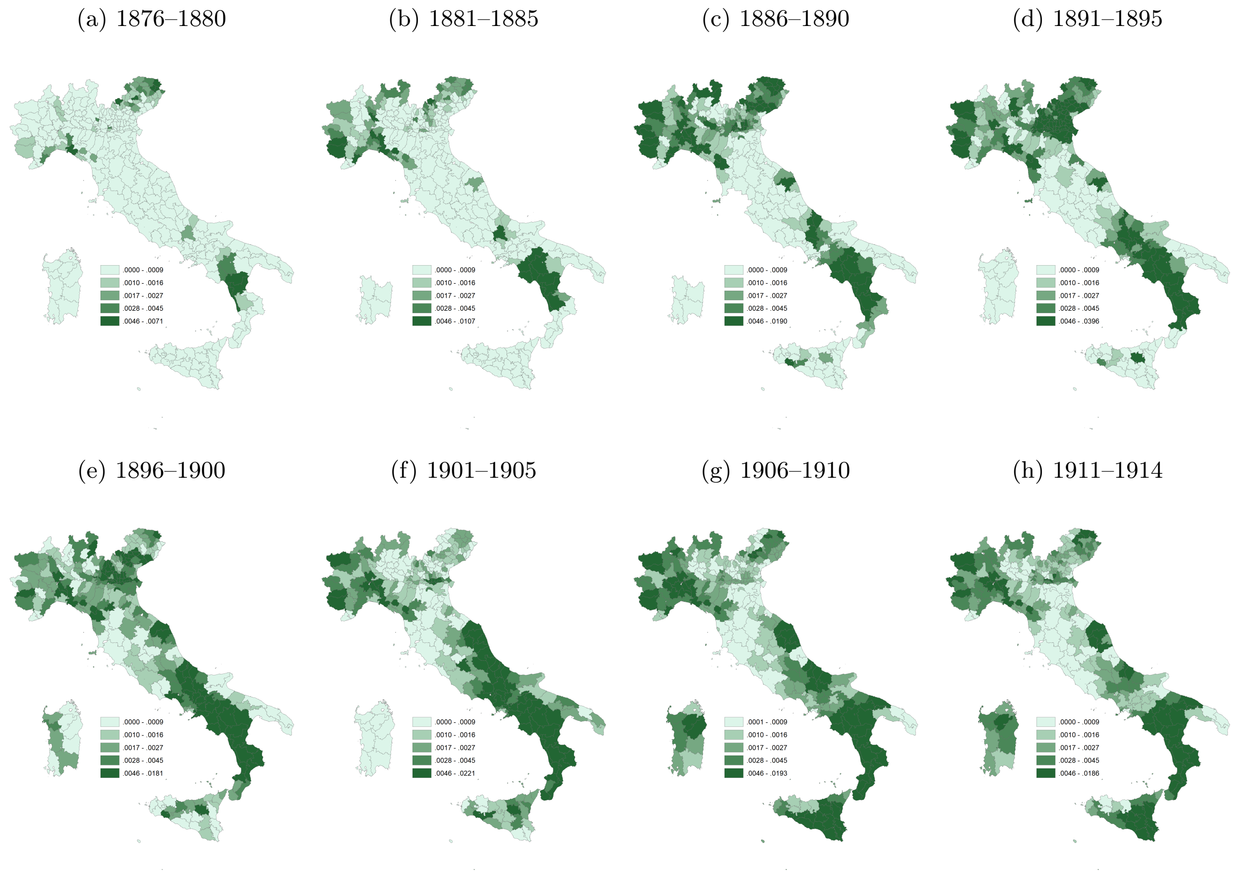


Figure A.1: District-level emigration rates to South America

Note: Each panel presents a district's average annual emigration rate to South America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914.

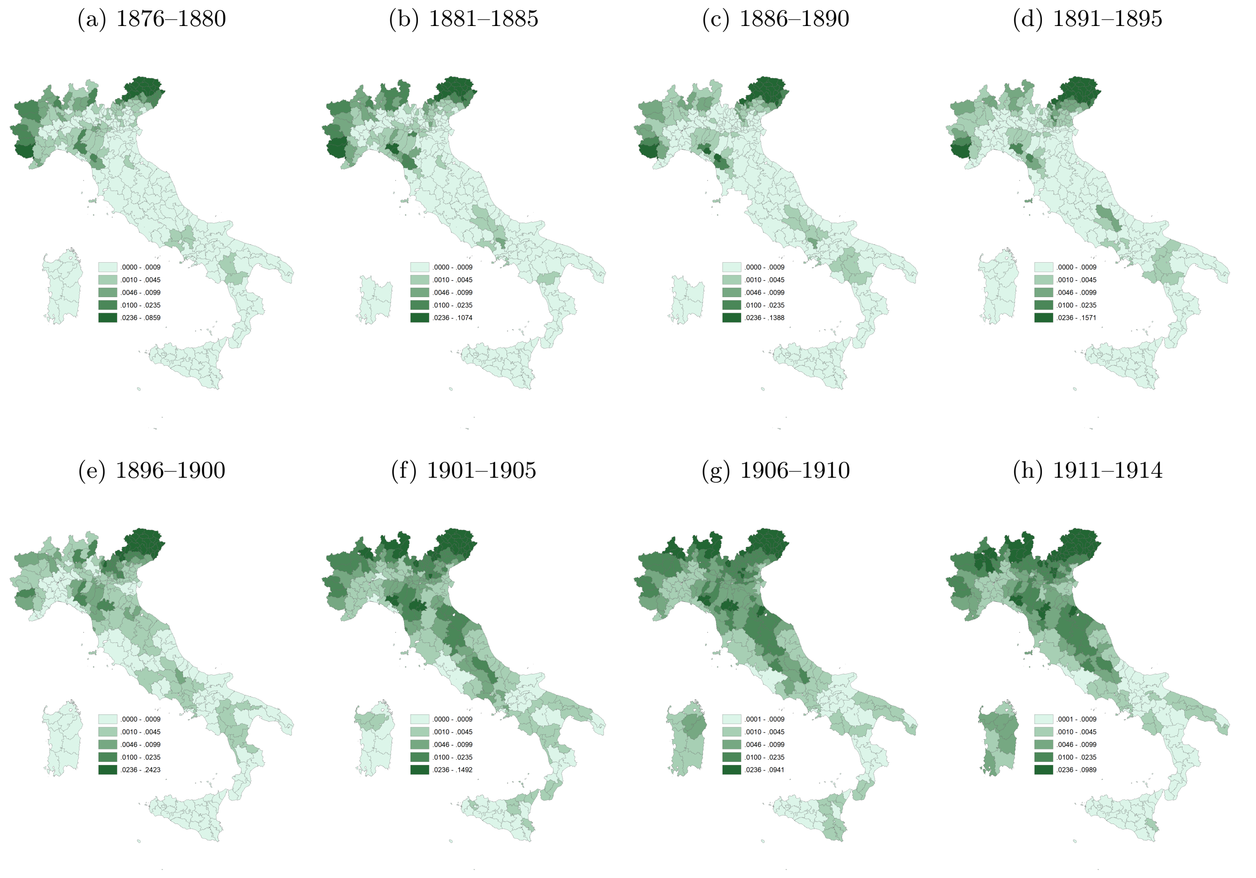


Figure A.2: District-level emigration rates to Europe

Note: Each panel presents a district's average annual emigration rate to Europe in the period in question. Scale is based on quintiles of emigration rates in 1911–1914.

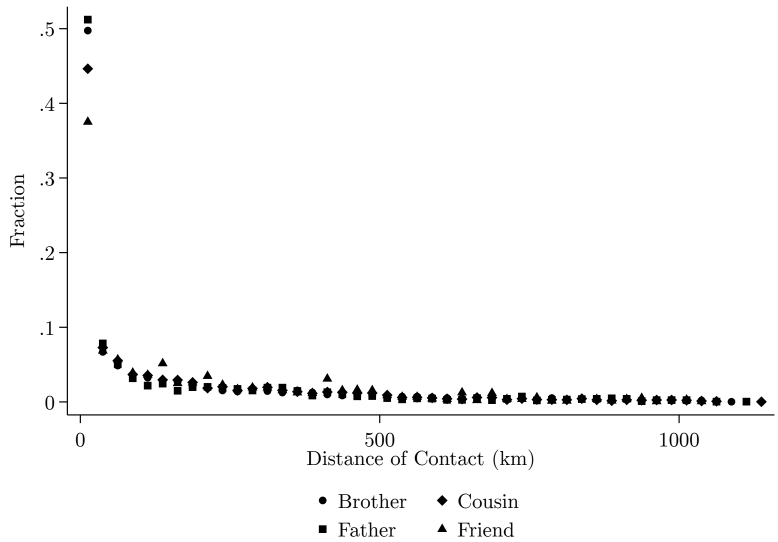


Figure A.3: Histogram of distance of connections in the US by connection type

Note: To construct this graph, we linked the geolocated Ellis Island passenger manifests (Spitzer and Zimran 2018) to the transcribed Castle Garden records (Battery Conservancy 2009) listing the name and relationship of the person being met in the US, and then linked that connection back to the geolocated Ellis Island passenger manifests. We retained cases in which this latter match had five or fewer matches. This Figure plots, for four different kinds of relationship, the minimum distance over the five or fewer matches between the migrant and the person who he reported meeting in the US.

Source: Our elaborations on the Ellis Island and Castle Garden passenger manifests.

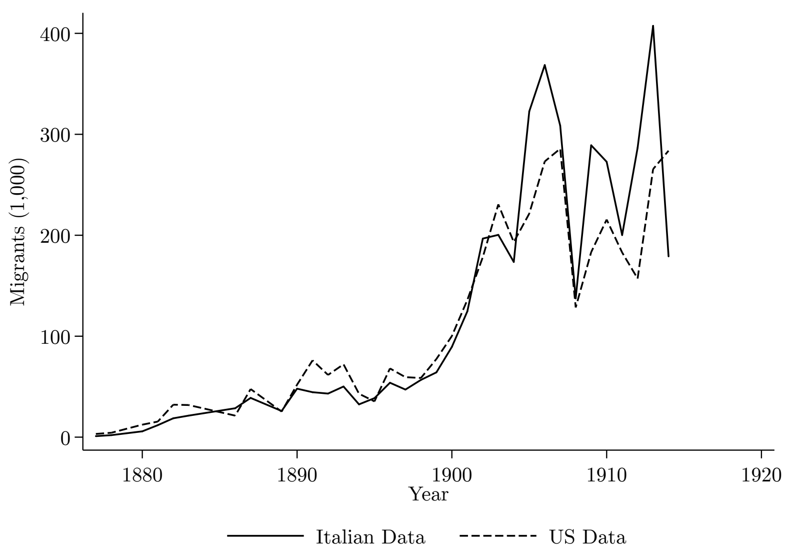


Figure A.4: Comparison of Italian emigration data and US immigration data

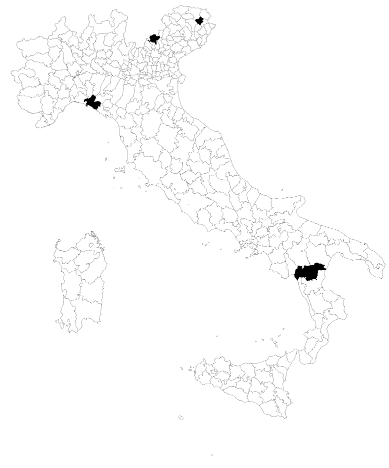


Figure A.5: Epicenters of mass migration to South America

Note: Epicenters are the districts marked in black. They are defined as described in text.

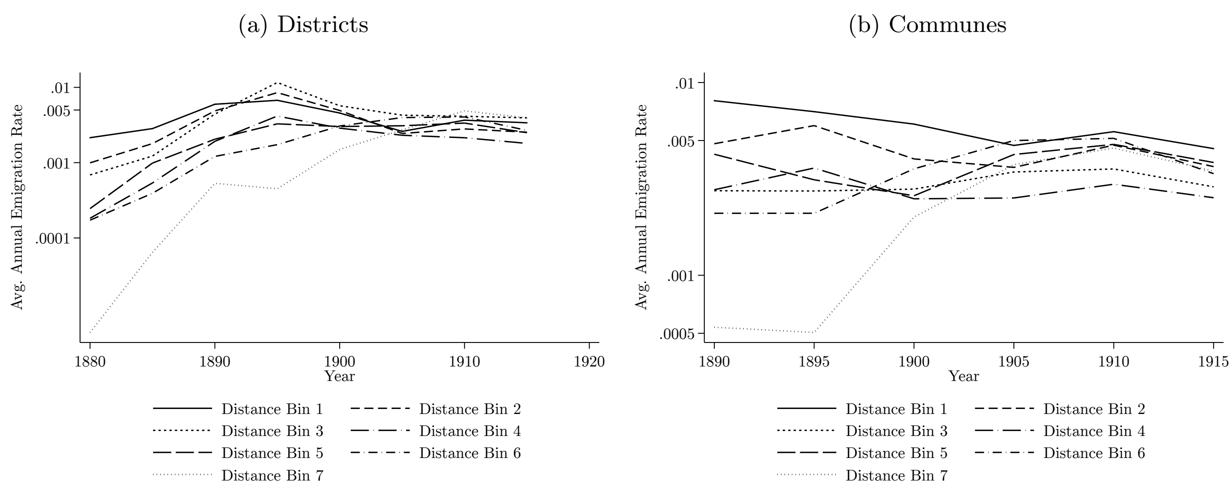


Figure A.6: Time series of migration to South America by distance from epicenter

Note: Distance bins are based on septiles of distance to epicenters. Panel (a) excludes districts more than 300 kilometers from an epicenter. Panel (b) excludes communes more than 350 kilometers from an epicenter.

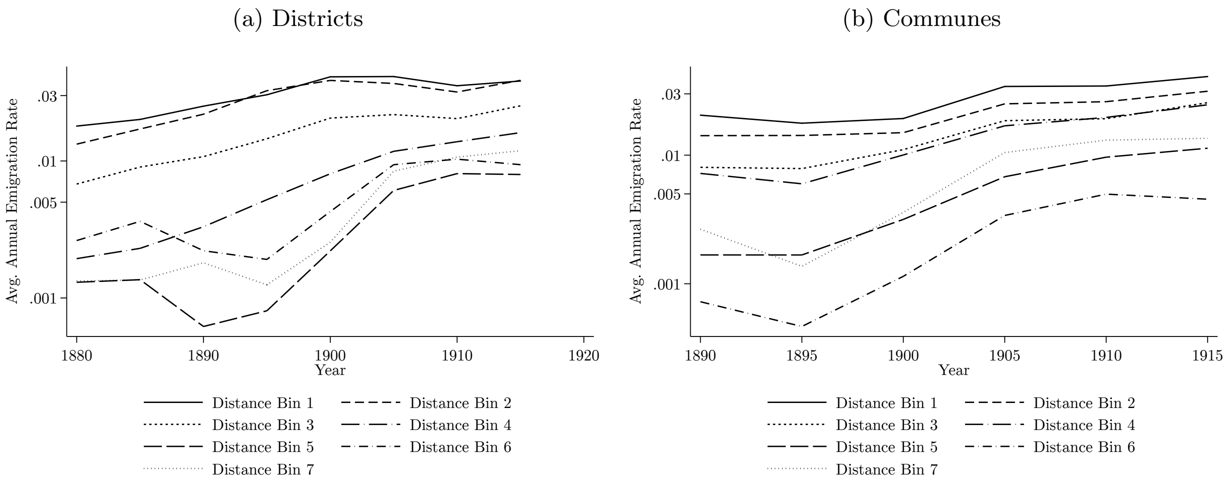


Figure A.7: Time series of migration to Europe by distance from epicenter

Note: Distance bins are based on septiles of distances to the border. Panel (a) excludes districts more than 300 kilometers from the border. Panel (b) excludes communes more than 250 kilometers from the border.

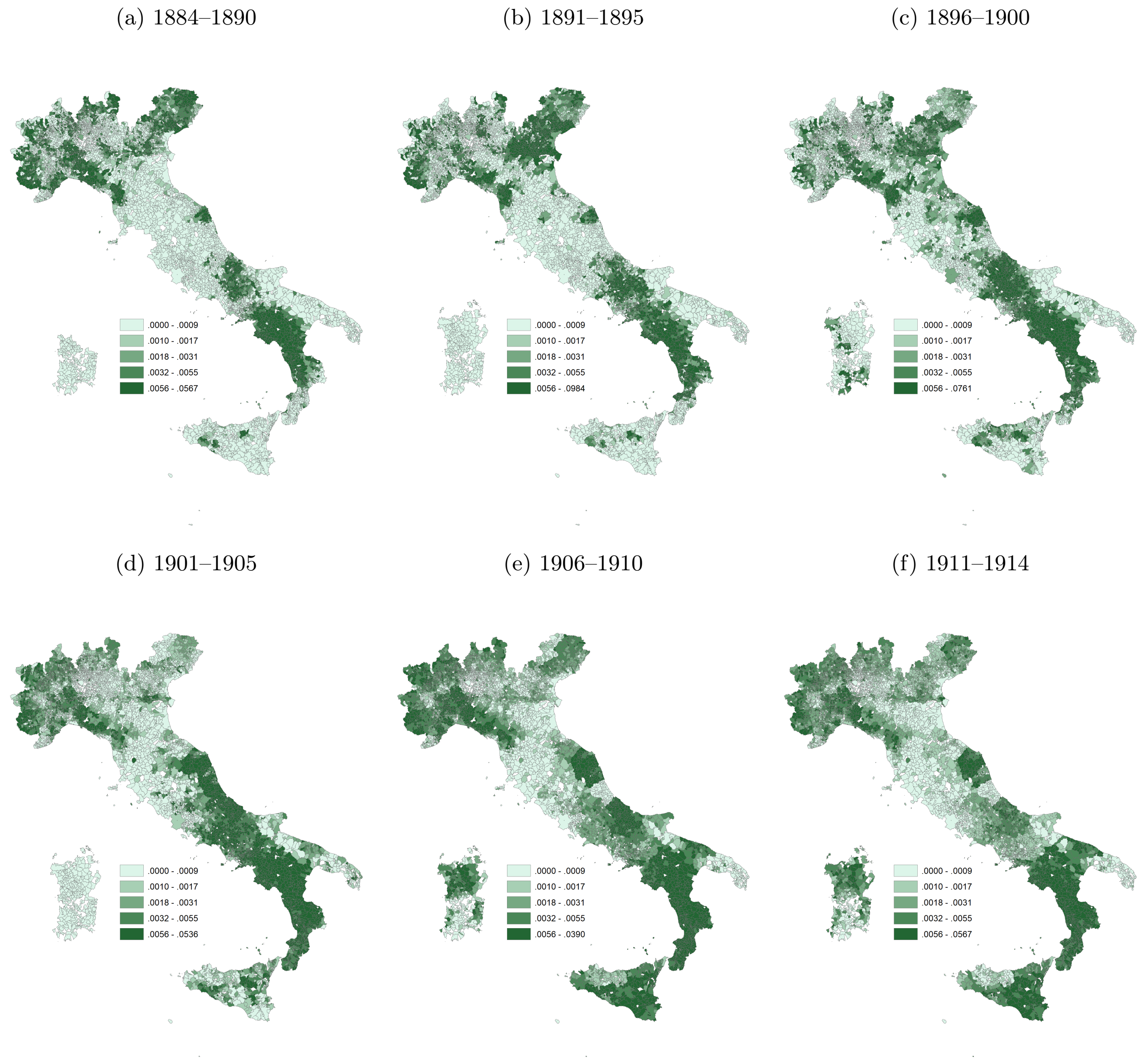
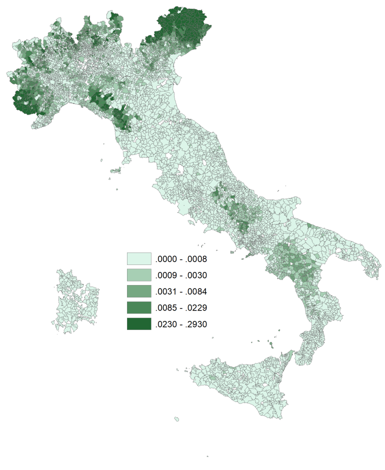


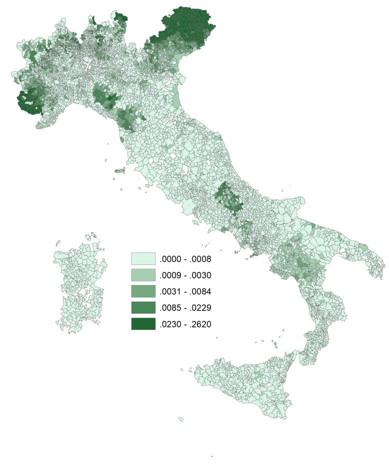
Figure A.8: Commune-level emigration rates to South America

Note: Each panel presents a commune's average annual emigration rate to South America in the period in question. Scale is based on quintiles of emigration rates in 1911–1914.

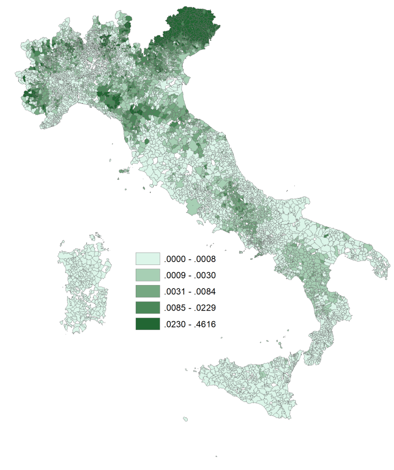
(a) 1884–1890



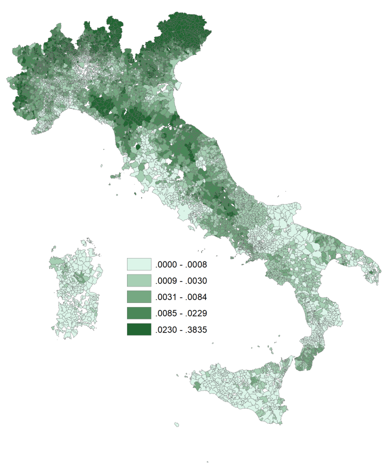
(b) 1891–1895



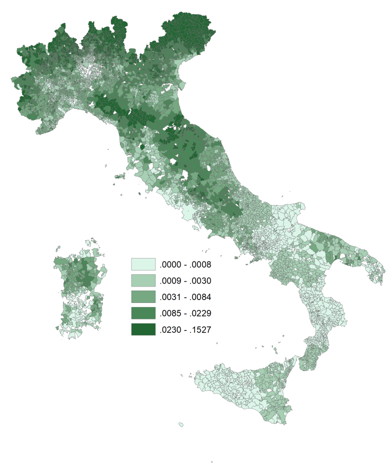
(c) 1896–1900



(d) 1901–1905



(e) 1906–1910



(f) 1911–1914

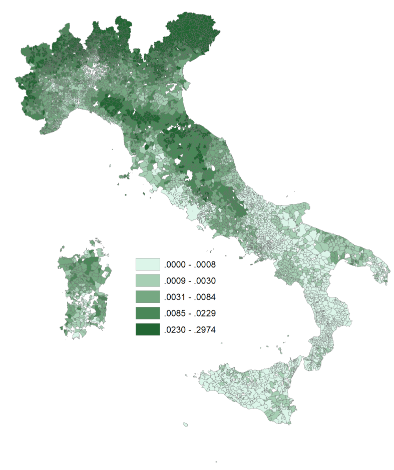


Figure A.9: Commune-level emigration rates to Europe

Note: Each panel presents a commune's average annual emigration rate to Europe in the period in question. Scale is based on quintiles of emigration rates in 1911–1914.

(a) Districts (b) Communes

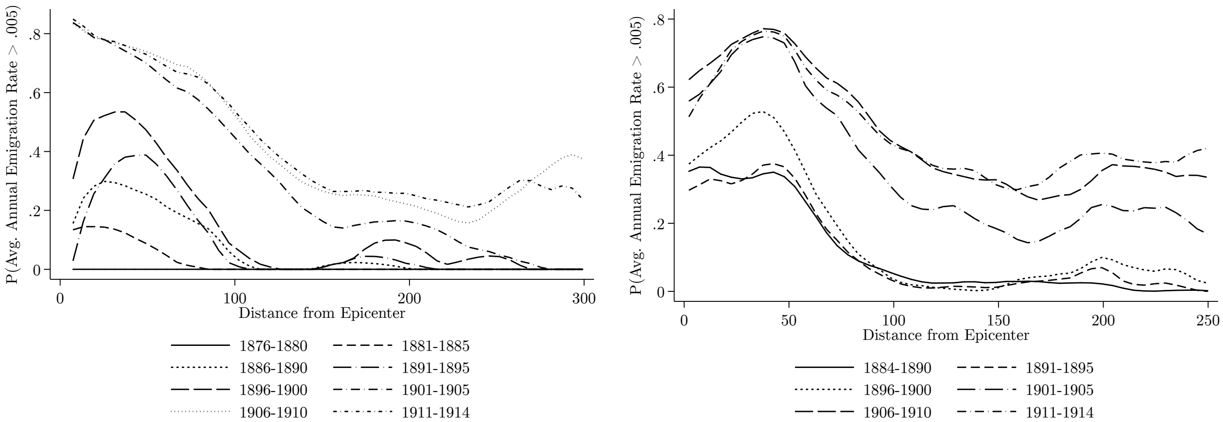


Figure A.10: Probability of achieving emigration rates to North America of at least 0.005 by distance to epicenter (km)

Note: These figures plot local polynomial regressions of an indicator equal to one for communes and districts with an average annual emigration rate in each half decade of at least 0.005 against distance from the epicenters of migration to North America. Epicenter districts are excluded.

(a) Districts (b) Communes

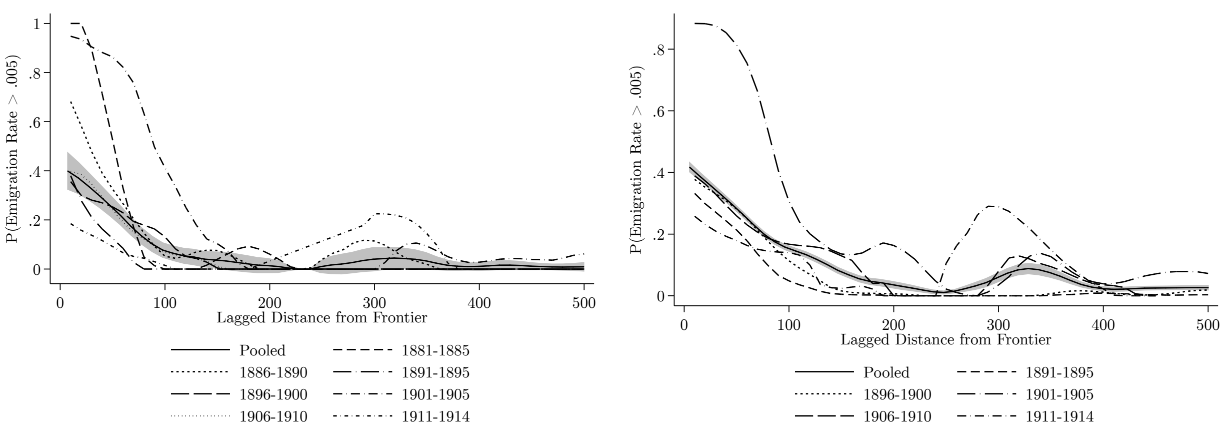


Figure A.11: Probability of achieving an emigration rate to North America of at least 0.005 by distance to the mass migration frontier (km)

Note: These figures present local polynomial regressions of an indicator equal to one for communes and districts with an average annual emigration rate in each half decade of at least 0.005 for the whole sample and for each half decade on the distance from a district that had ever achieved an average annual migration rate of at least 0.005 by the previous half decade, limiting the sample to communes in districts that had not yet achieved this threshold.

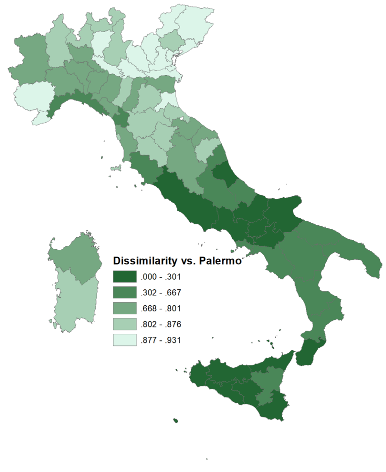


Figure A.12: Destination dissimilarity relative to the province of Palermo

Note: Dissimilarity indices based on emigration data for the entire study period.

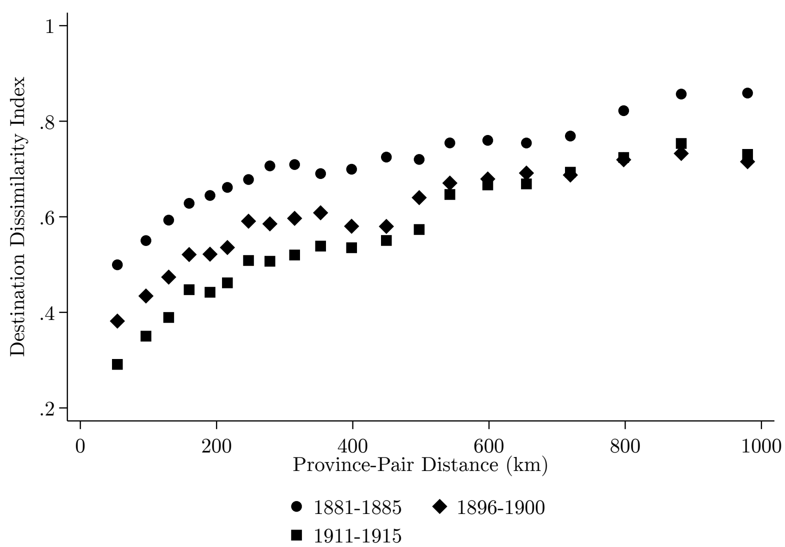


Figure A.13: Destination dissimilarity by distance and half decade

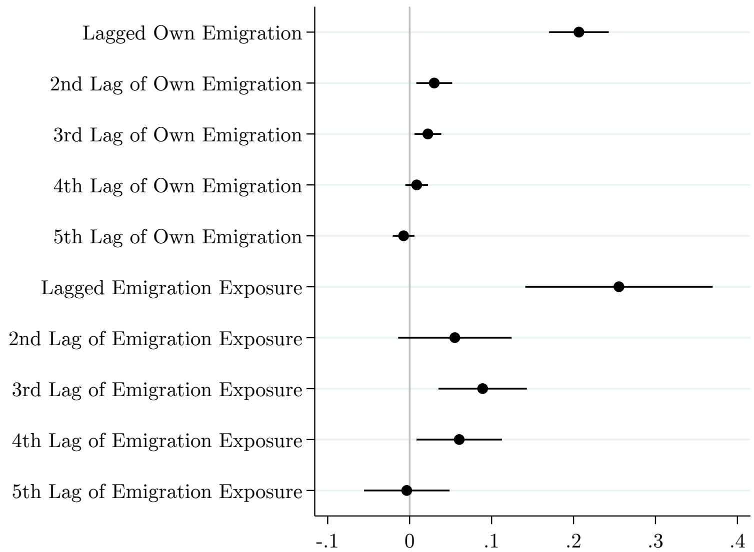


Figure A.14: Multiple lags of own emigration and emigration exposure

Note: These coefficients come from estimating equation (3) with multiple lags of $\log(e_{it})$ and $\log(e_{\neg it})$ on the right-hand side.

B A Model of Migration with Diffusion (For Online Publication)

Our model is a simplification of that proposed by Spitzer (2015) to estimate the impact of transitory income shocks on mass migration. It describes a diffusion process of migration, where the option to migrate is passed on between friends and relatives within and across communities. An individual can migrate only if he is linked. Linkage is formed either through having a contact with a close person who has already migrated before, or it could be created spontaneously, enabling an individual to become a pioneer migrant if he wishes. When a person migrates, he turns some of his friends and relatives who were not yet linked into linked potential migrants.

Mass migration in a community can be ignited by two ways: spontaneous pioneers who infect subsequent friends and relatives, or by relations with friends and relatives in neighboring communities. A key prediction of our hypothesis is that the second cause dominates the first, in the sense that if none of the neighboring communities have developed mass migration yet, then the probability of mass migration forming out of spontaneous linkage is very low. We assume that every individual is linked to a certain number of other individuals of the community, independently of their status.

The main state variables of the model (for commune i in year t) are

$$\mathfrak{S}_{it} = \{M_{it}, L_{it}, S_{it}, N_{it}\},$$

where M_{it} is the fraction of individuals in commune i in year t who have migrated, L_{it} is the fraction of individuals who are linked to a previous migrant but have not yet migrated (but because of being linked have the option to do so), S_{it} is the share of individuals who are not yet linked to a previous migrant (and thus cannot migrate) but are susceptible to being linked, and N_{it} is a measure of exposure of commune i to emigrants in all other communes. That is, N_{it} is a function of the M_{jt} for all $j \neq i$, like the emigration exposure measure above, taking the form

$$N_{it} = \frac{\sum_{j \neq i} M_{jt} P_j d_{ij}^{\pi}}{\sum_{j \neq i} P_j d_{ij}^{\pi}},$$

where P_j is the population of commune j. We assume that

$$M_{it} + L_{it} + S_{it} = 1$$

$$M_{it}, L_{it}, S_{it} \geq 0.$$

The first set of main parameters of the model is

$$\Theta_1 = \{\lambda, \delta, \alpha\},$$

where $\lambda > 0$ is the fraction of individuals in the same commune to which each individual is linked (i.e., the rate of intra-community transmission of the migration technology), $\delta > 0$ is the number of individuals in other communes to which individuals were linked (i.e., the rate of inter-community transmission of the migration technology), and $\alpha > 0$ is the rate at which individuals spontaneously gain the option to emigrate.[70]

Let m_{it} denote the rate of emigration of linked individuals from commune i in year t; that is, $e_{it} = m_{it} \times L_{it}$, where e_{it} is the emigration rate as measured in the reduced form analysis above. We will describe the determinants of m_{it} in more detail below. We assume that in each period, the timeline is as follows: first, individuals who were linked in the previous period decide whether or not to migrate; then, new links are created, caused by individuals who emigrated in the first part of the period (or spontaneous generation). The implied laws of motion for the state variables are then

$$\Delta M_{it} = m_{it} L_{it}$$

for the fraction migrating; that is, the change in the fraction migrating is equal to the fraction of linked individuals who migrate. For the fraction linked, the law of motion is

$$\Delta L_{it} = -m_{it} L_{it-1} + (\lambda \Delta M_{it} + \delta \Delta N_{it} + \alpha - \lambda\delta \Delta M_{it} \Delta N_{it} - \alpha\lambda \Delta M_{it} - \alpha\delta \Delta N_{it} - 2\alpha\lambda\delta \Delta M_{it} \Delta N_{it});$$

that is, those who migrate are lost from among the linked, and then new linked individuals are created, either spontaneously, from linkages to newly migrating individuals in the community, or from linkages to newly migrating individuals in other communities; note that the terms in parentheses after the first three cover cases in which individuals are linked both to a new migrant from within and outside the community, or become spontaneously and non-spontaneously linked simultaneously. Finally, the law of motion for the

[70]The underlying assumption regarding the community network is that links could be represented in a directed random graph, in which the number of other individuals linked to each person is fixed and known, and the links are completely independent from one another (and a person is equally likely to be linked to himself). For example, when a person migrates, he links two other persons, except that they may already be linked or have even moved. Thus, in a community of one-thousand individuals, the first migrant will link, in expectation, $2 \times \frac{999}{1000} = 1.998$ susceptible individuals. If 500 persons have already migrated or are already linked and there are 500 susceptible individuals, then the next migrant will link in expectation $2 \times \frac{500}{1,000} = 1$ individuals. Similarly, each individual is linked to a number of "weighted" individuals in neighboring communities, and has a rate at which he is likely to spontaneously become linked when an additional weighted neighbor migrates.

fraction susceptible is determined simply by those who newly enter the linked category, and is

$$\Delta S_{it} = -(\lambda \Delta M_{it} + \delta \Delta N_{it} + \alpha - \lambda\delta \Delta M_{it} \Delta N_{it} - \alpha\lambda \Delta M_{it} - \alpha\delta \Delta N_{it} - 2\alpha\lambda\delta \Delta M_{it} \Delta N_{it}).$$

We suppose that the migration probability m_{it} is determined by a local mean utility term η_i, reflecting the underlying permanent demand for migration from commune i; an observed local income shock w_{it}; an observed destination income shock v_t; and an unobserved idiosyncratic random term ε_{it}:

$$m_{it} = m(\eta_i, w_{it}, v_t, \varepsilon_{it}).$$

Assuming a logit demand function, this gives

$$m_{it} = \frac{\exp(\eta_i + \beta_0 w_{it} + \beta_1 v_t + \varepsilon_{it})}{1 + \exp(\eta_i + \beta_0 w_{it} + \beta_1 v_t + \varepsilon_{it})},$$

where the idiosyncratic term follows the distribution

$$\varepsilon_{it} \sim N(0, \sigma). \tag{B.1}$$

Note that if the migration probabilities m_{it} are known, the idiosyncratic error term can be written as

$$\varepsilon_{it} = \log\left(\frac{m_{it}}{1 - m_{it}}\right) - (\eta_i + \beta_0 w_{it} + \beta_1 v_t). \tag{B.2}$$

The key parameter that distinguishes the diffusion and internalist explanations for the delayed migration puzzle is α. This parameter determines the degree to which individuals are able to emigrate without linkage to a prior migrant. According to the diffusion hypothesis, this parameter is small, and the takeoff of emigration is determined by the interaction of suitable conditions and linkage to prior migrants. The internalist hypothesis, on the other hand, requires only suitable conditions to be present, and allows for a high α—that is, it does not admit a lack of nearby emigration connection as a meaningful deterrent to emigration.

C The Migration of Antonio Squadrito's Group in October 1903 (For Online Publication)

Antonio Squadrito was born around 1877 in the small Sicilian town of Gualtieri-Sicamino, near Messina. In 1898 he decided to migrate to the United States, among the first in his commune to do so. In New York he had a "distant relative from a northern province," and in his arrival manifest, his American contact was listed as an uncle living on 21st Street, but he paid for his travel by borrowing money from his father Giovanni. He arrived at the Battery in July 7, 1898, and his first job was in a Rhode Island quarry. Soon thereafter, several opportunities arose. He befriended another Italian who had a barbershop in Stonington, Connecticut, and joined him as an employee, gradually paying off his loan. Shortly after, his boss had to leave the business and forced Antonio to take a loan to purchase the shop. The shop prospered, and Antonio had his older married brother Giuseppe come in to help him. Giuseppe was followed by their father and two younger brothers. In June 1903 Antonio married Harriet H. Burtch-Gardiner, who, at 66, was 41 years his senior.[71] In the same summer, he travelled back to his hometown with the purpose of helping the migration of a large number of friends and relatives. By that time, five years after Antonio had first left Sicily, emigration was already widespread in Gualtieri-Sicamino.

Antonio collected a large group of individuals whose migration he facilitated (listed in Table C.1), mainly close and more distant relatives, from Gualitieri-Sicamino as well as from other neighboring places. Among them were his sister-in-law, her four year old niece, her brother, and her nephew, all from Gualtieri-Sicamino and destined for Stonington. The others had other destinations in the US, where they reported having relatives. A sixteen year old girl—a cousin from the neighboring commune of San Filipo—and five young men—all neighbors and family friends from Gualtieri-Sicamino—were traveling to Boston and to New York. Four farmer boys from Soccorso,[72] a detachment of Gualtieri-Sicamino, were on their way to the mines in Pennsylvania. They reported relatives in Philadelphia, but in reality they were illegally contracted laborers, and the uncle of one of them was the middleman who helped recruiting them. The entire group left for Messina en route to Napoli. From Napoli they embarked on the steamship Prinzess Irene on October 2, 1903, and arrived at Ellis Island on October 14. Broughton Brandenburg, the journalist and self-proclaimed immigration specialist who followed Squadrito's entourage and documented their migration, noted that this sort of group migration organized by a friend or relative was so common, that "The most notable feature was the ease with which one could detect that every seventh or eighth person had been to America before,

[71] Or perhaps it was her wealth, estimated at $60,000, that was inherited from her deceased husband, a whaling ship captain (Brandenburg 1904, p. 44).

[72] Brandenburg (1904) mistakenly referred to it as "Socosa."

and how had gathered around him a group of from two to thirty friends, relatives, and neighbors, going over in his care, just as our party was going in the care of Antonio Squadrito and myself" (Brandenburg 1904, p. 172).

In fact, the group was planned to be larger, as they had expected passengers from other communes to join them in Messina. These were Giuseppe Cardillo, accompanied by a few other people, and the Papalia family from Monforte San Giorgio, a small town situated about ten kilometers west of Gualtieri-Sicamino. Cardillo's hometown is unknown and so is the specific relation between the two families and the Squadritos. Eventually, according to Brandenburg (1904, p. 133), Cardillo's group decided to postpone their travel and the Papalias ended up taking the next steamer. Indeed, two weeks later, on October 28, Michele and Maria Papalia, originally from Monforte San Giorgio, and their five year old daughter Rosina were recorded arriving at Ellis Island on board the steamship Lahn, where they were listed as American citizens returning home to New York. All in all, the extended group that planned their joint voyage comprised of neighbors, friends, relatives, and other acquaintance from five different localities, at least four of which were within a short distance from one another.

How does this case fit the theoretical framework proposed in section 3 and Online Appendix 3? Clearly, it shows that the reality was more complex than the stylized story about a linear chain in which one individual links others in his geographic environment who depend on him, and leads them to the same destination. It is not clear, for example, how crucial the role played by Antonio Squadrito's relative was in enabling his own migration in 1898, and therefore it is impossible to tell whether or not he was a real pioneer. Even if he was linked by his relative, it is hard to tell whether this linkage conformed to the assumption about geographic proximity, because although he was a relative, according to Brandenburg he was from a "northern province" (Brandenburg 1904, p. 43). Furthermore, many in the group relied on additional contacts in the US. They were supported by Antonio, but he was not their sole sponsor, and it is probable that they would have migrated even without his help. Indeed, only a few were destined to join him in Stonington. Networks merged and diverged to different destinations, and it is unknown whether the emigration from Gualtieri-Sicamino could be traced back to a single local founding father or to several ancestors separately linked from other communes, and whether any of them were virtual pioneers. Nevertheless, those going to other destinations were still relying on other personal links, usually family members. Even those who were in reality contracted laborers, were recruited through a relative. If the case of Squadrito's group is indicative, then in a broad sense, the Italian transatlantic movement occurred within local networks based both on intra-communal and on short-distance inter-communal links. This is precisely the core insight that the theoretical framework that we propose is meant to capture.

Table C.1: Antonio Squadrito's Group, on board Prinzess Irene, arriving October 14, 1903

First Name	Last Name	Sex	Age	Relation to Antonio Squadrito	Place of Origin	Joining	Destination
Antonio	Squadrito	M	26		Gualtieri-Sicamino	Brothers, Giuseppe, Carmelo, and Gaetano	Stonington, CT
Carmela	Squadrito	F	32	Sister in law	Gualtieri-Sicamino	Husband, Giuseppe Squadrito (Antonio's brother)	Stonington, CT
Caterina	Squadrito	F	4	Niece	Gualtieri-Sicamino	Father, Giuseppe Squadrito	Stonington, CT
Giovanni	Pulejo	M	49	Brother in law, probably also a cousin	Gualtieri-Sicamino	Brother, Nicola	Boston, MA
Felice	Pulejo	M	16	Nephew	Gualtieri-Sicamino	Uncle, Nicola	Boston, MA
Concetta	Fomica	F	15	Cousin	San Filipo	Uncle, Stefano Senedile, Boston	Boston, MA
Antonio	Nastasia	M	16	Neighbor	Gualtieri-Sicamino	Uncle, Tommaso Trovato, Boston	Boston, MA
Gaetano	Mullura	M	16	Neighbor	Gualtieri-Sicamino	Uncle, Nicolo Puleo, Boston	Boston, MA
Nicola	Curro	M	27	Family friend	Gualtieri-Sicamino	Cousin, Angelo Ragusa, New York	New York, NY
Nunzio	Giunta	M	23	Fellow townsman	Gualtieri-Sicamino	Cousin, New York	New York, NY
Antonio	Genino	M	21	Fellow townsman	Gualtieri-Sicamino	Uncle, Giuseppe Maucino, Philadelphia	Philadelphia, PA
Salvatore	Niceta	M	20	Farm boy from detached village	Soccorso	Brother, Giuseppe Niceta, Philadelphia	Philadelphia, PA
Benedetto	Runzio	M	21	Farm boy from detached village	Soccorso	Cousin, Giuseppe Niceta, Philadelphia	Philadelphia, PA
Luciano	Sofia	M	17	Farm boy from detached village	Soccorso	Cousin, Giuseppe Niceta, Philadelphia	Philadelphia, PA
Salvatore	Damico	M	23	Farm boy from detached village	Soccorso	Brother in law, Antonio Salvatore, Philadelphia	Philadelphia, PA

Sources: Brandenburg (1904) and the Statue of Librety-Ellis Island Foundation

D Preparing Official Statistics for Analysis (For Online Publication)

The data that we collected from the *Statistica della Emigrazione Italiana per l'Estero* volumes and from the 1881 Italian census required considerable preparation before they could be used for analysis. At the commune level, the main difficulties are the changing of commune names over time, and the combination or division of communes to form other communes. A key source for this effort was the *Comuni e Loro Popolazione ai Censimenti dal 1861 al 1951*, published by ISTAT in 1960. This publication describes the changing borders of communes, allowing us to create consistently defined communes over the entire sample period.

Our analysis also requires knowing the geographic location of each commune. For communes that still exist (the vast majority), we were able to simply match the list of commune names to a GIS file of modern communes whose historical provinces could be determined using a shapefile of historic province boundaries provided by ISTAT. This was more difficult in the case of historic communes that were consistently defined throughout our study period but have since ceased to exist. For instance, the commune of Santo Stefano di Briga existed throughout our study period, but has since been incorporated into the commune of Messina. The best guess of geographic location that we are able to derive is thus to place Santo Stefano di Briga in the same place as Messina. This simplification is a possible source of error, but because most communes are quite small, the resulting error is likely to be small.

Another issue was the mapping of districts. To our knowledge, no shapefiles of Italian districts exist. We constructed the shapefile that we use by merging the polygons of all communes assigned to a particular district. For communes that were created after our study period, we determined the commune of which they were once a part, and assign the modern commune to the district of the historic commune from which it was split. Comparison of our resultant shapefile to a map that we were able to locate of historical districts shows that our generated shapefile is quite accurate.

Another issue arose from the fact that some northern provinces that were previously part of the Austro-Hungarian Empire (e.g., Veneto) had *distretti* instead of *circondari*. We treat both of these as districts, but the *distretti* are much smaller and were eventually eliminated, creating provinces with a single *circondario*. For the emigration data, we can reconstruct the district totals from the commune-level data. But for the census data, we must use province-level data on literacy and employment for these northern provinces.

E Suggestive Evidence on Destination Choice (For Online Publication)

The analysis of section 5.4 was limited by the fact that the data are only at the province-destination country level. In this appendix, we provide additional but suggestive evidence on the basis of a linkage between passenger manifests of Italians arriving at Ellis Island and US census records of 1910 and 1920.

The passenger manifests that we linked are described in detail by Spitzer and Zimran (2018).[73] We link them to the US census records of 1910 and 1920 according to a variant of the linkage algorithm used by Collins and Zimran (2019a,b). The main deviation from this algorithm is that we allow multiple passengers in the passenger list to link to a single individual in the census records, reflecting the possibility of repeat migration; in particular, we retain cases in which there were up to five passenger records linked to a single census record, but require that a passenger record had no more than one match in the census records. Given the difficulties in making a match in this case, we are not confident in the quality of our matches, which is why we consider the evidence in this appendix to only be suggestive. Ultimately, we are able to link 396,161 passengers (of 4,113,188 eligible) from the passenger lists to the 1920 census, and 268,773 passengers (of 2,806,481 eligible) from the passenger lists to the 1910 census, linking to 170,882 or 126,066 census records, respectively.

Based on the listed last place of residence provided in the passenger list, we assign each linked Italian man in the US census to a place of origin. In cases where he is linked to several passengers of different places of origin, we do this probabilistically; for instance, if an Italian man in the 1910 census is linked to one passenger from Palermo and one passenger from Messina, we consider this to be a half migrant from each place of origin.

In Figure E.1, we construct a migration matrix on the basis of the assigned province of origin and the state of destination and use this to construct a dissimilarity index for destination state in the US, plotting bin scatter plots over distance. Though the patterns are less stark than in Appendix Figure A.13, there is clear evidence that provinces more distant from one another had more dissimilar destination choices within the US. Given the low confidence that we have in our links, this exercise essentially amounts to asking whether individuals with names common to particular Italian places of origin tended to settle in similar places in the US.

In Figure E.2, we distill the level of geographic variation to the district of origin and the county of destination. A large number of empty bins requires us to treat these links and resulting figures with considerable

[73] They are also described by Bandiera, Rasul, and Viarengo (2013) and Spitzer, Tortorici, and Zimran (2019).

skepticism. Nonetheless, there are patterns in this figure that are further supportive of the destination similarity prediction of the diffusion hypothesis. In particular, the x-axis lists the top 25 counties in the US in 1910 by Italian-born male population. Each point indicates how much more likely an individual originating in a particular district who was linked from the Ellis Island passenger records to the 1910 or 1920 census was to live in that county than was an average Italian male in the US.[74] Two particular patterns in this Figure are important. The first is the presence of "spikes" (such as the Genovese in San Francisco) in this Figure, which indicate that the network linkage was likely an important part of the decision of where to migrate. The second is the persistence of these distributions between census years, indicating that new immigrant flows followed these network patterns. Indeed, the correlation between the district-county levels of over-representation in 1910 and 1920 is 0.77.

[74] For example, the point for Genova in San Francisco county for 1910 is derived as follows: 1.37 percent of Italian males in the US lived in San Francisco county in 1910; among individuals in the passenger records who originated in Genova district and were linked to the 1910 census, 10.93 percent lived in San Francisco county; the over-representation of Genovese in San Francisco is then calculated as $\frac{.1093}{.0137} - 1 = 6.99$. The Figure is limited to cases in which there were at least 25 linked individuals from a particular District in a county.

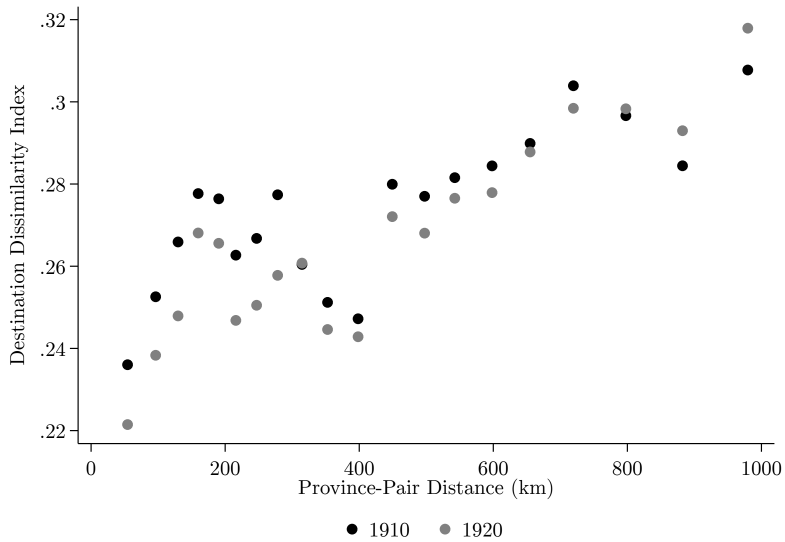
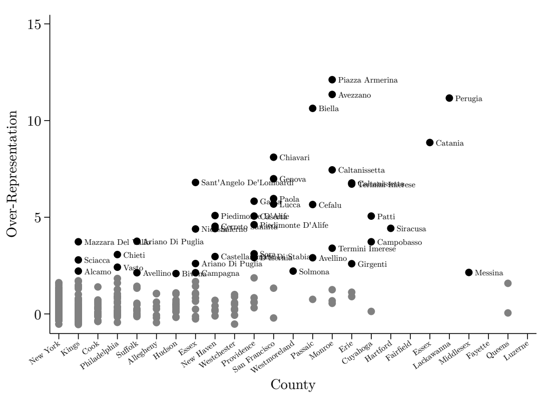


Figure E.1: Destination dissimilarity by distance census year

Note: Dissimilarity index calculated over US states and Italian provinces of origin.

(a) 1910



(b) 1920

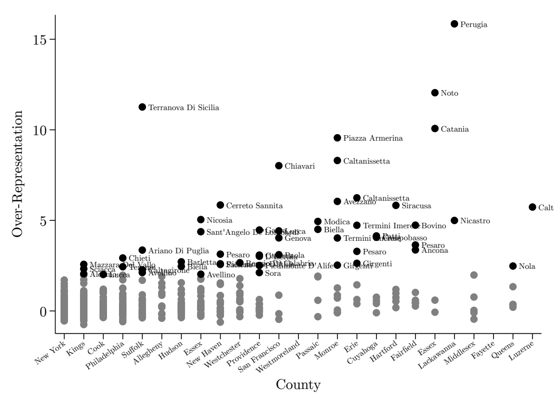


Figure E.2: Evidence of correlated destinations in the US

Note: Each dot is calculated as $\frac{s_{ic}}{s_c} - 1$, where s_{ic} is the probability that an individual from district i settles in county c and s_c is the probability that any Italian male settled in county c.

F Robustness to Choice of θ (For Online Publication)

The values of θ that we use in our main analysis are based on estimation of equation (3) by non-linear least squares without the inclusion of any controls. This yielded the estimated parameters of $\hat{\theta} = -2.28$ for the commune-year data and of $\hat{\theta} = -2.38$ for the commune-half decade data. In this appendix, we explore the robustness of the results to alternative values of θ.

The first robustness check involves performing our non-linear least squares estimation including controls in equation (3). The estimated parameters in this case are $\hat{\theta} = -2.52$ for the commune-year data and $\hat{\theta} = -2.30$ for the commune-half decade data. Given the similarity of these estimates to our benchmark estimates, we do not show results using these estimates.

As our next robustness check, we perform our non-linear least squares estimation including controls and province-year or province-half decade fixed effects in equation (3). This yields estimates of $\hat{\theta} = -3.21$ for the commune-year data and $\hat{\theta} = -1.45$ for the commune-half decade data. The greater magnitude of this parameter for the commune-year data is consistent with our expectations—over a half decade, the effect of a more distant commune on another commune's migration should be greater than in a single year. Tables F.1, F.2, and F.3 repeat the main results of section 6 with these alternative values of θ. Though the magnitudes of our estimates are affected by the use of these alternative parameters, the qualitative patterns that our estimates reveal are robust to the change in the estimated parameter.

Table F.1: Spatial contagion of US-bound migration, OLS, annual

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Own Emigration	0.474^a	0.463^a	0.452^a	0.268^a	0.249^a	0.218^a	0.243^a	0.204^a
	(0.013)	(0.013)	(0.011)	(0.016)	(0.012)	(0.010)	(0.011)	(0.036)
Lagged Emigration Exposure	0.236^a	0.203^a	0.207^a	0.278^a	0.252^a	0.153^a	0.300^a	0.204^a
	(0.016)	(0.016)	(0.016)	(0.028)	(0.024)	(0.013)	(0.040)	(0.043)
Latitude		0.076						
		(0.055)						
Longitude		−0.026						
		(0.041)						
log(Elevation)		0.070^a						
		(0.013)						
District Ag. Employment Share		0.645^b						
		(0.249)						
District Ind. Employment Share		0.970^a						
		(0.362)						
District Literacy Rate		0.149						
		(0.194)						
District Under 15 Fraction		2.936^a						
		(0.838)						
log(Rail Distance + 1)		0.012^a						
		(0.002)						
Lagged Emigration Exposure (SA)							−0.007	
							(0.030)	
Lagged Emigration Exposure (EU)							-0.072^c	
							(0.036)	
Interaction								−0.007
								(0.005)
Observations	145,126	144,113	144,113	145,081	144,068	144,059	144,068	144,068
R-squared	0.371	0.382	0.445	0.186	0.285	0.092	0.287	0.286
Province FE	Yes	Yes	Yes	No	No	No	No	No
Commune FE	No	No	No	Yes	Yes	Yes	Yes	Yes
Province-Year FE	No	No	No	No	No	Yes	No	No
Controls	No	No	Yes	No	Yes	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Standard errors clustered at the province level. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-year. Emigration rate regressors included as logs. Controls include year-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Interaction is the product of lagged own emigration rate and lagged emigration exposure to North America. R-squared is computed after absorbing fixed effects.

Table F.2: Spatial contagion of US-bound migration, OLS, half decade

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged Own Emigration	0.443^a	0.413^a	0.404^a	0.100^a	0.073^a	0.052^a	0.070^a	-0.476^a
	(0.018)	(0.017)	(0.017)	(0.017)	(0.018)	(0.016)	(0.018)	(0.089)
Lagged Emigration Exposure	0.498^a	0.410^a	0.643^a	0.806^a	1.108^a	1.055^a	1.147^a	0.781^a
	(0.128)	(0.137)	(0.113)	(0.207)	(0.123)	(0.116)	(0.148)	(0.152)
Latitude		0.140						
		(0.086)						
Longitude		-0.107						
		(0.071)						
log(Elevation)		0.096^a						
		(0.022)						
District Ag. Employment Share		1.080^a						
		(0.381)						
District Ind. Employment Share		1.649^a						
		(0.542)						
District Literacy Rate		0.399						
		(0.305)						
District Under 15 Fraction		5.525^a						
		(1.187)						
log(Rail Distance + 1)		0.011^a						
		(0.003)						
Lagged Emigration Exposure (SA)							-0.139	
							(0.146)	
Lagged Emigration Exposure (EU)							0.156	
							(0.126)	
Interaction								-0.080^a
								(0.013)
Observations	36,199	35,899	35,899	36,115	35,815	35,812	35,815	35,815
R-squared	0.320	0.347	0.421	0.087	0.225	0.055	0.226	0.243
Province FE	Yes	Yes	Yes	No	No	No	No	No
Commune FE	No	No	No	Yes	Yes	Yes	Yes	Yes
Province-Half Decade FE	No	No	No	No	No	Yes	No	No
Controls	No	No	Yes	No	Yes	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Standard errors clustered at the province level. Dependent variable is the log of the emigration rate to North America. Unit of observation is a commune-half decade. Emigration rate regressors included as logs. Controls include half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Interaction is the product of lagged own emigration rate and lagged emigration exposure to North America. R-squared is computed after absorbing fixed effects.

Table F.3: Spatial contagion regressions, IV

Variables	(1) 1 Yr	(2) 1 Yr	(3) 1 Yr	(4) 5 Yr	(5) 5 Yr	(6) 5 Yr
Panel A: OLS						
Lagged Own Emigration	0.265^{a}	0.245^{a}	0.218^{a}	0.107^{a}	0.080^{a}	0.052^{a}
	(0.015)	(0.011)	(0.010)	(0.016)	(0.018)	(0.015)
Lagged Emigration Exposure	0.217^{a}	0.211^{a}	0.149^{a}	0.527^{a}	0.833^{a}	1.035^{a}
	(0.022)	(0.023)	(0.013)	(0.175)	(0.155)	(0.118)
Panel B: Stage 0						
F-statistic of Instrument	17.757	10.714	9.273	16.272	22.682	3.591
Panel C: First Stage						
Lagged Own Emigration	0.353^{a}	0.289^{a}	0.125^{a}	0.075^{a}	0.061^{a}	0.023^{a}
	(0.019)	(0.015)	(0.005)	(0.010)	(0.007)	(0.003)
Lagged Instrumented Exposure	0.946^{a}	0.972^{a}	1.005^{a}	0.980^{a}	0.443^{a}	0.814^{a}
	(0.050)	(0.033)	(0.024)	(0.104)	(0.102)	(0.097)
Panel D: IV						
Lagged Own Emigration	0.310^{a}	0.251^{a}	0.214^{a}	0.105^{a}	0.005	0.021
	(0.033)	(0.016)	(0.010)	(0.026)	(0.031)	(0.018)
Lagged Emigration Exposure	0.102	0.191^{a}	0.169^{a}	0.552^{b}	1.980^{a}	2.200^{a}
	(0.065)	(0.034)	(0.024)	(0.259)	(0.435)	(0.454)
Controls	No	Yes	Yes	No	Yes	Yes
Province-Year FE	No	No	Yes	No	No	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: All regressions include commune and year fixed effects and year-specific functions of log distance from epicenter. Controls are year- or half decade-specific functions of latitude, longitude, elevation, agricultural employment share, industrial employment share, literacy rate, fraction under age 15, and distance to railroad. Standard errors clustered at the province level.

G Nickell Bias (For Online Publication)

Although dynamic panel estimates are unbiased, a natural concern is that the results of section 6 are asymptotically inconsistent as the number of communes (but not time periods) becomes large (Nickell 1981), and that this might affect the results. In an equation of the form of equation (3), Nickell (1981) shows that the asymptotic bias on the estimate of γ (the coefficient on the own lag of migration) is

$$\text{bias}(\hat{\gamma}) = -\frac{1+\gamma}{T-1}$$

and that the asymptotic bias on the estimate of β (the coefficient on any particular regressor, but in this case the coefficient on the lag of emigration exposure) is

$$\text{bias}(\hat{\beta}) = -\hat{\eta} \times \text{bias}(\hat{\gamma}). \tag{G.1}$$

In equation (G.1), $\hat{\eta}$ is the coefficient from a regression of $\log e_{it-1}$ on $\log e_{-it-1}$ in a first differenced regression; namely from estimating

$$\log(e_{it-1}) - \log(e_{it-2}) = \alpha_{t-1} + \eta \left[\log(e_{-it-1}) - \log(e_{-it-2})\right] + u_{it-1}$$

It is clear that the magnitude of these biases is increasing in γ, so we consider the case $\gamma = 1$, which would imply perfect persistence of shocks to emigration, as providing an upper bound of the possible bias. The upper bound (in terms of magnitude) on the bias of γ is thus -0.12, using $T = 18$, which is the average number of time periods in the regression. With an estimate of $\hat{\eta} = 1.30$, the upper bound on the bias of β is thus 0.15.[75] Thus the estimated β is too large to be solely the product of this inconsistency.

Unfortunately, we are not able to draw the same strong conclusion for the commune-half decade data because of the smaller number of observations per commune. In this case, the upper bound on the bias of γ is -0.50 with $T = 5$. With an estimate of $\hat{\eta} = 1.42$, the upper bound on the bias of β is thus 0.72, which is larger than our estimated β.[76] We can provide a somewhat tighter bound on the bias by using the estimated γ from column (1) of Table A.1 (which is not subject to this bias because it does not include commune fixed effects). The largest such estimate is 0.391, which produces an upper bound for the bias of β of -0.35, yielding an upper bound of the bias of γ of 0.49, which is approximately the value of our estimate. Thus, we

[75] This estimate is even smaller when province-year fixed effects are included.
[76] Again, this estimate is even smaller when province-half decade fixed effects are included.

cannot rule out that our results are driven by this bias, though it would have to be the case that commune fixed effects explained none of the persistence in a commune's emigration rate over time for our results to be completely explained by this bias.

H Additional Balancing Tests (For Online Publication)

Table H.1 extends the analysis of Table 6 by testing whether distance from epicenters can be used to construct an effective instrument for the 1881 characteristics. In Panel A, we regress a commune's characteristic x_i on its population- and distance-weighted neighbors' characteristics, defined as

$$x_{\neg i} = \frac{\sum_{j \neq i} x_j N_j d_{ij}^{\theta}}{\sum_{j \neq i} N_j d_{ij}^{\theta}}.$$

Naturally, the correlations are positive and strongly statistically significant—an almost mechanical relationship given that columns (1)–(4) are based on district-level data. In Panel B we estimate

$$x_i = \alpha_p + \beta \log(z_i) + u_i,$$

(where α_p are province fixed effects), finding that there is little relationship of these characteristics with distance from epicenters, a balancing test similar to those of Table 6. In Panel C, we use fitted values from Panel B, $\tilde{x}_j = \hat{\beta} \log(z_j)$, to compute

$$\tilde{x}_{\neg i} = \frac{\sum_{j \neq i} \tilde{x}_j N_j d_{ij}^{\theta}}{\sum_{j \neq i} N_j d_{ij}^{\theta}}$$

and estimate a first-stage regression of neighbors' characteristics $x_{\neg i}$ on $\tilde{x}_{\neg i}$. In Panel D, we estimate analogs of Panel A by instrumental variables. In no case is there a strong first and second stage of the instrumental variables estimate, even though there is a strong relationship in Panel A. Thus, the relationship of the instrumental variable with development as of 1881 does not indicate a threat to the assumption that neighbors' proximity to epicenters did not affect migration through the diffusion of push factors.

Table H.1: Epicenter distance is a poor instrument for predetermined characteristics

Variables	(1) Ag Share	(2) Ind Share	(3) Young Frac	(4) Lit Rate	(5) Rail Dist	(6) Elevation
Panel A: OLS						
Neighbors' Characteristic	1.404^a	1.436^a	1.470^a	1.316^a	1.495^a	1.597^a
	(0.078)	(0.080)	(0.047)	(0.057)	(0.043)	(0.039)
Panel B: Stage 0						
log(Distance to Epicenter)	−0.037	0.025	−0.002	0.037^b	0.107	0.044
	(0.024)	(0.016)	(0.003)	(0.016)	(0.545)	(0.216)
Panel C: First Stage						
Estimated Characteristic	0.380^c	0.094	0.060^a	0.142	0.945^a	0.321
	(0.228)	(0.423)	(0.019)	(0.205)	(0.307)	(0.208)
Panel D: IV						
Neighbors' Characteristic	−0.015	7.477	−0.598	3.241	−0.288	1.712^a
	(0.761)	(27.913)	(0.602)	(3.298)	(0.463)	(0.179)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable listed in column header. All regressions include province fixed effects. Standard errors clustered at the province level. Ag Share is 1881 share of district employment in agriculture. Ind Share is 1881 share of district employment in industry. Young Frac is 1881 share of district population under age 15. Lit Rate is the 1881 adult male literacy rate at the district level. Rail Dist is 1881 log distance of the commune to a rail line. Elevation is the log of elevation of the commune. Unit of observation is a commune.

References (For Online Publication)

Bandiera, Oriana, Imran Rasul, and Martina Viarengo (2013). "The Making of Modern America: Migratory Flows in the Age of Mass Migration." *Journal of Development Economics* 102, pp. 23–47.

Battery Conservancy (2009). "Castle Garden Database." Retrieved from `http://www.castlegarden.org`.

Brandenburg, Broughton (1904). *Imported Americans: The Story of the Experiences of a Disguised American and His Wife Studying the Immigration Question.* New York: Frederick A. Stokes Company.

Collins, William J. and Ariell Zimran (2019a). "Immigrants' Changing Labor Market Assimilation in the United States during the Age of Mass Migration." NBER Working Paper 26414.

——— (2019b). "The Economic Assimilation of Irish Famine Migrants to the United States." *Explorations in Economic History* 74, 101302.

Nickell, Stephen (1981). "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49:6, pp. 1417–1426.

Spitzer, Yannay (2015). "The Dynamics of Mass Migration: Estimating the Effect of Income Differences on Migration in a Dynamic Model with Diffusion." Mimeo., Hebrew University of Jerusalem.

Spitzer, Yannay, Gaspare Tortorici, and Ariell Zimran (2019). "International Migration Responses to the 1908 Messina-Reggio Calabria Earthquake." Mimeo., Vanderbilt University.

Spitzer, Yannay and Ariell Zimran (2018). "Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925." *Journal of Development Economics* 134, pp. 226–247.