Talents and Cultures: Immigrant Inventors and Ethnic Diversity in the Age of Mass Migration

Francesco Campo
University of Milan Bicocca

Mariapia Mendola
University of Milano Bicocca and IZA

Andrea Morrison
University of Pavia, ICRIOS-Bocconi University and Utrecht University

Gianmarco Ottaviano
Bocconi University, BAFFI-CAREFIN, CEP, CEPR, IGIER, and IZA

14 March 2022

Abstract
We investigate the importance of co-ethnic networks and diversity in determining immigrant inventors' settlements in the US by following the location choices of thousands of them across counties during the Age of Mass Migration. To do so, we combine a unique USPTO historical patent dataset on immigrants who arrived as adults with Census data, and exploit exogenous variation in both immigration flows and diversity induced by former settlements, WWI and the 1920s Immigration Acts. We find that co-ethnic networks play an important role in attracting immigrant inventors. Yet, we also find that immigrant diversity acts as an additional significant pull factor. This is mainly due to externalities that foster immigrant inventors' productivity. (JEL: F22, J61, O31)

Keywords: International Migration, Cultural Diversity, Innovation.

1. Introduction

During the Age of Mass Migration the United States became the world's prominent industrial nation (Hughes, 2004). Immigrant inventors played a crucial role in making the United States an innovation powerhouse, by bringing new knowledge from their countries of origin (Diodato et al., 2021; Moser et al., 2014) and contributing to the long-term technological development of the US

Acknowledgments: Previously circulated as "Immigrant Inventors and Diversity in the Age of Mass Migration". We thank the Editor Paola Giuliano and two anonymous referees for their insightful comments and suggestions. We received valuable comments from Dario Diodato, Michel Serafinelli, Marco Tabellini, Alexander Whalley and seminar participants at LSE, Utrecht, Bocconi, Bicocca and Bari. We are grateful to Samuel Bazzi for sharing US–frontier data with us, and to Tanguy Millardet, Navid Nobani, Elham Talebbeydokhti and Alessandro Vaccarino for invaluable support with data preparation. Morrison acknowledges the Marie Skłodowska-Curie Individual Fellowships (H2020-MSCA-IF-2017 Project number: 789505) 'GOTaM Cities' for financial support. The usual disclaimer applies.

E-mail: francesco.campo@unimib.it (Campo); mariapia.mendola@unimib.it (Mendola); andrea.morrison@unipv.it (Morrison); gianmarco.ottaviano@unibocconi.it (Ottaviano)

innovation system (Akcigit et al., 2017). This paper investigates the importance of ethnic networks and diversity as pull factors behind their consequential settlement decisions.

By 'immigrant inventor' we refer to an immigrant filing and obtaining at least one US patent in the early years since arrival when still registered as foreign national. We therefore exclude patentees arrived in their childhood and trained in the US as well as immigrants arrived as adults but patenting much later on during their professional life. The idea is that 'immigrant inventors' are more likely to have built their skill base before coming to the US and thus can be indeed considered as 'foreign talents', whereas the profiles of the two categories of patentees we exclude could arguably be deemed as at least partially 'local talents'. By analyzing the characteristics of US counties where immigrant inventors choose to reside, we study whether they are attracted or repulsed by the presence of immigrants from the same origin ('co-ethnic network'), immigrants from other origins ('between diversity') and the composition of immigrants from other origins ('within diversity'). We also investigate the mechanism behind attraction ('pull') or repulsion ('push') by identifying the dominant role played by consumption or production amenities.

To guide our empirical investigation we first develop a simple model of immigrant inventors' location decision across US counties in the wake of Roback (1982) and Ottaviano and Peri (2005, 2006). Taking the decision to migrate to the US as predetermined, in the model immigrant inventors freely select the county to work and live in based on both labor market and quality of life considerations. They are employed in a perfectly competitive innovation sector whose patents feed the production of a final aggregate good, which is itself produced under perfect competition and freely traded nationwide.[1] Immigrant inventors consume the final good as well as a bundle of locally supplied non-tradable services. These services are also used in the innovation sector to complement immigrant inventors' employment.[2]

The presence of other immigrants affects immigrant inventors' location choices through the supply of non-tradable services and two localized externalities. Specifically, it affects their productivity through a 'production amenity' and their utility through a 'consumption amenity'. In equilibrium immigrant inventors are indifferent between alternative counties as the net

1. For our purposes introducing trade costs for the final good (which might have been non-negligible in the period we study) would complicate the analysis of the model without altering its main insights.

2. Non-tradable services are aimed to capture in a simple way the fact that geographic locations provide different levels of access to financial and physical capital, technology, complementary institutions, and workers, which all impact the quality and productivity of the available jobs (Moretti, 2012). Moreover, many high–skilled occupations show agglomeration effects, where an individual worker's productivity is enhanced by being near to or working with many other workers in similar sectors or occupations (Glaeser and Resseger, 2010).

effect of the two externalities is capitalized in the price of local non-tradable services, which itself depends on the local density of immigrant inventors. For instance, should immigrant inventors favour a certain county over the average county, their higher density in the former would drive the local price of non-tradable services above the national average. For them to be nonetheless happy with locating there, it must be that they enjoy a localized production amenity or a localized consumption amenity that compensates them for more expensive services. If their productivity is above the national average, this means that they are compensated by a production amenity; if their productivity is below the national average, this means that they are compensated by a consumption amenity.

The model's empirical implications are twofold in terms of assessing the mechanism through which other immigrants affect the location decisions of immigrant inventors. First, all the rest given, if immigrant inventors flock to (away from) counties where other immigrants are concentrated, this means that they are attracted (repelled) by an overall amenity (disamenity). Second, if in the case of amenity in those counties immigrant inventors are more (less) productive, this means that they are attracted by a production (consumption) amenity. Analogously in the case of disamenity, if immigrant inventors are more (less) productive, it means that they are deterred by an immigrant consumption (production) disamenity.

We test these implications exploiting the US experience during the Age of Mass Migration from 1870 to 1920, when more than 30 million people migrated to the US mainly from different parts of Europe with high variation in the number and the mix of immigrants both across US counties and over time (Abramitzky and Boustan, 2017; Hatton and Williamson, 1998; Bandiera et al., 2013). Typically the international mobility of talented individuals is particularly difficult to study as it requires data on the cross-country location choices of migrants in the upper tail of the skill distribution, which are available only in very rare and specific settings (see, e.g., Kleven et al. (2013)). What makes this period of massive inflows of foreigners particularly attractive for our purposes is that it provides us with a unique opportunity to identify global talents thanks to fast track naturalization. The reason is that the 1802 naturalization law, which would be in place for over 100 years, allowed any foreigner (i.e. free white male) who had been in residence for five years to be admitted to citizenship, which implies that a patentee registered as a foreign citizen in a patent record by the United States Patent and Trademark Office (USPTO) can be arguably classified as 'immigrant inventors'.[3]

Our data are drawn from two sources. For the outcome variables of immigrant inventors (presence and productivity), we exploit an original dataset compiled by Diodato et al. (2021), which identifies immigrant inventors in

3. A concrete example can be found in Mendola et al. (2020).

historical patent documents of the United States Patent and Trademark Office (USPTO). The dataset has been generated through a text-mining algorithm, analogous to the one described in Petralia et al. (2016), and a semi-automated procedure that extracts detailed information, from digitalized patent records, on both country of origin and county of residence for inventors arrived in the US between 1870 and 1940. It contains about 43,000 patents granted to about 20,000 immigrants together with the patentees' counties of residence as reported in the patent records.

For the explanatory variables related to county characteristics, the foregoing pieces of information are matched with NHGIS IPUMS county-level decennial census files (Manson et al., 2019) between 1870 and 1930.[4] We focus on counties that in each census year have at least $2,500$ inhabitants (which is the IPUMS threshold used to distinguish 'urban' from 'rural' counties) and at least one foreign-born resident. This generates a balanced panel of 1911 counties for census years 1870 to 1930. For each county c we obtain the shares of immigrants in the local population by country of origin e. We then use these shares to compute our main variables explaining immigrant inventors' outcomes in county c from country e: the share of all immigrants from country e in the population of county c, the share of immigrants from countries other than e in total population of county c, and the dispersion of county c's immigrants across countries of origin other than e. The first explanatory variable is meant to capture the role of co-ethnic networks (McKenzie and Rapoport, 2007). The second and third explanatory variables are meant to capture between diversity and within diversity respectively (Ottaviano and Peri, 2005, 2006; Ager and Brückner, 2013).

We then assess the impacts of our explanatory variables measured in each census year on immigrant inventors' outcomes in the subsequent decade. This implies that the last census year we consider for the explanatory variables is 1930 and the last decade we consider for the outcomes is 1930-1940. We exploit variation in co-ethnic networks and diversity across counties and ethnicities over time.[5] The unit of analysis is the sub-population cell defined by county of residence c and ethnicity e and we study how within-cell changes in co-ethnic networks and diversity affect within-cell changes in immigrant inventors' outcomes. We include county-by-ethnicity fixed effects, absorbing time-invariant ethnic-specific local pull factors, plus state-by-year fixed effects, accounting for state-specific shocks, and county-by-ethnicity

4. The decennial Census files are available on IPUMS NHGIS site: https://data2.nhgis.org/main.

5. In USPTO data immigrant status is identified from foreign nationality. Differently, in census data it is identified from foreign birthplace. Accordingly, the co-ethnic network of immigrant inventors with ethnicity e consists of all immigrants born in the foreign country the immigrant inventors were national of when they were granted their first US patent.

(linear) time trends, in order to account for any cell-specific trajectory over time in immigrant inventors' outcomes and explanatory variables.

As immigrants are not randomly assigned across counties, but rather self-select according to individual and local factors, OLS estimates would be biased if unobserved (county or ethnicity) time-varying factors simultaneously affected immigrants' local presence, ethnic composition and innovation activity. Moreover, local innovation shocks, as well as the inflows of immigrant inventors, may affect immigration and ethnic diversity if their economic impact results in significant labour demand shifts at county level and these are serially correlated. We deal with these potential biases in two ways. First, as immigrants tend to geographically cluster along ethnic lines, for our baseline regressions we construct a set of shift-share instrumental variables for each potentially endogenous explanatory variable following the canonical approach based on pre-existing immigrant settlements (Card, 2001). Second, as a robustness check, we exploit the quasi-experimental variation provided by the breakout of WWI and the Immigration Acts passed in 1921 and 1924. These acts restricted the number of new immigrants through quotas based on their birthplace and de facto ended the Age of Mass Migration (King, 2009; Ager and Hansen, 2017; Tabellini, 2020). Discrimination by birthplace exogenously changed the ethnic mix of immigrants.

We find that co-ethnic networks play an important role in attracting immigrant inventors. However, between diversity and within diversity also act as significant pull factors, with the dominant driving force identified in production rather than consumption amenities. These findings are robust to checks of instruments' validity (Goldsmith-Pinkham et al., 2020) and to the inclusion of potential confounding factors such as counties' population (Ager and Brückner, 2013) and exposure to the American frontier (Bazzi et al., 2020). We also test whether the estimated effects are heterogeneous according to population size, and find that these are mainly driven by counties in the third tercile of baseline distribution of population (more than 18,000 inhabitants). The analysis on the mechanisms shows that ethnic diversity, both 'between' and 'within', positively affects the degree of skills heterogeneity (both at industry and occupational level) among migrants' population. The production mechanism we uncover is consistent with different, non-mutually exclusive explanations highlighted in the literature (Peri, 2016). Immigrant inventors could benefit from the heterogeneous set of skills and ideas associated with immigrant diversity as long as these were complementary to their own skills in knowledge production. A more diverse environment could promote the circulation of ideas and knowledge spillovers, as well as a better understanding of the state of technology. Diversity may be conducive to an environment that is more tolerant toward creative destruction and thus more fertile for inventors to grow their own innovations. Differently from other immigrants, inventors may be less exposed to the costs of navigating diversity thanks to better communication and cognitive skills that lower linguistic and cultural

barriers (Giuliano, 2007; Algan and Cahuc, 2010). We further rule out a series of alternative interpretations of our findings by considering other mechanisms through which diversity may attract inventors, including inter–group connections, as proxied by inter-ethnic marriage and residential contact, cultural proximity and natives' attitudes as inferred by migrant ethnic groups' salience on newspapers.

Our findings speak to the literature on the impact of immigrant diversity on economic productivity and growth (Alesina et al., 2016; Suedekum et al., 2014; Docquier et al., 2018; Bahar et al., 2020). At the local level this issue has been studied, among others, by Ottaviano and Peri (2005, 2006) and Ager and Brückner (2013). At the firm or team level, it has been investigated, among others, by Ozgen et al. (2014), Boeheim et al. (2012), Kahane et al. (2013) and Kemeny (2017). Differently from this literature, we are interested in whether immigrant diversity attracts or deters global talents to choose where to live. Our analysis also speaks to studies that have gathered evidence on the regional distribution and location choices of immigrants, showing that co-ethnic networks, wages and economic prosperity playing a prominent role among pull factors (McKenzie and Rapoport, 2007). These studies do not target the specific role of diversity in attracting global talents. Finally, our analysis complements existing works investigating immigrant patentees from a historical perspective, which mainly focus on the impact on the receiving economies' technological trajectories (see, e.g., Arkolakis et al. (2019); Moser et al. (2014); Moser and San (2019); Diodato et al. (2021); Akcigit et al. (2017); Ottinger (2020)). Differently from these works, we take the reverse angle and investigate the impact of receiving economies' characteristics on immigrant patentees' location choices. Moreover, while these works consider the entire population of immigrant patentees regardless of when they started patenting, we focus instead on a subset of them at the top of the skill distribution, who fit the notion of global talents. Our findings are relevant for today's advanced economies that have became major receivers of migrants' flows and, in the a long term perspective, have started thinking about immigration in terms of not only the level but also composition.

The rest of the paper is organized as follows. Section 2 provides a brief account of the historical context. Section 3 presents the model that informs our empirical analysis. Section 4 introduces our dataset. Section 5 describes our empirical strategy. Section 6 discusses our findings. Section 7 presents the robustness checks. Section 8 offers some concluding remarks.

2. Historical Context

Immigration to the US during the Age of Mass Migration (1870-1920) is remarkable for many reasons. First, it is estimated that more than 30 million people migrated, which makes this period the one with the highest inflow

of immigrants in US history (Hatton and Williamson, 1998). Mass migration ended by the 1920s, when country specific quotas were enforced (more details below). By this time, the share of immigrants had reached its highest peak at 14% of the total US population.

Second, immigration originated prevalently from Europe. However, differently from previous inflows, immigrants were sourced from a wide variety of countries and also from different regions within each country. Diversity was spurred by several consecutive waves of immigration. These started in the early nineteenth century with the migration of northern Europeans, prevalently from Ireland, Germany and England. By 1880 the composition of inflows shifted towards Germans and Scandinavians. By the end of these first waves the immigrant stock in the US consisted prevalently of northern and western Europeans. Towards the turn of the century a new wave of immigration brought to the US mainly eastern and southern Europeans, who quickly reached a share of the total stock of immigrants similar to the previous immigrant waves (roughly around 40% of the foreign born population) (Abramitzky and Boustan, 2017).

Third, the newly formed immigrant communities in the US were highly clustered in space, and formed ethnic enclaves in cities and regions (Abramitzky and Boustan, 2017). For example, Germans were the largest group in the lower Mid-West, while Scandinavians represented the largest group in the upper Mid-West. Italians tended to cluster in East Coast counties and cities like New York, Boston and Rhode Island, while they were almost absent in many counties of Wisconsin and Minnesota. Clustering was strong also within urban areas, where immigrant communities tended to form ethnic enclaves. However, there were differences in location patterns by ethnicity as well as by immigrant wave. The early waves of immigrants showed stronger patterns of concentration, forming urban ghettos closely delimited in specific neighbourhoods where they reproduced the life-style of their countries, if not regions, of origin. Subsequent waves tended to be more dispersed. Immigrants from different ethnic groups followed own localization patterns and became more or less dispersed. For example, Germans represented a rather heterogeneous community, divided along religious and regional lines (e.g. catholic and protestant; Bavarian like Rudolf Eickemeyer and Prussian like Charles Steinmetz). They were also rather diversified in terms of occupations and class structure. All these differences, on top of the large size of the German immigrant population, favored a more diffused urban distribution, which was not the case for other ethnic communities (Bergquist, 1984).

Fourth, although the vast majority of immigrants were unskilled and of humble origin, a non-negligible part consisted of skilled workers and professionals. Differently from contemporary waves of migration, during the Mass Migration immigrants were both positively and negatively selected (Hatton and Williamson, 1998). Moreover, differences in skills and professional experience were significant across immigrant groups from different countries of

origin. German and British tended to be more skilled than natives in specific trades, whereas Italians were usually negatively selected often proceeding from poorer southern Italian regions (Abramitzky and Boustan, 2017). Therefore immigrants contributed to the growing US economy by providing unskilled labor but also relevant skills and know-how for the US industry and agriculture (Sequeira et al., 2020), threby shaping future US comparative advantage (Ottinger, 2020). Immigrants also made a major contribution in terms of scientific and technological discoveries, being overrepresented among inventors and patentees (Khan, 2005; Khan and Sokoloff, 2004; Akcigit et al., 2017). This can be explained by the strong incentive given to invention and technological innovation in the US (Khan, 2005). On the one hand, the US patenting system was relatively inexpensive compared to European countries, which lowered the entry barriers to independent inventors without a large financial endowment. On the other hand, inventive activity in those days required less physical capital and formal education than today and it was therefore primarily carried out by independent inventors (Hughes, 2004), who played a key role in supplying with high-quality innovation the market for technology, even after the emergence of corporate R&D laboratories in the early 20th century (Nicholas, 2010).

Among immigrant inventors, a variety of profiles and backgrounds can be singled out. A first group includes the foreign born who migrated to the US during their childhood or immediately after. These immigrants learned their trade, built their skills and developed all their professional experience in the US. They include both unskilled workers like John F. O'Connor and remarkable scientists and entrepreneurs like Elihu Thomson. John F. O'Connor arrived in the US from Ireland when he was a child. He was the typical inventor who learned on the job the secrets of his trade and, through trial and error, produced several ameliorations of the railroad gearing (Khan, 2005; McFadyen, 1936). His contribution is notable also because he became one of the greatest patentees of his time. Elihu Thomson's history is well known. Of British origins, he moved to the US at the age of five. Thomson made several contributions in the fields of electricity, power transmission and related fields. Despite he was a reluctant entrepreneur, he was a founder of Thomson-Houston Electric Company, which after merging with Edison General Electric became General Electric.

A second group includes inventors whose formal training or professional experience started in Europe, though their major achievements and contributions (also measured in terms of patents) materialized after migrating to the US. In this group notable and well known examples are Alexander Graham Bell, Charles Steinmetz and Nikola Tesla. Their inventions in the fields of electricity, radio transmission and communication revolutionized the understanding of these phenomena and crucially contributed to the development of the emerging electric and telecommunication industries in the US. Our analysis focuses on this second group of inventors: skilled immigrants who arrived in the US as adults with a baggage of relevant work or intellectual experience.

The 1802 naturalization law, which would be in place for over 100 years in the US, allowed any foreigner (i.e. free white male) who had been in residence for five years to be admitted to citizenship. As discussed by Ueda (1992), naturalization was used as an inducement policy to promote more immigration, "to attract immigrants and absorb them into local life" with administrative procedures being "extremely loose and casually administered" for much of the 19th century (p. 737). Immigration, however, raised political opposition over time (Tabellini, 2020). In 1907, to investigate the socio-economic impact of immigrants, the US Congress established an Immigration Commission, which eventually recommended the introduction of restrictions. Starting in 1914 WWI led to an abrupt stop to immigration from Europe, shutting down arrivals from enemy countries such as Germany and the Austro-Hungarian Empire. For instance, with respect to the previous decade, in the 1910s inflows from Germany fell twice as much as those from Great Britain (Tabellini, 2020). Nonetheless, sizeable inflows started over when the conflict ended in 1918. In 1917 the Congress approved a literacy test for all new immigrants arriving in the US. However, this measure did not significantly limit new arrivals. A permanent quota system was then designed in 1921 based on 'national origin' and enshrined in the Immigration Acts in 1921 and 1924.

The shift to a more restrictive immigration policy was advocated by increasing anti-immigration sentiments, especially against recent immigrant flows from Southern and Eastern Europe (Goldin, 1994). As these flows had gained momentum with the beginning of the XX century, the first Immigration Act approved established that the yearly number of new immigrants from any given country should not exceed 3% of the stock of co-nationals already living in the US according to the 1910 census. In 1924 the second Immigration Act revised the quota to 2% and the reference year for its calculation to 1890, thus imposing stricter restrictions on the inflow of Southern and Eastern Europeans as their immigrant communities were much smaller in 1890 than in 1910. The result was a substantial slowdown in immigrant flows from those parts of Europe. For instance, the flow of Italian immigrants halved, going from above 1 million in the 1910-19 decade to 528,000 in the following decade. Immigrants from Northern Europe, on the other hand, were little affected by the quotas given their large presence in 1890 and the significant slowdown in their arrivals from 1900 onward. The quota system thus introduced a regulatory time discontinuity that is heterogeneous across nationalities. It constrained the inflows from Southern and Eastern Europe while leaving those from North Europe largely unaffected as long as quotas were much less binding for them.

3. Location Choice Model

To guide the ensuing empirical analysis, this section develops a simple model of immigrant inventors' location choices, in which local co-ethnic networks and

ethnic diversity affects both their productivity and their quality of life through localized externalities. In doing so, we build on Ottaviano and Peri (2006) in the wake of Roback (1982), highlighting which variables considered exogenous to the inventors' location choices will need to be instrumented in the empirical investigation.

We assume that inventors choose their locations among a large number of counties. Inter-county commuting costs are prohibitive so that inventors' counties of work and residence coincide. We ignore intra-county commuting costs to concentrate on the inter-county distribution of inventors as this is what we observe in the data. Inventors differ in terms of country of origin, which places them in E different ethnic groups ('ethnicities') indexed $e = 1, ..., E$ including natives.

Focusing on a generic county c, we use L_{ec} to denote the number of inventors from ethnic group e who work in that county c. There the different dimensions of multi-ethnicity relevant for ethnic group e are defined by a vector m_{ec} of variables measuring the composition of ethnicities in the local population. The ethnic group's viewpoint, emphasized here as m_{ec}, is meant to capture both the diversity and the co-ethnic network variables we will use in the empirical analysis. These variables are assumed to be exogenous to inventors' location choices and, as such, will need to be instrumented. They affect their production or consumption through external effects that can be positive or negative. To provide a conceptual framework within which to assess the nature and the sign of those effects is the model's purpose.

Inventors' preferences are defined over the consumption of goods G and services S. Goods have no ethnic dimension and are freely traded across counties.[6] Their price is set at the national level and taken as given at the county level. Differently, services are non-tradable and differentiated by ethnicity, which will allow us to determine whether co-ethnic networks mainly work through market or non-market interactions. The utility of an inventor of ethnicity e in county c is given by:

$$U_{ec} = \Lambda(m_{ec})\, S_{ec}^{1-\lambda} G_{ec}^{\lambda} \tag{1}$$

with $0 < \lambda < 1$, where S_{ec} and G_{ec} are services and goods consumption respectively, and $\Lambda(m_{ec})$ captures the 'utility effect' of multi-ethnicity m_{ec}. If the first derivative $\Lambda'_e(m_{ec})$ is positive, multi-ethnicity is a local 'consumption amenity'; if negative, it is a local 'consumption disamenity'. We assume that inventors choose the county that offers them the highest indirect utility. Given

6. The assumption of national prices for traded goods may look too strong, especially in our period of observation. However, the introduction of trade costs would only complicate the analysis without affecting its main insights on how multi-ethnicity affects inventors' location choices.

(1), utility maximization yields:

$$q_{ec}S_{ec} = (1-\lambda)w_{ec}L_{ec},\ p_cG_{ec} = \lambda w_{ec}L_{ec} \tag{2}$$

where q_{ec} and p_c are the prices of local services and goods respectively, while w_{ec} is the inventors' wage. Substituting (2) in (1) gives an inventor's indirect utility:

$$V_{ec} = (1-\lambda)^{1-\lambda}\lambda^{\lambda}\,\Lambda(m_{ec})\,\frac{w_{ec}}{q_{ec}^{1-\lambda}p_c^{\lambda}}. \tag{3}$$

Goods are supplied by perfectly competitive firms exploiting inventions through a linear technology. Inventions are themselves supplied by perfectly competitive labs employing inventors together with co-ethnic services. Specifically, the number of inventions generated by labs employing inventors of ethnicity f together with their co-ethnic services is determined by the following technology:

$$I_{fc} = \Phi_f(m_{fc})\,S_{fc}^{1-\varphi}L_{fc}^{\varphi} \tag{4}$$

with $0 < \varphi < 1$. In (4) $\Phi_f(m_{fc})$ captures the 'productivity effect' associated with multi-ethnicity modelled as a shift in total factor productivity. If the first derivative $\Phi'_f(m_{fc})$ is positive, multi-ethnicity is a local 'production amenity'; if negative, it is a local 'production disamenity'. Assuming a one-to-one linear technology and homogenous inventions, the supply of goods associated with innovations by inventors of ethnicity f is $G_{fc} = I_{fc}$ with county-level output $G_c = \sum_{f=1}^{E} G_{fc}$. As for services, for each ethnic group they are offered by members of the group other than inventors, again through a one-to-one linear technology. Due to the assumption on the technology, the local supply of services of ethnicity f is given by the number N_{fc} of these members, which is assumed to be exogenous to the inventors' location choices.

Given perfect competition among both firms and labs, profit maximization requires:

$$q_{fc}S_{fc} = (1-\varphi)p_cG_{fc},\ w_{fc}L_{fc} = \varphi p_cG_{fc}, \tag{5}$$

which implies marginal cost pricing so that neither firms nor labs make profits in equilibrium. As goods are freely traded, their price is the same in all counties and we can set $p_c = 1$ by choosing goods as unit of value.

A location equilibrium is defined as a set of prices $(q_{ec}, w_{ec}, c = 1, ..., C, e = 1, ..., E)$ such that in all counties inventors maximize their utilities given their budget constraints, firms and labs maximize profits given their technological constraints, and the makets for inventors, goods and services clear. Moreover, no firm or lab has any incentive to exit or enter. This is granted by conditions (5), which with $p_c = 1$ jointly imply:

$$q_{ec}^{1-\varphi}w_{ec}^{\varphi} = (1-\varphi)^{1-\varphi}\varphi^{\varphi}\Phi(m_{fc}) \tag{6}$$

Lastly, in equilibrium no inventor has any incentive to change location. This is the case when inventors are indifferent between alternative counties as these

offer the same level v_e of indirect utility exogenous to county c:

$$V_{ec} = v_e \tag{7}$$

for all $c = 0, ..., C$ with V_{ec} determined by (3).

Given $p_c = 1$, conditions (6) and (7) together with (3) determine the equilibrium wage of inventors and the equilibrium price of their co-ethnic services. Then, (5) and (4) can be used to express the equilibrium average productivity of immigrant inventors as a function of the wage obtaining:

$$\frac{I_{ec}}{L_{ec}} = \Theta_{Ie}\Phi(m_{ec})^{\frac{1-\lambda}{1-\lambda\varphi}}\Lambda(m_{ec})^{-\frac{1-\varphi}{1-\lambda\varphi}}, \tag{8}$$

where Θ_{Ie} is a bundling parameter.[7] Finally, (2) and (5) can be used together with market clearing for co-ethnic services to find the equilibrium number of immigrant inventors:

$$L_{ec} = \Theta_{Le}\Phi(m_{ec})^{\frac{\lambda}{1-\lambda\varphi}}\Lambda(m_{ec})^{\frac{1}{1-\lambda\varphi}}N_{ec}, \tag{9}$$

where Θ_{Le} is another bundling parameter.[8]

Equations (8) and (9) will guide our empirical analysis. They capture the equilibrium relation of the dimension of multi-ethnicity relevant for the location and the productivity of immigrant inventors of a given ethnic group. They must be estimated together in order to empirically assess whether and why multi-ethnicity acts as a pull or push factor. For instance, let's say we observe that L_{ec} increases with m_{ec} so that immigrant inventors of a given nationality are more present where there is more multi-ethnicity. As (9) shows that L_{ec} is an increasing function of $\Lambda(m_{ec})$ and $\Phi(m_{ec})$, their higher presence could be due to a consumption amenity $\Lambda'(m_{ec}) > 0$ but also to a production amenity $\Phi'(m_{ec}) > 0$, which does not allow us to identify the channel through which m_{ec} operates. However, as (8) implies that I_{ec}/L_{ec} increases with $\Phi(m_{ec})$ and decreases with $\Lambda(m_{ec})$, if we also observe that I_{ec}/L_{ec} increases (decreases) with m_{ec}, then it must be that the effect of $\Phi'(m_{ec}) > 0$ ($\Lambda'(m_{ec}) > 0$) dominates. Hence, we can conclude that immigrant inventors' location choices are driven by a dominant production (consumption) amenity associated with multi-ethnicity. Vice versa, if we observe that L_{ec} decreases and I_{ec}/L_{ec} increases (decreases) with m_{ec}, then immigrant inventors' location choices are driven by a dominant consumption (production) disamenity.

Moreover, when estimated together, (8) and (9) also allow us to assess whether the co-ethnic network operates mainly through market (N_{ec}) or non-market ($\Lambda(m_{ec})$ and $\Phi(m_{ec})$) interactions. In the former case, a larger (smaller)

7. Specifically, we have $\Theta_{Ie} \equiv \varphi^{-1}(\theta_\Lambda)^{-\frac{1-\varphi}{1-\lambda\varphi}}(\theta_\Phi)^{\frac{1-\lambda}{1-\lambda\varphi}}$ with $\theta_\Lambda \equiv (1-\lambda)^{1-\lambda}\lambda^\lambda v_e^{-1}$ and $\theta_\Phi \equiv (1-\varphi)^{1-\varphi}\varphi^\varphi$.

8. Specifically, we have $\Theta_{Le} \equiv \varphi(1-\lambda\varphi)^{-1}(\theta_\Lambda)^{\frac{1}{1-\lambda\varphi}}(\theta_\Phi)^{\frac{\lambda}{1-\lambda\varphi}}$.

co-ethnic network is associated with a larger (smaller) number of immigrant inventors by (9), but it is immaterial for their productivity by (8) despite co-ethnic services entering both consumption and production. Therefore, if co-ethnic networks affect immigrant innovators' productivity, they must operate through non-market interactions.

4. Data Description

Our dataset draws from two sources. For immigrant inventors, we exploit an original dataset compiled by Diodato et al. (2021) from the United States Patent and Trademark Office (USPTO) between 1870 and 1940. For county variables, we rely on NHGIS IPUMS decennial Census files between 1870 and 1930 (Manson et al., 2019).

4.1. Patents Data

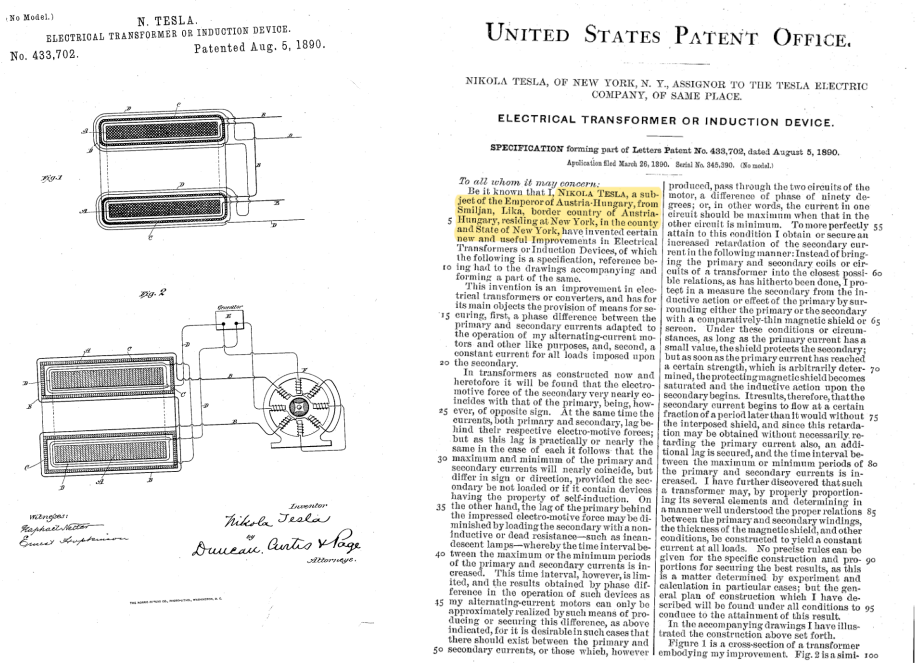
The dataset in Diodato et al. (2021) identifies migrant inventors in historical USPTO patent documents through a text-mining algorithm, analogous to the one described in Petralia et al. (2016), and a semi-automated procedure, which extracts detailed information on both country of origin and US county of residence of inventors migrated to the US from 1870 to 1940.[9]

As an illustration, consider the patent record with document number 433,702 reported in Figure 1. This record refers to a patent granted to Nikola Tesla, the great Serbian inventor, and its Tesla Electric Company in August 1890. The patent's abstract (highlighted) identifies Tesla's nationality, Austria-Hungary Empire, and his county of residence in the US, New York. These pieces of information are used to classify Tesla as an 'immigrant inventor', that is, a patentee from a foreign country e who resides in a US county c.[10] The automated algorithm identifies patents that can be attributed to an immigrant inventor based on keywords related to nationality. These include 'subject of' or 'citizen of', which is the patents' wording usually associated with the description a foreign inventor's country of origin. Such keywords should appear in combination with words such as 'residing at', which indicate where the immigrant inventor is located in the US. This first step leads to the identification of about 20,000 inventors with foreign nationality but living in the

9. A replication example of how the algorithm works is provided here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3ZLC8E.

10. Tesla arrived to the United States in 1884 from Europe after having studied in Graz (Austria) and started working almost immediately at Edison's premises. He soon left Edison and begun his career as an independent inventor, which brought him fame and recognition, yet did not made him rich. Tesla is considered as one of the greatest immigrant inventor, because of his contribution to AC electricity transmission and to many other technological fields.

US. In a second step, the algorithm has been trained to search for all the patents belonging to the same group of inventors, making it possible to keep track of the patenting activity of inventors arrived as foreign nationals who eventually obtain US citizenship through naturalization. By tracking inventors' county of residence at the time the patent is granted, the dataset also includes patentees moving across counties in the US. As a final step, a semi-automated procedure is used to double-check all patents identified as granted to immigrants. The end result is a database containing about 43,000 patents granted to about 20,000 immigrants together with their nationality and county of residence as reported in the patent records. With this information we compute the number of immigrant inventors with nationality e located in county c of state s in census year t, which we denote by L_{ecst}. Using corresponding number of patents I_{ecst}, we also compute their average productivity I_{ecst}/L_{ecst}.



(No Model.)

N. TESLA.

ELECTRICAL TRANSFORMER OR INDUCTION DEVICE.

No. 433,702. Patented Aug. 5, 1890.

Fig. 1

Fig. 2

Witnesses: Raphael Netter, Ernest Hopkinson

Inventor Nikola Tesla

by Duncan, Curtis & Page Attorneys.

UNITED STATES PATENT OFFICE.

NIKOLA TESLA, OF NEW YORK, N. Y., ASSIGNOR TO THE TESLA ELECTRIC COMPANY, OF SAME PLACE.

ELECTRICAL TRANSFORMER OR INDUCTION DEVICE.

SPECIFICATION forming part of Letters Patent No. 433,702, dated August 5, 1890.

Application filed March 26, 1890. Serial No. 345,390. (No model.)

To all whom it may concern:

Be it known that I, NIKOLA TESLA, a subject of the Emperor of Austria-Hungary, from Smiljan, Lika, border country of Austria-Hungary, residing at New York, in the county and State of New York, have invented certain new and useful Improvements in Electrical Transformers or Induction Devices, of which the following is a specification, reference being had to the drawings accompanying and forming a part of the same.

This invention is an improvement in electrical transformers or converters, and has for its main objects the provision of means for securing, first, a phase difference between the primary and secondary currents adapted to the operation of my alternating-current motors and other like purposes, and, second, a constant current for all loads imposed upon the secondary.

In transformers as constructed now and heretofore it will be found that the electromotive force of the secondary very nearly coincides with that of the primary, being, however, of opposite sign. At the same time the currents, both primary and secondary, lag behind their respective electro-motive forces; but as this lag is practically or nearly the same in the case of each it follows that the maximum and minimum of the primary and secondary currents will nearly coincide, but differ in sign or direction, provided the secondary be not loaded or if it contain devices having the property of self-induction. On the other hand, the lag of the primary behind the impressed electro-motive force may be diminished by loading the secondary with a non-inductive or dead resistance—such as incandescent lamps—whereby the time interval between the maximum or the minimum periods of the primary and secondary currents is increased. This time interval, however, is limited, and the results obtained by phase difference in the operation of such devices as my alternating-current motors can only be approximately realized by such means of producing or securing this difference, as above indicated, for it is desirable in such cases that there should exist between the primary and secondary currents, or those which, however produced, pass through the two circuits of the motor, a difference of phase of ninety degrees; or, in other words, the current in one circuit should be maximum when that in the other circuit is minimum. To more perfectly attain to this condition I obtain or secure an increased retardation of the secondary current in the following manner: Instead of bringing the primary and secondary coils or circuits of a transformer into the closest possible relations, as has hitherto been done, I protect in a measure the secondary from the inductive action or effect of the primary by surrounding either the primary or the secondary with a comparatively-thin magnetic shield or screen. Under these conditions or circumstances, as long as the primary current has a small value, the shield protects the secondary; but as soon as the primary current has reached a certain strength, which is arbitrarily determined, the protecting magnetic shield becomes saturated and the inductive action upon the secondary begins. It results, therefore, that the secondary current begins to flow at a certain fraction of a period later than it would without the interposed shield, and since this retardation may be obtained without necessarily retarding the primary current also, an additional lag is secured, and the time interval between the maximum or minimum periods of the primary and secondary currents is increased. I have further discovered that such a transformer may, by properly proportioning its several elements and determining in a manner well understood the proper relations between the primary and secondary windings, the thickness of the magnetic shield, and other conditions, be constructed to yield a constant current at all loads. No precise rules can be given for the specific construction and proportions for securing the best results, as this is a matter determined by experiment and calculation in particular cases; but the general plan of construction which I have described will be found under all conditions to conduce to the attainment of this result.

In the accompanying drawings I have illustrated the construction above set forth.

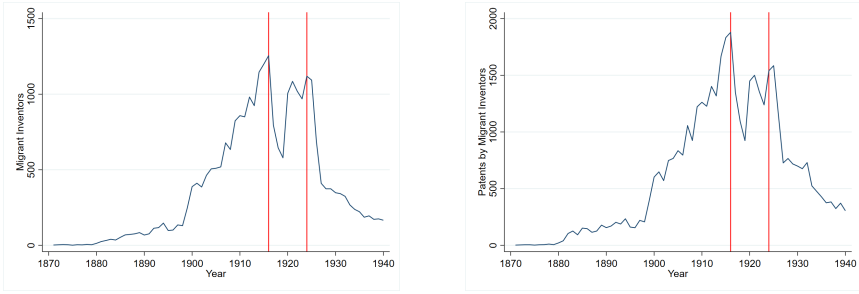
Figure 1 is a cross-section of a transformer embodying my improvement. Fig. 2 is a simi-

Figure 1: Original Patent Document. The figure reports an example of a historical patent document of the United States Patent and Trademark Office (USPTO), with highlighted the information codified by the text analysis.

Two remarks are in order. First, the dataset identifies immigrant inventors based on foreign nationality rather than foreign birthplace. This is different from census data as we will discuss in the next section. Second, as already discussed in Section 2, at that time naturalization was relatively easy and fast after five years of residence. This entails that patents granted to applicants

recorded as foreigner nationals by the USPTO tend to refer to recently arrived foreign-trained immigrants given that US-trained immigrants were likely to be already naturalized before patenting. In this respect, our dataset captures the technology-savvy talents at the top of the immigrant skill distribution.

Figures 2(a) and 2(b) report the number of immigrant inventors active in the US and their patenting activity from 1880 to 1940. The number of patents granted to foreign nationals steadily increases during the Age of Mass Migration. The outbreak of WWI first and then the introduction of immigration quotas in 1922 and 1924 (highlighted by red lines) is associated with a reduction in the number of immigrant inventors and their patents after 1920.



(a) Migrant inventors (b) Patents by migrant inventors

Figure 2: Patents by and number of migrant inventors by year. 1880-1940

Table 1 reports descriptive statistics on patenting activity and immigrant inventors by decade and nationalities. It considers 15 nationality groups (consistent with boundary changes across countries of origin occurred during the period) with totals reported in the last column. It shows that inventors from Great Britain and Ireland outperform all other nationalities with more than 16,000 patents. They are followed by Scandinavians and Germans (over 5,000 patents), Eastern Europeans (about 4,000 patents) and Austro-Hungarians (about 3,200 patents). Looking at the number of patents per inventor in our data (Figure 3), about one third of immigrant inventors are granted only one patent and the vast majority of them are granted less than ten over their career. This is consistent with qualitative evidence that in the period under consideration inventing activity was primarily an independent endevor.

Figure 4 presents a map of the distribution of immigrant inventors across US counties between 1880 and 1940, standardized by the county's population in 1930. Figure 5 depicts the parallel distribution of immigrant inventors' patents.

Table 1: Patents and number of migrant inventors in US by nationality. 1880-1930

	1880-90		1890-00		1900-1910		1910-1920		1920-30		1930-1940		1880-1940	
Nationality	Pat.	Inv.	Pat.	Inv.	Pat.	Inv.	Pat.	Inv.	Pat.	Inv.	Pat.	Inv.	Pat.	Inv.
Asia	0	0	7	5	59	39	285	185	245	144	21	14	621	390
Australia and New Zealand	0	0	1	1	6	4	9	8	18	11	16	3	52	28
Austro-Hungarian Emp.	25	3	91	41	396	257	1,363	896	1,017	532	285	99	3,240	1,855
Benelux	8	5	19	9	133	71	184	98	86	47	29	6	461	238
Canada	27	20	108	54	405	216	541	256	572	242	229	76	1,912	877
Eastern Europe	16	8	62	45	393	268	1,377	811	1,528	898	502	143	3,996	2,213
France	26	11	56	29	278	130	281	143	257	118	85	22	994	459
Germany	124	60	305	171	1,325	699	2,065	927	1,014	431	316	108	5,203	2,420
Great Britain and Ireland	876	313	1,422	699	3,537	1,721	4,431	2,019	3,795	1,345	1,871	416	16,271	6,656
Greece	0	0	3	2	25	14	77	59	118	94	15	9	240	179
Italy	9	6	51	25	289	195	743	510	751	428	312	66	2,195	1,244
Portugal	0	0	0	0	3	3	13	9	26	22	1	1	43	35
Rest Of America	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Scandinavia	65	46	340	203	1,601	741	2,311	1,140	1,479	678	700	180	6,623	3,046
Spain	5	5	9	5	39	19	54	35	86	48	5	5	198	117
Switzerland	47	17	45	26	277	142	385	183	286	128	205	40	1,318	546
Total	0	0	0	0	0	0	0	0	0	0	0	0	43,367	20,303

Data source: Diodato et al. (2021). Each row displays the number of patents and inventors by nationality and decade from 1880 until 1940. Last two columns report the same information for the whole period under consideration. Bottom row aggregates data across all ethnicities.

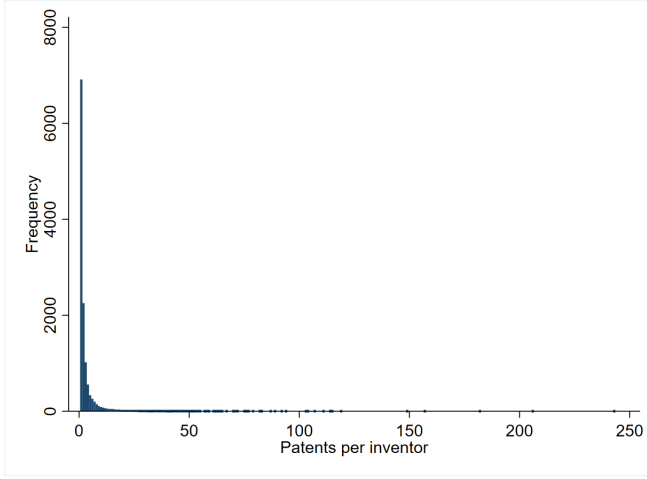


Figure 3: Number of patents per inventor. 1880-1940

4.2. Census Data

We match our historical patent data on immigrant inventors with US Census data between 1870 and 1930.[11] In particular, we employ NHGIS IPUMS county-level decennial census files (Manson et al., 2019) to construct measures of the

11. As it will be discussed in Section 5, we will investigate the impact of county variables observed at census frequency on immigrant inventors' outcomes in the subsequent decades. Accordingly, the last census year we consider is 1930, that is, the one related to the last decade covered by our patent data 1930-1940.

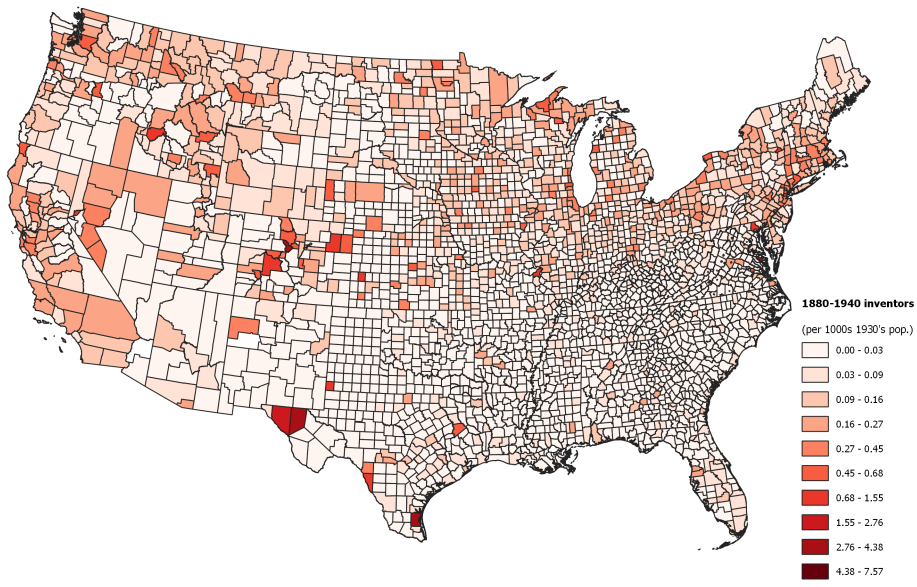


Figure 4: Migrant inventors by county (on 1000s 1930's pop.). 1880-1940

different dimensions of our model's local multi-ethnicity m_{ec}.[12] We consider county boundaries in 1990 and focus on counties where in each census year there are at least 2,500 residents and at least one foreign-born resident.[13] This gives a balanced panel of 1,911 counties for the years 1880–1930.[14] Differently from USPTO data on immigrant inventors, here immigrant status is identified based on foreign birthplace rather than foreign nationality. Accordingly, the co-ethnic network of immigrant inventors with ethnicity e consists of all immigrants born in the foreign country the immigrant inventors are nationals of when they are granted their first US patent.

The key variable we recover from the IPUMS files is the number of members of ethnic group e located in county c of state s in census year t, which we denote by N_{ecst}. Then, assigning natives to group $e = 1$, we calculate the local population as $P_{cst} = \sum_e N_{ecst}$ and the total number of local immigrants as $M_{cst} = \sum_{e \neq 1} N_{ecst}$. Finally, we compute the (%) share of group e's immigrants

12. These files are available at: https://data2.nhgis.org/main.

13. We adopt the crosswalk, developed by Eckert et al. (2018), between 1990's and historical counties' boundaries.

14. Census data on 1870 counties' ethnic composition are exclusively employed to compute initial ethnicities shares for shift-share 2SLS analysis. Moreover, in Section 7.1 and Appendix Section A we make use of 1870 local economic and demographic features to perform a battery of test on shift-share instruments' validity.

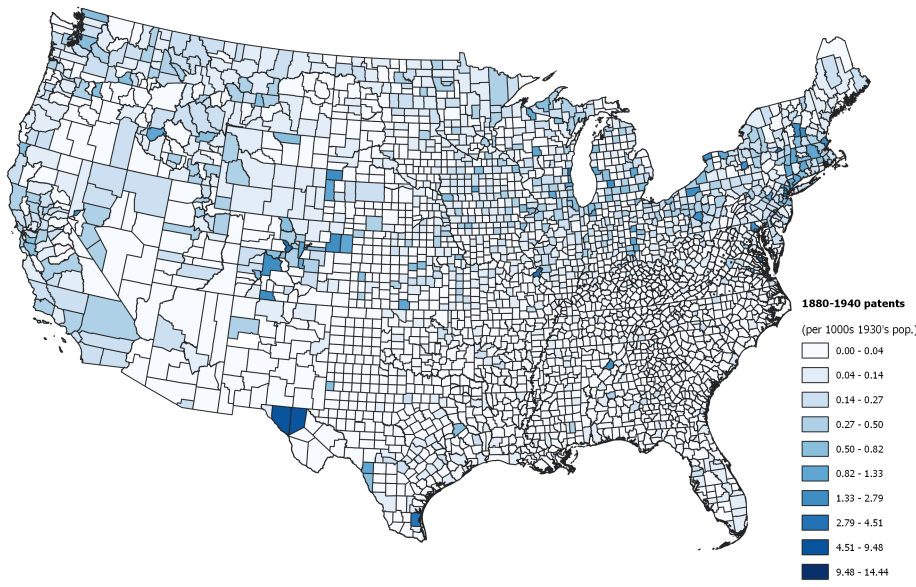


Figure 5: Migrant inventors' patents by county (on 1000s 1930's pop.). 1880-1940

in the local population as $s_{ecst} = N_{ecst}/P_{cst}$; the (%) share of immigrants from all other groups as $s_{-ecst} = (M_{-ecst} - N_{ecst})/P_{cst}$, where M_{-ecst} is the stock of immigrants of all ethnicities except e; and the dispersion within immigrant population across ethnic groups other than e by the Theil index:

$$Theil_{-ecst} = \sum_{i \neq e} \frac{N_{icst}}{M_{-ecst}} \ln\left(\frac{M_{-ecst}}{N_{icst}}\right). \tag{10}$$

We use s_{ecst} to capture group e's co-ethnic network, s_{-ecst} to capture the diversity 'between' natives and the rest of the immigrant population, and $Theil_{-ecst}$ to capture the diversity across ethnicities 'within' the immigrant population.[15]

Table 2 reports the shares of ethnic groups in the US population between 1870 and 1930. In the last two rows it also reports the overall immigration share and the Theil index for immigrants only. The overall immigration share

15. The Theil index aggregates ethnic groups' shares using a logarithmic weight that decreases with the shares. This implies a decreasing marginal contribution to diversity of each group's relative size. Most studies in the literature use the fractionalization index (i.e. the complement to one of the Herfindal index) as a measure of local ethnic diversity (see, e.g., Alesina et al. (2016); Docquier et al. (2018)). When we use this alternative index instead of the Theil index, our empirical analysis delivers similar results (available upon request).

Table 2: Immigration shares (%) and within–diversity in US Census data 1870-1930

Birthplace	1870	1880	1890	1900	1910	1920	1930
Asia	0.16	0.20	0.17	0.26	0.00	0.00	0.00
Australia and New Zealand	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Austro-Hungarian Emp.	0.14	0.14	0.48	0.76	1.81	1.41	1.10
Benelux	0.11	0.05	0.17	0.17	0.17	0.17	0.15
Canada	1.28	1.44	1.57	1.55	2.56	1.99	2.06
Eastern Europe	0.02	0.07	0.52	1.07	1.80	2.64	2.28
France	0.30	0.21	0.18	0.13	0.13	0.14	0.11
Germany	4.40	3.95	4.45	3.50	2.70	1.59	1.32
Great Britain and Ireland	6.83	5.56	4.99	3.66	2.78	2.04	1.76
Greece	0.00	0.00	0.00	0.01	0.11	0.17	0.13
Italy	0.02	0.05	0.29	0.64	1.45	1.52	1.47
Portugal	0.00	0.00	0.02	0.05	0.07	0.07	0.05
Rest Of America	0.10	0.14	0.16	0.17	0.26	0.48	0.06
Scandinavia	0.61	0.83	1.49	1.48	1.49	1.24	1.04
Spain	0.00	0.00	0.01	0.00	0.02	0.04	0.04
Switzerland	0.20	0.10	0.16	0.15	0.13	0.11	0.09
All migrants	14.18	12.74	14.67	13.60	15.48	13.62	11.68
Within migrants diversity (Theil)	1.40	1.49	1.78	2.00	2.13	2.19	2.12

Data source: NHGIS IPUMS county-level decennial census files (Manson et al., 2019). Each row indicates the (%) share, out of U.S. total population, of immigrants by foreign birthplace and decade from 1880 until 1930. Last two rows report, respectively, the (%) share of foreign-born population and the Theil index of diversity within the foreign-born population.

peaks in 1910 (15.48%) with a sharp decline after WWI and the introduction of immigration quotas in 1920s. Although immigrants from Great Britain and Ireland, Germany and Scandinavia account for most of the immigrant population at the beginning of the Age of Mass Migration, the table shows that their shares start to decline at the end of 19th century when a sizeable number of immigrants start to arrive from Southern Europe (especially Italy), Eastern Europe and the Austro-Hungarian Empire. This leads to a substantial increase in the diversity of the immigrant population with the Theil index increasing from 1.4 in 1870 to 2.19 in 1920.

Figures 6 and 7 display the cross-county distribution of the average overall immigration share and the average the Theil index in the period 1880-1930.

5. Empirical Strategy

In operationalizing (8) and (9) we express the salient features of multi-ethnicity m_{ec} in terms of group e's co-ethnic network, between diversity s_{-ecst} and within diversity $Theil_{-ecst}$. We then exploit variation in co-ethnic networks and diversity across our $1,911$ counties and 15 ethnicities over time. In particular, we look at the impacts of local co-ethnic networks and diversity in each census year on the change in immigrant inventors' presence and productivity in the subsequent decade. This implies that the last census year we consider for the explanatory variables is 1930 and the last decade we consider for the outcome variables is 1930-1940. Clearly, as we move along, we have to carefully deal with the confounding factors our location choice model abstracts from.

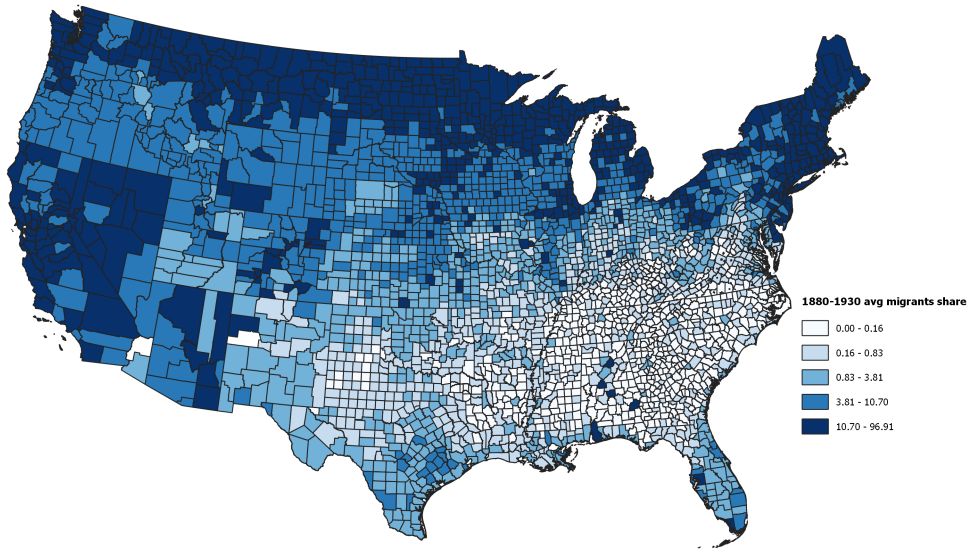


Figure 6: Immigration share by county (1880-1930 county average)

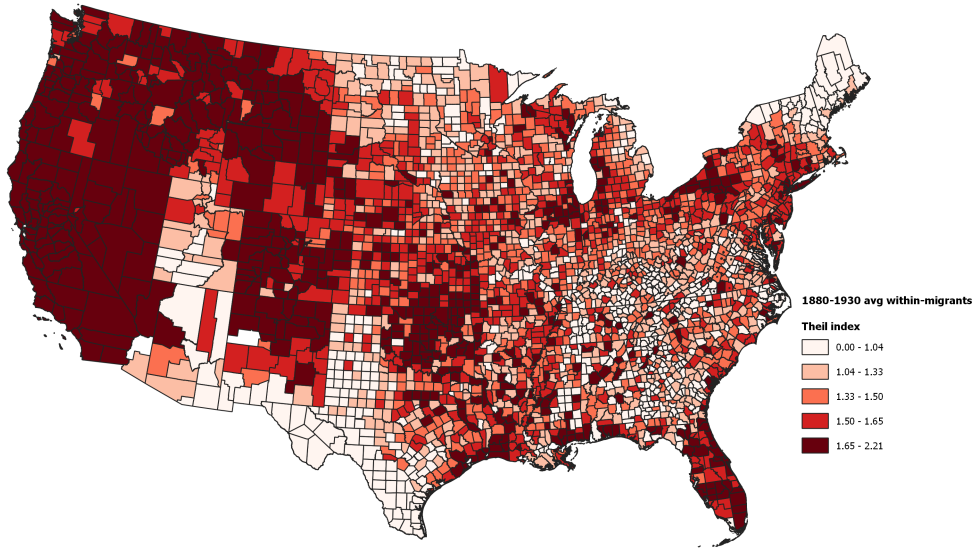


Figure 7: Within–migrants diversity Theil Index (1880-1930 county average)

Our specification is the following:

$$\ln(Y_{ecst}) = \alpha_0^y + \beta_1^y s_{ecst} + \beta_2^y s_{-ecst} + \beta_3^y Theil_{-ecst} + \delta_{st}^y + \mu_{ec}^y + t\pi_{ec}^y + \varepsilon_{ecst}^y \tag{11}$$

where s_{ecst}, s_{-ecst} and $Theil_{-ecst}$ measure group e's co-ethnic network, between– and within–diversity respectively, as described in Section 4.2.[16] As addressed by the model described in Section 3, we estimate Equation (11) in parallel using as outcome variable Y_{ecst} either the (log) number of group e's inventors L_{ecst} (expression (9)) or their (log) average patenting productivity $T_{ecst} = I_{ecst}/L_{ecst}$ (expression (8)).[17] [18] We control for unobserved heterogeneity by including ethnicity-by-county fixed effects μ_{ec}, which absorb all time-invariant characteristics for ethnic group e in county c, so that identification comes from decennial variations within ethnicity-county cells. Moreover, we introduce state-by-year fixed effects δ_{st} to adjust for state-specific shocks[19] and ethnicity-by-county time-linear trends $t\pi_{ec}$ to account for any cell-specific linear trajectories over time. Finally, ε_{ecst} is an idiosyncratic component. Standard errors are clustered at the level of the unit of analysis as to consider the correlation over time within ethnicity-county cells.[20]

The main coefficients of interest are β_1^y, β_2^y and β_3^y for $Y \in \{L, T\}$ related to the role of ethnic networks, between– and within–diversity, respectively. According to the model, if the estimated coefficients were positive (negative) in both regressions (11) for $Y \in \{L, T\}$, then that variable would act as a pull (push) factor through a dominant production amenity (disamenity). Differently, if the estimated coefficient were positive (negative) for $Y = L$ and negative (positive) for the $Y = T$, then the variable would act as pull (push) factor through a dominant consumption amenity (disamenity). Finally, a positive

16. We also consider a modified version of the empirical model by including a weighted specification of the Theil diversity index, with weights equal to either linguistic or religious distance (Spolaore and Wacziarg, 2016). Results are qualitatively similar to those reported in the main analysis (see results in Appendix B.2).

17. Since patent counts might be a rough indicator of productivity, we use an alternative indicator developed by Kelly et al. (2021) based on patent quality, which measures a patent's novelty and impact. See Appendix section B.3 for more details and results that confirm our baseline findings.

18. To deal with the large amount of zeros in both outcome variables, we consider the log of the outcome variable plus one. We also performed estimates considering the inverse hyperbolic sine transformation - which handles the transformation of null values - of the outcome variables. Results are qualitatively similar and available on request.

19. The inclusion of state–by–year fixed effects captures state–specific institutional and policy changes, such as the introduction of compulsory education between 1850 and 1917 (Bandiera et al., 2013), which may be potentially an extra pull factor for inventors and talented individuals. Moreover, state–by–year fixed effects also adjust for spatial spillovers at the state–level. Moreover, in Appendix Table B.6 we report results with ethnicity–by–year fixed effects as an extra robustness check. Their inclusion does not affect our results.

20. Appendix Table B.7 reports estimates with standard errors clustered at the county level. Results are unaffected.

estimate for β_1^L and a zero estimate for β_1^T would reveal that co-ethnic networks act as a pull factor through market rather than non-market interactions.

5.1. Identification

As immigrants are not randomly assigned across localities but self-select into specific locations according to individual and regional characteristics (Card, 2001), OLS estimation of (11) could be biased if unobserved (county or ethnicity) time-varying factors simultaneously affected immigration, ethnic composition and immigrant inventions. On the one hand, technological shocks to local productivity may attract or repel both immigrants and natives, but may disproportionately affect the location choices of the former if these are more mobile than the latter (Kerr et al., 2016). This confounding factor would generate an upward bias in the estimated correlation between diversity and inventors' outcomes (Ottaviano and Peri, 2005, 2006; Ager and Brückner, 2013). On the other hand, it has been argued that low-skilled immigration in the US changed the scale of production by stimulating labor complementary inventions (Acemoglu, 2010; Doran and Yoon, 2018). Conversely, innovations may have fostered labor-saving technological change, hence reducing diversity through the displacement of low-skilled immigrants. This reverse causality channel would generate a downward bias in the estimated relation between diversity and inventors' outcomes. However, the presence of immigrant inventors may also promote local productivity and growth (Kerr et al., 2016). In this case, their location choices would affect the location choices of other immigrants by stimulating the local economy (Abramitzky et al., 2019; Romer, 1990; Zucker et al., 1998; Jaffe et al., 2001; Kerr and Lincoln, 2010; Hunt, 2011). This additional channel of reverse causality would then lead to an upward bias in the estimated correlation between diversity and inventors' outcomes.

We address these issues by adopting a 2SLS approach and constructing a set of shift-share instrumental variables for each endogenous variable in our model following the widely used methodology based on pre-existing immigrant settlements (Card, 2001). Then, to check the robustness of our findings, in Section 7.3 we will also investigate an alternative methodology exploiting the quasi-experimental variation provided by the breakout of WWI and the introduction of immigration quotas in the early 1920s as discussed in Section 2 (King, 2009; Ager and Hansen, 2017; Tabellini, 2020).

The shift-share approach developed by (Card, 2001) - and then extensively used in the immigration literature - exploits the tendency of new immigrants to choose areas where previous immigrants of the same origin have settled in order to benefit from local co-ethnic networks. We rely on this logic to construct instruments for our key variables s_{ecst}, s_{-ecst} and $Theil_{-ecst}$. As explained in Section 4.2, their building blocks are the numbers of members of the different ethnic groups e located in county c of state s in census year t, which we denoted by N_{ecst}.

Specifically, we take 1870 as reference year and, similarly to Docquier et al. (2018), we define the predicted change in the stock of members of ethnic group e (native group included) in county c between census years $t-1$ and t as:

$$\Delta\widehat{N}_{ecst} = s^{US}_{ecs,1870} \times \Delta N_{e,-s,[t-1;t]} \qquad t = 1880, .., 1930 \tag{12}$$

We adopt a leave-out version of the aggregate 'shift' component (Adao et al., 2019), $\Delta N_{e,-s,[t-1;t]}$, which is the change in the stock of immigrants from group e between $t-1$ and t in the whole US excluding state s where county c is located. Removing the state-specific component makes sure that state-level shocks, which might affect the aggregate shifts in migration flows, do not enter the definition of the instrument. Then (12) apportions the aggregate shift component across counties according to their shares $s^{US}_{ecs,1870}$ of the total number of group members who were already in the US in 1870. Next, we compute the predicted stock of immigrants from e in county c for census year t as their stock in 1870 plus the cumulated sum of the predicted changes until t:

$$\widehat{N}_{ecst} = N_{ecs,1870} + \sum_{\tau \leq t} \Delta\widehat{N}_{ecs\tau} \qquad t = 1880, .., 1930. \tag{13}$$

Finally, we compute the shift-share predicted measures of group e's co-ethnic network, between and within diversity replacing $\widehat{N}_{ecst}$ in the definitions of s_{ecst}, s_{-ecst} and $Theil_{-ecst}$ respectively.

6. Results

Tables 3 and 4 present OLS and 2SLS results based on the estimation strategy described above. In particular, Table 3 reports first stage estimates for the shift-share instrumental variables. Columns 1 to 3 refer to first stage estimates for each of the three endogenous variables in our model. Both outcome and explanatory variables are standardized by subtracting sample mean and then dividing by standard deviation. First-stage coefficients can therefore be interpreted as standard deviation changes in the endogenous variables induced by a standard deviation change in each instrument. Point estimates show that each shift-share instrument strongly predicts the corresponding endogenous variables, while being weakly correlated with the remainder of endogenous variables. The values for the weak instrument test for multiple endogenous variables by Sanderson and Windmeijer (2016) are above 10, which is the commonly adopted rule-of-thumb threshold for a robust first stage (Stock and Yogo, 2002).[21]

21. In case of multiple endogenous variables, the Sanderson and Windmeijer F-statistic represents a more appropriate test, if compared to the widely-used Cragg-Donald or Kleibergen-Paap statistics, to check whether a particular endogenous regressor is weakly

Table 3: First-Stage Results: shift-share instruments

	(1) Network s_{ecst}	(2) Within diversity $Theil_{-ecst}$	(3) Between diversity s_{-ecst}
shift-share $\widehat{s}_{ecst}$	0.3456***	-0.0501***	-0.0214***
	(0.0277)	(0.0069)	(0.0055)
shift-share $\widehat{Theil}_{-ecst}$	0.0050	0.0948***	0.0301***
	(0.0057)	(0.0105)	(0.0056)
shift-share $\widehat{s}_{-ecst}$	-0.0549***	-0.4340***	0.3051***
	(0.0104)	(0.0232)	(0.0188)
Observations	171,990	171,990	171,990
Ethnicity by County FE	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes
S&W Weak identification test	203.7	168.6	138.8

This table reports first-stage estimates for each of the endogenous variables in eq. (11). Column 1 shows the first-stage results for ethnicity e's network variable (s_{ecst}), Column 2 for Theil index of diversity within county c's foreign-born population from countries/areas other than e ($Theil_{-ecst}$), Column 3 for the share of foreign-born population (as a fraction of c's total population) from countries/areas other than e (s_{-ecst}).

All of three shift-share instruments are constructed by using the building blocks $\widehat{N}_{ecst} = N_{ecs,1870} + \sum_{\tau \leq t} \Delta \widehat{N}_{ecs\tau}$. $N_{ecs,1870}$ represents the stock of migrants from e in county c in 1870, and $\Delta \widehat{N}_{ecs\tau} = s^{US}_{ecs,1870} \times \Delta N_{e,-s,[\tau-1;\tau]}$, where $s^{US}_{ecs,1870}$ is the 1870 share of total migrants in the US from e living in c, and $\Delta N_{e,-s,[\tau-1;\tau]}$ is the change in the stock of migrants from e between $\tau - 1$ and τ in the US excluding state s where c is located.

Both outcome and explanatory variables are standardized by subtracting sample mean and then dividing by standard deviation. First-stage coefficients can therefore be interpreted as s.d. changes in the endogenous variables induced by a s.d. change in the corresponding instrument.

All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1). Last row reports the values for the weak instrument test for multiple endogenous variables (Sanderson and Windmeijer, 2016).

Table 4 reports OLS and 2SLS results for specification (11). In panel A the outcome variable is the (log) stock of immigrant inventors L_{ecst} from ethnic group e living in county c between t and $t+1$. Columns 1 and 2 report OLS estimates with and without cell-specific linear time trends. Using the same specifications, Columns 3 and 4 show 2SLS results with shift-share IVs. We find a positive and significant impact of co-ethnic networks on immigrant inventors' location choices. This could be explained by co-ethnic hiring or lower screening costs, which is in line with a production-related amenity (Akcigit et al., 2017). Both OLS and 2SLS also reveal positive and significant impacts of both between

identified as it is constructed by netting out the linear projections of the remaining endogenous variables.

and within diversity on immigrant inventors' location choices. Hence, both co-ethnic networks and diversity appear to act as pull factors.

Specifically, if we focus on 2SLS estimates, in the baseline specification with linear time trends, a one percentage point (p.p.) increase in the within diversity Theil Index is associated with a rise in the stock of immigrant inventors roughly equal to 40%. If we consider a one standard deviation increase in the Theil index (equivalent to 0.28 p.p.), the effect is equal to about 11%.[22] As for between diversity, after adjusting for cell-specific linear time trends, the shift-share results in Column 4 reveal that a one p.p. increase in the population share of immigrants other than e is associated with a 5% increase in the number of immigrant inventors belonging to group e. Scaling by one standard deviation increase in between diversity (3.64 p.p.), this effect corresponds to 18.5% (corresponding to a 0.02 deviation from the sample mean). Finally, the same specification in Column 4 indicates that a one p.p. increase in the share of immigrants from their own country is associated with a 9% increase in the number of immigrant inventors. If we again rescale by one standard deviation change in the network variable (0.8 p.p.), this effect is equal to 7.2% (corresponding to a 0.007 additional inventors compared with the sample mean).

OLS point estimates for all three explanatory variables, although positive and significant, are always significantly lower than 2SLS results. Based on the discussion at the beginning of Section 5.1, this downward bias may be due to omitted time-varying variables that drive co-ethnic networks and diversity in one direction and inventors in the opposite direction. Reverse causation may also be at work as long as inventions fostered labor-saving technological change, hence reducing co-ethnic networks and diversity through the displacement of low-skilled immigrants. The inclusion of cell-specific linear time trends leads to a sizeable increase in the magnitude of point estimates. This suggests the presence of cell-specific trends in explanatory variables and outcome moving in opposite directions (Wooldridge, 2016).

These findings imply that co-ethnic networks and diversity attract immigrant inventors. However, as discussed in Section 5, they are not enough to assess which mechanism is at work, namely, whether immigrant inventors are attracted by production or consumption considerations. This is why in panel B we re-estimate the specification in equation (11) with (log) immigrant inventors' productivity $T_{ecst} = I_{ecst}/L_{ecst}$ as outcome variable in line with our model's equation (8). We find positive effects of both between and within diversity on immigrant inventors' patenting productivity together with a positive co-ethnic network effect. In light of our location choice model, these findings suggest that what attracts immigrant inventors are dominant positive production

22. At the sample mean (= 0.11 immigrant inventors per ethnicity-by-county cell), the effect equals 0.012 additional inventors per cell.

amenities. After adjusting for linear time trends, the 2SLS shift-share results in Column 4 imply that a one standard deviation increase in within diversity leads to a 5.3% rise in immigrants inventors' productivity, while the effect of a one standard deviation increase in between diversity leads to a rise in their productivity by about 8.3%.[23] As for co-ethnic networks, a one standard deviation increase is associated with a 4.16% rise in immigrants inventors' productivity (corresponding to 0.002 additional patents per inventor with respect to sample mean).

To summarize, the positive and significant coefficients estimates in both panels of Table 4 reveal that co-ethnic networks, between and within diversity act as pull factors on immigrants inventors and this happens through a dominant production amenity channel. We further explore the potential mechanisms underlying our findings in Appendix section B. We first consider the role of diversity in affecting inter–group connections and we use inter–ethnic marriage and residential contact as proxies (Giuliano and Tabellini, 2020). Secondly, we check whether diversity has any impact on natives' attitudes toward immigrants. In particular, we follow the approach developed by Fouka et al. (2021) and use ethnic-specific mentions on historical newspapers to measure how any ethnic group in our analysis is salient at the local level. Third, we test the effect of ethnic diversity on skills heterogeneity among immigrants at both industry and occupational level. Specifically, while inter-group connections and natives' attitude may be a relevant channel for the consumption amenity, skills diversity may be a potential driver of the production amenity channel. Results in Tables B.1 and B.2 rule out both inter–group connections and ethnic groups' local salience as potential mechanisms behind our findings. On the other hand, results reported in Table B.3 suggest that ethnic and skills diversity are positively correlated among immigrant population, indicating that ethnic diversity may well work through the production amenity channel. Finally, we check whether our findings are driven by migrant inventors choosing to settle in counties with a more diverse set of culturally close immigrants. We operationalize this test by including in baseline specification (11) a weighted version of the Theil index, with weights equal to either linguistic or religious distance between ethnic group e and the rest of ethnic groups in the county. Results in Table B.4 show that the estimated coefficients for the weighted versions of Theil index are significant and not different from the ones for baseline estimates in Table 4. This indicated that cultural proximity is not a relevant mediating factor in the relationship between diversity and immigrant inventors' location choice and productivity.

23. With respect to the sample mean of 0.06 patents per immigrant inventor by cell, those effects imply about 0.003 and 0.004 additional patents per inventors, respectively.

Table 4: Diversity, migrant inventors' location choice and productivity. OLS and 2SLS estimates

A) Dep. var: log(number of immigrant inventors)				
	(1)	(2)	(3)	(4)
	OLS		Shift-Share IV	
	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0184***	0.0266***	0.0288**	0.4242***
	(0.0019)	(0.0026)	(0.0140)	(0.0716)
Between Diversity: s_{-ecst}	0.0012***	0.0052***	0.0111***	0.0518***
	(0.0002)	(0.0005)	(0.0022)	(0.0065)
Network: s_{ecst}	0.0101***	0.0320***	0.0235***	0.0908***
	(0.0015)	(0.0039)	(0.0047)	(0.0111)
Observations	171,990	171,990	171,990	171,990
R-squared	0.6482	0.7195		
B) Dep. var: log(immigrant inventors' productivity)				
	(1)	(2)	(3)	(4)
	OLS		Shift-Share IV	
	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0139***	0.0157***	0.0032	0.1923***
	(0.0019)	(0.0023)	(0.0152)	(0.0556)
Between Diversity: s_{-ecst}	0.0007***	0.0027***	0.0016	0.0237***
	(0.0002)	(0.0004)	(0.0018)	(0.0044)
Network: s_{ecst}	0.0046***	0.0146***	0.0088***	0.0529***
	(0.0011)	(0.0024)	(0.0033)	(0.0079)
Observations	171,990	171,990	171,990	171,990
R-squared	0.5011	0.6302		
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends		Yes		Yes

Estimates in Panel A consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Panel B is the (log of 1+) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c. Columns 1 and 2 display OLS estimates, while Columns 3 and 4 report 2SLS results employing shift-share instruments.

All regressions include ethnicity by county fixed effects and state by year fixed effects. Estimates in Columns 2 (OLS) and 4 (shift-share 2SLS) also adjust for ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

7. Robustness Checks

In what follows, we perform two types of robustness checks: one directed at checking on instruments' validity (Goldsmith-Pinkham et al., 2020); another one addressing the concern about possible omitted variables at the county level that may be relevant during the period of US history covered by our analysis.

7.1. Tests for Instrument Validity

The shift-share instruments employed in this analysis rely on 1870 local co-ethnic settlements (share component) to apportion nation-wide changes in the stock of immigrants from country/area e between two decades (shift component). This type of design is not immune from threats to identification strategy and a recent strand of literature has analysed the conditions for the validity of the shift-share instruments (Borusyak et al., 2018; Adao et al., 2019; Goldsmith-Pinkham et al., 2020). While we can plausibly assume the leave-out shift component to be independent from unobserved local innovations, country shares at the baseline, $s^{US}_{ecs,1870}$, may correlate with initial local characteristics as the distribution of immigrants across counties is not likely to be random with respect to factors such as population size or economic development. Exclusion restrictions would be violated in the presence of any unobserved county-level shock affecting both immigrants' pre-settlements and immigrant inventors' outcomes. For example, productivity or labour demand shocks in one or more counties in 1870 may simultaneously attract immigrant workforce and stimulate innovation. If these shocks were serially correlated, the validity of the shift-share IVs would be compromised.

Jaeger et al. (2018) discuss in details the issue of serial correlation when using shift-share instruments and suggest that estimates of the impact of immigration might be conflated, both in the short and long run, in case of serial correlation of migration inflows from the same countries of origin. However, as we report in section 2, the composition by country of origin of inflows to US rapidly changed at the end XIX century, with a marked shift from Northern to Southern and Eastern European migrants. Moreover, the shocks to aggregate migration flows generated by WWI and then the introduction of Quotas during 1920s contributed even more to exogenous changes in the ethnic composition of newcomers.[24] The particular frame provided by the Age of Mass Migration therefore reduces concerns related to serial correlation of migration inflows and represents a valid setting for the analysis with shift-share instruments (Abramitzky et al., 2019; Tabellini, 2020; Giuliano and Tabellini, 2020).

24. In section 7.3 we perform additional tests using an alternative instrument which leverages, similarly to Ager and Hansen (2017) and Tabellini (2020), on the exogenous shocks induced by WWI and the introduction of Quotas.

Goldsmith-Pinkham et al. (2020) propose a novel approach to unpack the sources of variation behind shift-share IVs *à la* Card (2001) and Bartik instruments in general. In particular, given E ethnic groups, and thus E initial ethnic shares, according to Rotemberg (1983) they decompose the Bartik IV into a weighted sum of E just-identified instrumental variable estimators that use each initial share as a separate instrument. The Rotemberg weights sum to 1 and depend on the covariance between the e-th instrument's fitted value and the endogenous variable.[25]

In our setting, the Rotemberg weights are useful to understand which ethnicities contribute the most to the overall variation in the shift-share instruments. In Appendix Section A we replicate the complete battery of tests suggested by Goldsmith-Pinkham et al.. In particular, after computing ethnic-specific Rotemberg weights, we check whether initial ethnicity shares correlate with local characteristic in 1870 (see Appendix Table A.4).[26] We find that the initial concentration of Eastern Europeans, the ethnicity with highest Rotemberg weights for all shift-share IVs,[27] is not significantly correlated with any of the county-level characteristics in 1870, while the distribution of immigrants from the rest of ethnicities is significantly associated with population size, the share of workers in manufacturing sector, output per worker in farming and manufacturing sectors, and illiteracy rate.

In what follows, we account for these potential confounding factors at the baseline year by estimating again specification (11) while including interactions between year dummies and a set of 1870 county–level variables. Firstly, we control for (log) 1870 population, which is the local feature with the strongest correlation with initial ethnicity shares (see Appendix Table A.4). Secondly, we add as controls the remaining 1870 local characteristics, which significantly correlate with ethnicity shares (i.e. the share of workers and the (log) average wage in manufacturing sector, (log) output per worker in farming sectors, the illiteracy rate and a dummy for water transportation access in 1860). Finally, we include the full vector of 15 ethnic local shares in 1870.

Table 5 reports shift-share 2SLS results for both immigrant inventors' location choice (panel A) and productivity (panel B). Column 1 reports baseline estimates as in Column 4 of Table 4 for a direct comparison. The effects of diversity variables, as well as those of co-ethnic networks, are robust to the inclusion of the interactions between year dummies and county characteristics in 1870. Point estimates for both outcomes are indeed still positive and significant, although smaller than the baseline estimates.

25. The e-th instrument's fitted value is equal to the shift-share predicted change in local immigrant stocks $\Delta\widehat{N}_{ecst} = s^{US}_{ecs,1870} \times \Delta N_{e,-s,[t-1;t]}$.

26. The set of county-level variables, employed for the estimates in Appendix Table A.4, are available in IPUMS NHGIS Census files.

27. Eastern Europeans account from more than half of positive Rotemberg weights for all three of shift-share instrumental variables.

Table 5: Shift-share 2SLS results – control for 1870 county characteristics

A) Dep. var: log(number of immigrant inventors)				
	(1) $log(L)_{ecst}$	(2) $log(L)_{ecst}$	(3) $log(L)_{ecst}$	(4) $log(L)_{ecst}$
	Baseline			
Within Diversity: $Theil_{-ecst}$	0.4242***	0.3056***	0.2187***	0.2547***
	(0.0716)	(0.0633)	(0.0624)	(0.0630)
Between Diversity: s_{-ecst}	0.0518***	0.0395***	0.0310***	0.0241***
	(0.0065)	(0.0055)	(0.0056)	(0.0054)
Network: s_{ecst}	0.0908***	0.0788***	0.0709***	0.0597***
	(0.0111)	(0.0103)	(0.0102)	(0.0090)
Observations	171,990	170,820	170,820	171,990
B) Dep. var: log(immigrant inventors productivity)				
	(1) $log(T)_{ecst}$	(2) $log(T)_{ecst}$	(3) $log(T)_{ecst}$	(4) $log(T)_{ecst}$
	Baseline			
Within Diversity: $Theil_{-ecst}$	0.1923***	0.1370***	0.0913*	0.1377**
	(0.0556)	(0.0523)	(0.0538)	(0.0573)
Between Diversity: s_{-ecst}	0.0237***	0.0179***	0.0136***	0.0181***
	(0.0044)	(0.0040)	(0.0043)	(0.0047)
Network: s_{ecst}	0.0529***	0.0472***	0.0434***	0.0474***
	(0.0079)	(0.0076)	(0.0078)	(0.0081)
Observations	171,990	170,820	170,820	171,990
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes
1870 (log) pop × Year		Yes	Yes	
1870 controls × Year			Yes	
1870 ethnicities shares × Year				Yes

2SLS estimates in Panel A consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for 2SLS estimates in Panel B is the (log of 1+) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c.
Column 1 presents baseline 2SLS estimates as in Column 4 of Table 4, while the remainder of columns introduce 1870 county-level controls interacted with year dummies. In Column 2 we control for log-population, in Column 3 we add the share of workers and (log) average wage in manufacturing sector, (log) output per worker in farming sector, the illiteracy rate and a dummy for waterboard access in 1860. In Column 4, we adjust for initial ethnicity shares ($s^{US}_{ecs,1870}$).
All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

7.2. Population Size and Frontier Exposure

During our period of analysis the population was still very unevenly distributed across the US. This was partly due to uneven local growth, but also by the westward movement of European settlers from the original Atlantic coast (XVII century) to the Far West (XIX century) until the expansion of the American frontier ended with the annexation of the last remaining western territories as states in 1912. This expansion represented a crucial structural change in both population dynamics and culture. Bazzi et al. (2020) find that after more than a century counties with higher 'frontier exposure' (as measured by the number of years spent on the frontier) still show a higher degree of 'individualism' (as measured by negative attitudes towards redistribution, public spending and other social policies such as the Affordable Care Act and the minimum wage). They explain this pattern in terms of the selective migration to the frontier of people with higher self-reliance. The same would apply in general to counties with low population density.

As long as immigrants may fit the self-reliant type, immigrant inventors and all sorts of other immigrants may have congregated in frontier or low-density counties for reasons unrelated to co-ethnic networks and diversity. Conversely, one could argue that immigrants and, in particular, immigrant inventors tend to prefer populous and urbanised areas. In the former case, the fact that low population density and frontier exposure are associated with both a more diverse population and more immigrant inventors implies a potential positive bias in our estimates. In the latter case, the bias would be negative. So far, we have not included population size as a control in specification (11) as diversity itself may affect population growth via, for instance, output growth (Ager and Brückner (2013)). If this were so, population would be a 'bad control' as it would be directly affected by the treatment variable.

In order to check the robustness of our earlier results, we now introduce a time-varying control for population to compare counties with similar demographic size. To control for frontier exposure, we use the same data as in Bazzi et al. (2020) to identify for each census year the counties with population density below two inhabitants per square mile. Differently from them, however, we do not consider the time invariant number of years a county was on the frontier, but rather the time-varying number of years since the county 'crossed' the frontier. If, for example, a certain county crossed the frontier in 1860, then 40 years have elapsed since its frontier exposure in 1900.

Table 6 reports the results for immigrant inventors' presence and productivity with log-population size and our time-varying frontier exposure as additional controls. The two tables confirm our findings in Section 6 about the positive effects of co-ethnic networks, between diversity and within diversity on immigrant inventors' outcomes.

We also check whether the effects of co-ethnic networks and diversity are heterogeneous across population size classes by separately considering counties

in different terciles of population in 1880.[28] While OLS estimates remain positive for all terciles, most of the action in terms of co-ethnic networks and diversity causing immigrant inventors' presence (panel A) and productivity (panel B) seems to take place in the third tercile (Columns 5 and 6) consisting of counties with population above about 18,000 residents. In this tercile the point estimates for both co-ethnic networks and diversity are all positive and significant also in 2SLS regressions. Differently, in the second tercile (Columns 3 and 4) with county population between about 10,000 and 18,000 thousands inhabitants, the point estimates for both between and within diversity are positive but not significantly different from zero. Lastly, in the first tercile there is no evidence of any casual effects on either immigrant inventors' outcomes.

28. Appendix Sections B.7, B.8 and B.9 present further heterogeneous tests by, respectively, six NBER patents technological sectors, 1880 population density terciles and three macro-regions. Table B.9 (Appendix Section B.7) shows that the effect is consistent across sectors and hence our findings are not driven by innovation in any particular technological area. Coherently with heterogeneous effects by population terciles in Table 7, Table B.10 (Appendix Section B.8) indicates that the significant effects mainly result from counties in mid and top terciles of the 1880 population density distribution. Finally, Table B.11 (Appendix Section B.9) gathers 2SLS estimates (shift-share IVs) disaggregated by three macro-regions (Northeast & Midwest, South and West), and reveals that, for both inventors' location choice and productivity, most of the significant associations with local co-ethnic networks and diversity stem from counties in Northeast & Midwest and South.

Table 6: OLS and 2SLS estimates - conditional on population size and frontier exposure

A) Dep. var: log(number of immigrant inventors)				
	(1)	(2)	(3)	(4)
	OLS		2SLS	
	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0247***	0.0245***	0.3863***	0.3858***
	(0.0025)	(0.0025)	(0.0659)	(0.0660)
Between Diversity: s_{-ecst}	0.0046***	0.0046***	0.0527***	0.0526***
	(0.0005)	(0.0005)	(0.0066)	(0.0066)
Network: s_{ecst}	0.0315***	0.0315***	0.0919***	0.0916***
	(0.0039)	(0.0039)	(0.0111)	(0.0111)
$log(pop)_{cst}$	0.0261***	0.0265***	-0.1523***	-0.1517***
	(0.0075)	(0.0075)	(0.0244)	(0.0244)
Years since exposure to frontier		0.0005***		0.0003
		(0.0002)		(0.0002)
Observations	171,990	171,990	171,990	171,990
B) Dep. var: log(immigrant inventors productivity)				
	(1)	(2)	(3)	(4)
	OLS		2SLS	
	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0152***	0.0152***	0.1738***	0.1739***
	(0.0023)	(0.0023)	(0.0515)	(0.0515)
Between Diversity: s_{-ecst}	0.0025***	0.0025***	0.0241***	0.0242***
	(0.0004)	(0.0004)	(0.0045)	(0.0045)
Network: s_{ecst}	0.0145***	0.0145***	0.0534***	0.0534***
	(0.0024)	(0.0024)	(0.0079)	(0.0079)
$log(pop)_{cst}$	0.0070	0.0070	-0.0745***	-0.0746***
	(0.0048)	(0.0047)	(0.0182)	(0.0182)
Years since exposure to frontier		0.0001		-0.0000
		(0.0002)		(0.0002)
Observations	171,990	171,990	171,990	171,990
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

Estimates in Panel A consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Panel B is the (log of 1+) number of patents, granted between t and $t+1$, per inventors from ethnicity e and living in county c. Columns 1 and 2 display OLS estimates, while Columns 3 and 4 report 2SLS results employing shift-share instruments.

Columns 1 and 3 include county's log-population. Columns 2 and 4 add a control for frontier exposure, i.e. the number of years since the county was on either the eastern or western frontier.

All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table 7: OLS and 2SLS estimates - heterogeneous effects by 1880 county population size

	A) Dep. var: log(number of immigrant inventors)					
	(1)	(2)	(3)	(4)	(5)	(6)
	1st tercile $pop_{c1880} <= 9798$		2nd tercile $9806 >= pop_{c1880} <= 18831$		3rd tercile $pop_{c1880} >= 18854$	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0067***	-0.6576	0.0041*	0.1048	0.0572***	0.3761***
	(0.0021)	(0.9864)	(0.0021)	(0.0942)	(0.0089)	(0.0689)
Between Diversity: s_{-ecst}	0.0007**	-0.0520	0.0020***	0.0182	0.0179***	0.0915***
	(0.0003)	(0.0819)	(0.0007)	(0.0189)	(0.0018)	(0.0078)
Network: s_{ecst}	0.0044*	-0.0286	0.0175***	0.0500**	0.0774***	0.1394***
	(0.0026)	(0.0861)	(0.0061)	(0.0219)	(0.0087)	(0.0186)
Observations	48,690	48,690	54,900	54,900	68,400	68,400
	B) Dep. var: log(immigrant inventors productivity)					
	(1)	(2)	(3)	(4)	(5)	(6)
	1st tercile $pop_{c1880} <= 9798$		2nd tercile $9806 >= pop_{c1880} <= 18831$		3rd tercile $pop_{c1880} >= 18854$	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0061**	-1.0565	0.0062**	0.2122	0.0294***	0.1688***
	(0.0024)	(1.4558)	(0.0027)	(0.1446)	(0.0079)	(0.0534)
Between Diversity: s_{-ecst}	0.0005	-0.0834	0.0012	0.0330	0.0080***	0.0360***
	(0.0003)	(0.1213)	(0.0009)	(0.0270)	(0.0012)	(0.0060)
Network: s_{ecst}	0.0017	-0.0663	0.0143***	0.0655**	0.0316***	0.0686***
	(0.0020)	(0.1260)	(0.0052)	(0.0304)	(0.0051)	(0.0119)
Observations	48,690	48,690	54,900	54,900	68,400	68,400
Ethnicity by County FE	Yes	Yes	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes	Yes	Yes

Estimates in Panel A consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Panel B is the (log of 1+) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c. Columns 1, 3 and 6 display OLS estimates, while Columns 2, 4 and 6 report 2SLS results employing shift-share instruments.
Columns 1 and 2 consider counties in the bottom tercile as for 1880 population, Columns 3 and 4 counties in the second tercile, Columns 5 and 6 counties in the top tercile.
All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends.
Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

7.3. WWI and Quota shocks

We supplement the shift-share analysis with an alternative approach that leverages the exogenous variation in migrant inflows generated by the two events that, as discussed in Section 2, put an end to the Age of Mass Migration: WWI and quotas. In doing so, we follow Ager and Hansen (2017) and Tabellini (2020). Ager and Hansen (2017) allocate the negative immigration shock ('missing migrants') induced by the quotas at national level across local labor markets according to their shares of quota-affected nationalities in 1920 just before restrictions were introduced. While specifying a similar city-level measure of quota exposure, Tabellini (2020) also exploits the outbreak of WWI to construct an analogous measure of 'missing migrants' based on the 1910 geographic distribution of immigrants born in countries that were not part of the Allies during the conflict.

Combining the two approaches we first construct the following ethnicity-by-county measure of 'WWI exposure' during the 1910s as in Tabellini (2020):

$$WWI.exp._{ecs,1920} = s^{US}_{ecs,1910} \times Enemy_e \times Imm_{e,00-10} \tag{14}$$

where $Enemy_e$ is a dummy equal to 1 for enemy countries (Germany and the Austro-Hungarian Empire), $Imm_{e,00-10}$ is the average yearly migration inflow from country e to the US from 1900 to 1910, and $s^{US}_{ecs,1910}$ is county c's share of the total number of ethnic group e's members already in the US in 1910.[29] Though WWI curbed immigration from all origins, arrivals from enemy countries were completely shut down. Hence, (14) tells that counties with a higher share $s^{US}_{ecs,1910}$ of enemy immigrants in 1910 were more exposed to the negative aggregate WWI immigration shock $Enemy_e \times Imm_{e,00-10}$.

We then define an ethnicity-by-county measure of 'quota exposure' during the 1920s as in Ager and Hansen (2017):

$$Q.exp._{ecs,1930} = s^{US}_{ecs,1920} \times \max\left(\frac{Imm_{e,00-14} - Q_e}{Imm_{e,00-14}}, 0\right) \tag{15}$$

where $s^{US}_{ecs,1920}$ is county c's share of the total number of ethnic group e's members already in the US in 1920, $Imm_{e,00-14}$ is the yearly migration inflow from country e to the US from 1900 to 1914, Q_e is the number of immigrants from country e allowed to enter the US by the corresponding quota between 1922 and 1930 as per Census Statistical Abstract 1931. The ratio in (15) measures the quota exposure for foreign-group e in the US as a whole and

29. $Imm_{e,00-10}$ is built by using the micro-data from 1920 IPUMS Full-Count Census file (Ruggles et al., 2003), which reports the migrant's year of arrival to US. We collapse this information at national level to obtain estimates of yearly inflows by migrants' birthplace from 1900 to 1914. $Imm_{e,00-10}$ in (14) takes the simple average over the period 1900-1910, whereas $Imm_{e,00-14}$ in (15) takes the simple average over the period 1900-1914.

Table 8: Quota exposure by foreign nationality

Ethnicity	(1) Avg yearly inflow 1900-1914	(2) Avg yearly quota 1922-1930	(3) Quota exposure
Asia	9,243	2,022	0.78
Australia and New Zealand	454	537	0
Austro-Hungarian Emp.	75,026	14,571	0.81
Benelux	6,546	3,419	0.48
Canada	26,253	Unrestricted	0
Eastern Europe	139,383	29,762	0.79
France	4,093	4,449	0
Germany	23,976	54,086	0
Great Britain and Ireland	52,498	69,830	0
Greece	8,186	1,162	0.86
Italy	78,037	16,823	0.78
Portugal	3,882	1,156	0.70
Rest Of America	18,720	0	0
Scandinavia	34,956	25,471	0.27
Spain	1,718	405	0.76
Switzerland	2,537	2,596	0

Column 1 indicates the average number of arrivals by birthplace between 1900 and 1914 (source: 1920 IPUMS Full-Count Census micro-data (Ruggles et al., 2003)). Column 2 reports the average quota by nationality between 1922 and 1930, i.e. the maximum number of new arrivals to US allowed by 1921 and 1924's Immigration Acts (source: Census Statistical Abstract 1931). Column 3 displays the values of aggregate quota exposure by ethnicity as defined in (15) (Ager and Hansen, 2017).

ranges between 0 and 1. It equals 0 when the quota for country e is higher than the actual average yearly inflow between 1900 and 1914. It equals 1 in the extreme case in which immigration from country e is totally banned. It takes values between 0 and 1 when the quota is lower than the actual average yearly inflow.

Table 8 reports the quota exposure and its components by ethnicity. For illustrative purposes, it is useful to consider the quota exposure for Italian and German immigrants. The former experienced large inflows from 1900 to 1914 with about 78,000 average yearly arrivals, but the average yearly quota introduced in the early 1920s allowed less than 17,000 new arrivals per year from 1922 to 1930. As a result, the quota for Italians was binding and their quota exposure is very high (0.8). Conversely, from 1900 to 1914 German inflows were much smaller with only about 24,000 average yearly arrivals. The corresponding quota of about 54,000 new arrivals for 1922-1930 was not binding so that Germans' quota exposure is nil (0).

The rationale for using $WWI.exp._{ecs,1920}$ and $Q.exp._{ecs,1930}$ to build instruments for s_{ecst}, s_{-ecst} and $Theil_{-ecst}$ is that counties with higher shares of WWI- or quota-affected ethnic groups are expected to experience lower growth in the stocks of immigrants from those ethnic groups. We proceed

as follows. We first run a stage-zero regression of the change in the stock of immigrants from e to c on $WWI.exp._{ecs,1920}$ and $Q.exp._{ecst}$:

$$\Delta N_{ecst} = a_0 + a_1 1920 \times WWI.exp._{ecs,1920} + a_2 1930 \times Q.exp._{ecs,1930} + \delta_{st} + \mu_{ec} + \varepsilon_{ecst} \tag{16}$$

where ΔN_{ecst} is the change in the stock of ethnic group e in c between $t-1$ and t. Exposure measures $WWI.exp._{ecs,1920}$ and $Q.exp._{ecst}$ are interacted with a year dummy in order to check whether they are significant predictors in the affected years only. We include ethnicity-by-county and state-by-year fixed effects (μ_{ec} and δ_{st}). The estimated coefficients from (16) allow us to predict WWI and quota induced changes over time in the immigrant stocks across ethnicity-by-county cells as:

$$\begin{aligned} \Delta WWI\text{-}\widehat{N}_{ecs1920} &= \widehat{a}_1 WWI.exp._{ecs,1920}, \\ \Delta Q\text{-}\widehat{N}_{ecs1930} &= \widehat{a}_2 Q.exp._{ecs,1930}. \end{aligned}$$

We then obtain the predicted post-WWI and post-quota stocks by adding these predicted changes to the stocks in 1910 and 1920 respectively:

$$\begin{aligned} WWI\text{-}\widehat{N}_{ecs1920} &= N_{ecs1910} + \Delta WWI\text{-}\widehat{N}_{ecs1920}, \\ Q\text{-}\widehat{N}_{ecs1930} &= N_{ecs1920} + \Delta Q\text{-}\widehat{N}_{ecs1930}. \end{aligned}$$

We compute the WWI and quota predicted measures of group e's co-ethnic network, between and within diversity by replacing $WWI\text{-}\widehat{N}_{ecs1920}$ and $Q\text{-}\widehat{N}_{ecs1930}$ in the definitions of s_{ecst}, s_{-ecst} and $Theil_{-ecst}$ while using the shift-share prediction $\widehat{N}_{ecst}$ for natives ($e = 1$). We finally define our instruments by taking the difference between predicted measures and the corresponding value of the variables in the previous decade. For example, the Quota IV for within-diversity is equal to:

$$Q\text{-}\Delta\widehat{Theil}_{-ecs1930} = Q\text{-}\widehat{Theil}_{-ecs1930} - Theil_{-ecs1920}. \tag{17}$$

Similarly to Tabellini (2020) and in order to accommodate instruments specification, we adopt a first-differenced version of baseline model in (11). First stage equations include interactions between WWI and Quota IVs and, respectively, 1920 and 1930 year dummies as these instruments are meant to be significant predictors of the change in endogenous variables for the post-shocks decades only.

Table 9 displays both stage-zero and first-stage regression results for the WWI- and quota-based IVs. The stage-zero estimates in Column 4 shows that, consistently with Ager and Hansen (2017) and Tabellini (2020), the ethnicity-by-county measures of WWI and quota exposure have significant negative effects on the change in the local immigrant stocks in the post-WWI and quota decades. On the one hand, during the 1920s (see the interaction with the 1930 time dummy), more quota-exposed ethnicity-county cells exhibit significantly

smaller changes in immigrant stocks than less exposed cells. On average, a percentage point increase in quota exposure reduces the change in the stock of immigrants by 2,706 units between 1920 and 1930. On the other hand, a 'missing migrant' predicted by the WWI-exposure variable corresponds to a reduction of 1.16 actual immigrants between 1910 and 1920. The variations induced by WWI and the quotas provide a strong enough prediction for the first difference in all the endogenous variables. First-stage estimates in Columns 2 to 4 highlight that these IVs are positively and significantly associated with the correspondent endogenous variables, and the Sanderson and Windmeijer (2016) Weak Instrument tests return again values above the 10-threshold for robust first stage regressions.

We report 2SLS estimates using WWI and quota IVs in Table 10. Columns 1 and 2 show the results for immigrant inventors' location choice, respectively, without and with the inclusion of ethnicity-by-county fixed effects, which reflect the first differencing of linear time-trends in specification (11). Columns 3 and 4 replicate the same estimates with immigrant inventors' productivity as outcome variable. All specifications yield a positive and significant correlation between diversity, as well as co-ethnic networks, and both outcome variables.

Table 9: First Stage Results: quota and WWI instruments

	(1) Stage-zero	(2)	(3)	(4)
		1st stage regressions		
	ΔN_{ecst}	$\Delta Theil_{-ecst}$	Δs_{-ecst}	Δs_{ecst}
$1920 \times WWI.exp._{ecs,1920}$	-1.1664***			
	(0.3214)			
$1930 \times Q.exp._{ecs,1930}$	-270614.5328**			
	(126,261.5296)			
$1920 \times WWI$-$\Delta\widehat{Theil}_{-ecs1920}$		0.0354***	0.0135***	0.0032
		(0.0036)	(0.0016)	(0.0022)
$1930 \times Q$-$\Delta\widehat{Theil}_{-ecs1930}$		0.0615***	-0.0150***	0.0090***
		(0.0039)	(0.0036)	(0.0026)
$1920 \times WWI$-$\Delta\widehat{s}_{-ecs1920}$		0.0153***	0.1092***	0.0203***
		(0.0029)	(0.0033)	(0.0039)
$1930 \times Q$-$\Delta\widehat{s}_{-ecs1930}$		0.0274***	0.1083***	0.0203***
		(0.0021)	(0.0064)	(0.0037)
$1920 \times WWI$-$\Delta\widehat{s}_{ecs1920}$		0.0084***	0.0272***	0.0741***
		(0.0022)	(0.0030)	(0.0081)
$1930 \times Q$-$\Delta\widehat{s}_{ecs1930}$		0.0077***	0.0287***	0.0632***
		(0.0014)	(0.0037)	(0.0111)
Observations	171,795	171,795	171,795	171,795
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
First differences model Yes	Yes	Yes	Yes	Yes
S&W Weak identification test		101	252.7	30.43

Column 1 displays stage-zero regression of the change in the stock of migrants from country/area e in county c on the measures of WWI and quota exposure, as defined in (14) and (15), interacted with 1920 and 1930 dummies, respectively. Coefficients from stage-zero regression are then used to construct WWI and Quota instruments according to the procedure described in Section 7.3. Column 2 shows the first-stage results for 10 years-difference in ethnicity e's network variable (s_{ecst}), Column 3 for 10 years-difference in Theil index of diversity within county c's foreign-born population from countries/areas other than e ($Theil_{-ecst}$), Column 4 for 10 years-difference in the share, as a fraction of c's total population, of foreign-born population from countries/areas other than e (s_{-ecst}).
All regressions include ethnicity by county and state by year fixed effects. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).
In first-stage regressions in Columns 2 to 4, both outcome and explanatory variables are standardized by subtracting sample mean and then dividing by standard deviation. First-stage coefficients can therefore be interpreted as s.d. changes in the endogenous variables induced by a s.d. change in the corresponding instrument.
Last row reports the values for the weak instrument test for multiple endogenous variables (Sanderson and Windmeijer, 2016).

Table 10: 2SLS estimates - WWI and quota IVs

	(1) Immigrant inventors location choice	(2) Immigrant inventors location choice	(3) Immigrant inventors productivity	(4) Immigrant inventors productivity
	$\Delta log(L)_{ecst}$	$\Delta log(L)_{ecst}$	$\Delta log(T)_{ecst}$	$\Delta log(T)_{ecst}$
Within Diversity: $\Delta Theil_{-ecst}$	0.7004***	0.4200***	0.4343***	0.2599***
	(0.1436)	(0.0546)	(0.1288)	(0.0508)
Between Diversity: Δs_{-ecst}	0.0118***	0.0239***	0.0047*	0.0134***
	(0.0029)	(0.0027)	(0.0025)	(0.0022)
Network: Δs_{ecst}	0.1025***	0.0927***	0.0183	0.0294***
	(0.0310)	(0.0206)	(0.0147)	(0.0098)
Observations	171,795	171,795	171,795	171,795
Ethnicity by County FE		Yes		Yes
Year by State FE	Yes	Yes	Yes	Yes
First differences model Yes	Yes	Yes	Yes	Yes

All columns present 2SLS estimates employing WWI and Quota instruments as defined in Section 7.3. Columns 1 and 2 consider as outcome variable the 10 years-difference in (log) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable in Columns 3 and 4 is the 10 years-difference in (log) number of patents per inventor from ethnicity e and living in county c.

All specifications correspond to a first-differenced version of the baseline model in (11), and include state by year fixed effects. Estimates in Columns 2 and 4 also adjust for ethnicity by county fixed effects. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

8. Conclusions

We have investigated the importance of ethnic networks and diversity in determining immigrant inventors' settlements exploiting a decennial dataset on immigrant inventors arrived to US during the Age of Mass Migration from 1880 to 1940. The dataset contains about 43,000 patents granted to about 20,000 immigrants together with the patentees' counties of residence and ethnicity as reported in USPTO records. These pieces of information are matched with NHGIS IPUMS county-level decennial census files between 1870 and 1930. Exploiting variation across $1,900$ counties and 15 ethnicities over time, we have looked at the impact of local co-ethnic networks and diversity in each census year on the change in immigrant inventors' presence and productivity in the subsequent decade.

We have found that co-ethnic networks as well as 'between' and 'within' diversity act as significant pull factors for immigrant inventors. A model of immigrant inventors' location choices has allowed to identify the main mechanism through which those pull factor operate in externalities that foster inventors' productivity. Our findings are robust to checks of instruments' validity and to the inclusion of several control variables, including counties' population and exposure to the American frontier. Though based on historical evidence, they are nonetheless relevant for today's advanced economies where the migration share has risen significantly in recent years but the rhetoric around it has turned to be mostly negative. Immigration has become the bogeyman of many politicians in several receiving countries who have gained national prominence by stirring anti–immigrant sentiment. Yet, when it comes to immigrants in the top tail of the skill distribution, even anti–immigration politicians tend to fudge their rhetoric as the bulk of evidence points to global talents boosting innovation and productivity (Kerr et al., 2016).

References

Abramitzky, R., Ager, P., Boustan, L. P., Cohen, E., and Hansen, C. W. (2019). The effects of immigration on the economy: Lessons from the 1920s border closure. Technical report, National Bureau of Economic Research.

Abramitzky, R. and Boustan, L. (2017). Immigration in american economic history. *Journal of economic literature*, 55(4):1311–45.

Acemoglu, D. (2010). When does labor scarcity encourage innovation? *Journal of Political Economy*, 118(6):1037–1078.

Adao, R., Kolesár, M., and Morales, E. (2019). Shift-share designs: Theory and inference. *The Quarterly Journal of Economics*, 134(4):1949–2010.

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

Ager, P. and Hansen, C. W. (2017). Closing heaven's door: Evidence from the 1920s u.s. immigration quota acts. *Discussion Papers 17-22, University of Copenhagen. Department of Economics.*

Akcigit, U., Grigsby, J., and Nicholas, T. (2017). The rise of american ingenuity: Innovation and inventors of the golden age. *National Bureau of Economic Research.*, No. w23047.

Alesina, A., Harnoss, J., and Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, page 1–38.

Algan, Y. and Cahuc, P. (2010). Inherited trust and growth. *American Economic Review*, 100(5):2060–92.

Arkolakis, C., Peters, M., and Lee, S. K. (2019). European immigrants and the united states' rise to the technological frontier. Technical report, Society for Economic Dynamics.

Bahar, D., Rapoport, H., and Turati, R. (2020). Birthplace diversity and econoimc complexity: Cross-country evidence. *Research Policy*, page 103991.

Bandiera, O., Rasul, I., and Viarengo, M. (2013). The making of modern america: Migratory flows in the age of mass migration. *Journal of Development Economics*, 102:23–47.

Bazzi, S., Fiszbein, M., and Gebresilasse, M. (2020). Frontier culture: The roots and persistence of "rugged individualism" in the united states. *Econometrica*, 88(6):2329–2368.

Bergquist, C. W. (1984). *Labor in the Capitalist World-economy*, volume 7. Sage Publications, Inc.

Berkes, E. and Gaetani, R. (2020). The geography of unconventional innovation geography of unconventional innovation. *The Economic Journal*.

Boeheim, R., Horvath, G., and Mayr, K. (2012). Birthplace diversity of the workforce and productivity spill-overs in firms. *WIFO Working Papers*, 438.

Borusyak, K., Hull, P., and Jaravel, X. (2018). Quasi-experimental shift-share research designs. Technical report, National Bureau of Economic Research.

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, 19(1):22–64.

Diodato, D., Morrison, A., and Petralia, S. (2021). Migration and invention in the Age of Mass Migration. *Journal of Economic Geography*. lbab032.

Docquier, F., Turati, R., Valette, J., and Vasilakis, C. (2018). Birthplace diversity and economic growth: Evidence from the us states in the post-world war ii period. *mimeo*.

Doran, K. and Yoon, C. (2018). Immigration and invention: Evidence from the quota acts. *Unpublished manuscript, https://www3. nd. edu/˜ kdoran/Doran_Quotas. pdf*.

Eckert, F., Gvirtz, A., Liang, J., and Peters, M. (2018). A consistent county-level crosswalk for us spatial data since 1790. Technical report, Working Paper.

Fouka, V., Mazumder, S., and Tabellini, M. (2021). From immigrants to americans: Race and assimilation during the great migration. *The Review of Economic Studies*.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from US daily newspapers. *Econometrica*, 78(1):35–71.

Giuliano, P. (2007). Living arrangements in western europe: Does cultural origin matter? *Journal of the European Economic Association*, 5(5):927–952.

Giuliano, P. and Tabellini, M. (2020). The seeds of ideology: Historical immigration and political preferences in the united states. Technical report, National Bureau of Economic Research.

Glaeser, E. L. and Resseger, M. G. (2010). The complementarity between cities and skills. *Journal of Regional Science*, 50(1):221–244.

Goldin, C. (1994). *The regulated economy: A historical approach to political economy*, chapter The political economy of immigration restriction in the united states,, page 223–258. University of Chicago Press.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624.

Hatton, T. J. and Williamson, J. G. (1998). *The age of mass migration: Causes and economic impact*. Oxford University Press on Demand.

Hughes, T. P. (2004). *American genesis: a century of invention and technological enthusiasm, 1870-1970*. University of Chicago Press.

Hunt, J. (2011). Which immigrants are most innovative and entrepreneurial? distinctions by entry visa. *Journal of Labor Economics*, 29(3):417–457.

Jaeger, D. A., Ruist, J., and Stuhler, J. (2018). Shift-share instruments and the impact of immigration. Technical report, National Bureau of Economic Research.

Jaffe, A. B., Lerner, J., and Stern, S. (2001). *Innovation policy and the economy*. Mit Press.

Kahane, L., Longley, N., and Simmons, R. (2013). The effects of coworker heterogeneity on firm-level output: Assessing the impacts of cultural and

language diversity in the national hockey league. *The Review of Economics and Statistics*, 95(1):302–314.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*.

Kemeny, T. (2017). Immigrant diversity and economic performance in cities. *International Regional Science Review*, 40(2):164–208.

Kerr, S. P., Kerr, W., Özden, C., and Parsons, C. (2016). Global talent flows. *Journal of Economic Perspectives*, 30(4):83–106.

Kerr, W. and Lincoln, W. F. (2010). The supply side of innovation: H-1b visa reforms and us ethnic invention. *Journal of Labor Economics*, 28(3):473–508.

Khan, B. Z. (2005). *The Democratization of Invention: patents and copyrights in American economic development, 1790-1920*. Cambridge University Press.

Khan, Z. and Sokoloff, K. L. (2004). Institutions and democratic invention in 19th-century america: Evidence from" great inventors," 1790-1930. *American Economic Review*, 94(2):395–401.

King, D. (2009). *Making Americans: Immigration, race, and the origins of the diverse democracy*. Harvard University Press.

Kleven, H. J., Landais, C., and Saez, E. (2013). Taxation and international migration of superstars: Evidence from the european football market. *American economic review*, 103(5):1892–1924.

Logan, T. D. and Parman, J. M. (2017). The national rise in residential segregation. *The Journal of Economic History*, 77(1):127–170.

Manson, S., Schroeder, J., Van Riper, D., and Ruggles, S. (2019). Ipums national historicla geographical information system: Version 14.0 [database]. ipums. *Institute for Social Research and Data Innovation. University of Minnesota*.

McFadyen, A. (1936). Meet the champion inventors. *Popular Science*, 128(1).

McKenzie, D. and Rapoport, H. (2007). Network effects and the dynamics of migration and inequality: theory and evidence from mexico. *Journal of development Economics*, 84(1):1–24.

Moretti, E. (2012). *The New Geography of Jobs*. Houghton Mifflin Harcourt.

Moser, P. and San, S. (2019). Immigration, science, and invention: Evidence from the 1920s quota acts. *Unpublished manuscript*.

Moser, P., Voena, A., and Waldinger, F. (2014). German jewish émigrés and us invention. *The American Economic Review*, 104(10):3222–3255.

Nicholas, T. (2010). The role of independent invention in us technological development, 1880–1930. *The Journal of Economic History*, 70(1):57–82.

Ottaviano, G. and Peri, G. (2006). The economic value of cultural diversity: Evidence from u.s. cities. *Journal of Economic Geography*, 6(1):9–44.

Ottaviano, G. I. and Peri, G. (2005). Cities and cultures. *Journal of Urban Economics*, 58(2):304–337.

Ottinger, S. (2020). Immigrants, industries, and path dependence. Technical report, UCLA mimeo.

Ozgen, C., Peters, C., Niebuhr, A., Nijkamp, P., and Poot, J. (2014). Does cultural diversity of migrants employees affect innovation? *International Migration Review*, 48(9):377–416.

Peri, G. (2016). Immigrants, productivity, and labor markets. *Journal of Economic Perspectives*, 30(4):3–30.

Petralia, S., Balland, P.-A., and Rigby, D. L. (2016). Unveiling the geography of historical patents in the united states from 1836 to 1975. *Scientific data*, 3(1):1–14.

Roback, J. (1982). Wages, rents, and the quality of life. *Journal of political Economy*, 90(6):1257–1278.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102.

Rotemberg, J. (1983). Instrument variable estimation of misspecified models.

Ruggles, S., King, M. L., Levison, D., McCaa, R., and Sobek, M. (2003). Ipums-international. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 36(2):60–65.

Sanderson, E. and Windmeijer, F. (2016). A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221.

Sequeira, S., Nunn, N., and Qian, N. (2020). Immigrants and the making of america. *The Review of Economic Studies*, 87(1):382–419.

Spolaore, E. and Wacziarg, R. (2016). Ancestry, language and culture. *The Palgrave Handbook of Economics and Language*, page 174.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression. Technical report, National Bureau of Economic Research.

Suedekum, J., Wolf, K., and Blien, U. (2014). Cultural diversity and local labour markets. *Regional Studies*, 48(1):173–191.

Tabellini, M. (2020). Gifts of the immigrants, woes of the natives: Lessons from the age of mass migration. *The Review of Economic Studies*, 87(1):454–486.

Ueda, R. (1992). American national identity and race in immigrant generations: Reconsidering hansen's" law".

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education.

Zucker, L. G., Darby, M. R., and Armstrong, J. (1998). Geographically localized knowledge: spillovers or markets? *Economic Inquiry*, 36(1):65–86.

Appendix A: Shift-share IVs and identification: Rotemberg weights (Goldsmith-Pinkham et al., 2020)

We here perform the complete battery of tests suggested by Goldsmith-Pinkham et al. (2020) with the aim of analysing the main sources of variation driving shift-share IVs in our paper. In particular, we compute ethnic-specific Rotemberg weights by making use of the replication code provided by authors.[30] More in detail, we adopt the same procedure as for the estimates of the inverse elasticity of labor supply (section 6) as it resembles our baseline panel specification with two-way fixed effects. In this case, we include the ethnicity shares interacted with year fixed effects as underlying instruments, and then aggregate Rotemberg weights at the ethnicity level.

In Appendix Tables A.1, A.2 and A.3 we report, for each endogenous variable in our analysis, the complete set of statistics advised by Goldsmith-Pinkham et al. to study the type of variation driving the identification with shift-share instruments. Panel A reports the share and sum of both positive and negative Rotemberg weights. Panel B reports the correlation matrix, across ethnicities, between the Rotemberg weight (α_e), the nation-wide change in immigrants stock ($g_e = N_{e,-s,[t-1;t]}$), the coefficients from just-identified 2SLS estimates for location choice ($\beta(inv)_e$) and productivity ($\beta(prod)_e$), the first-stage F-statistic of the ethnicities share (F_e), and the standard deviation in the 1870 ethnicities shares across counties ($\text{Var}(s^{US}_{ecs,1870})$). Panel C reports the top five ethnicities according to Rotemberg weights together with the corresponding values of nation-wide change in immigrants stock, g_e, and coefficients from just-identified 2SLS regressions, $\beta(inv)_e$ and $\beta(prod)_e$.

Panel C of Table A.1 shows that top-5 Rotemberg weight ethnicities for the shift-share IV for Theil index account for almost all positive weights (1.692/1.740). Eastern Europe approximately receives one third of positive weights and thus largely contributes to variation in the instrument, followed by Canada ($\alpha_e = 0.4$), Great Britain and Ireland ($\alpha_e = 0.35$), Germany ($\alpha_e =$ 0.0.19) and Asia ($\alpha_e = 0.14$). This suggests that the validity of shift-share 2SLS estimates is particularly sensitive to deviations from the identifying assumption related to variation across counties in initial share of Eastern Europeans and Canadians. Panel B indicates that Rotemberg weights are strongly correlated with the variation across counties in the initial ethnicities shares, $\text{Var}(s^{US}_{ecs,1870})$, and even more with the nation-wide changes in immigrants stock, g_e. Both the 'shift' and the 'share' components therefore explain a substantial amount of the variation in this instrument.

Panel C shows that top-5 Rotemberg weight ethnicities for the shift-share IV for Theil index account for almost all positive weights (1.692/1.741). Eastern Europe approximately receives one third of positive weights and thus largely

30. This is available at: https://github.com/paulgp/bartik-weight.

contributes to variation in the instrument, followed by Canada ($\alpha_e = 0.4$), Great Britain and Ireland ($\alpha_e = 0.35$), Germany ($\alpha_e = 0.0.19$) and Asia ($\alpha_e = 0.14$). This suggests that the validity of shift-share 2SLS estimates is particularly sensitive to deviations from the identifying assumption related to variation across counties in initial share of Eastern Europeans and Canadians. Panel B indicates that Rotemberg weights are strongly correlated with the variation across counties in the initial ethnicities shares, $\text{Var}(s^{US}_{ecs,1870})$, and even more with the nation-wide changes in immigrants stock, g_e. Both the 'shift' and the 'share' components therefore explain a substantial amount of the variation in this instrument.

We display the same set of statistics regarding the shift-share IVs for between diversity, in Table A.2, and co-ethnic networks, in Table A.3. Again, Eastern Europe receives the highest Rotemberg weight for both IVs, and accounts for an even larger share of total positive weights than in the case of shift-share Theil index IV. Panel B of Table A.2 shows that Rotemberg weights for shift-share between diversity present fairly similar levels of correlation with variation in the initial local ethnicities shares, $\text{Var}(s^{US}_{ecs,1870})$, and the nation-wide change in immigrants stock, g_e. The same correlation matrix for the shift-share co-ethnic networks IV, in panel B of Table A.3, instead highlight that Rotemberg weights are mostly correlated with variation in the initial local ethnicities shares. The 'share', rather then the 'shift', component hence mainly contributes to the variation in this instrument.

We finally check whether initial ethnicities shares, $s^{US}_{ecs,1870}$, correlate with county-level characteristics in 1870. It is indeed likely that immigrants' concentration at the baseline depends from local factors, such as population size or economic development. As illustrated in Section 7.1, there might be productivity or labour demand shocks in one or more counties which at the baseline affect immigrants' concentration and stimulate, or disincentive, innovation at the same time. If these shocks were serially correlated, identifying assumptions would be violated.

In Appendix Table A.4, we focus on the subset of ethnicities among top-5 according to Rotemberg weights for shift-share IVs, and regress each of county-level initial ethnicities shares on the vector of 1870 local characteristics that are available in NHGIS Census county-level file (Manson et al., 2019), including (log) population, (log) farming and manufacturing output per worker, share of population in manufacturing sector, (log) wage in manufacturing and share of illiterate population. We also check whether initial ethnicities shares are correlated with the share of high-skilled individuals, i.e. scientists, college professors and engineers, who are identified by means of occupational (1950 classification) information provided in 1870 IPUMS full-count Census micro-data (Ruggles et al., 2003). Moreover, in order to proxy for a county's market access, we consider NHGIS indicators for water and railroad transportation in 1860. All estimates account for state fixed effect. Standard errors are clustered at the state level. Most important, the result in column 6 reveals that the

concentration of Eastern Europeans, the ethnicity with highest Rotemberg weights for all shift-share IVs, is not significantly correlated with any of the county-level characteristics. The distribution of immigrants from the rest of ethnicities, except Asian, on the other hand, is positively and significantly associated with (log) population size. The initial shares of immigrants from Austro-Hungarian Empire – the second-ranked ethnicity as for Rotemberg weights for between diversity and co-ethnic networks IVs – Canada, Germany and Great Britain & Ireland also display a significant and positive correlation with the share of workers in manufacturing sector. We also detect a significant association between i) the illiteracy rate and the initial distribution of migrants from Austro-Hungarian Empire and Germany; ii) the (log) output per worker in farming sector and the initial distribution of Germans; iii) the (log) average wage in manufacturing sector and the 1870 share of immigrants from Benelux; iv) the indicator for water transportation access and the initial shares of migrants from Benelux and Germany. In section 7.1, we deal with these confounding factors at the baseline by performing a series of estimates in which we control for local characteristics and ethnicities shares in 1870 interacted with year fixed effects.

Table A.1: Rotemberg weights - shift-share IV for Theil index (within-diversity)

Panel A: Negative and positive weights

	Sum	Mean	Share
Negative	-0.740	-0.106	0.298
Positive	1.740	0.249	0.702

Panel B: Correlations of ethnicity aggregates

	α_e	g_e	$\beta(inv)_e$	$\beta(prod)_e$	F_e	$\text{Var}(z_e)$
α_e	1					
g_e	0.625	1				
$\beta(inv)_e$	-0.110	-0.121	1			
$\beta(prod)_e$	-0.115	-0.124	0.997	1		
F_e	0.469	0.334	-0.238	-0.255	1	
$\text{Var}(z_e)$	0.341	0.541	0.339	0.336	-0.342	1

Panel C: Top 5 Rotemberg weight ethnicities

	$\hat{\alpha}_e$	g_e	$\hat{\beta(inv)}_e$	$\hat{\beta(prod)}_e$
Canada	0.407	3.36e+05	-3.074	-0.743
Eastern Europe	0.597	4.61e+05	-36.904	-0.873
Germany	0.193	-1.38e+04	42.981	9.645
Great Britain and Ireland	0.352	-8.02e+04	17.668	4.059
Asia	0.143	-1.05e+04	2.275	0.168

This table reports the complete set of diagnostic statistics (Goldsmith-Pinkham et al., 2020) for shift-share IV for Theil index (within diversity). Panel A reports the share and sum of both positive and negative Rotemberg weights. Panel B reports the correlation matrix, across ethnicities, between the Rotemberg weight (α_e), the nation-wide change in immigrants stock ($g_e = N_{e,-s,[t-1;t]}$), the coefficients from just-identified 2SLS estimates for location choice ($\beta(inv)_e$) and productivity ($\beta(prod)_e$), the first-stage F-statistic of the ethnicities share (F_e), and the standard deviation in the 1870 ethnicities shares across counties ($\text{Var}(s^{US}_{ecs,1870})$). Panel C reports the top five ethnicities according to Rotemberg weights together with the corresponding values of nation-wide change in immigrants stock, g_e, and coefficients from just-identified 2SLS regressions, $\beta(inv)_e$ and $\beta(prod)_e$.

Table A.2: Rotemberg weights - shift-share IV for the share of immigrants other than ethnicity e (between-diversity)

Panel A: Negative and positive weights

	Sum	Mean	Share
Negative	-0.107	-0.053	0.088
Positive	1.107	0.092	0.912

Panel B: Correlations of ethnicity aggregates

	α_k	g_e	$\beta(inv)_e$	$\beta(prod)_e$	F_e	$\text{Var}(z_e)$
α_e	1					
g_e	0.488	1				
$\beta(inv)_e$	-0.089	-0.295	1			
$\beta(prod)_e$	-0.159	-0.333	0.962	1		
F_e	0.270	0.512	-0.187	-0.149	1	
$\text{Var}(z_e)$	0.539	0.005	-0.174	-0.088	-0.137	1

Panel C: Top 5 Rotemberg weight ethnicities

	$\hat{\alpha}_e$	g_e	$\beta(\hat{inv})_e$	$\beta(\hat{prod})_e$
Austro-Hungarian Emp.	0.221	2.14e+05	0.417	0.069
Benelux	0.036	22815.894	0.065	0.025
Canada	0.176	3.36e+05	0.059	0.014
Eastern Europe	0.555	4.61e+05	0.331	0.008
Germany	0.065	-1.38e+04	-1.063	-0.238

This table reports the complete set of diagnostic statistics (Goldsmith-Pinkham et al., 2020) for shift-share IV for the share of immigrants other than ethnicity e (between-diversity). Panel A reports the share and sum of both positive and negative Rotemberg weights. Panel B reports the correlation matrix, across ethnicities, between the Rotemberg weight (α_e), the nation-wide change in immigrants stock ($g_e = N_{e,-s,[t-1;t]}$), the coefficients from just-identified 2SLS estimates for location choice ($\beta(inv)_e$) and productivity ($\beta(prod)_e$), the first-stage F-statistic of the ethnicities share (F_e), and the standard deviation in the 1870 ethnicities shares across counties ($\text{Var}(s^{US}_{ecs,1870})$). Panel C reports the top five ethnicities according to Rotemberg weights together with the corresponding values of nation-wide change in immigrants stock, g_e, and coefficients from just-identified 2SLS regressions, $\beta(inv)_e$ and $\beta(prod)_e$.

Table A.3: Rotemberg weights - shift-share IV for ethnicity's e co-ethnic network

Panel A: Negative and positive weights

	Sum	Mean	Share
Negative	-0.037	-0.019	0.034
Positive	1.037	0.086	0.966

Panel B: Correlations of ethnicity aggregates

	α_k	g_e	$\beta(inv)_e$	$\beta(prod)_e$	F_e	$\text{Var}(z_e)$
α_e	1					
g_e	0.085	1				
$\beta(inv)_e$	-0.174	0.352	1			
$\beta(prod)_e$	-0.204	0.320	0.992	1		
F_e	-0.004	-0.153	-0.244	-0.240	1	
$\text{Var}(z_e)$	0.543	-0.088	-0.237	-0.198	-0.368	1

Panel C: Top 5 Rotemberg weight ethnicities

	$\hat{\alpha}_e$	g_e	$\hat{\beta(inv)}_e$	$\hat{\beta(prod)}_e$
Austro-Hungarian Emp.	0.220	2.14e+05	4.916	0.816
Benelux	0.033	22815.894	0.842	0.324
Canada	0.164	3.36e+05	0.749	0.181
Eastern Europe	0.518	4.61e+05	4.169	0.099
Germany	0.061	-1.38e+04	-13.435	-3.015

This table reports the complete set of diagnostic statistics (Goldsmith-Pinkham et al., 2020) for shift-share IV for for ethnicity's e co-ethnic network. Panel A reports the share and sum of both positive and negative Rotemberg weights. Panel B reports the correlation matrix, across ethnicities, between the Rotemberg weight (α_e), the nation-wide change in immigrants stock ($g_e = N_{e,-s,[t-1;t]}$), the coefficients from just-identified 2SLS estimates for location choice ($\beta(inv)_e$) and productivity ($\beta(prod)_e$), the first-stage F-statistic of the ethnicities share (F_e), and the standard deviation in the 1870 ethnicities shares across counties ($\text{Var}(s^{US}_{ecs,1870})$). Panel C reports the top five ethnicities according to Rotemberg weights together with the corresponding values of nation-wide change in immigrants stock, g_e, and coefficients from just-identified 2SLS regressions, $\beta(inv)_e$ and $\beta(prod)_e$.

Table A.4: Test for correlation between 1870 ethnicities shares and local characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	ethnicity 1870 share – $s^{US}_{ecs,1870}$:						
	Austro-Hung. Emp.	Benelux	Canada	Eastern Europe	Germany	Great Britain & Ireland	Asia
1870 county characteristics:							
(log) population	0.0798***	0.0615**	0.0400***	0.0992	0.0720**	0.0676**	0.0567
	(0.0276)	(0.0255)	(0.0145)	(0.0885)	(0.0276)	(0.0291)	(0.0568)
(log) p.c. farming output	-0.0056	0.0099	-0.0011	-0.0350	-0.0196*	-0.0181	0.0053
	(0.0122)	(0.0134)	(0.0032)	(0.0333)	(0.0105)	(0.0116)	(0.0075)
(log) p.c. manufacturing output	-0.0061	-0.0040	-0.0004	-0.0205	-0.0090	-0.0092	-0.0182
	(0.0066)	(0.0064)	(0.0043)	(0.0143)	(0.0075)	(0.0073)	(0.0169)
share manufacturing workers	0.0149*	0.0121	0.0127**	0.0272	0.0199**	0.0200**	0.0075
	(0.0084)	(0.0074)	(0.0052)	(0.0242)	(0.0082)	(0.0082)	(0.0075)
(log) manufacturing wage	-0.0137	-0.0221*	-0.0139	-0.0134	-0.0155	-0.0146	0.0013
	(0.0172)	(0.0128)	(0.0095)	(0.0264)	(0.0142)	(0.0128)	(0.0057)
share illiterates	-0.0012*	-0.0010	0.0004	0.0003	-0.0014*	-0.0005	-0.0007
	(0.0007)	(0.0008)	(0.0005)	(0.0009)	(0.0007)	(0.0006)	(0.0010)
share high-skilled	0.0223	-0.0080	0.0090	0.0325	0.0134	0.0331	0.0553
	(0.0315)	(0.0217)	(0.0096)	(0.0314)	(0.0244)	(0.0325)	(0.0631)
water transport access	0.0418	0.0524*	0.0018	-0.0276	0.0214*	-0.0056	-0.0275
	(0.0293)	(0.0284)	(0.0074)	(0.0363)	(0.0113)	(0.0094)	(0.0269)
railroad access	0.0077	0.0456	-0.0098	-0.0468	0.0118	-0.0061	-0.0384
	(0.0150)	(0.0470)	(0.0095)	(0.0506)	(0.0177)	(0.0159)	(0.0377)
Observations	1,895	1,895	1,895	1,895	1,895	1,895	1,895
R-squared	0.0800	0.0807	0.4046	0.0811	0.1810	0.2440	0.2584
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

We here consider the group of top-5 Rotemberg weights for shift-share IVs. Each column reports the regression of 1870 ethnicity e's share ($s^{US}_{ecs,1870}$) on the vector of 1870 local characteristics (source: NHGIS Census county-level file (Manson et al., 2019)), including (log) population, (log) farming and manufacturing output per worker, share of population in manufacturing sector, (log) wage in manufacturing, share of illiterate population, share of high-skilled population (scientists, college professors and engineers) and dummies for 1860 water transport and railroad access. All estimates account for state fixed effects. Standard errors are clustered at the state level (*** p<0.01, ** p<0.05, * p<0.1).

Appendix B: Mechanisms and other robustness tests

B.1. Mechanisms

In this section we investigate more in detail the potential mechanisms behind the link between ethnic diversity and immigrant inventors' location choice and productivity. First, we test whether diversity affects inter-group connections as proxied by marriage and residential contact between individuals from different ethnicities (Giuliano and Tabellini, 2020). Higher inter-group contact may grant easier access to a more diversified set of local non-tradable goods and services and therefore affect inventors' utility function through a consumption amenity (sub–section B.1.1).

Secondly, we investigate the relationship between diversity and attitudes toward ethnic minorities. The presence of migrants from other ethnicities may indeed reduce the level of attention, and possibly the backlash, by natives toward a particular ethnic group. As in Fouka et al. (2021), we rely on historical local newspapers contents to define the number of mentions for each ethnic group and to quantify their degree of salience at the county level (sub–section B.1.2).

We then turn to the impact of ethnic diversity on skills heterogeneity. As shown by Ottinger (2020), exposure to specialized knowledge in countries of origin affected manufacturing specialization in US locations where migrants' network settled in mid-XIX century, and entailed spatial patterns that persisted over following decades. Hence, given migrants' peculiar background, higher ethnic diversity may result in higher variety of skills (Ager and Brückner, 2013), which is likely to affect inventors' production function through a production amenity. A more heterogeneous set of skills may indeed entail a major exchange and combination of knowledge from complementary fields of production and expertise, and thus boost the creation of new ideas and productivity (sub–section B.1.3).

We finally take into account cultural distance among ethnic groups. In fact, while so far we have assumed that all ethnicities have the same degree of diversity among them. Nevertheless, each pair of ethnic group features a different level of proximity along several cultural dimensions. This would affect our findings if immigrant inventors chose to settle in areas with a more diverse set of culturally close migrants. To illustrate, consider a German inventor choosing between two counties with the same characteristics and level of birthplace diversity but differing in the average cultural distance among immigrants other than Germans. If she attaches some value to cultural proximity with the rest of migrants, she will choose to settle in the county where diversity stems from a set of more culturally close migrants (e.g. Northern European as opposed to Southern Europeans or Asians). We perform this test by considering a modified version of the baseline model as in (11) including a weighted specification of Theil index of within-diversity with weights equal to

either linguistic or religious distance (Spolaore and Wacziarg, 2016) between group j and each of the rest of ethnic groups in county c (section B.2).

B.1.1. Inter-group connections. We follow Giuliano and Tabellini (2020) in the adoption of two proxies for inter-group connections: inter-marriage and residential contact. In both cases, we rely on IPUMS full-count Census micro-data (Ruggles et al., 2003) for their definition. As for inter-marriage, we exploit the information on spouse's birthplace and define the rate of out-group marriage, MR^J_{ecst}, that is the (%) share of migrants from ethnicity e living in county c at time t who are married to either US natives ($J = N$) or migrants from other ethnicities ($J = M$).[31] We then specify the following econometric model:

$$MR^J_{ecst+10} = \alpha_0 + \beta_1 s_{ecst} + \beta_2 s_{-ecst} + \beta_3 Theil_{-ecst} + \delta_{st} + \mu_{ec} + t\pi_{ec} + \varepsilon_{ecst} \tag{B.1}$$

which exactly resembles, on the RHS, the baseline model in (11) and, assuming that diversity exerts its effect with a 10 years delay, takes as outcome variable the rate of out-group marriage in $t + 10$.

Similarly to Logan and Parman (2017) and Giuliano and Tabellini (2020), the construction of the residential contact measures exploits the fact that census enumeration was performed door-to-door and the order in which households are listed in the records is likely their order on the street. A household's neighbours are hence the households appearing before and after in the records. We therefore consider household heads only and compute for each ethnicity the number of households neighbouring with different-ethnicities households. We include households having at least one observed neighbour and two households are classified as neighbouring only if living in the same county. We adopt the same model specification as in (B.1) with outcome variable NB^J_{ecst}, i.e. the (%) share of households from ethnicity e living in county c at time t who neighbour with either US natives ($J = N$) or migrants from other ethnicities ($J = M$).

Table B.1 reports shift-share 2SLS estimates on the impact of diversity on inter-group marriage and residential contact with both natives and migrants from other ethnicities.[32] Results reveal no significant association and seemingly rule out inter-group contact as a potential channel driving our findings.

31. As IPUMS full-count Census micro-data also provide information about parents' birthplace, we are able to identify second generation immigrants (i.e. born in US from foreign-born father - or mother when father's birthplace is missing). In the definition of both inter-group connections variables we therefore include in each ethnicity both first and second generation immigrants.

32. IPUMS full-count Census data are not available for 1890 as original records went destroyed in a fire. The number of observation is therefore reduced and we consider our explanatory variables from each decade from 1890 until 1930, while the outcomes defined in this section from 1900 until 1940.

Table B.1: Diversity and inter-ethnic contact: marriage and residential contact. 2SLS estimates

	(1)	(2)	(3)	(4)
	Marriage		Residential contact	
	Natives	Different ethnicity immigrants	Natives	Different ethnicity immigrants
Within Diversity: $Theil_{-ecst}$	13.9343	4.0797	-24.4760	10.0170
	(26.0502)	(12.1886)	(32.7595)	(20.8876)
Between Diversity: s_{-ecst}	0.7593	0.2823	-1.9583	0.7455
	(1.8145)	(0.8441)	(2.2878)	(1.4632)
Network: s_{ecst}	0.1695	-0.0347	-2.1700	-0.8343
	(1.5631)	(0.7268)	(1.9950)	(1.2863)
Observations	143,325	143,325	143,325	143,325
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

All regressions represent 2SLS estimates using shift-share instrumental variables. Estimates in Columns 1 and 2 respectively consider as outcome variable the share of migrants from ethnicity e, living in county c in $t+1$, who are married to US natives and immigrants from different ethnicities. In Columns 3 and 4 the outcome variables are respectively the share of household heads from ethnicity e, living in county c in $t+1$, neighbouring with US natives and immigrants from different ethnicities. All columns display 2SLS results employing shift-share instruments.
All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.1.2. Attitudes toward immigrants. Although the analysis in Section B.1.1 reveals no significant effect on inter-group connections, diversity might still affect migrant inventors' location choice through its impact on natives' attitude toward immigrants. A change in the size and ethnic composition of immigrant population may in fact reduce the local salience and, as a consequence, the hostility by natives toward a particular ethnic minority. To test these implications, we follow Fouka et al. (2021) and infer ethnic-specific local salience from historical newspapers contents which, as documented in Gentzkow and Shapiro (2010), respond to readers' demand and hence plausibly reflect the population degree of attention and attitudes toward immigrants.

As in Fouka et al., we extrapolate data on ethnic-specific mentions on newspapers from the site Newspapers.com. We trained an algorithm to perform, for all decades between 1880 and 1940, the research of a list of keywords related to each of the fifteen ethnicities considered in the paper.[33] The algorithm then

33. The ethnic-specific keywords we include as input in the Newspapers.com search engine are: australian, zealander (Australian & New Zealand); albanian, austrian, bohemian, croatian, czech, hungarian, macedonian, serbian, slovak, slovenian, yugoslav (Austro-Hungarian Empire); belgian, dutch, luxembourgish (Benelux); canadian (Canada); danish, finnish, norwegian, swedish (Scandinavia); armenian, bulgarian, latvian, lithuanian, polish, romanian, russian, soviet (Eastern Europe); french (France); german (Germany); british, english, irish, scottish, welsh (Great Britain & Ireland); greek (Greece); italian

scrapes all the locations found by Newspapers.com search engine together with the corresponding number of mentions. This procedure yields a dataset with the county-level number of newspapers mentions per ethnicity in each decade. We then build a measure of county-level relative salience by dividing the number of mentions for ethnicity e in newspapers in county c between t and $t+10$ by the total number of mentions for all ethnicities in the same location and during the same period.[34]

We again adopt the baseline specification as in (11) and test whether co-ethnic networks and diversity at time t affect the relative frequency of newspapers mentions for ethnicity e in county c between t and $t+10$. Table B.2 reports both OLS and shift-share 2SLS estimates. While focusing on 2SLS results, we find no significant correlation between both co-ethnic networks and diversity and the relative ethnic-specific salience as inferred from newspaper mentions. The outcome of these tests therefore suggests that the degree of attention (regardless this is positive or negative) that any ethnicity achieves on media is not a major mechanism behind our findings.

(Italy); chinese, japanese, korean (Asia); portuguese (Portugal); spanish (Spain); swiss (Switzerland).

34. Fouka et al. normalize the number of articles containing words such as "immigrant" or "quota" with the number of articles containing the word "and", as a way of computing the total number of articles published in a location. Our algorithm cannot perform such exercise as Newspapers.com engine does not allow the search of very common words.

Table B.2: Diversity and relative ethnic-specific local salience as inferred from newspapers mentions. OLS and 2SLS estimates

	(1)	(2)	(3)	(4)
	Relative frequency of ethnic-specific mentions on newspapers			
	OLS		Shift-Share IV	
Within Diversity: $Theil_{-ecst}$	-0.0198	0.0000	-0.4057	-0.8722
	(0.0319)	(0.0386)	(0.2817)	(0.7668)
Between Diversity: s_{-ecst}	-0.0060**	0.0029	-0.0470	-0.0421
	(0.0024)	(0.0034)	(0.0302)	(0.0589)
Network: s_{ecst}	0.0182*	-0.0057	0.0005	-0.0605
	(0.0108)	(0.0158)	(0.0379)	(0.0620)
Observations	171,990	171,990	171,990	171,990
R-squared	0.9249	0.9514		
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends		Yes		Yes

All Estimates consider as outcome variable the (%) share of mentions for ethnicity e in newspapers in county c between t and $t+10$ over the total number of mentions for all ethnicities in county c during the same period. Columns 1 and 2 display OLS estimates, while Columns 3 and 4 report 2SLS results employing shift-share instruments.

All regressions include ethnicity by county fixed effects and state by year fixed effects. Estimates in Columns 2 (OLS) and 4 (shift-share 2SLS) also adjust for ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.1.3. Skills diversity. The analysis on the link between ethnic and skills diversity unfolds along two dimensions: industry and occupation. We identify industry and occupation (1950 classification) of employment from IPUMS full-count Census micro-data (Ruggles et al., 2003), and, while adopting the same econometric specification in (B.1), we build two measures of skills diversity. The first one is the Theil index of skills (industry/occupation at the first digit) diversity computed among all migrants except those from ethnicity j. This allows us to test whether an increase in ethnic diversity affect skills dispersion in the rest of migrant population in the county. Secondly, we define the Theil index of industry and occupational diversity for each ethnicity in a county and check whether exposure to higher ethnic diversity affects the within-ethnicity skills dispersion or specialization.

Results for 2SLS estimates in Table B.3 show that higher ethnic diversity, both between and within, is positively and significant associated with higher skills diversity in all migrants population except ethnicity j both at industry (Column 1) and occupational level (Column 3). Ethnic diversity, on the other hand, is negatively associated with within-ethnicity skills dispersion, i.e. with increasing industry (Column 2) and occupational (Column 4) specialization. Yet, the latter effects are not precisely estimated. At the same time, the share of same ethnicity migrants negatively and significantly affects within-ethnicity skill dispersion. This is partly in line with the findings in Ottinger (2020) and indicates that the size of co-ethnic networks might be relevant for spatial patterns of specialization.

Table B.3: Ethnic and skills diversity. 2SLS estimates

	(1)	(2)	(3)	(4)
	Industry diversity		Occupational diversity	
	Rest of migrant population	Within ethnicity	Rest of migrant population	Within ethnicity
Within Diversity: $Theil_{-ecst}$	1.1156***	-0.4797	0.7438***	-0.6183
	(0.3681)	(0.3640)	(0.2752)	(0.4020)
Between Diversity: s_{-ecst}	0.0808***	-0.0322	0.0453**	-0.0439
	(0.0267)	(0.0254)	(0.0195)	(0.0282)
Network: s_{ecst}	0.0759***	-0.0452*	0.0425**	-0.0731***
	(0.0246)	(0.0231)	(0.0179)	(0.0259)
Observations	143,325	143,325	143,325	143,325
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

All regressions represent 2SLS estimates using shift-share instrumental variables. Columns 1 considers as outcome variable the Theil index of industry (at the first digit) diversity among immigrant population with the exclusion of ethnicity e, while Column 2 considers the same diversity index computed among immigrants from ethnicity e. The estimates in Column 3 and 4 perform the same empirical exercise with occupational (at the first digit) diversity. All columns display 2SLS results employing shift-share instruments.

All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.2. Cultural distance

We now check whether inter-ethnic cultural proximity plays a significant role in the decision by immigrant inventors to settle in more diverse areas. More in detail, we want to test whether in choosing among a set of local diversity bundles - with each differing not only in the level of diversity but also in the in the average cultural distance from own ethnicity - inventors value more (i.e. choose to locate or are more productive) those with higher cultural proximity. We operationalize this test by considering a cultural distance-weighted version of the within-diversity Theil index defined in (10), as follows:

$$WTheil^{J}_{-ecst} = \sum_{i \neq e} \frac{\theta^{J}_{ei} N_{icst}}{\sum_{i \neq e} \theta^{J}_{ei} N_{icst}} \ln\left(\frac{\sum_{i \neq e} \theta^{J}_{ei} N_{icst}}{\theta^{J}_{ei} N_{icst}}\right). \tag{B.2}$$

We here assign, to each ethnicity $i \neq e$, the cultural distance weight between e and i, θ^{J}_{ei}. These weights are derived from bilateral linguistic ($J = L$) and religious ($J = R$) distance data provided by (Spolaore and Wacziarg, 2016). θ^{J}_{ei} is set to be between 0 and 1 and is defined as:

$$\theta^{J}_{ei} = \frac{d^{J}_{ei} - d^{J}_{min}}{d^{J}_{max} - d^{J}_{min}} \tag{B.3}$$

where d^{J}_{ei} is the value of Spolaore and Wacziarg's cultural distance indicator between ethnicities i and e, while d^{J}_{min} and d^{J}_{max} are respectively the minimum and maximum values of the same variables among all pairs of ethnicities.[35] The cultural distance weights are set in such a way that the contribution of each ethnic group is inversely proportional to culturally proximity to ethnicity e. For instance, if two ethnic groups speak the same language, they will not contribute to each other's Theil index weighted according to linguistic distance.

We present 2SLS estimates with cultural distance-weighted Theil index in Table B.4. The results show that both between and weighted-within-diversity are positively associated with immigrant inventors' location choice and patenting productivity. The effect is robust to the adoption of both language (Columns 1 and 2) and religion (Columns 3 and 4) as proxies of cultural distance. Furthermore, point estimates do not significantly differ from those in baseline estimates in Column 4 of Table 4. This suggests that cultural proximity is not a relevant factor for immigrant inventors' settlements and productivity when choosing among different local diversity bundles.

35. In cases of ethnic categories consisting of multiple countries of birth (e.g. Eastern Europeans), we take weighted averages of cultural distance variables. For generic ethnicity e, these weights are based on the average stock of migrants from each country belonging to e between 1880 and 1940. For example, among Eastern European countries, higher weights are attached to Russia and Poland, as migrants from these countries represented most of Eastern European residents in US during the Age of Mass Migration.

Table B.4: 2SLS Estimates with cultural-distance-weighted Theil index

	(1)	(2)	(3)	(4)
	Linguistic distance		Religious distance	
	location choice: $log(L)_{ecst}$	productivity: $log(T)_{ecst}$	location choice: $log(L)_{ecst}$	productivity: $log(T)_{ecst}$
Weighted-Within Diversity: $WTheil^{J}_{-ecst}$	0.3888***	0.1668***	0.4543***	0.2320***
	(0.0649)	(0.0508)	(0.0665)	(0.0524)
Between Diversity: s_{-ecst}	0.0477***	0.0211***	0.0518***	0.0255***
	(0.0058)	(0.0039)	(0.0059)	(0.0041)
Network: s_{ecst}	0.0874***	0.0508***	0.0892***	0.0534***
	(0.0108)	(0.0077)	(0.0109)	(0.0078)
Observations	171,990	171,990	171,990	171,990
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

All estimates consider a weighted version of the Theil index of ethnic within-diversity. The weight in Columns 1 and 2 is the linguistic distance index between ethnicity e and each of other foreign ethnicities entering the definition of the Theil index in county c at time t. In Columns 3 and 4 we consider as weight the religious distance between two ethnicities, while in Columns 5 and 6 the index of genetic distance. All columns display 2SLS results employing shift-share instruments.

All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.3. Estimates with patent quality (Kelly et al., 2020) as an outcome

In order to test whether the effect of diversity on immigrant inventors' location choice is related to a production amenity, we have so far employed the number of patents per inventor as a measure of productivity. However, patents count per inventor might represent a rough metric as it does not account for patent quality, which rather depends on its novelty and influence for subsequent innovations. Patents number of citations may be therefore a more suitable measure for this purpose. Nevertheless, the probability of a citation being recorded is very low for patents filed before 1945 (Berkes and Gaetani, 2020) and we cannot rely on consistent citations data for the period under consideration in this paper.

Kelly et al. (2021) develop a novel indicator of a patent quality that is based on textual similarity between each pair of US patents from 1840 onward. The idea is that influential paper are those containing more words or phrases, and ultimately concepts, which are not found in previous patents (novelty) but which are frequent in subsequent ones (impact). In particular, they devise the following measure of patent significance:

$$q_j^\tau = \frac{FS_j^\tau}{BS_j^\tau} \tag{B.4}$$

where BS_j^τ is a measure of backward similarity, i.e. the degree of textual similarity between patent j and patents filed within τ years before its registration. This quantifies patent's novelty and the lower it is the higher the quality of the patent as it contains a set of words which are relatively novel with respect to patents issued in the previous τ years. Conversely, FS_j^τ measures the degree of textual similarity with patents issued in the subsequent τ years and therefore its impact on later creation of ideas.

We match data on patents' quality from replication files in Kelly et al. with our data on patents by immigrant inventors. More in detail, we consider a time window equal to 20 years for forward similarity ($\tau = 20$), which is the largest time-frame provided by authors and it is plausibly the most appropriate choice for the aim of capturing patents' quality beyond short term impact. On the other hand, we select $\tau = 5$ as for the backward similarity indicator since this is the only one provided by authors in replication file. We then calculate the ratio between 20-year-forward and 5-years-backward similarity for each patent and take the average at the ethnicity-by-county level. We finally define the log of the average quality of patents at ethnicity-by-county level, $\log(q_{ect})$, as dependent variable.

Results in Table B.5 are consistent with the empirical evidence on immigrant inventors' productivity in Table 4. In fact, co-ethnic networks and both between and within diversity are positively and significantly associated with patent quality.

Table B.5: Diversity, migrant inventors' location choice and patents' quality. OLS and 2SLS estimates

	(1)	(2)	(3)	(4)
	Dep. var: log(patent innovativeness)			
	OLS		Shift-Share IV	
Within Diversity: $Theil_{-ecst}$	0.0195***	0.0293***	0.0021	0.3570***
	(0.0027)	(0.0038)	(0.0203)	(0.0883)
Between Diversity: s_{-ecst}	0.0014***	0.0047***	0.0070***	0.0434***
	(0.0003)	(0.0006)	(0.0027)	(0.0071)
Network: s_{ecst}	0.0138***	0.0312***	0.0301***	0.1125***
	(0.0015)	(0.0038)	(0.0047)	(0.0144)
Observations	171,990	171,990	171,990	171,990
R-squared	0.4852	0.5723		
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends		Yes		Yes

The outcome variable in all estimates is the (log) average patent innovativeness by inventors from ethnicity e living in county c between t and $t+1$. Columns 1 and 2 display OLS estimates, while Columns 3 and 4 report 2SLS results employing shift-share instruments.

All regressions include ethnicity by county fixed effects and state by year fixed effects. Estimates in Columns 2 (OLS) and 4 (2SLS, shift-share IVs) also adjust for ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.4. Estimates with ethnicity-year fixed effects

Table B.6: Estimates with ethnicity-year fixed effects

	(1)	(2)	(3)	(4)
	location choice: $log(L)_{ecst}$		productivity: $log(T)_{ecst}$	
	OLS	2SLS	OLS	2SLS
Within Diversity: $Theil_{-ecst}$	0.0249***	0.4716***	0.0140***	0.2110***
	(0.0026)	(0.0782)	(0.0023)	(0.0598)
Between Diversity: s_{-ecst}	0.0051***	0.0554***	0.0026***	0.0251***
	(0.0005)	(0.0070)	(0.0004)	(0.0047)
Network: s_{ecst}	0.0339***	0.0912***	0.0153***	0.0532***
	(0.0039)	(0.0134)	(0.0024)	(0.0095)
Observations	171,990	171,990	171,990	171,990
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Year by Ethnicity FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

Estimates in Columns 1 (OLS) and 2 (2SLS, shift-share IVs) consider as outcome variable the (log) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Columns 3 (OLS) and 4 (2SLS) is the (log) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c.

All regressions include ethnicity by county, state by year and ethnicity by year fixed effects, plus ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.5. Estimates with standard errors clusterized at the county level

Table B.7: Estimates with standard errors clusterized at the county level

	(1)	(2)	(3)	(4)
	location choice: $log(L)_{ecst}$		productivity: $log(T)_{ecst}$	
	OLS	2SLS	OLS	2SLS
Within Diversity: $Theil_{-ecst}$	0.0266***	0.4242**	0.0157***	0.1923**
	(0.0058)	(0.1817)	(0.0031)	(0.0961)
Between Diversity: s_{-ecst}	0.0052***	0.0518***	0.0027***	0.0237***
	(0.0014)	(0.0188)	(0.0005)	(0.0089)
Network: s_{ecst}	0.0320***	0.0908***	0.0146***	0.0529***
	(0.0042)	(0.0191)	(0.0024)	(0.0102)
Observations	171,990	171,990	171,990	171,990
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

Estimates in Columns 1 (OLS) and 2 (2SLS, shift-share IVs) consider as outcome variable the (log) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Columns 3 (OLS) and 4 (2SLS) is the (log) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c.

All regressions include ethnicity by county and state by year, plus ethnicity by county time-linear trends. Standard errors clustered at the county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.6. Estimates with control for high-skilled population

We here re-estimate the baseline model with the introduction of a control for the share of high-skilled population. In particular, we rely on IPUMS full-count Census micro-data from 1880 to 1930 and exploit the information on occupation (1950 classification) to identify individuals at the upper tail of skill distribution.[36] In particular, we consider college professors, scientists and engineers and compute the share of high-skilled population in county c at time t. As shown in Table B.8, the relationship between diversity, co-ethnic networks and both immigrant inventors' location choice and productivity is not affected by the inclusion of this additional time-varying control.

Table B.8: Estimates with control for high-skilled population

	(1)	(2)	(3)	(4)
	location choice: $log(L)_{ecst}$		productivity: $log(T)_{ecst}$	
	OLS	2SLS	OLS	2SLS
Within Diversity: $Theil_{-ecst}$	0.0339***	0.6762***	0.0190***	0.2998***
	(0.0033)	(0.1186)	(0.0028)	(0.0821)
Between Diversity: s_{-ecst}	0.0069***	0.0781***	0.0034***	0.0353***
	(0.0007)	(0.0112)	(0.0005)	(0.0069)
Network: s_{ecst}	0.0440***	0.1405***	0.0201***	0.0811***
	(0.0049)	(0.0170)	(0.0030)	(0.0118)
Share of high-skilled population	0.0123	-0.0334	0.0043	-0.0144
	(0.0167)	(0.0269)	(0.0161)	(0.0209)
Observations	143,325	143,325	143,325	143,325
Ethnicity by County FE	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes

Estimates in Columns 1 (OLS) and 2 (2SLS, shift-share IVs) consider as outcome variable the (log) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Columns 3 (OLS) and 4 (2SLS) is the (log) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c.

All regressions include ethnicity by county and state by year fixed effects, ethnicity by county time-linear trends and adjust for the share of high-skilled population (college professors, scientists and engineers). Standard errors clustered at the county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

36. IPUMS full-count 1890's Census micro-data are not available as original records went destroyed in a fire. We are not therefore able to define the share of high-skilled population in that year and the estimates in Table B.8 thus leverage on data from 1880 to 1930 with the exclusion of 1890. As a result, the number of observations drops from 171,990 to 143,325.

B.7. Estimates by NBER technological macro-sectors

To check whether our findings are driven by immigrant inventors and innovation in one or more particular technological areas, we here perform 2SLS shift-share estimates by NBER macro-sectors (namely: chemical, computers & communications, drugs & medical, electrical & electronic, mechanical and others). 2SLS shift-share results in Table B.9 show that the estimated effects of co-ethnic networks and diversity on immigrant inventors' location choice and productivity is positive and significant across all technological sectors.

Table B.9: 2SLS estimates by NBER technological macro-sectors

	A) Dep. var: log(number of immigrant inventors)					
	(1) Chemical	(2) Computers & communications	(3) Drugs & medical	(4) Electrical & electronic	(5) Mechanical	(6) Others
Within Diversity: $Theil_{-ecst}$	0.1134***	0.0408***	0.0348***	0.1431***	0.2759***	0.3263***
	(0.0286)	(0.0125)	(0.0134)	(0.0273)	(0.0501)	(0.0576)
Between Diversity: s_{-ecst}	0.0150***	0.0043***	0.0043***	0.0162***	0.0324***	0.0401***
	(0.0026)	(0.0010)	(0.0013)	(0.0025)	(0.0045)	(0.0053)
Network: s_{ecst}	0.0161***	0.0033**	0.0037**	0.0129***	0.0526***	0.0639***
	(0.0042)	(0.0016)	(0.0016)	(0.0038)	(0.0077)	(0.0091)
Observations	171,990	171,990	171,990	171,990	171,990	171,990
	B) Dep. var: log(immigrant inventors productivity)					
	(1) Chemical	(2) Computers & communications	(3) Drugs & medical	(4) Electrical & electronic	(5) Mechanical	(6) Others
Within Diversity: $Theil_{-ecst}$	0.0620**	0.0379***	0.0325**	0.1078***	0.1381***	0.1795***
	(0.0248)	(0.0131)	(0.0163)	(0.0257)	(0.0401)	(0.0416)
Between Diversity: s_{-ecst}	0.0091***	0.0037***	0.0036***	0.0126***	0.0148***	0.0202***
	(0.0021)	(0.0011)	(0.0013)	(0.0023)	(0.0032)	(0.0033)
Network: s_{ecst}	0.0120***	0.0027*	0.0031**	0.0098***	0.0292***	0.0361***
	(0.0036)	(0.0016)	(0.0015)	(0.0030)	(0.0051)	(0.0062)
Observations	171,990	171,990	171,990	171,990	171,990	171,990
Ethnicity by County FE	Yes	Yes	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes	Yes	Yes

All regressions represent 2SLS estimates using shift-share instrumental variables. Estimates in Panel A consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Panel B is the (log of 1+) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c.

All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.8. Estimates by 1880 population density terciles

Table B.10: OLS and 2SLS estimates - heterogeneous effects by 1880 county population density

A) Dep. var: log(number of immigrant inventors)						
	(1)	(2)	(3)	(4)	(5)	(6)
	1st tercile		2nd tercile		3rd tercile	
	$pop.den._{c1880}$ <= 18.9		18.9 >= $pop.den._{c1880}$ <= 38.7		$pop.den_{c1880}$ >= 38.7	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$	$log(L)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0092***	0.1280	0.0032	0.0979*	0.0547***	0.2854***
	(0.0022)	(4.5780)	(0.0028)	(0.0583)	(0.0084)	(0.0780)
Between Diversity: s_{-ecst}	0.0008**	0.0171	0.0026***	0.0237**	0.0177***	0.1114***
	(0.0003)	(0.3841)	(0.0009)	(0.0099)	(0.0019)	(0.0105)
Network: s_{ecst}	0.0069***	0.0482	0.0180**	0.0476***	0.0808***	0.1675***
	(0.0026)	(0.3913)	(0.0076)	(0.0174)	(0.0093)	(0.0217)
Observations	57,690	57,690	50,760	50,760	63,450	63,450
B) Dep. var: log(immigrant inventors productivity)						
	(1)	(2)	(3)	(4)	(5)	(6)
	1st tercile		2nd tercile		3rd tercile	
	$pop.den._{c1880}$ <= 18.9		18.9 >= $pop.den._{c1880}$ <= 38.7		$pop.den._{c1880}$ >= 38.7	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$	$log(T)_{ecst}$
Within Diversity: $Theil_{-ecst}$	0.0094***	2.2193	0.0028	0.1460**	0.0312***	0.1474**
	(0.0023)	(19.7492)	(0.0036)	(0.0709)	(0.0072)	(0.0589)
Between Diversity: s_{-ecst}	0.0008**	0.1905	0.0021**	0.0277**	0.0076***	0.0408***
	(0.0003)	(1.6552)	(0.0010)	(0.0121)	(0.0012)	(0.0080)
Network: s_{ecst}	0.0048*	0.2221	0.0140**	0.0512***	0.0309***	0.0728***
	(0.0025)	(1.6845)	(0.0056)	(0.0186)	(0.0052)	(0.0140)
Observations	57,690	57,690	50,760	50,760	63,450	63,450
Ethnicity by County FE	Yes	Yes	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes	Yes	Yes

Estimates in Panel A consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in Panel B is the (log of 1+) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c. Columns 1, 3 and 6 display OLS estimates, while Columns 2, 4 and 6 report 2SLS results employing shift-share instruments.
Columns 1 and 2 consider counties in the bottom tercile as for 1880 population density (inhabitants per sq. mile), Columns 3 and 4 counties in the second tercile, Columns 5 and 6 counties in the top tercile.
All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

B.9. Estimates by macro-regions

Table B.11: 2SLS estimates - heterogeneous effects by macro-regions

	(1)	(2)	(3)	(4)	(5)	(6)
	Location choice			Productivity		
	Northeast & Midwest	South	West	Northeast & Midwest	South	West
Within Diversity: $Theil_{-ecst}$	0.2869***	0.4430**	-0.4370	0.1363***	0.5267**	0.0671
	(0.0675)	(0.1999)	(0.3509)	(0.0508)	(0.2165)	(0.3534)
Between Diversity: s_{-ecst}	0.0716***	0.0334**	-0.0318	0.0230***	0.0386***	0.0395
	(0.0087)	(0.0134)	(0.0589)	(0.0059)	(0.0145)	(0.0626)
Network: s_{ecst}	0.1077***	0.1653	-0.0266	0.0492***	0.1370	0.0770
	(0.0133)	(0.1131)	(0.0885)	(0.0085)	(0.1016)	(0.0881)
Observations	92,430	67,320	12,240	92,430	67,320	12,240
Ethnicity by County FE	Yes	Yes	Yes	Yes	Yes	Yes
Year by State FE	Yes	Yes	Yes	Yes	Yes	Yes
Ethn. by County time-linear trends	Yes	Yes	Yes	Yes	Yes	Yes

This table reports 2SLS estimates (shift-share IVs) disaggregated by three macro-regions (Northeast & Midwest, South and West).. Estimates in Columns 1 to 3 consider as outcome variable the (log of 1+) number of inventors from ethnicity e, living in county c, who are granted at least one patent between t and $t+1$. The outcome variable for estimates in columns 4 to 6 is the (log of 1+) number of patents (granted between t and $t+1$) per inventor from ethnicity e and living in county c.

All regressions include ethnicity by county fixed effects, state by year fixed effects and ethnicity by county time-linear trends. Standard errors clustered at ethnicity-by-county level in parentheses (*** p<0.01, ** p<0.05, * p<0.1).