How Cultural Diversity Drives Innovation

Surnames and Patents in U.S. History*

Max Posch† Jonathan Schulz‡ Joseph Henrich§

December 20, 2024

Abstract

This paper examines the impact of cultural diversity on innovation. Focusing on the historical U.S. (1850-1940), we develop a novel measure of cultural diversity based on surnames and combine this with patent-based measures of innovation. Leveraging quasi-random variation in counties' surname compositions—stemming from the interplay of historical immigration flows—we find that rising cultural diversity increased both the quantity and quality of innovation within counties and for individual inventors: e.g., a one standard deviation increase in a county's cultural diversity raised patenting rates by 107% and individual inventor productivity by 20%. Examining mechanisms, we provide evidence suggesting that greater surname diversity fosters faster innovation both by expanding the range of ideas, skills and perspectives available for recombination and by fostering the diverse social interactions that facilitate idea sharing. The results suggest that the free flow of information between diverse minds drives innovation and contributed to the rise of the U.S. as a global innovation leader.

Keywords: Innovation, cultural diversity, surnames, immigration, collective brain

JEL Classification: O33, R11, N92, J15, Z13

---

*Max Posch and Jonathan Schulz share co-first authorship. For comments and helpful discussions, we thank Ran Abramitzky, Alberto Alesina, Mike Andrews, Pablo Balan, Michael Clemens, Tyler Cowen, Klaus Desmet, Dan Fetter, Vicky Fouka, Martin Fiszbein, Oded Galor, Erik Hornung, Chad Jones, Ross Mattheis, Petra Moser, Nathan Nunn, Tzachi Raz, Slava Savitskiy, and seminar and workshop participants at Bonn, CERGE-EI, Cologne, Copenhagen, Louvain, Lund, NBER Summer Institute, NYU and UBC. We thank Enrico Berkes for generous data sharing. Research reported in this publication was supported by the John Templeton Foundation under Award Number 62161.

†Department of Economics, University of Exeter. (email: m.posch@exeter.ac.uk)

‡Department of Economics, George Mason University. (email: jonathan.schulz77@gmail.com)

§Department of Human Evolutionary Biology, Harvard University. (email: henrich@fas.harvard.edu)

> It is hardly possible to overrate the value [...] of placing human beings in contact with persons dissimilar to themselves, and with modes of thought and action unlike those with which they are familiar. [...] Such communication has always been, and is peculiarly in the present age, one of the primary sources of progress.
>
> John Stuart Mill
> *Principles of Political Economy*

1 Introduction

At least since John Stuart Mill (1871), scholars from various disciplines have posited that social interactions among diverse individuals stimulate innovation and creativity, primarily because such interactions foster the exchange and recombination of ideas (e.g., Jacobs, 1969; Glaeser et al., 1992; Weitzman, 1998; Muthukrishna and Henrich, 2016; Jones, 2023). Indeed, research from the history of technology (Usher, 2013), economics (Acemoglu et al., 2016; Mokyr, 2002; Akcigit et al., 2013) and related fields (Youn et al., 2015) finds that many, if not most, innovations arise from the recombination of existing ideas. From this perspective, greater cultural diversity may fuel faster innovation by supplying a broader range of know-how and perspectives. Here, within the historical U.S. context, we test this hypothesis using a new fine-grained measure of cultural diversity and two patent-based measures of innovation.

To conceptualize cultural diversity, drawing on the field of cultural evolution (Boyd and Richerson, 1985; Henrich, 2016), we define 'culture' as information stored in people's heads that they acquired through social learning—by observing, listening to and copying others (Bisin and Verdier, 2001; Giuliano and Nunn, 2021). This definition includes beliefs, values, interests, heuristics, motivations and ways of thinking as well as some forms of human capital, such as know-how, skills and tacit knowledge (Nunn, 2021). These are all 'culture' because they are learned from other people and thus may persist over generations. Following this approach, we consider cultural traits to be a kind of 'information,' so 'culturally-transmitted information' refers to traits acquired from others.

Such information tends to cluster in (extended) family networks across generations because individuals often learn from older relatives, particularly from same-sex relatives such as fathers, grandfathers and uncles (for males). Sons in the U.S., for example, are far more likely to patent if their fathers patented and they typically invent in the same or a closely related technology subclass as their fathers (Bell et al., 2019). Similar patterns

emerge from studies of the intergenerational transmission of entrepreneurship (Lindquist et al., 2015).[1]

Because surnames, like many cultural traits in males, are transmitted patrilineally, they offer a natural lens through which to identify fine-grained clusters of culturally-transmitted information. Exploiting this parallel, we explore the unique contribution of surname-based cultural diversity in driving innovation and compare it with more conventional diversity measures, such as those based on ancestry or birth country (Ottaviano and Peri, 2006; Ager and Brückner, 2013; Alesina et al., 2016; Docquier et al., 2020; Fulford et al., 2020).

To construct our measure of cultural diversity, we collect all surnames reported in the full-count U.S. Census data from 1850 to 1940 and compute the entropic diversity of surnames across counties, which are presumed to be the primary locations for social interactions during this period. Counties might not capture all social interactions, especially in today's interconnected world, but we will show that they provide a reasonable approximation within the pre-1950 historical context.

We link county-level surname diversity to two patent-based indicators of innovation. First, we connect it to the universe of patents filed in the United States, as recorded in the Comprehensive Universe of U.S. Patents (Berkes, 2018). Second, we integrate this patent data with an indicator for breakthrough patents created by Kelly et al. (2021). This measure identifies patents that are both highly novel and notably impactful. Breakthrough patents, or what we term simply 'breakthroughs,' are characterized by their relatively lower textual similarity to earlier patents and higher similarity to later ones.

To estimate the causal effect of surname diversity on innovation, we adapt an instrumental variable (IV) strategy from Burchardi et al. (2019) to surnames. We isolate quasi-random variation in U.S. counties' surname compositions arising from the interactions of two historical forces: (i) the staggered arrival of immigrants with different surnames and (ii) the changing attractiveness of counties to immigrants over time. Importantly, immigration does not always increase surname diversity: its impact depends on how common or rare the incoming surnames already are in a county. For example, if many immigrants with the surname "Corleone" arrive, it might increase diversity if "Corleones" are rare, or decrease it if they are already common within that county.

Using data from 1900 to 1940, we estimate the effect of surname diversity on our patent-based innovation measures at four levels of analysis: (i) counties, (ii) surname-

[1] Ample evidence across diverse societies reveal a central role for older same-sex relatives, particularly parents, in the process of cultural transmission (Henrich and Broesch, 2011; Hewlett, 2016; Schniter et al., 2023).

counties (e.g., all "Smiths" living in Brooklyn), (iii) inventors and (iv) patents. Across the first three levels, we find that greater surname diversity within counties leads to faster innovation based on both all patents (innovation quantity) and just breakthroughs (innovation quality). All specifications include year or state-year fixed effects and, when appropriate, county, county-surname or inventor fixed effects. The latter two, respectively, absorb any unobserved and time-invariant characteristics of surname groups or individual inventors—such as their abilities, interests or knowledge—that may potentially affect the estimates. To account for scale effects, we hold population size constant using either county population sizes or an instrumental variable approach based on immigration flows. To address remaining identification concerns, we include county-specific linear trends in some specifications, which allow us to estimate the impact of surname diversity shocks on changes from the *growth rate* of local patenting rather than on absolute levels.

The estimated effects on the quantity and quality of innovation are large and economically relevant. At the county level, for example, a one standard deviation increase in surname diversity, roughly corresponding to the growth in surname diversity in New York County between 1870 to 1940, raises patent rates by 107% and breakthrough patent rates by 80% relative to their respective sample means. These estimates correspond to 1.76 times and 0.76 times the estimated effect of population size, respectively. At the individual inventor level, the same one standard deviation increase in surname diversity boosts inventors' productivity by roughly half a patent (a 20% increase over the sample mean) and increases the probability of achieving a breakthrough by 10 percentage points over a five-year period. Over a 50-year career, stretching from 1870 to 1920, these estimates imply that the same inventor would obtain five additional patents and one breakthrough if they lived in New York County instead of 300 miles further north in Clinton County, NY. To confirm the stability of these patterns before 1900, we interact surname diversity with each decade in a county-level OLS analysis. The results suggest a persistent relationship between surname diversity and innovation that extends well back into the 19th century.

Since our IV strategy relies on historical immigration patterns to isolate quasi-random variation in surname diversity, a potential concern is that immigration might have a direct effect on innovation that is not channeled through surname diversity. We mitigate this concern in two ways. First, we orthogonalize our push-pull instruments by each county's history of immigration. Second, we use exogenous variation in the number of recent immigrants, following the approach pioneered by Terry et al. (2024). We find that, although immigration promotes innovation (confirming prior work, e.g., Sequeira et al., 2020; Arkolakis et al., 2020), including this as a control has minimal influence on our estimates of the impact of surname diversity on innovation. These findings reinforce our

interpretation that surname diversity, rather than immigration itself, drives our results and bolsters confidence in our empirical strategy.

At all three levels of analysis, our results hold across a broad range of robustness checks, including placebo tests of dynamic changes (e.g., regressing prior-period innovation on current surname diversity), separate analyses of major U.S. regions (e.g., the South and West), log-transformations of the dependent variables, and alternative formulations of surname diversity, including using weights based on cultural distance. Further, by including a squared term, we also find no clear evidence for a concave relationship between surname diversity and innovation. Crucially, our findings are robust when including more conventional measures of cultural diversity in these specifications, such as those based on birth country, ancestral originating country and race. These checks demonstrate robustness against a variety of challenges.

At the level of the individual patent, our fourth level of analysis, we show that greater surname diversity among co-inventors predicts both increased technological complexity for a patent, based on the number of technology codes, and a higher likelihood of becoming a breakthrough patent. This specification includes fixed effects for inventor team size, counties, state-periods, patent-technology-periods and, in its most stringent version, county-periods. To establish causality, we employ our IV for county-level surname diversity to extract quasi-random variation in patent-level surname diversity. The first-stage result confirms that greater surname diversity at the county level leads to greater surname diversity at the patent level. The second-stage result reveals large, positive effects: a one standard deviation increase in patent-level surname diversity adds approximately four technology codes (a 150% increase relative to the sample mean) to a patent and increases the likelihood of becoming a breakthrough by roughly 90 percentage points, most of which is at the extensive margin (going from one-inventor to two-co-inventor patents). By linking greater county-level surname diversity directly to more social interaction among diverse co-inventors at the level of the patent, these findings open the door to our discussion of mechanisms.

Our mechanism section considers how and why cultural diversity influences innovation. Beginning with the stylized fact that many, if not most, innovations represent recombinations of existing ideas, cultural diversity can shape rates of innovative recombination both directly, by expanding the set of ideas, perspectives and domains of know-how that can be recombined—the 'informational channel'—and indirectly, by influencing the willingness of individuals to interact, exchange information and potentially collaborate—the 'social-behavioral channel'. Conceptually, both the informational and social-behavioral channels must be important since neither a population of diverse minds that never interact nor

a group of cognitive clones who freely interact but all share the same information will generate recombinations (Henrich and Muthukrishna, 2024). Below, we offer evidence suggesting that rising surname diversity both enriches the local informational diversity—the raw materials for recombinative innovation—and fosters more social interaction among diverse minds.

Empirically, considering the informational channel, we find that surnames cluster with educational levels, occupations, ancestry and the technological categories of inventions (i.e., patent technology codes). Therefore, counties and inventor teams will tend to have greater informational diversity when they have higher surname diversity. Consistent with this, using our county-level specifications, we show that greater surname diversity leads to greater occupational and patent technology code diversity. These findings offer *prima facie* evidence that rising surname diversity also increases the information available for recombination. However, keep in mind that beyond these observables surnames likely also capture unobserved dimensions of informational diversity, including distinct interests, beliefs, thinking styles, preferred metaphors, know-how, tacit knowledge, attentional biases, intuitions and more. Like occupations and technology codes, a large body of research across several disciplines shows how individuals acquire many cognitive and motivational traits by learning from same-sex parents, family members and their local social networks (Bell et al., 2019; Christakis and Fowler, 2009; Clark, 2014; Güell et al., 2015; Henrich, 2016; Barone and Mocetti, 2021; Schniter et al., 2023).

For the social-behavioral channel, there are two alternative hypotheses. Research on ethnic fractionalization (Alesina and La Ferrara, 2005) and the 'paradox of diversity' (Schimmelpfennig et al., 2022) suggests that greater cultural diversity might lead people to withdraw from unfamiliar or less comfortable social exchanges, preferring interactions with their extended families and co-ethnics. This effect would result in less diverse social interactions and potentially less innovation. Alternatively, as a population diversifies with newcomers who bring distinct expertise, interests and skills, heterogeneous social interactions may increase for two reasons: (1) individuals may find it harder to satisfy their needs solely within their small, homogeneous and relatively decreasing groups, and (2) the spread of groups with distinct information may create more opportunities for valuable positive-sum exchanges. Consistent with research on contact theory (e.g., Allport, 1954; Bursztyn et al., 2024) and the impact of market integration on prosocial behavior (Henrich et al., 2010; Enke, 2022; Agneman and Chevrot-Bianco, 2023; Rustagi, 2024), the potential for mutually beneficial engagement with outsiders may foster greater openness to social interactions with diverse individuals, which in turn may spur innovative recombination. Moreover, surname diversity (or the lack of it) partially captures the extent to which

large family groups are present. Research on family ties connects intensive kin networks with lower trust in strangers and increased moral parochialism, both of which can inhibit interactions and collaborations among diverse individuals (Alesina and Giuliano, 2014; Schulz et al., 2019; Enke, 2019; Henrich, 2020).

To map out the social-behavioral channel, we construct measures of the strength of family ties and residential segregation for both surnames and ancestral-countries. The family ties measure reflects the tendency of individuals to form strongly homophilic networks with relatives (Raz, 2024). The measures of residential segregation assess the frequency of adjacent neighbors with either the same surname or ancestral-country relative to what is expected based on random assortment. Counties where people more freely assort provide greater opportunities for culturally diverse individuals to connect and share information.

Using our county-level specifications, we find that greater surname diversity weakens family ties and reduces residential segregation for both surnames and ancestral countries. Interestingly, this holds even when controlling for ancestral-country and birth-country diversity. Together with our patent-level results, these findings suggest that increased fine-grained cultural diversity fosters more social interactions among heterogeneous individuals, thereby stimulating innovation.

This interpretation is further supported by an analysis of the impact of surrounding counties on innovation—that is, spillovers across county borders. We find that local proximity dominates the process, as the diversity of surrounding counties has a limited impact on the innovation within focal counties.

In sum, our analyses provide evidence of how rising cultural diversity fueled U.S. innovation, as captured by patents and breakthroughs, across four levels of analysis—counties, county surnames, inventors, and patents—from the mid-19th to the mid-20th centuries. Our exploration of mechanisms suggest that the fine-grained cultural diversity measured using surnames likely affects innovation through both informational and social-behavioral mechanisms. These results support the hypothesis that social interactions among diverse individuals foster innovation through recombination.

1.1 Contributions and Related Literature

We offer four main contributions. First, methodologically, we introduce a new fine-grained measure of cultural diversity based on surnames, demonstrate its relationship to informational diversity (e.g., occupations and patent technology codes) and compare it with more conventional cultural diversity measures in explaining innovation. Second, building

on the push-pull method from the immigration literature (Burchardi et al., 2019), we develop instrumental variables for surname diversity as well as county population size and ancestral-country diversity. Third, we show how rising cultural diversity powerfully fueled innovation in the historical United States during the country's rise to the technological frontier. Fourth, our analysis of mechanisms highlights the role of diverse social interactions—even at the level of the inventor and patent—and explores the influence of rising surname diversity on the likelihood of diverse social interactions.

These contributions relate to several lines of work in economics. To start, our findings complement important research on the impact of population diversity on innovation and economic growth (Ottaviano and Peri, 2006; Kerr, 2008; Ashraf and Galor, 2013; Ager and Brückner, 2013; Alesina et al., 2016; AlShebli et al., 2018; Arbatli et al., 2020; Docquier et al., 2020; Fulford et al., 2020; Ashraf et al., 2021; Fiszbein, 2022; Galor et al., 2024). In our conceptual framework, population diversity can fuel innovation if it captures differences in the information that individuals bring to social interactions, leading to the recombination of ideas and more novel creations.

Our results enrich this literature on population diversity and innovation in two ways. First, we demonstrate the importance of a particular form of cultural diversity in driving innovation during the era when the U.S. rose to global dominance. Second, by comparing surname diversity with both ancestral- and birth-country diversity, we highlight the larger impact of surname diversity and the distinct role of the social-behavioral channel. These results suggest that the impact of particular forms of diversity may depend on how they operates via the informational and social-behavior channels. Although certain forms of diversity may increase the potential for innovation via the informational channel, they can reduce or entirely short-circuit these effects if they suppress productive interactions among diverse individuals.

Our paper informs work connecting innovation to cities, population density, agglomeration, transportation and geographic proximity (Jacobs, 1985; Feldman and Audretsch, 1999; Carlino et al., 2007; Agrawal et al., 2008; Kerr, 2010; Glaeser, 2011; Packalen and Bhattacharya, 2015; Akcigit et al., 2017; Berkes and Gaetani, 2021; Atkin et al., 2022; Andersson et al., 2023; Chiopris, 2024) as well as an emerging literature examining how social institutions, organizations and psychological dispositions (e.g., trust) spur innovation by facilitating diverse social interactions (Algan and Cahuc, 2014; de la Croix et al., 2018; Nguyen, 2021; Buggle, 2020; Cinnirella et al., 2022; Atkin et al., 2022; Andrews, 2023; Andrews and Lensing, 2024). In our theoretical account, these factors increase the likelihood of diverse individuals meeting and potentially exchanging ideas, leading to new recombinations.

Our paper contributes to these research lines in two ways. First, while we focus on cultural diversity, our results indicate that population size independently fuels innovation alongside diversity (we report the coefficients on population size throughout). Second, consistent with this literature's emphasis on the potential for face-to-face meetings, our exploration of the social-behavioral mechanism suggests that greater surname diversity affects innovation partly through its impact on the potential for social interaction among diverse minds. Rising surname diversity at the county level leads to greater diversity at the patent level, resulting in patents that are more complex and more likely to be breakthroughs.

Our use of the quasi-random variation in surname diversity induced by immigration naturally connects our research to the literature examining the impact of immigration on innovation (Moser et al., 2014; Abramitzky and Boustan, 2017; Moser and San, 2020; Sequeira et al., 2020; Tabellini, 2020; Arkolakis et al., 2020; Abramitzky et al., 2023; Lissoni and Miguelez, 2024; Terry et al., 2024). In our conceptual framework, immigration can powerfully drive innovation when migrants bring relevant information and can socially interact in their new communities.

We contribute to the literature on immigration and innovation by showing that immigration operates on innovation partly through its influence on population size and cultural diversity. However, our results also suggest that something else is afoot in addition to these mechanisms: perhaps immigrants are self-selected on traits like individualism or trust (Knudsen, 2024; Gorodnichenko and Roland, 2016) or the experience of moving itself makes individuals more social or creative (Maddux et al., 2010; Bernstein et al., 2022).

Finally, our focus on surname diversity connects our paper to the literature on kinship and family ties within economics (Alesina and Giuliano, 2010, 2014, 2015; Buonanno and Vanin, 2017; Schulz et al., 2019; Enke, 2019; Henrich, 2020; Schulz, 2022; Bahrami-Rad et al., 2022; Ghosh et al., 2023). We contribute to this literature by showing how rising surname diversity, holding ancestral-country diversity constant, leads to less residential segregation, weaker family ties and patents with more diverse inventors.

Our presentation rolls out as follows: Section 2 describes the data and the operationalization of our key variables. Section 3 lays out the construction of the instruments for cultural diversity and population size, highlighting our identification assumption. Section 4 presents the estimates of the causal effect of county-level surname diversity on innovation at three nested levels of analysis—county, county-surname and inventor—and reports robustness checks across these levels. Section 5 presents our patent-level results, and Section 6 sheds light on potential mechanisms by offering evidence for both the informational and social-behavioral channels. Section 7 concludes with a review and

consideration of policy implications.

2 Data and Measurement

We collect data on individuals' surnames, locations, migration and patents. Summary statistics are presented in Tables B1 and B2, and additional details are in Appendix A.

2.1 Surname diversity

To measure surname diversity across counties and over time, we draw on the surname data provided by the full-count Integrated Public Use Microdata Series (Ruggles et al., 2021). We use the nine waves from 1850 to 1940 which contain the variable `namelast` of all individuals and county identifiers. Appendix A details how we clean the surname variable and harmonize county boundaries. Following Burchardi et al. (2019), we also obtain the variables age and `yrimmig` (the year of immigration) to estimate surname entropy for the mid-decades 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the mid-decade. The average surname population share within counties is 0.045% (s.d. = 0.131), the 25th percentile is 0.003%, and the 75th percentile is 0.040%. Figure B1 depicts the distribution.

We operationalize surname diversity using Shannon entropy. To motivate this measure and conceptually link it to recombinative innovations, consider a subpopulation consisting of N individuals partitioned into K surname groups, each of size N_k such that $\sum_{k=1}^{k=K} N_k = N$. Each group carries unique knowledge, labeled as $s_k \in \{a, b, c, d, ...\}$ and $s_k \neq s_h$ for all $k \neq h$. When individuals from different surname groups meet, the likelihood of recombinative innovation increases. Information theory (Shannon, 1948) tells us that the average informational content (or the innovation potential) of such a population in which people randomly meet is

$$E = -\sum p_k \log_2 p_k \tag{1}$$

where $p_k = \frac{n_k}{N}$ is the probability that a person with group affiliation k is drawn and $\log_2 p_k$ is the informational content embedded in this individual (expressed in bits).

Shannon entropy is a central concept in information theory and is widely used in many scientific disciplines. The term $-\log_2 p_k$ is the self-information of subgroup k and captures the level of surprise (or the informational content of a specific outcome). The negative log reflects that rarely-encountered groups carry more surprise (or more information) compared to more frequent encounters. To arrive at Shannon entropy, the self-information

is weighted by the probability of its occurrence and summed over all possible outcomes. For example, if the population only consists of one group k, the outcome of a draw is entirely predictable, resulting in an entropy of 0, and thus, no recombinations can arise through social interaction. In contrast, entropy for a population with a fixed number of groups is maximized if all groups are of equal size. In such a population, the average individual engaging in a random social interactions is most likely to observe someone different from themselves. A random draw will thus have more informational content (in expectation), which is reflected by higher entropy.

In economics, a fractionalization index is usually used to capture diversity. For our purposes, however, Shannon's entropy offers the conceptually clearest way to represent information diversity and has favorable mathematical properties (Carcassi et al., 2021). In particular, the fractionalization underweights the importance of rare surnames (p_k vs. $\log_2 p_k$), i.e., under the assumption that surnames carry unique pieces of information, rare surnames are more "valuable"—they carry a higher expected surprise. Empirically, the two measures are highly correlated in our setting ($\rho = 0.87$, Table B5).

Figure B2 displays surname diversity across U.S. counties in the year 1940, presenting both the raw data and the residual variation that is orthogonal to log population size.[2] Clear geographical patterns emerge from the data. Counties in California and most of the Northeast exhibit high surname diversity, whereas Utah and the Southern states show considerably lower diversity, indicating greater homogeneity in surnames. Some of these differences had persisted for nearly a century by 1940, as illustrated in Figure B4. The figure also indicates that surname diversity increased across all US census regions between 1850 and 1940. Table B3 reports surname diversity over time of the counties with the highest and lowest surname diversity in 1940.

Finally, we estimate the most likely ancestral origin country for each surname, denoted as $o(k)$, to account for ancestral-country specific factors in our empirical design and develop a measure of ancestral-country diversity. We achieve this by drawing on the universe of immigrants to the U.S. from 1850 to 1940 and identifying the ancestral country for each surname as the one where the majority of immigrants bearing that name were born. Our methodology is detailed in Appendix A. We achieve an accuracy rate of 68% among immigrants whose country of birth is recorded. As reported in Table B6, there is a moderate correlation between ancestral-country diversity and surname diversity, with conditional correlation coefficients ranging from 0.26 to 0.41.

[2]Figure B3 illustrates the variation in surname diversity conditional on log county population from 1850 to 1930.

2.2 Innovation

To measure innovation, we rely on patent data from the Comprehensive Universe of U.S. Patents (CUSP) data set compiled by Berkes (2018).[3] For each patent, the data set provides inventor names and location of residence, filing and issuing years of patents, and the U.S. Patent and Trademark Office technology classifications. Using this data, we construct two kinds of innovation outcomes, one based on the number of patents and the second based on 'breakthrough' patents, which we call 'breakthroughs'. The latter was developed by Kelly et al. (2021) and identifies patents that are both highly novel and impactful based on their low textual similarity to earlier patents but high similarity to later ones. We use this measure of patent novelty and impact rather than the number of citations an individual patent has received because the U.S. Patent and Trademark Office did not consistently begin to record patent citations until after 1947.

We construct innovation outcome variables at four levels of aggregation: (i) county, (ii) surname-county, (iii) inventor, and (iv) patent. The county-level outcomes measure the number of (breakthrough) patents filed by inventors residing in county i during period t, normalized by the county population.[4] For patents filed by multiple inventors, possibly from different counties, we divide the patent count by the number of inventors. The county-surname-level outcomes measure the same quantities but are aggregated and normalized at the level of surnames within counties. The inventor-level outcomes measure the number of patents and breakthroughs filed by inventor j during period t. Finally, the patent-level outcomes are an indicator equal to one if a patent was a breakthrough patent and zero otherwise, as well as the number of technology codes on the patent. The latter measures the complexity of the patent and has been interpreted as a metric of recombination. Details on our patent-based measures can be found in Appendix A.

[3] Although patents have been widely used in economics and other disciplines to study innovation, important concerns remain (Griliches, 1990; Moser, 2013; Lerner and Seru, 2022). These include the fact that many innovations are not patented, industries have variable patenting tendencies, types of inventions have different patentability, and increased patenting in a specific technology category could inhibit innovation rates. Seeking to address these concerns, prior work has demonstrated that results using patents per capita parallel those using alternative measures, including using patent citations (Terry et al., 2024; Acemoglu et al., 2016), patents with novel technology codes (Lerner and Wulf, 2007), the presence of 'creative' (Gomez-Lievano et al., 2017) or 'supercreative' (Bettencourt et al., 2007) occupations, exhibits and prizes at World Fairs (Dowey, 2017; Moser, 2013; Squicciarini and Voigtländer, 2015) and economic productivity (e.g., Alesina et al., 2016; Sequeira et al., 2020; Terry et al., 2024). Overall, while far from perfect, current evidence suggests that patents offer a valuable proxy for innovative activity.

[4] In our main analysis of the causal impact of surname diversity on innovation, we follow Terry et al. (2024) in normalizing patent outcomes by start-of-panel population values, i.e., population in 1900. This choice accounts for the likelihood that population growth, potentially endogenous due to innovative regions attracting more people, could impact the results. In the OLS estimation of the relationship between diversity and patents back until 1850, we normalize the outcomes by contemporaneous county population.

3 Constructing an Instrument for Surname Diversity

We will estimate the effect of surname diversity on innovation at four levels: (i) counties, (ii) surname-counties, (iii) inventors, and (iv) patents. For example, we will use the following county-level equation:

$$Y_i^t = \beta_1 \text{ Surname diversity}_i^t + \beta_2 \text{ Population}_i^t + \alpha_{s(i),t} + \alpha_i + \varepsilon_i^t \tag{2}$$

where $\text{Surname diversity}_i^t$ denotes the surname diversity in county i in period t. Y_i^t represents the outcome of interest, typically the number of patents or breakthrough patents filed in county i in the five-year period starting in year t, normalized by the county's population. The coefficient β_1 is our main interest. Population_i^t denotes county-level population, which we include to control for scale effects. The parameters $\alpha_{s(i),t}$ and α_i are state-period and county fixed effects, respectively. These fixed effects enable the estimation of β_1 from changes within the same county over time, while controlling for both persistent and time-varying differences across states. The error term is denoted by ε_i^t. In our most restrictive model specifications, we include county-specific linear trends, represented by $\alpha_i \times t$. These trends control for any inherent county-specific trend in Y_i^t, allowing us to focus only on the variation in the growth rate of patenting over time within each county.

The main concern with the OLS estimate of β_1 is that reverse causality or unobserved factors that co-determine surname diversity and innovation might bias the estimates. For instance, a highly innovative county may attract a more diverse set of migrants, which would increase surname diversity (Abramitzky and Boustan, 2017). Similarly, highly-skilled individuals, who may be more prone to innovate, might prefer more diverse counties. In both instances, a positive correlation between surname diversity and innovation could be observed even if no causal relationship exists. On the other hand, some historical evidence suggests that immigrants tended to settle in lower-income (i.e., less innovative) areas during this period, as discussed by Sequeira et al. (2020). This possibility would bias our estimates downward.

Therefore, we employ an instrumental variable (IV) strategy to estimate the causal effect of surname diversity on innovation. We observe that migration, along with births, deaths and marriages, plays a significant role in changing counties' surname diversity. Thus, we leverage an IV approach from the immigration literature that allows us to isolate quasi-exogenous variation in surname diversity and to estimate the local average treatment effect (LATE) of immigration-induced changes in surname diversity on innovation as opposed to changes stemming from births, deaths, marriage, or domestic migration.

It is important to recognize that immigration does not uniformly increase surname diversity: its impact depends on the existing local surname distribution. For example, if families with the surname "Smith" immigrate to a county with few or no residents named "Smith," surname diversity increases. Conversely, if they move to a county where the "Smiths" are already common, diversity decreases. Table B4 underscores this observation by listing counties with the highest and lowest immigrant shares in 1900. Notably, counties with minimal immigrant presence (less than 1%), such as Harris County, Georgia (surname diversity: 8.55) and Jackson County, Texas (8.37), exhibit surname diversity comparable to counties with high immigrant shares (over 50%), like Logan County, North Dakota (7.5) and Webb County, Texas (8.56). These findings suggest that the mere presence of immigrants does not necessarily drive surname diversity. Consistent with this, controlling for immigration does not significantly alter our estimates (see Section 4.4).

To construct the instrument for surname diversity, we adapt the historical push-pull approach of Burchardi et al. (2019) to predict the number of residents with surname k in county i in period t, $N^t_{k,i}$. This approach isolates quasi-random variation in surname stocks across counties by predicting these stocks using historical migration patterns. The instrument for surname diversity is then computed by calculating entropy based on these predicted surname stocks.

To understand how the push-pull interaction terms affect surname stocks and diversity, consider a historical example adapted from Bursztyn et al. (2024). Beginning in the 1910s, many "Garcias" (Mexican individuals) immigrated due to the Mexican Revolution, representing a substantial "push". Meanwhile, Detroit (Wayne County, MI) attracted large numbers of immigrants because of the booming automobile industry—a large "pull" for Detroit.[5] The push-pull interaction increased the stock of Garcias in Detroit from the early 1910s (Garcia push 1915 × Detroit pull 1915), which contributed to greater surname diversity, as relatively few Garcias lived in Detroit before the 1910s. In contrast, after World War I, the "Johnsons," "Smiths," "Andersons," "Browns," and "Millers" once again became the most common immigrants, a "push" from British surnames. The push-pull interaction thus increased the stock of these surnames in Detroit from the early 1920s (e.g., Brown push 1925 × Detroit pull 1925). However, these increases did not necessarily contribute to greater diversity, as many "Johnsons," "Smiths," "Andersons," "Browns," and "Millers" were already present.

[5]Figure B5 shows how the timing of migrations varies for common surnames, highlighting the spike in "Garcias" and other Mexican surnames during the 1910s and 1920s. Figure B6 indicates that US counties became attractive to immigrants at various points in time; for instance, Detroit (Wayne, MI) gained appeal starting in the mid-1910s.

Step 1: Predicting Surname Stocks Burchardi et al. (2019) show that a reduced form model of immigration driven by push (such as economic or political conditions in the immigrants' origin countries) and pull shocks (like economic opportunities in the destination counties), combined with a leave-out strategy, can identify variation in ancestry shares that is plausibly exogenous to local factors and ancestral country-destination county specific factors.

We tailor this approach to surnames. Formally, our model is given by:

$$N_{k,i}^{t} = \delta_{o(k),i} + \delta_{k,r(i)} + \sum_{\tau=1880}^{t-1} a^{\tau} \underbrace{I_{k,-r(i)}^{\tau}}_{\text{Push}} \underbrace{\frac{I_{i,-o(k)}^{\tau}}{I_{-o(k)}^{\tau}}}_{\text{Pull}} + \sum_{\tau=1880}^{t-1} d^{\tau} \frac{I_{i}^{\tau}}{I^{\tau}} + u_{i,k}^{t}, \tag{3}$$

where $N_{k,i}^{t}$ denotes the stock of people (in thousands) with surname k residing in county i in year t; $I_{k,-r(i)}^{\tau}$ is the push factor, given by the total number of immigrants (in thousands) with surname k who arrive in the U.S. in the period ending in year τ (1880, 1895, 1900, 1905, 1910, 1915, 1920, 1925, 1930) and settle *outside* the region containing county i. The variable $\frac{I_{i,-o(k)}^{\tau}}{I_{-o(k)}^{\tau}}$ is the pull factor, which captures the relative attractiveness of a specific county i during the period ending in τ. It is given by the share of immigrants settling in county i, excluding immigrants who likely are from the same origin country as those with surname k. $\delta_{o(k),i}$, denotes origin country-county fixed effects, removing any time-invariant factors that make specific counties more attractive to immigrants from the origin country that is most strongly associated with individuals with surname k (e.g., Swedish immigrants tended to settle in counties in Minnesota). $\delta_{k,r(i)}$ are surname-region fixed effects. They remove time-invariant unobserved factors that may make specific census regions (i.e., one of the nine census divisions defined by the Census Bureau) more attractive to migrants with certain surnames. Our surname setting enables us to also include $\sum_{\tau=1880}^{t-1} d^{\tau} \frac{I_{i}^{\tau}}{I^{\tau}}$, i.e., the relative share of migrants who settle in a county in each period τ. This term captures the time-varying relative attractiveness of a county in the past. It isolates the push-pull instruments from county-level conditions that drew migrants in each period τ up to $t-1$, which may still affect innovation in period t.

Like Burchardi et al. (2019), we adopt a leave-out strategy. That is, we exclude immigrants settling in the census region r where county i is located from the push factor, denoted by $-r(i)$, and we remove immigrants with surnames associated with the same origin country o as those with surname k from the pull factor, denoted by $-o(k)$. This approach ensures that our estimates are not driven by confounding factors that vary at the origin country-destination county level such as geographic similarity of the two regions.

For instance, consider migrants from Denmark, a leading dairy producer at the time, who settled in areas suitable for dairy farming. If these Danes and their descendants continue to innovate in the dairy industry, and this innovation pulls more Danes to dairy counties in numbers large enough to affect aggregate flows from Denmark, then failing to exclude Danish migrants when constructing the pull factor for, say, the "Jensens" may confound the analysis. By removing immigrants from the same originating country when forming the pull factor, we address concerns that such confounding factors bias our estimates.

From this estimation, we derive predicted surname stocks

$$\hat{N}_{k,i}^{t} = \sum_{\tau=1880}^{t-1} \hat{b}^{\tau} \left(I_{k,-r(i)}^{\tau} \frac{I_{i,-o(k)}^{\tau}}{I_{-o(k)}^{\tau}} \right)^{\perp} \tag{4}$$

where $\hat{b}^{\tau}$ are the coefficients estimated from equation (3) and $\left(I_{k,-r(i)}^{\tau} \frac{I_{i,-o(k)}^{\tau}}{I_{-o(k)}^{\tau}} \right)^{\perp}$ are the residuals of a regression of the push-pull interactions on all the controls in (3), which ensures that the instruments are orthogonal to these controls.

Step 2: Constructing the Instrument for Surname Diversity We calculate the instrument for surname diversity using the predicted surname stocks $\hat{N}_{k,i}^{t}$:

$$\widehat{\text{Surname diversity}}_{i}^{t} = -\sum_{k} \left(\frac{\hat{N}_{k,i}^{t}}{\sum_{k} \hat{N}_{k,i}^{t}} \log_2 \left(\frac{\hat{N}_{k,i}^{t}}{\sum_{k} \hat{N}_{k,i}^{t}} \right) \right). \tag{5}$$

The data allow us to construct the instrument for surname diversity for eight periods: 1900, 1905, 1910, 1915, 1920, 1925, 1930 and 1940. These periods will be part of our main causal analysis.

Equation (2) controls for the population size of counties. Therefore, we also construct an instrument for population size by summing up the predicted stocks in each county for each period:

$$\widehat{\text{Population}}_{i}^{t} = \sum_{k} \hat{N}_{k,i}^{t}. \tag{6}$$

Identifying Assumption The identifying assumption is given by:

$$I_{k,-r(i)}^{\tau} \frac{I_{i,-o(k)}^{\tau}}{I_{-o(k)}^{\tau}} \perp \varepsilon_{i}^{t} \quad \forall k, \tau < t. \tag{7}$$

It requires that any factor affecting a temporary change in a county's innovation in t (ε_i^t) does not systematically correlate with earlier immigration of individuals with a given surname k to other regions within the US ($I_{k,-r(i)}$) interacted with immigrants form different origin countries than k simultaneously settling in that destination county i ($I_{i,-o(k)}/I_{-o(k)}$). If this condition is satisfied, the predicted stocks of surnames used to instrument for surname diversity are exogenous to innovation, and so is the instrument for surname diversity.[6]

Predictive Power of the Instrument Table 1 reports the estimates of equation (3) for each period from 1900 to 1940. Following Burchardi et al. (2019), to facilitate the interpretation of coefficients as the marginal effect of migrations in that period, we sequentially orthogonalize each instrument with respect to the previous instruments. The push-pull instruments predict surname stocks well, with R^2 values ranging from 0.58 to 0.62 in columns without any controls. The F-statistic for the Wald test of the joint nullity of the push-pull instruments is highly significant in all periods. The coefficients on the instruments hardly change when we add the fixed effects and past immigration controls. The results indicate that push-pull instruments from most periods are significant predictors of the stock of surnames in subsequent years. For example, the estimates reported in Column 16 suggest that most of the instruments from 1880 to 1930 are significant predictors of the stock of surnames in 1940.[7]

Using the predicted surname stocks across counties and periods derived from estimating equation (3), we construct the instruments for surname diversity, equation (5), and population, equation (6). Figure B9 provides maps displaying this instrument,

[6]Our approach differs from the standard shift-share approach discussed in the literature. Typically, shift-share instruments are used as excluded instruments in a two-stage least squares regression, with identification achieved either by a large number of uncorrelated shifts (Borusyak et al., 2022) or by exogenous shares (Goldsmith-Pinkham et al., 2020). In contrast, we use a series of shift-share terms to predict stocks of surnames across U.S. counties. We leverage how the timing of historic waves of immigration (push) coincided with the timing of county-level attractiveness to immigrants (pull)—an arrangement that can be viewed as a series of historical quasi-experiments. We then use these predicted stocks to construct predicted surname diversity, which serves as an instrument for realized surname diversity in a new regression equation with a different set of controls. Despite these differences, our approach shares key features with the "many uncorrelated shifts" approach. For example, the number of surnames is very large (e.g., 454,566 in the 1940 regression, Columns 15-16 of Table 1). The fixed effects in our specification ensure that the push-pull instruments leverage only the residual variation in the shifts after absorbing time-invariant county-origin country and surname-destination region factors. The leave-out strategy removes immigration from a surname's origin-country, which help mitigate issues related to clustered (by origin country) and serially correlated shifts. Finally, including the full history of county immigration share controls approximates controlling for the sum of shares.

[7]Qualitatively, our results parallel those of Burchardi et al. (2019), who estimate the push-pull factor at the level of originating countries (not surnames).

$\widehat{\text{Surname diversity}}_i^t$ in equation (5), for each 5-year period from 1900 to 1940, net of county and state-period fixed effects. Figure B10 displays the same variation but additionally nets out county-specific trends. The maps show substantial variation in the instrument over time and across counties.

4 County-level Surname Diversity and Innovation

Using the quasi-exogenous variation captured by our instrument, we provide estimates of the causal effect of county-level surname diversity on innovation at three levels—counties, surname-counties and inventors—and then assess the robustness of the results. We begin with our county-level findings.

4.1 County-level Estimates

Table 2 reports our main county-level findings. Our baseline IV estimates of β_1 in equation (2) are in Columns 2 and 5 of Panel C, respectively. They estimate the impact of surname diversity in year t on the number of patents and breakthrough patents filed over the subsequent five years relative to other counties in the same state. The coefficients are positive and statistically significant for both patents and breakthrough patents. They suggest that a one standard deviation increase in a county's surname diversity increases the flow of patents by 1.13 per 1,000 people, a 107% increase with respect to the sample mean (1.05). The estimated effect on breakthrough patents is 0.104 per 1,000 people, an 80% increase with respect to the sample mean (0.13). The difference in surname diversity between average entropy in 1940 and Loving, TX, the county with the lowest surname diversity (see Table B3), implies $1.287 \times (9.71 - 6.01) = 4.762$ per 1,000 residents more patents and $0.115 \times (9.71 - 6.01) = 0.426$ per 1,000 residents more breakthroughs, corresponding to roughly 129 additional patents and 12 additional breakthroughs.[8] When we compare these coefficients to those of population size, we find that the coefficient of surname diversity is about 1.76 times that of population size for patents and about 0.76 times for breakthrough patents. This suggests that the causal effects of scale (based on county population size) and diversity (based on surnames) are of the same order of magnitude.[9]

[8] The coefficients are taken from unreported estimates of regressions that use unstandardized surname diversity as independent variables.

[9] We obtain similar magnitudes when estimating Poisson regressions. For patents, the estimated coefficient is 0.7641, which implies a treatment effect of $\exp(0.7641) - 1 = 1.15$, suggesting a 115% increase; for breakthrough patents, the estimated coefficient is 0.7801, implying a treatment effect of $\exp(0.7801) - 1 = 1.18$,

Table 1: Surname-county level estimates of the effects of push-pull instruments on surname stocks

	No. of people in county i with surname k in year:															
	1900		1905		1910		1915		1920		1925		1930		1940	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
$I^{1880}_{k,-r(i)} \times \frac{I^{1880}_{-o(k),i}}{I^{1880}_{-o(k)}}$	3.263***	3.151***	3.271***	3.140***	3.812***	3.639***	3.846***	3.668***	4.257***	4.058***	4.418***	4.182***	4.889***	4.596***	5.016***	4.704***
	(0.104)	(0.106)	(0.112)	(0.112)	(0.117)	(0.117)	(0.097)	(0.097)	(0.106)	(0.106)	(0.125)	(0.126)	(0.102)	(0.099)	(0.098)	(0.095)
$I^{1895}_{k,-r(i)} \times \frac{I^{1895}_{-o(k),i}}{I^{1895}_{-o(k)}}$	0.722**	0.931***	1.081***	1.289***	1.443***	1.652***	1.805***	2.033***	2.038***	2.299***	2.869***	3.183***	3.183***	3.539***	3.245***	3.640***
	(0.295)	(0.289)	(0.359)	(0.358)	(0.315)	(0.311)	(0.174)	(0.170)	(0.188)	(0.184)	(0.201)	(0.195)	(0.248)	(0.225)	(0.314)	(0.293)
$I^{1900}_{k,-r(i)} \times \frac{I^{1900}_{-o(k),i}}{I^{1900}_{-o(k)}}$			-0.516	-1.381	2.135	0.952	-0.160	-0.937	1.199	0.429	-4.712	-5.265*	-3.889	-4.545	-2.471	-2.631
			(3.166)	(3.111)	(3.616)	(3.550)	(2.855)	(2.699)	(3.101)	(2.915)	(3.275)	(3.090)	(3.505)	(3.309)	(3.207)	(2.980)
$I^{1905}_{k,-r(i)} \times \frac{I^{1905}_{-o(k),i}}{I^{1905}_{-o(k)}}$					14.507***	15.342***	18.806***	19.327***	21.604***	22.245***	27.458***	27.682***	30.517***	30.812***	33.261***	33.180***
					(0.941)	(0.996)	(1.125)	(1.159)	(1.306)	(1.335)	(1.384)	(1.431)	(1.171)	(1.120)	(1.094)	(1.058)
$I^{1910}_{k,-r(i)} \times \frac{I^{1910}_{-o(k),i}}{I^{1910}_{-o(k)}}$							18.276***	17.835***	20.650***	20.248***	27.962***	27.438***	31.176***	30.101***	33.057***	31.915***
							(3.054)	(3.050)	(3.365)	(3.356)	(3.468)	(3.442)	(3.365)	(3.365)	(2.942)	(2.967)
$I^{1915}_{k,-r(i)} \times \frac{I^{1915}_{-o(k),i}}{I^{1915}_{-o(k)}}$									8.119***	8.409***	14.490***	14.751***	16.782***	16.942***	19.092***	19.215***
									(1.488)	(1.415)	(1.965)	(1.873)	(1.432)	(1.315)	(1.611)	(1.493)
$I^{1920}_{k,-r(i)} \times \frac{I^{1920}_{-o(k),i}}{I^{1920}_{-o(k)}}$											-1.490	-1.515	-1.018	-0.982	2.825*	2.644*
											(2.056)	(1.972)	(1.085)	(1.004)	(1.490)	(1.377)
$I^{1925}_{k,-r(i)} \times \frac{I^{1925}_{-o(k),i}}{I^{1925}_{-o(k)}}$													26.776***	27.354***	32.424***	32.829***
													(1.214)	(1.286)	(1.499)	(1.564)
$I^{1930}_{k,-r(i)} \times \frac{I^{1930}_{-o(k),i}}{I^{1930}_{-o(k)}}$															-34.025***	-29.284***
															(3.464)	(3.320)
Constant	0.010***		0.008***		0.007***		0.004***		0.003***		0.001		-0.002***		0.000	
	(0.000)		(0.000)		(0.000)		(0.000)		(0.001)		(0.001)		(0.000)		(0.000)	
F-stat	824	942	547	607	516	590	499	500	506	517	349	337	620	640	698	681
Observations	5,346,601	5,346,601	6,604,833	6,604,833	7,012,154	7,012,154	7,284,327	7,284,327	7,364,850	7,364,850	8,067,190	8,067,190	8,143,606	8,143,606	8,980,212	8,980,212
R^2	0.604	0.705	0.581	0.681	0.600	0.695	0.604	0.692	0.607	0.696	0.578	0.662	0.624	0.705	0.613	0.695
Origin country-County fixed effects		✓		✓		✓		✓		✓		✓		✓		✓
Surname-Census division fixed effects		✓		✓		✓		✓		✓		✓		✓		✓
I^{τ}_{i}/I^{τ} controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: This table reports OLS estimates for the specification described in equation (3), corresponding to step 1 of the instrument construction. An observation is a surname-county in a period from 1900 to 1940. The table includes the F-statistic for the Wald test for the joined nullity of the push-pull instruments. Standard errors clustered at the surname level. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table 2: County-level estimates of the effect of surname diversity on innovation

	Patents per 1,000 people (mean = 1.05, sd = 1.73)				Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.322***	1.125***	1.081***	1.338***	0.148***	0.107***	0.099***	0.099***
	(0.169)	(0.202)	(0.263)	(0.174)	(0.016)	(0.015)	(0.017)	(0.018)
Population		0.674***	0.645***	0.736***		0.139***	0.125***	0.099***
		(0.106)	(0.106)	(0.189)		(0.020)	(0.018)	(0.029)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.956***	0.698***	0.667**	0.939***	0.111***	0.046**	0.043*	0.061***
	(0.179)	(0.235)	(0.283)	(0.236)	(0.014)	(0.021)	(0.024)	(0.019)
Population (IV)		0.546***	0.521***	0.205		0.138***	0.122***	0.051
		(0.147)	(0.161)	(0.205)		(0.026)	(0.025)	(0.035)
Panel C: Instrumental-variable estimates								
Surname diversity	1.388***	1.161***	1.128***	1.392***	0.161***	0.104***	0.104***	0.101***
	(0.171)	(0.205)	(0.262)	(0.191)	(0.018)	(0.018)	(0.019)	(0.019)
Population		0.664***	0.640***	0.731***		0.166***	0.137***	0.118**
		(0.137)	(0.120)	(0.236)		(0.028)	(0.021)	(0.054)
Sanderson-Windmeijer *F*-stat	129	162; 426	147; 227	149; 29	129	162; 426	147; 227	149; 29
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.689***	0.693***	0.723***	0.729***		-0.160***	-0.231***	-0.103**
	(0.061)	(0.075)	(0.073)	(0.074)		(0.033)	(0.042)	(0.041)
Population (IV)		-0.009	-0.076*	-0.142**		0.837***	0.948***	0.550***
		(0.041)	(0.040)	(0.063)		(0.053)	(0.073)	(0.107)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (2) as well as IV first-stage estimates. An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in t. In Columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In Columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

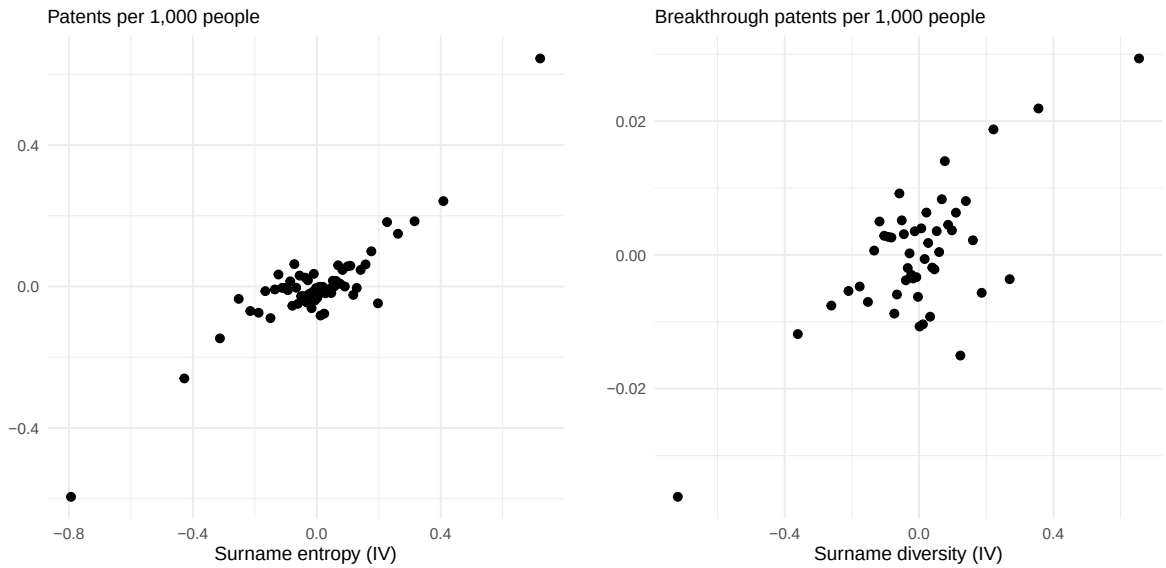


Figure 1: Reduced-from relationships of surname diversity (IV) and innovation outcomes
Notes: County-level data from 1900 to 1940. Observations are residualized by population (IV), county fixed effects and state-period fixed effects.

Panels A–C show our OLS, reduced-form and IV estimates, respectively. Panel D reports the first stage of the IV. Across specifications, the IV estimates are stable and similar to the OLS estimates, suggesting biases cancel out. These estimates hold even when we control for county-specific time trends, which allows us to focus on variation in the growth rate of patents within the same counties over time. Similarly, across specifications, the first stage has a strong F-statistic for both surname diversity and population size.

To visualize these effects, Figure 1 plots the reduced-form relationships between the instrument for surname diversity and both patent outcomes. They show that these relationships are strong, approximately linear and not driven by a small set of observations. Figure B11 shows that these relationships hold—even sharpen—when we additionally control for county-specific trends. Figure B12 depicts the first-stage relationship for our IV.

While data for our IV approach is only available from 1900 to 1940, Figure B13 presents OLS estimates for each decade extending back to 1850, the earliest census wave for which surname data are available. By regressing both patents and breakthroughs per capita on surname diversity (interacted with decadal dummies) while controlling for population size and including fixed effects for counties and state-periods, the figure shows the stability of this relationship across time. Beginning in the mid-19th century, we consistently find large, positive, and remarkably stable coefficients over time for both patents and breakthroughs.

or a 118% increase.

4.2 County-surname-level Estimates

A potential concern with interpreting our findings so far is that surname-specific traits, such as abilities, interests, or knowledge, might drive innovation rather than the diversity of these traits. Clark (2014) and Barone and Mocetti (2021), for example, find that rare surnames are proxies for the vertical transmission of traits, and these traits might affect innovation.

We address this concern by estimating the following county-surname-level specification:

$$Y_{i,k}^{t} = \beta_1 \text{ Surname diversity}_i^t + \beta_2 \text{ Population}_i^t + \alpha_{i,k} + \alpha_{s(i),t} + \alpha_{k,t} + \varepsilon_{i,k}^{t} \tag{8}$$

where, as before, Surname diversity$_i^t$ and Population$_i^t$ are county i's surname diversity and population size in year t. $Y_{i,k}^{t}$ is now the number of patents or breakthroughs filed by people with surname k residing in county i in the five-year period starting in t (in terms of 1,000 residents with the same surname in the year 1900). For example, 18,351 individuals with the surname 'Johnson' resided in Cook County (IL) in 1900 and filed 69 patents and 1 breakthrough patent between 1900 and 1904. Therefore, while the surname diversity remains defined at the county-period level, the innovation outcomes vary at the surname-county-period level.[10]

Crucially, we can now include county-surname fixed effects as well as surname-period fixed effects (denoted by $\alpha_{i,k}$ and $\alpha_{k,t}$, respectively), which absorb any traits specific to individuals with a particular surname in a given county or time period (e.g., traits specific to all 'Johnsons' in Cook County or in 1940). The remaining parameters and variables are as in equations (2) and the coefficient of interest remains β_1. Standard errors are clustered on states and each observation is weighted by its relative share of the county's population, making these estimates comparable to the county-level estimates in Table 2.

Table 3 reports our main surname-county level estimates. We find positive and statistically significant effects with effect sizes comparable to those in Table 2. In our most stringent specifications (columns 4 and 8), the estimates for β_1 indicate that a one standard deviation increase in surname diversity results in a 133% increase in patents (relative to the sample mean) and a 91% increase in breakthroughs. By comparison, a standard deviation increase in the size of a county's population results in increases of 145% and 170% for patents and breakthroughs, respectively. Crucially, the surname-fixed effects ensure that the estimates are orthogonal to any unobserved surname-specific characteristics.

[10] Here, we switch the unit of analysis, resulting in an increase in the number of observations. For each county-period, the number of observations corresponds to the number of surnames in that county during that time period. We focus on surname-county observations that are balanced over the full panel from 1900 to 1940. See Appendix A for details on how we constructed the sample.

Table 3: County-surname-level estimates of the effect of surname diversity on innovation

	Patents per 1,000 people (mean = 1.02, sd = 13.31)				Breakthrough patents per 1,000 people (mean = 0.10, sd = 2.23)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.173***	0.930***	0.948***	1.143***	0.128***	0.087***	0.088***	0.061***
	(0.194)	(0.174)	(0.178)	(0.182)	(0.031)	(0.026)	(0.027)	(0.015)
Population		3.702***	3.720***	3.679***		0.617***	0.624***	0.728***
		(0.600)	(0.613)	(0.597)		(0.107)	(0.108)	(0.100)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.742***	0.521***	0.530***	0.675***	0.080***	0.046*	0.046*	0.044
	(0.160)	(0.177)	(0.182)	(0.161)	(0.021)	(0.024)	(0.025)	(0.030)
Population (IV)		3.072***	3.089***	0.958		0.472***	0.479***	0.112
		(0.946)	(0.971)	(0.971)		(0.155)	(0.158)	(0.280)
Panel C: Instrumental-variable estimates								
Surname diversity	1.382***	1.207***	1.227***	1.358***	0.149***	0.122***	0.123***	0.091**
	(0.251)	(0.245)	(0.252)	(0.248)	(0.035)	(0.035)	(0.036)	(0.043)
Population		3.060***	3.071***	1.478		0.474***	0.480***	0.170
		(0.551)	(0.568)	(1.509)		(0.100)	(0.102)	(0.424)
Sanderson-Windmeijer *F*-stat	178	201; 50	201; 49	215; 300	178	201; 50	201; 49	215; 300
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.537***	0.532***	0.532***	0.520***		-0.040**	-0.040**	-0.021***
	(0.040)	(0.041)	(0.041)	(0.046)		(0.016)	(0.016)	(0.004)
Population (IV)		0.058*	0.060*	-0.028		0.981***	0.982***	0.674***
		(0.030)	(0.031)	(0.059)		(0.142)	(0.144)	(0.049)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (8) as well as IV first-stage estimates. An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname entropy and population in t. In Columns 1 to 4, the dependent variable is number of patents filed by individuals with surname n residing in county in i in the five-year period starting in t divided by surname population size in 1900. In Columns 5 to 8, the dependent variable is number of breakthrough patents filed by individuals with surname n residing in county in i in the five-year period starting in t divided by surname population size in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Intuitively, this suggests that the "Winklers" get more innovative when they are in more culturally diverse places.

4.3 Inventor-level Estimates

To shed light on the impact of surname diversity on innovation at the individual-level, we extend our analysis to the inventor level. The patent data for this era does not include unique inventor IDs. This poses challenges related to identifying inventors due to inconsistencies in name strings and the presence of common names such as "John Smith." Since we only observe the name and the location of residence, this is particularly difficult for mobile inventors. Despite these obstacles, we have assigned unique inventor IDs for approximately 47% of the patents filed between 1900 and 1944 based on name similarities

and their county of residence (see Appendix A for details). This effectively restricts the sample to non-moving inventors. While data on mobile inventors would be useful, these data still allow us to study how the productivity of inventors who remain in the same county changes over time as their county's diversity changes.

Using this dataset, we estimate the following inventor-level specification:

$$Y_j^t = \beta_1 \text{ Surname diversity}_{i(j)}^t + \beta_2 \text{ Population}_{i(j)}^t + \alpha_j + \alpha_{s(i),t} + \varepsilon_j^t, \tag{9}$$

where Y_j^t is the number of patents and breakthroughs filed by inventor j residing in county i during the five-year period starting in t. The specification includes inventor fixed effects (α_j) and state-period fixed effects ($\alpha_{s(i),t}$), which absorb unobserved inventor-specific traits and time-varying factors at the state level. By including inventor fixed effects, we ensure that our results are not influenced by any time-invariant personal characteristics of the inventors. As a result, the regression relies solely on variation arising from immigration-induced, quasi-random changes in the surname diversity of the inventor's county of residence. Standard errors are clustered at the county level.

Table 4 presents our main inventor-level results. Across specifications, we find positive and statistically significant effects of surname diversity on both patents and breakthrough patents. The IV estimates in Panel C indicate that a one standard deviation increase in county-level surname diversity leads inventors to file about half an additional patent during the five-year period starting in t, on average. This increase represents a 20% rise relative to the sample mean (2.24). The effect on breakthrough patents is around one-tenth of an additional breakthrough, roughly a 40% increase relative to the sample mean (0.26). The estimates are very similar when we additionally control for population size, which interestingly does not affect patenting at the inventor level. The first stage has strong F-statistics for both surname diversity and population size.

4.4 Robustness and Extensions

We test the robustness of our findings using nine different approaches at all three levels of analysis. First, to assess the impact of immigration on our findings, we explicitly control for it. Second, as a placebo check, we regress past innovation on future surname diversity. Third, to assess the role of broad geographic heterogeneity, we conduct our main analyses separately in each of the four major U.S. Census regions. Fourth, we consider alternative forms for both our innovation outcomes and key predictor variables. For innovation, we assess the impact of using log transformations. For surname diversity, we evaluate five alternative formulations. Fifth, we probe the robustness of our results by dropping the

Table 4: Inventor-level estimates of the effect of surname diversity on innovation

	Patents (mean = 2.24, sd = 1.69)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity	0.220***	0.219***	0.043*	0.043*
	(0.069)	(0.069)	(0.024)	(0.023)
Population		-0.016		-0.011
		(0.039)		(0.012)
Panel B: Reduced-form estimates				
Surname diversity (IV)	0.077***	0.078***	0.017***	0.016**
	(0.022)	(0.022)	(0.006)	(0.006)
Population (IV)		0.003		-0.008
		(0.038)		(0.010)
Panel C: Instrumental-variable estimates				
Surname diversity	0.463***	0.438***	0.104**	0.094**
	(0.153)	(0.147)	(0.042)	(0.040)
Population		-0.045		-0.019
		(0.044)		(0.012)
Sanderson-Windmeijer *F*-stat	44	49; 691	44	49; 691
Panel D: First-stage estimates				
	Surname diversity		Population	
Surname diversity (IV)	0.167***	0.179***		0.020
	(0.025)	(0.025)		(0.025)
Population (IV)		0.095***		0.867***
		(0.018)		(0.056)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,944	209,944	209,944	209,944

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9) as well as IV first-stage estimates. An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in year t, and the dependent variables are number of patents filed by the inventor in the five-year period starting in t (Columns 1–2) and number of breakthrough patents filed by the inventor in the five-year period starting in t (Columns 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

controls from our zero stage. Sixth, we drop the most surname diverse counties from the sample. Seventh, we explore the role of non-linear effects in surname diversity. Eighth, we assess the potential impact of schooling on our findings. Finally, we examine robustness to incorporating more conventional measures of cultural diversity.

4.4.1 Sensitivity to Immigration

Because our instrument is constructed from immigration data, a potential concern is that immigration, operating through channels not related to either population size or surname diversity, may influence patenting. Previous research suggests that immigrants significantly fuel innovation due to factors like higher skills, entrepreneurial spirit or unique patentable knowledge (Moser et al., 2014; Abramitzky and Boustan, 2017; Sequeira et al., 2020; Terry et al., 2024).

We address this concern in two ways. Fundamentally, we orthogonalized the push-pull instruments in equation (3) using counties' historical immigration patterns. Thus, addressing this concern lies at the core of our identification strategy. Here, to confirm the effectiveness of this approach, we additionally control for the number of recent immigrants arriving in each county during each period. Our instrument is based on changes in surname diversity induced by immigration, which depends on the existing surname distribution in a county. Thus, this approach allows us to separate the effects of recent immigration from those of surname diversity.

To address the endogeneity of immigration, we adapt the shift-share approach from Terry et al. (2024) to our context. We predict migration inflow for each surname into each county between t and $t-1$ based on the proportion of people with that surname already residing there in $t-1$. This method estimates each surname's migration inflow into a county by the end of t. To mitigate endogeneity concerns about the surname shares in $t-1$, we use *predicted*, not actual, previous-period surname stocks, derived from the historical push-pull approach. We then estimate total migration into a county by summing these predicted inflows, excluding data from the period ending in 1940 due to the lack of immigration year information necessary for calculating predicted immigration.

Table B7 presents these results for our county-level specification (2). In line with previous work, we find a positive effect of immigration on innovation across specifications. Crucially, for our purposes, we also find that controlling for immigration has little impact on the estimates for surname diversity. For example, the IV estimate for patents per 1,000 people is 1.087 without controlling for immigration (Column 1 of Panel C), compared to 1.025 with this control (Column 2). The estimates for breakthrough patents are 0.083 and 0.072 in Columns 5 and 6, respectively. We observe similar patterns when controlling

for immigration in the county-surname (Table B8) and inventor-level analysis (Table B9). Taken together, these findings suggest that the orthogonalization of our instrumental variable to the influence of immigration worked as expected and confirms that our estimates are not confounded by immigration-related factors other than surname diversity.

4.4.2 Placebo Test and Dynamic Effects

Tables B10, B11 and B12 show the dynamic effects of changes in surname diversity on our innovation outcomes at the county, county-surname and inventor level, respectively. Two findings stand out. First, looking back, we do not observe a positive effect of current surname diversity in period t on past innovation in the previous period, $t-5$ (Columns 1 and 5). This lack of evidence for reverse causality strengthens our confidence in the identification strategy. Second, looking forward, we find that surname diversity impacts innovation not only in period t (0 to 5 years after we observe surnames) but also in period $t+5$ (5 to 10 years after our observation of surnames) at both the county and county-surname levels.

4.4.3 Estimates for Major U.S. Regions

Tables B13, B14 and B15 report the estimates for the four major U.S. census regions: the Midwest, Northeast, South, and West. We observe that across specifications, the region-specific estimates are uniformly positive and mostly statistically significant. At the county level, the range of IV coefficients for patents across regions (Panel C, Column 1) spans from 0.79 to 1.74, and for breakthrough patents (Column 3), from 0.07 to 0.19. The larger coefficients for the West are particularly interesting, possibly reflecting the region's unique historical and immigration context. The estimates for the South further refute the idea that our results are driven by high immigration flows since the South experienced relatively low rates of immigration during this period.

4.4.4 Alternative Forms of Key Variables

We consider the sensitivity of our results to alternative forms of measuring both innovation and surname diversity. For innovation, we justify our decisions to normalize (using population in 1900) and winzorize our patent-based measures on statistical grounds, as detailed in Section 2 and Appendix A. Nevertheless, we examine the effects of using log-transformations (without winsorization) for patents and breakthroughs to address the right skewness present in these measures, and we also consider the impact of removing the normalization. To avoid dropping observations with zeroes, we add 1 to the patent

variables. Tables B16 and B17 show the results for log population-normalized outcomes and Tables B18, B19 and B20 report the estimates for log non-normalized patents and breakthroughs. In sum, our results are robust to these transformations of the dependent variables.[11]

For surname diversity, Section 2 explains our choice of an entropic diversity measure based on conceptual, statistical and parsimony considerations. Nevertheless, it is important to explore how critical this choice is to our results. Conceptually, diversity is composed of three parts: (1) 'surname variety,' or a function of the number of distinct surnames in a county, (2) 'dispersion,' which indicates how individuals are distributed across surnames (the opposite of 'balance' or 'concentration') and (3) 'disparity,' which measures the differences between surnames (e.g., Stirling, 2007). Our measure combines 'variety' and 'dispersion' but does not include a metric to capture the 'disparity' or cultural distance among surnames. That is, our entropy measure treats each surname as an independent category equidistant from other surnames. This means that our approach does not account for the fact that the Smiths and Browns might share more information due to a shared cultural background than the Smiths and Garcias.

To address this, we create a distance-weighted measure of surname diversity by incorporating a proxy for cultural distance based on genetic distance (following Spolaore and Wacziarg, 2009). As explained in Appendix A, we assigned each surname to an ancestral-country and then used the genetic distance between these countries and the U.S. to assign a d value to each surname. Specifically, we compute $-\sum dp \log p$, where d is the genetic distance, our proxy for cultural distance. The resulting weighted measure of surname diversity is highly correlated with our baseline entropy measure ($\rho \approx 0.87$).

Table B21 reports the estimates for our county-level equation (2) using this distance-weighted entropy measure as the main endogenous variable. We find that the distance-weighted entropy measure increases both of our innovation measures across specifications (Columns 1–4 and 6–9). However, when we include our more parsimonious baseline entropic measure of surname diversity (Columns 5 and 10), the coefficients for distance-weighted diversity shrink, lose significance and even switch directions (in 5 out of 6 estimates). We observe similar patterns at the county-surname and inventor level (Tables B22 and B23).

These results suggest that adding a measure of disparity, at least at the level of the ancestral country, does not improve our results. Next, we break our surname diversity measure down into its two components: variety and dispersion. Tables B24, B25 and

[11] For log-transformed breakthroughs, the coefficients at the county-surname level are positive but sometimes imprecise because 99.4% of observations have zero breakthroughs.

B26 report OLS comparisons between our baseline surname diversity and its components. At the county level, we find that when our conceptually-motivated entropy measure of surname diversity is included, the coefficients on two components are small, mostly poorly estimated and sometimes go in the wrong direction. We find broadly similar patterns at the county-surname and inventor levels. Taken together, these results indicate that our entropic surname diversity—the interactions of variety and dispersion—captures the relevant relationship between surname diversity and innovation in an effective way. Neither of its components is the dominant driver of the results, suggesting that a high degree of variety is much more effective in fostering innovation when paired with a high degree dispersion.

Finally, while we prefer our entropic measure for the reasons outlined above, we also report results from using a more traditional measure of diversity in economics: a fractionalization index, calculated as $1 - \sum p$. Tables B27 and B29 demonstrate that the county and inventor-level results are robust to using surname fractionalization rather than surname entropy. Table B28 suggests that while the county-surname-level OLS estimates remain robust, the IV becomes weak, rendering the reduced-form and IV estimates unreliable.

4.4.5 Excluding Controls from Zero-stage Estimation

A potential concern with our empirical strategy arises from the inclusion of controls and fixed effects in the zero-stage regression used to predict surname stocks. Specifically, while our push-pull interaction terms are orthogonalized with respect to these controls, the estimated coefficients α^τ from equation (3) could still be functions of the included controls and fixed effects. This dependency may introduce endogeneity into our predicted surname stocks $N^t_{k,i}$, as these stocks are computed using the estimated coefficients α^τ. Consequently, the exogeneity of our instrument—the predicted surname entropy measure—might be compromised if the coefficients capture effects related to the controls rather than solely the exogenous variation from the push-pull factors.

To explore the importance of this concern, we conduct a robustness check by re-estimating the zero-stage regression without any controls or fixed effects. By removing these additional variables, we ensure that the estimated coefficients α^τ are solely determined by the push-pull interactions. This approach isolates the variation in surname stocks that is attributable only to the shocks captured by the push and pull factors.

The odd-numbered columns in Table 1 report these estimates. As expected, the R^2 statistics drop by about 0.10. However, the coefficients on the push-pull instruments remain stable. We then recalculate the predicted surname stocks $N^t_{k,i}$ using these new coefficients and reconstruct our predicted surname entropy measure. Re-estimating our

county-, surname-county and inventor-level specifications with this alternative instrument, we find that the results remain consistent with our original findings (Tables B30, B31 and B32. The estimated effects are similar in magnitude and statistical significance, suggesting that our conclusions are not driven by the inclusion of controls in the zero-stage regression.

4.4.6 Dropping Most Surname Diverse Counties, Rare and Frequent Surnames

A potential concern may be that outliers drive our results. While all regressions rely on within-state variation, one might be concerned that the results largely reflect the effects of large, diverse urban areas. To assess this possibility, we replicate our estimates in subsamples that drop the 5%, 10% and 20% most diverse counties (measured in 1900). These estimates are reported in Tables B33, B34 and B35. We find that the estimates in these subsamples are very similar to our baseline ones.

A related concern may be that locally rare and frequent surnames may affect our results. To tackle this concern, we recompute the entropy and population measures after excluding the top and bottom 5% of surnames in each county. The results, presented in Tables B36, B37 and B38 are very similar to our main findings, suggesting that our findings are not unduly driven by the inclusion of locally rare and frequent surnames.

4.4.7 Surname Diversity Squared

Some scholars have suggested that cultural diversity may bear a concave relationship with innovation because greater cultural diversity, at a certain point, may induce less social interaction (Schimmelpfennig et al., 2022), thus reducing innovation indirectly via the social-behavioral channel. To address this potential non-linearity, we add a squared term for surname diversity to our regressions at the county, county-surname and inventor level. As reported in Tables B39, B40 and B41, we do not find clear evidence of any non-linear effects of surname diversity on innovation. If anything, the result hints at a convex relationship.

4.4.8 Education

Given that Americans experienced rising levels of formal schooling over our study period, and that education is linked to innovation, we consider how including education and educational diversity in our analysis impacts our key effects. Indeed, while the cultural diversity captured by surname diversity provides a rich source of ideas, more schooling may facilitate the transformation of these ideas into patentable innovations. Notably, both

Alesina et al. (2016) and Terry et al. (2024) show that skill levels of immigrants influence rates of innovation.

To assess the importance of formal education in our setting, we estimate simple OLS specifications based on the cross-section of U.S. counties in the year 1940—the only census wave in our sample that reports years of schooling. Table B42 presents the estimates for specifications that include years-of-schooling diversity (Panel A) and average years of schooling (Panel B). As a benchmark, we first regress our two measures of innovation on surname diversity and population size (Panel A, Columns 1 and 4). Then, in Columns 2 and 5 of Panel A, we regress these innovation variables on years of schooling diversity, and in Columns 3 and 6 of Panel A, we examine the horse race between the two diversity measures.

We find that years of schooling diversity is significantly related to our innovation outcomes; however, the coefficients are about half the size to those on surname diversity. In the horse race, the coefficients on surname diversity remain stable, while the ones on educational diversity shrink by another half.

Next, turning to Panel B, we regress our patent outcomes on average years of schooling in Columns 1 and 4. In Columns 2 and 5, we additionally control for surname diversity, and in Columns 3 and 6 we examine the interaction between surname diversity and years of schooling.

Across all specifications, we find a significant association between surname diversity and innovation. Schooling is also related to patenting; however, its inclusion affects the coefficients on surname diversity to a limited degree. Notably, the interaction between the two independent variables is highly significant. These findings suggest that surname diversity fuels patenting, especially among more highly educated individuals. Indeed, the magnitude of the relationship between schooling and innovation almost doubles with an increase in surname diversity by one standard deviation.

4.4.9 Surname vs. Other Measures of Cultural Diversity

How does our surname diversity measure compare to more conventional indicators of cultural diversity, which have typically been based on birth-countries or even broader categories such as continental origins (e.g., "Asians")? One might argue that surnames are merely a noisy proxy for the cultural diversity captured by ancestral origin countries. To examine this, we conduct two "horse race" analyses that directly comparing surname diversity with alternative diversity measures.

To this end, we draw on the Census data to construct measures of diversity based on birthplace, race, and occupation. We include occupational diversity because, although

occupation is not traditionally considered a form of culture, individuals' occupational choices are influenced by learning from same-sex parents and broader community networks. Additionally, as mentioned in Section 2, we infer the ancestral country for each surname using Census data from 1850 to 1940 to predict the most likely country of origin. This completes our set of four alternative diversity measures.

Table B6 presents the conditional correlations between surname diversity and these alternative measures, controlling for population and various geographical and temporal fixed effects. For instance, ancestral-country diversity is moderately correlated with surname diversity, with conditional correlation coefficients ranging from 0.26 to 0.41, depending on the fixed effects and population control variables included.

To assess the predictive power of surname diversity relative to these alternative measures, we estimate a series of OLS regressions, the results of which are reported in Table 5. Panels A and B show the estimates for patents and breakthroughs, respectively. Replicating Columns 3 and 7 from Table 2, Column 1 establishes a baseline model that includes our county-level surname diversity measure and county population size.

Birth- and Ancestral-country Diversity In Columns 2 and 4, we replace surname diversity with birth-country diversity and ancestral-country diversity, respectively. The standardized coefficients on these alternative diversity measures are positive and statistically significant, though notably smaller than the coefficient on surname diversity in the baseline model (Column 1). This positive relationship aligns with prior research (Ottaviano and Peri, 2006; Ager and Brückner, 2013; Alesina et al., 2016; Docquier et al., 2020; Fulford et al., 2020). In Columns 3 and 5, we reintroduce surname diversity into the models. When surname diversity is included, the relationships between innovation and both birthplace and ancestral-country diversity shrink dramatically and cannot be distinguished from zero at conventional levels of precision. By contrast, the coefficient on surname diversity remains large and robust, comparable to that in the baseline model. If surname diversity were merely a noisy proxy for ancestral or birth country diversity, we would expect its coefficient to attenuate when these alternative measures are included; yet, this is not what we observe.

Racial Diversity Columns 6 and 7 incorporate racial diversity into the analysis. The inclusion of racial diversity does not materially affect the coefficients on surname diversity, which remain stable and statistically significant.

Occupational Diversity In Column 8, we examine the relationship between occupational diversity and innovation. Consistent with our expectations, occupational diversity exhibits a strong positive relationship with innovation, although the coefficient is approximately half the magnitude of that on surname diversity in the baseline model. In Column 9, we add surname diversity to the regression. The coefficient on surname diversity remains substantial and statistically significant, with only a slight reduction compared to the baseline. This finding is consistent with evidence that occupation is one among several traits transmitted across generations.

Immigrant Shares by Birth Country We also consider the population shares of immigrants from each of the 59 birth countries to account for the role of immigration from specific origins. By including these immigrant shares, we aim to control for the potential impact that immigrants from particular countries may have on innovation outcomes. When we include these shares in the regression models in Column 10, we find that the coefficients on surname diversity remain virtually unaffected.

Combined Diversity Measures Finally, Columns 11 and 12 include all diversity measures simultaneously in the model. The coefficients on surname diversity remain stable and significant, similar to that in the baseline model. In contrast, the coefficients on birth country, ancestral-country, and racial diversity become statistically insignificant or negative. Occupational diversity continues to show positive relationships with innovation, remaining significant except in Column 11 but loses significance in Column 12, where county-specific linear trends are included. However, the coefficients on occupational diversity are substantially smaller—ranging from about 40% to 10%—than those on surname diversity.

IV Estimates for Ancestral-Country Diversity To extend our analysis beyond the correlations observed in Table 5, we construct an IV for ancestral-country diversity using the same methodology applied to surname diversity, but substituting inferred ancestral countries for surnames. This approach yields an IV for ancestral-country diversity that parallels our instrument for surname diversity, allowing us to employ a similar IV strategy.

To generate predicted stocks of individuals from each ancestral country, we estimate a variant of equation (3), where the dependent variable is the number of individuals from a specific ancestral country residing in a given U.S. county. The push and pull factors are redefined at the ancestral country–county level. We adjust the leave-out strategy by excluding immigrants from countries neighboring the ancestral country from the pull

Table 5: County-level horse race between surname diversity and other diversities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A:	Patents per 1,000 people											
Surname diversity	1.081***		1.076***		1.134***		1.078***		0.978***	1.099***	1.065***	1.356***
	(0.263)		(0.264)		(0.262)		(0.249)		(0.228)	(0.240)	(0.230)	(0.152)
Birth-country diversity		0.302***	-0.047								-0.042	0.032
		(0.084)	(0.121)								(0.131)	(0.143)
Ancestral-country diversity				0.233***	-0.110**						-0.080*	-0.150**
				(0.083)	(0.054)						(0.044)	(0.056)
Racial diversity						0.007	-0.162				-0.150	-0.470***
						(0.201)	(0.119)				(0.112)	(0.140)
Occupational diversity								0.507***	0.260***		0.259***	0.155
								(0.157)	(0.091)		(0.092)	(0.100)
Population	0.645***	0.853***	0.647***	0.813***	0.652***	0.847***	0.651***	0.888***	0.687***	0.659***	0.697***	0.747***
	(0.106)	(0.111)	(0.107)	(0.109)	(0.107)	(0.110)	(0.107)	(0.121)	(0.110)	(0.096)	(0.100)	(0.189)
R^2	0.810	0.800	0.812	0.798	0.810	0.799	0.812	0.803	0.813	0.814	0.815	0.877
Panel B:	Breakthrough patents per 1,000 people											
Surname diversity	0.099***		0.100***		0.105***		0.102***		0.088***	0.104***	0.095***	0.099***
	(0.017)		(0.019)		(0.019)		(0.018)		(0.019)	(0.018)	(0.021)	(0.020)
Birth-country diversity		0.039***	0.007								0.004	0.038
		(0.013)	(0.014)								(0.015)	(0.026)
Ancestral-country diversity				0.019***	-0.012						-0.008	-0.018**
				(0.007)	(0.009)						(0.008)	(0.008)
Racial diversity						-0.004	-0.020				-0.023	-0.052
						(0.030)	(0.025)				(0.025)	(0.038)
Occupational diversity								0.060***	0.038***		0.037***	0.015
								(0.011)	(0.009)		(0.008)	(0.009)
Population	0.125***	0.144***	0.125***	0.141***	0.126***	0.144***	0.125***	0.149***	0.130***	0.121***	0.127***	0.094***
	(0.018)	(0.018)	(0.018)	(0.018)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.017)	(0.017)	(0.028)
R^2	0.678	0.675	0.679	0.674	0.678	0.675	0.679	0.677	0.680	0.681	0.682	0.755
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Immigrant shares by 59 birth countries										✓	✓	✓
County-specific linear trends												✓
Observations	21,984	21,873	21,873	21,984	21,984	21,873	21,873	21,871	21,871	21,873	21,871	21,871

Notes: The table reports OLS estimates of regressions of innovation outcomes on surname diversity and other diversities. An observation is a county in a period from 1900 to 1940. In Panel A (Panel B), the dependent variable is number of patents and breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered on states and reported in parentheses. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

factor (see Appendix A for details). The estimates from these zero-stage regressions are reported in Table B43. Consistent with our findings for surnames, the push-pull instruments are highly significant predictors of ancestral-country stocks.

Using these predicted ancestral-country stocks across counties and time periods, we construct the IV for ancestral-country diversity. We then re-estimate equation (2), incorporating both surname diversity and ancestral-country diversity as independent variables.

Table B44 presents the results of these IV regressions. The coefficients on surname diversity remain large, stable, and precisely estimated, with magnitudes comparable to those reported in Table 2. In contrast, the coefficients on ancestral-country diversity are small, imprecisely estimated, or even negative in some specifications.

These results support the OLS horse race findings in Table 5, suggesting that surname diversity is not simply an imprecise proxy for ancestral-country diversity. Below, in our discussion of mechanisms, we will delve deeper into the pathways through which these different types of cultural diversity affect innovation.

5 Patent-level Surname Diversity and Innovation

We extend our analysis from the impact of county-level surname diversity on innovation to estimate the effect of patent-level surname diversity on patent quality and complexity.[12] We estimate:

$$Y_p^t = \beta \text{ Surname diversity of patent}_p^t + \alpha_{n(p)} + \alpha_{i(p)} + \alpha_{s(p),t} + \alpha_{f(p),t} + \varepsilon_p^t, \tag{10}$$

where Surname diversity of patent$_p^t$ is the diversity of surnames of the inventors who filed the patent p in the five-year period starting in t. Y_p^t represents either the breakthrough indicator or the number of technology codes assigned to the patent. The latter has been used as a measure of the complexity of a patent and the degree of recombination (Strumsky et al., 2011; Akcigit et al., 2013; Fiszbein, 2022; Youn et al., 2015).[13] The specification includes fixed effects for inventor team size ($\alpha_{n(p)}$) because larger teams often have higher surname diversity and tackle more complex and impactful projects. It includes county ($\alpha_{i(p)}$) and state-period ($\alpha_{s(p),t}$) fixed effects to absorb geographic and time-varying differences in team diversity, patent quality and complexity. It also includes fixed effects for patent technology fields interacted with period fixed effects ($\alpha_{f(p),t}$).[14] Different fields may have varying team compositions, and patents in these fields can differ in impact and complexity, which may change over time. In our most stringent specifications, we include county-period fixed effects ($\alpha_{i(p),t}$), absorbing any county-specific, time-varying unobserved factors. If multiple inventors reside in different counties, the patent is assigned to each county and weighted by $\frac{1}{n}$, where n is the number of co-inventors. Standard errors are clustered at the county level.

To identify the causal effect of patent-level surname diversity on patent outcomes, we instrument for patent-level surname diversity using our IV for county-level surname diversity. This approach allows us to estimate the local treatment effect of *patent* surname diversity on the quality and complexity of innovation for patents whose surname diversity is influenced by county-level variation.

Table 6 reports these estimates. Panels A (OLS), B (reduced-form) and C (IV) indicate that greater surname diversity at the patent level increases the likelihood of a patent becoming a breakthrough and having more technology codes, suggesting these are more complex recombinations. The IV effect size is large, implying that a one-standard deviation

[12] In our sample, 162,424 patents (about 11%) have more than one patentee, and 11,963 of these have a single surname (Table B45). These percentages remained fairly stable over the historical period we consider (Figure B14).

[13] There are 159,402 patent technology codes in our sample from 1900 to 1944.

[14] These are six broad technological fields: chemical, computers, drugs, electrical, mechanical or other.

Table 6: Patent-level estimates of the effect of surname diversity on innovation

	Breakthrough patent indicator (mean = 0.17, sd = 0.38)			Technologies per patent (mean = 2.49, sd = 1.82)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity of patent	0.015***	0.010***	0.010***	0.032***	0.020***	0.019***
	(0.001)	(0.001)	(0.001)	(0.005)	(0.005)	(0.005)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.010**	0.008**		0.039***	0.032**	
	(0.004)	(0.004)		(0.015)	(0.013)	
Panel C: Instrumental-variable estimates						
Surname diversity of patent	1.127*	0.908*		4.341**	3.731*	
	(0.583)	(0.520)		(2.178)	(2.022)	
Sanderson-Windmeijer *F*-stat	10	9		10	9	
Panel D: First-stage estimates						
	Surname diversity of patent					
Surname diversity (IV)	0.009***	0.008***				
	(0.003)	(0.003)				
Team size fixed effects	✓	✓	✓	✓	✓	✓
County fixed effects	✓	✓		✓	✓	
State-Period fixed effects	✓	✓		✓	✓	
Patent technology field-Period fixed effects		✓	✓		✓	✓
County-Period fixed effects			✓			✓
Observations	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459

Notes: The table reports least squares, reduced-form and IV estimates for the specifications described in equation (10). An observation is a patent with inventors residing in a given county in a five-year period from 1900 to 1940. If there are multiple inventors who reside in different counties, the patent is assigned to all counties and the observation is weighted by 1 / number of inventors. The endogenous variables is surname diversity of patent. The dependent variables are breakthrough patent indicator in Columns 1 to 3 and the number of technology codes per patent in Columns 4 to 6. Columns 2 and 5 add six broad patent technology fields (chemical, computers, drugs, electrical, mechanical and 'other') fixed effects interacted with period fixed effects. Columns 3 and 6 additionally include country-period fixed effects. Standard errors are clustered clustered at the county level. All diversity variables are standardized to have a mean of zero and a unit variance. The sources and construction of all variables are detailed in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

increase in surname diversity at the patent level delivers roughly a 91 percentage point increase in the likelihood of breakthroughs and an increase of roughly four additional technology codes per patent (the mean number of codes per patents is 2.49), according to Columns 2 and 5. We further confirm that the OLS estimates remain robust when we incorporate county-period fixed effects, thus comparing patents within the same county and five year period (Columns 3 and 6).

Panel D shows the first stage of the IV. The estimates suggest that a one standard deviation increase in predicted county-level surname diversity increases the diversity of co-inventor surnames at the patent level by around 0.01 standard deviations. This establishes that higher surname diversity in a county generates more diverse interactions

among patentees, at least in terms of their surnames. This relatively small effect arises, in part, from the fact that most patents during this period have only a single inventor.

Notably, the IV estimates here are substantially larger than the OLS estimates. Part of this gap might be due to the moderate strength of the IV. With an F-statistic of 9–10, the IV is at the lower bound of what is commonly accepted as a sufficiently strong instrument. Additionally, we suspect a downward bias affects the OLS estimates due to the cumulative nature of innovation. It is well established that patents build on prior patents, and particularly on breakthroughs (Acemoglu et al., 2016; Youn et al., 2015; Akcigit et al., 2013). Breakthroughs often integrate ideas from diverse domains, leading to 'refinements' and 'reuse patents' among similarly diverse or even the same sets of co-inventors who build on the insights found in the breakthrough. This creates reverse causality running from breakthroughs to both patent surname diversity and technology codes. Such refinements and reuse patents rarely become breakthroughs themselves and tend to focus on specific domains of the breakthrough, thus, the downward bias for both outcomes. Our IV taps a source of plausibly exogenous variation in surname diversity, removing this bias and delivering larger estimates. Without the IV, the observed correlation between patent-level surname diversity and breakthrough likelihood and patent complexity seems small.

One might worry that our entropic measure of surname diversity might not apply well to inventor teams. To address this, Table B46 reports the relationship between county-level surname diversity and three alternative measures of patent-level diversity, separately capturing variety (log number of distinct surnames among patentees, $\log N$) and dispersion ($(-\sum p \log p)/\log N$) as well as our measure that includes disparity (genetic distance-weighted surname entropy: $-\sum dp \log p$). The estimates in Panel C (IV) indicate that shifts in county-level surname diversity result in similar changes across these alternatives. When we include team size fixed effects (even-numbered columns), the estimates imply that surname diversity decreases the likelihood that inventors on a same-sized team share the same surname.[15]

Finally, we investigate whether our main patent-level findings replicate when using genetic distance-weighted surname diversity and ancestral-country diversity. First, our estimates hardly change when using patent- and county-level genetic distance-weighted surname diversity, although the instrument has become weaker (Table B48). Second, Panel C of Table B49 shows that the instrument for county-level ancestral-country diversity

[15] A comparison with the estimates reported in the even- and odd-numbered columns of Tables B46 and B47, showing the results with and without team size fixed effects for all patents and those with 2 or more co-inventors, respectively, suggests that shifts in county-level diversity may (1) increase the number of 2-person teams (with different surnames) but (2) for larger teams (2+ co-inventors), primarily operate through greater team diversity (not size).

has no effect on patent-level outcomes. If anything, the coefficient for the number of technologies per patent is negative, though the magnitude is small. We also do not observe an effect on the ancestral-country or surname diversity of patents (Panel D), suggesting that greater ancestral-country diversity within counties does not lead to more diverse collaborations among inventors. Lastly, Panel A reports that ancestral-country diversity is positively correlated with patent-level outcomes, implying that once a collaboration is established, patents filed by more diverse teams, in terms of ancestral backgrounds, tend to be more complex and are more likely to be high impact. However, this correlation disappears for the production of breakthroughs in Panel B when we add patent-level surname diversity.

Taken together, these results establish an important set of links: they connect surname diversity at the county-level to more diverse interaction at the patent-level and then connect the more diverse teams to more complex inventions that are more likely to become breakthroughs. This opens the door to probing the the underlying mechanisms more deeply.

6 Mechanisms

Our hypothesis posits that greater surname diversity increases the production of both patents and breakthrough patents because innovation often emerges through the recombination of ideas that occurs when diverse individuals interact. Above, we connected county-level surname diversity to patent-level diversity and to innovation quality. To further flesh out the underlying mechanisms, we use census and patent data to explore the extent to which surname diversity affects the informational component—different people must possess distinct ideas—and social-behavioral component—individuals must be willing to interact and share their ideas. Jointly, these two components should facilitate the creation of innovative recombinations.

6.1 Informational Mechanism

Our approach assumes that surnames in the U.S. (1850-1940) capture significant informational diversity within populations, which arises from social learning and the intergenerational transmission of information. Conceptually, this aligns with previous work on population diversity and growth that uses markers like country of birth or ancestry to measure such diversity (e.g., Alesina et al., 2016; Fulford et al., 2020). Here, as proof of concept, we show that at least some forms of knowledge and skills cluster in surnames

during our study period. As noted above, we argue that surnames also capture other informational differences (not observed here), such as those related to techniques, languages, perspectives, intuitions and interests.[16]

6.1.1 Surnames Cluster Information

To assess whether surnames capture relevant information in our context, we calculate Herfindahl-Hirschman Indices (HHIs) that capture how strongly surnames cluster in four domains: (1) education, (2) occupations, (3) birthplaces and (4) technology patent classes.[17] The construction of the concentration measure for each domain proceeds in two steps. First, using occupation as an example, we calculate HHIs for each surname across all occupations. This gives us a measure of how strongly each surname clusters in occupations. We normalize this measure such that a value of zero implies a uniform distribution of a surname across occupations and a value of one implies that a surname only occurs within a single occupation. Second, we average the surname-specific HHI across all surnames, weighted by the number of people with a given surname. This averaged index reveals the overall surname concentration across occupations based on the full population. Following this protocol, we construct the HHIs for the other informational domains and apply the same approach to groupings based on the birth country and ancestral country, replacing surnames with more conventional measures of cultural diversity.

Table 7 reports the surname concentration indices for the different domains, samples, and years in the first row. Here, we use only adult men (ages 25-60), but Table B50 reports the results when all individuals are included. In these tables, all surname concentration indices are well above zero, indicating that surnames are concentrated in years of schooling (Column 1), occupations (Columns 2 to 5), countries of birth (Columns 6 and 7), regions of

[16] Research from across the social sciences indicates that surnames should capture substantial and important informational differences. Three lines of evidence are relevant. First, across populations, people rely on different languages, thinking styles, decision heuristics, preferences, metaphors, attentional biases and ritual practices (Nisbett, 2003; Falk et al., 2018; Henrich, 2020). Work in cognitive science, for example, indicates that speaking and thinking in different languages has consequences for people's perceptions, attention and reasoning (Blasi et al., 2022). So, culturally-evolved cognitive differences are common. Second, research on social mobility indicates that surnames reflect unique skills and socialization (Clark, 2014; Güell et al., 2015; Barone and Mocetti, 2021) and much cultural transmission flows from same-sex parents to their children (Hewlett, 2016; Schniter et al., 2023). Third, as explained above, analyses of U.S. patents and Swedish entrepreneurship reveal the significance of cultural transmission from both same-sex parents. The latter two points indicate that some of these cognitive differences will occur among family groupings.

[17] There are 307 and 407 patent technology classes in our 1880-9 and 1949 samples, respectively, including categories like "Geometrical Instruments", "Stoves and Furnace", and "Chemistry: Electrical and Wave Energy."

Table 7: Surnames cluster in education, occupations, birthplaces, and patent technologies

	Years of schooling	Occupation				Country of birth		Region of birth	USPTO tech class	
Sample:	All			Immigrants				Germans	Inventors	
Year:	1940	1880	1940	1880	1940	1880	1940	1880	1880-9	1940-9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Surname	0.089	0.167	0.056	0.121	0.073	0.415	0.246	0.270	0.092	0.068
Country of birth	0.086	0.173	0.052	0.117	0.064					
Country of ancestry	0.081	0.158	0.049	0.104	0.050				0.004	0.003
U.S. county of residence	0.102	0.231	0.088	0.183	0.084	0.273	0.123	0.248	0.014	0.017

Notes: This table reports normalized Herfindahl-Hirschman Indices (HHI), where larger values indicate greater concentration. The indices are calculated as the average HHI of the variable in the header computed for each value of the variables on the left. For example, we calculate the normalized Herfindahl index of occupations for each surname and then average over all surnames using the number of individuals with a given surname as weights. Column 2 (1 and 3) includes all male individuals aged 25-60 in the 1880 (1940) census. Columns 4 and 6 (5 and 7) include all male immigrants aged 25-60 in the 1880 (1940) census. Column 8 restricts the sample to German male immigrants aged 25-60 in 1880. We use the 32 subnational regional origins for German immigrants recorded by the Census (the variable `bp1d` with codes 45301 to 45362). Column 9 (10) includes all inventors of patents issued from 1880 to 1889 (1940 to 1949). We use the main technology class on the patent.

birth within Germany (Column 8),[18] and patent technology classes (Columns 9 and 10).[19] To illustrate, two people with the same surname in 1880 have a roughly 17% (above chance) probability of holding the same occupation out of 249 possible occupations (Column 2). For example, "Smith" and "Johnson", the most frequent surnames in the 1880 census, show a distinct pattern. The Smiths outnumber the Johnsons by a factor of about 1.65 times. Yet, among individuals who reported "blacksmith" as their occupation, there are 2.46 times more Smiths than Johnsons. "Smith," of course, has a long-standing association with metal-working and this result illustrates this pattern in 1880. Similarly, two same-surname immigrants have about a 42% probability of having the same country of birth (Column 6) and, if they are German, a 27% probability of being from the same inner-German region (out of 32 regions). Finally, surnames cluster on educational attainment (Column 1) and on the fine-grained technology classes applied to patents (Columns 9 and 10). The latter finding is consistent with Bell et al.'s 2019 analysis of father-son patenting patterns in the U.S. at the dawn of the 21st century.

To put these concentration indices for surnames into context, we compare them to parallel HHI indices for country of birth (row 2), country of ancestral origin (row 3)

[18]This domain is restricted to German regions because fine-grained subregional birthplace data are available for this country only.

[19]The patent data set does not allow us to uniquely identify the universe of inventors. Hence, we cannot detect all inventors who file multiple patents in the same technology category, which could bias the concentration upwards. We still report this statistic because this bias is likely small, given the low level of regional clustering in this variable (row 2), where we would expect a similar upward bias if regional mobility among inventors is not very high.

and residence county (row 4). We calculate these in the same fashion as with surnames. Compared to country of birth and ancestry, surnames are about equally indicative of education and occupation; however, for patent technology classes, surnames are substantially more indicative.[20] In 1880, occupations are relatively more concentrated in counties than in surnames, but this difference narrows markedly by 1940. By contrast, surnames are substantially more indicative of countries and regions of birth compared to immigrants' residence counties.

Building on the clustering of important informational types demonstrated in Table 7, Table 8 uses equation (2) to connect our county-level surname diversity measures to occupational and patent technology code diversity. Column 1 and 3 of Panel C show that a one standard deviation increase in surname diversity leads to roughly a one-third standard deviation increase in occupational and technology code diversity. This finding is robust to additionally controlling for ancestral-country (Columns 2 and 4) and birth-country diversity (Table B51). Taken together, these results supports the role of the informational channel in driving innovation. However, as noted previously and shown in Table 5, surname diversity captures more than just occupational diversity.

6.2 Social-Behavioral Mechanisms

Distinct forms of cultural diversity could potentially produce different indirect effects on innovation by either fostering or suppressing social interactions among diverse minds. As described above, some existing research leads us to suspect that greater diversity might lead to greater segmentation, fragmentation and polarization in a population (Alesina and La Ferrara, 2005). This would inhibit many forms of social interaction and informational exchanges, thereby reducing the chances for innovative recombinations (Andrews, 2023; Andrews and Lensing, 2024). Alternatively, as surname diversity increases and the relative sizes of informationally distinct surname groups shrink, the rate of diverse social

[20] A potential concern is that, although surnames may often be nested within coarser categories like country of origin, regional birthplace and race, there have been historical processes that muddy this hierarchical nesting. For example, many formerly enslaved Africans carry the European surnames of their enslavers (Cook et al., 2022). Surname diversity may thus underestimate the diversity stemming from African heritage. To address this, we construct a more fine-grained measure and check how it relates to our main indicator. This measure creates additional 'surname categories' based on race-surname combinations. For example, the number of white "Jacksons" enters the diversity indicator as a separate category from the number of black "Jacksons". Similarly, we calculate a surname diversity indicator that further differentiates along country of birth. Table B5 shows that the main surname diversity indicator in 1940 is almost perfectly correlated with those more fine-grained diversity measures. Furthermore, we obtain similarly high correlation coefficients between the main surname diversity indicator and indicators that are based on (i) phonetically uncorrected surnames, (ii) surnames of men only, (iii) surnames of household heads only, and (iv) surnames of whites only.

interactions may increase for two reasons: (i) people may be less able to satisfy their needs by interacting solely within their own small and relatively diminishing groups, and (ii) the spread of different groups with distinct information may create new avenues for mutually beneficial exchanges. Such fruitful engagements with outsiders may foster greater openness to social interactions with diverse others, which in turn (and often incidentally) may promote the recombination of ideas. Both hypotheses could apply to different forms of diversity, with large ethnic or racial groupings operating differently from smaller groups, such as families, lineages or clans.

Based on prior work, greater surname diversity may result in greater social interaction among diverse minds. This is because surname diversity partially captures the extent to which large family groups are present. A growing body of research suggests that strong family ties and large residential kinship groups influence both people's psychology—leading to less trust and openness toward strangers—and their social behavior—resulting in less interaction with outgroup members (Enke, 2019; Alesina and Giuliano, 2014; Schulz et al., 2019; Schulz, 2022; Bahrami-Rad et al., 2022).

To better understand the potential indirect impacts of surname diversity on innovation, we construct three variables likely to influence the rate of diverse social interactions at the county level: (i) the residential segregation of surnames, (ii) the residential segregation of ancestral-country groups and (iii) the strength of family ties.

The first two measures, calculated in parallel fashion, aim to capture the residential clustering for two types of groups, one based on surnames and the other based on ancestral countries. For each, we constructed a measure of next-door neighbor segregation closely following the methodology of Logan and Parman (2017). Constructing these measures involves two steps: first, quantifying the degree of segregation for each group (surname or ancestral country) in each county during specific periods; second, averaging these degrees across all groups within a county to obtain a county-level measure of segregation (see Appendix A). This gives us two measures of segregation: (1) for surnames, our measure captures the cluster of same-surname pairs at the street-level, indicating segregation *beyond* what would be expected from random choices of residence; and (2) for ancestral countries, our measure captures the cluster of same-ancestral-country pairs at the street-level. Figure B7 depicts the geographic distributions of both types of residential segregation in 1940.

Crucially, these segregation measures are constructed to account for the distributions of surnames or ancestral countries at the county level. Positive values signify segregation beyond what would occur by chance, while negative values imply that neighbors are more diverse than what chance alone would predict, indicating a *preference for diversity*.

Assessing the degree of segregation in relation to a county's pattern against that expected by chance allows us to capture people's settlement preferences. Thus, an association between surname diversity and the segregation of either surnames or ancestral countries is not merely driven by more diverse counties being necessarily less segregated. Rather, it reflects differences in fine-grained settlement preferences while holding background diversity constant. Both surname segregation and ancestral-country segregation measures are highly negatively correlated with surname diversity ($\rho = -0.71$ and $\rho = -0.54$, respectively), indicating that individuals in counties with low surname diversity tend to live in closer geographic proximity to others with the same surname and country of ancestry.

The third measure, the strength of family ties, captures the size, homogeneity, and stability of households as social units (Raz, 2024). It represents the first principal component of four county-level variables: (i) the divorce-to-marriage ratio, (ii) the share of elderly people living without a relative, (iii) the share of people living with at least one person who is not their relative, and (iv) the mean size of families (Appendix A). Higher values for the strength of family ties indicate a preference for surrounding oneself with relatives. Figure B8 illustrates the geographic distribution of the strength of family ties in 1940. This variation is fairly strongly negatively correlated with surname diversity ($\rho \approx -0.50$), suggesting that that individuals in counties with low surname diversity tend to have strong kin-based networks.

To analyze the effect of surname diversity on the potential for diverse social interactions, we re-estimate our county-level specification (2), replacing the dependent variable by our three proxy measures. Table 8 reports the results in Columns 3, 5 and 7. The IV estimate in Column 3 of Panel C indicates that a one standard deviation increase in surname diversity reduces the residential segregation of surname groups by 0.41 standard deviations (below what would be expected by chance alone). For the segregation of ancestral-country groups, the impact of surname diversity is even larger (Column 5, Panel C): a one standard deviation increase in surname diversity leads to a one standard deviation reduction in residential segregation of ancestral-country groups. Similarly, as shown in Column 7 of Panel C, a one standard deviation increase in surname diversity results in the strength of family ties weakening by 0.32 standard deviations. In all cases, the OLS and reduced-form estimates in Panels A and B are similar to the IV estimates. These results suggest that an increase in surname diversity is associated with more diverse social interactions within counties.

We also consider what happens when we add our measure of ancestral-country diversity to these specifications. These estimates are presented in the even-numbered columns in Table 8. Crucially, we find that the coefficients on surname diversity remain negative,

Table 8: Effects on proxy measures of informational and social-behavioral mechanisms

	Occupational diversity		USPTO tech codes diversity		Residential segregation of surname groups		Residential segregation of ancestral-country groups		Strength of family ties	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Least-squares estimates										
Surname diversity	0.311***	0.335***	0.303***	0.362***	-0.387***	-0.432***	-1.079***	-1.320***	-0.202***	-0.157***
	(0.034)	(0.038)	(0.048)	(0.058)	(0.081)	(0.096)	(0.049)	(0.047)	(0.041)	(0.046)
Ancestral-country diversity		-0.044**		-0.096***		0.078		0.417***		-0.080***
		(0.020)		(0.037)		(0.049)		(0.019)		(0.027)
Population	-0.052**	-0.051**	0.059**	0.062**	0.044	0.040	0.091***	0.073***	-0.088**	-0.085**
	(0.021)	(0.021)	(0.029)	(0.029)	(0.033)	(0.032)	(0.025)	(0.021)	(0.036)	(0.035)
Panel B: Reduced-form estimates										
Surname diversity (IV)	0.288***	0.278***	0.233***	0.247***	-0.314***	-0.352***	-0.824***	-0.878***	-0.207***	-0.237***
	(0.030)	(0.031)	(0.039)	(0.044)	(0.069)	(0.074)	(0.061)	(0.064)	(0.035)	(0.038)
Ancestral-country diversity (IV)		0.011		-0.015		0.037		0.054***		0.031**
		(0.009)		(0.017)		(0.025)		(0.014)		(0.015)
Population (IV)	-0.122***	-0.119***	-0.029	-0.034	0.129**	0.146***	0.455***	0.479***	-0.089**	-0.079**
	(0.029)	(0.029)	(0.034)	(0.035)	(0.052)	(0.053)	(0.068)	(0.070)	(0.040)	(0.039)
Panel C: Instrumental-variable estimates										
Surname diversity	0.377***	0.382***	0.350***	0.457***	-0.407***	-0.586***	-1.031***	-1.340***	-0.322***	-0.467***
	(0.038)	(0.051)	(0.055)	(0.110)	(0.084)	(0.136)	(0.059)	(0.067)	(0.047)	(0.081)
Ancestral-country diversity		-0.010		-0.193		0.345*		0.597***		0.299**
		(0.063)		(0.156)		(0.183)		(0.076)		(0.119)
Population	-0.125***	-0.121**	0.002	0.045	0.114*	-0.006	0.506***	0.298***	-0.240***	-0.337***
	(0.042)	(0.047)	(0.055)	(0.072)	(0.069)	(0.087)	(0.097)	(0.076)	(0.067)	(0.089)
Sanderson-Windmeijer *F*-stat	149; 29	40; 75; 120	163; 18	38; 43; 60	145; 25	39; 63; 86	145; 25	39; 63; 86	158; 28	55; 86; 124
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	21,982	21,982	10,353	10,353	13,727	13,727	13,727	13,727	21,968	21,968

Notes: The table reports OLS, reduced-form and IV estimates for the the specification described in equation (2) but with occupational diversity, USPTO technology codes diversity, residential segregation of surname groups, residential segregation of ancestral-country group and strength of family ties as the outcome variables. Even columns add ancestral-country diversity to the specification. An observation is a county in a period from 1900 to 1940. Standard errors are clustered at the state level. All variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

large, well-estimated and stable. However, in contrast, the results for ancestral-country diversity are not consistent across Panels A, B, and C. The IV estimates in Columns 6, 8 and 10 of Panel C are positive and well-estimated, suggesting that greater ancestral-country diversity may result in greater segregation and stronger family ties—precisely the opposite of the impact of surname diversity. However, the coefficients in the OLS and reduced-form regressions are small, sometimes poorly estimated, and sometimes flip signs. We observe similar patterns when we incorporate birth-country diversity instead of ancestral-country diversity (Table B51). Thus, unlike for surname diversity, we do not find consistent evidence that ancestral- or birth-country diversity have the same impacts on social interaction.

The above results expand upon our earlier patent-level findings in two ways. First, Table 6 shows that greater surname diversity at the county level results in greater surname diversity at the patent level and a larger number of technology codes per patent. The former indicates that more surname-diverse counties spur more diverse interactions among inventors, while the latter suggests that such interactions encourage more complex

recombinations. The increased likelihood of more diverse co-inventors is not surprising given the results presented above: more surname-diverse counties are more residentially intermixed—both in terms of surnames and ancestral countries—and have weaker family ties, thus enabling inventors to interact more outside their family and ethnic group. Second, by contrast, our patent-level results do not suggest that greater ancestral-country diversity at the county level results in more heterogeneous interactions at the patent-level (based on both ancestral countries and surnames, Table B49). Consistent with this, when surname diversity is included in the county-level model (Table 8), we do not observe robust relationships between ancestral-country diversity and either family ties or residential segregation.

6.2.1 Local Proximity and Surname Diversity

Our hypothesis proposes that most innovations arise from the recombination of ideas, often sparked by social interactions among diverse minds. Thus, we expect that people acquire ideas from others they frequently interact with, often serendipitously, in their daily lives (Atkin et al., 2022). If true, physical proximity and local diversity will play a big role (Jaffe et al., 1993; Carlino and Kerr, 2015). To explore this, we study the impact of the surname diversity found in neighboring counties on local innovation. If local social interactions are the most important, then the impact of neighboring counties should be small or negligible relative to that of a focal county's surname diversity. Thus, we compute the surname diversity of people residing in surrounding counties at successively further distances from each focal county. Specifically, for each county i at time t, we pool the individuals in surrounding counties within 100 miles, compute their surname diversity and construct a separate instrument for these individuals, excluding i itself. We repeat this exercise for individuals between 100 and 200 miles and between 200 and 300 miles.[21]

Table 9 presents the results. For patents per capita, as shown in Columns 1 to 3 of Panel C, the IV coefficients for surname diversity within the focal county are large and do not decrease in magnitude when we include the surname diversity of surrounding counties. The effect size of surname diversity in neighboring counties within a 100-mile radius is about half the size of the coefficient for own-county surname diversity, and the coefficients for surname diversities in more distant locals are small and poorly estimated. For breakthrough patents (Columns 4 to 6, Panel C), the IV coefficients for own-county surname diversity remain stable with the addition of surrounding counties. Here, the coefficients for surname diversity of surrounding counties at any distance are small

[21] We use the NBER's County Distance Database to compute these areas for each county.

Table 9: Spillover analysis

	Patents per 1,000 people (mean = 1.05, sd = 1.73)			Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.281***	1.282***	1.282***	0.093***	0.093***	0.094***
	(0.177)	(0.178)	(0.177)	(0.018)	(0.018)	(0.018)
Surname diversity (< 100 miles)	0.492*	0.476*	0.472*	0.038	0.031	0.031
	(0.249)	(0.238)	(0.237)	(0.041)	(0.042)	(0.042)
Surname diversity (100 < 200 miles)		0.113	0.121		0.056	0.057
		(0.236)	(0.236)		(0.081)	(0.081)
Surname diversity (200 < 300 miles)			-0.040			0.019
			(0.175)			(0.040)
Population	0.731***	0.730***	0.730***	0.098***	0.097***	0.097***
	(0.189)	(0.189)	(0.189)	(0.029)	(0.028)	(0.028)
Population (< 100 miles)	-0.017	-0.028	-0.078	0.080*	0.070	0.065
	(0.166)	(0.177)	(0.183)	(0.045)	(0.042)	(0.041)
Population (100 < 200 miles)		-0.137	-0.206		-0.095**	-0.103**
		(0.205)	(0.222)		(0.040)	(0.046)
Population (200 < 300 miles)			-0.300*			-0.039
			(0.173)			(0.045)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.891***	0.888***	0.887***	0.058***	0.056***	0.056***
	(0.239)	(0.239)	(0.239)	(0.018)	(0.018)	(0.018)
Surname diversity (IV, < 100 miles)	0.342**	0.337**	0.341**	0.009	0.008	0.007
	(0.161)	(0.155)	(0.157)	(0.028)	(0.027)	(0.027)
Surname diversity (IV, 100 < 200 miles)		0.127	0.132		0.065**	0.064**
		(0.155)	(0.156)		(0.030)	(0.031)
Surname diversity (IV, 200 < 300 miles)			-0.019			-0.028
			(0.107)			(0.027)
Population (IV)	0.223	0.224	0.223	0.050	0.050	0.050
	(0.206)	(0.206)	(0.206)	(0.035)	(0.035)	(0.034)
Population (IV, < 100 miles)	0.037	0.021	-0.043	0.082**	0.066*	0.060*
	(0.135)	(0.141)	(0.146)	(0.036)	(0.034)	(0.032)
Population (IV, 100 < 200 miles)		-0.097	-0.181		-0.085**	-0.094**
		(0.180)	(0.191)		(0.035)	(0.040)
Population (IV, 200 < 300 miles)			-0.358**			-0.034
			(0.135)			(0.038)
Panel C: Instrumental-variable estimates						
Surname diversity	1.069***	1.062***	1.062***	0.103***	0.101***	0.098***
	(0.271)	(0.271)	(0.282)	(0.021)	(0.021)	(0.022)
Surname diversity (< 100 miles)	0.510	0.567**	0.564**	-0.015	-0.017	-0.019
	(0.312)	(0.273)	(0.266)	(0.056)	(0.054)	(0.054)
Surname diversity (100 < 200 miles)		-0.099	-0.088		0.099	0.138
		(0.784)	(0.922)		(0.139)	(0.165)
Surname diversity (200 < 300 miles)			-0.054			-0.188
			(0.941)			(0.175)
Population	0.646***	0.648***	0.648***	0.135***	0.134***	0.133***
	(0.110)	(0.107)	(0.105)	(0.019)	(0.018)	(0.018)
Population (< 100 miles)	0.125	0.130	0.133	0.033	0.037	0.038
	(0.127)	(0.128)	(0.126)	(0.033)	(0.035)	(0.033)
Population (100 < 200 miles)		-0.133	-0.128		-0.059**	-0.058*
		(0.139)	(0.140)		(0.027)	(0.029)
Population (200 < 300 miles)			0.036			0.027
			(0.161)			(0.030)
Sanderson-Windmeijer *F*-stat: Surname diversity	169; 137	147; 104; 28	203; 157; 25; 26	169; 137	147; 104; 28	203; 157; 25; 26
Sanderson-Windmeijer *F*-stat: Population	30; 463	39; 775; 159	42; 621; 231; 413	30; 463	39; 775; 159	42; 621; 231; 413
County fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓
Observations	21,936	21,936	21,936	21,936	21,936	21,936

Notes: The table reports OLS, reduced-form and IV estimates of regressions of innovation outcomes on surname diversity and population. The unit of observation is a county-period from 1900 to 1940 (including the midyears). The table sequentially adds surname diversity and population in areas within 100 miles (excluding i), 100 miles to 200 miles, 200 miles to 300 miles of county i, and the associated instruments. For each specification we report the first-stage Sanderson-Windmeijer F-stats for all endogenous variables. Standard errors are clustered on states and reported in parentheses. All independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels. First-stage estimates are reported in Table B52.

and poorly estimated. Overall, consistent with the importance of the social-behavioral mechanism, these findings suggest that the causal link between surname diversity and innovation is geographically localized, suggesting that meeting and interacting face-to-face on a day-to-day basis may be an important ingredient for innovation during our study period. The results also offer empirical support for our choice of the county as our primary unit of analysis.

7 Conclusion

The United States rose to dominate global innovation during the period from 1850 to 1940. To better understand the factors driving this process, we consider the impact of America's changing cultural diversity on innovation. Our core hypothesis is that many, if not most, innovations arise from the recombination of existing ideas, approaches and techniques that come together when diverse minds meet, exchange ideas, and sometimes collaborate. To create a fine-grained proxy for this diversity, we use surnames drawn from the complete U.S. Census to calculate an entropic measure of variation. To measure innovation, we use both patents and a text-based measure of breakthrough patents. For causal identification, we employ an instrumental variable approach that uses historical immigration flows (push-pull interactions) to extract plausibly exogenous shifts in surname diversity across space and time. Our IV analyses suggest a causal connection between surname diversity and innovation at four levels: the county, surname-county, inventor and patent. At the county level (1900–1940), we estimate that a standard deviation increase in surname diversity leads to roughly 107% more patents per capita and 80% more breakthrough patents per capita relative to their means, corresponding to 1.76 times and 0.76 times the effect of population size, respectively. These relationships are remarkably stable over time and likely extend well back into the 19th century. At the inventor level, a standard deviation increase in surname diversity raises inventors' productivity by roughly five patents and one breakthrough patent over a 50-year career. At the patent level, a one-standard deviation increase in surname diversity delivers roughly a 91 percentage point increase in the likelihood of a patent becoming a breakthrough and roughly 150% more technology codes per patent.

To explore the causal pathways running from surname diversity to innovation, we close by offering evidence supporting the connection between surname diversity and (1) informational diversity, which opens the door directly to more innovative recombinations, and (2) the potential for heterogeneous social interaction, which increases the likelihood that diverse individuals will meet and interact. Focusing on the informational channel,

we show that both surnames and ancestral countries cluster with education, occupations, originating countries and the technological categories of patents, but both birth- and ancestral-country diversity do not have an effect on innovation beyond the impact captured by surname diversity. We also show that greater surname diversity results in greater occupational and patent technology code diversity. Focusing on the social-behavioral channel, we demonstrate that greater surname diversity at the county level results in weaker family ties and less residential segregation among people with both different surnames and ancestral countries, thus creating conditions conducive to more diverse social interactions.

Broadly, our results highlight the need for researchers to more deeply consider how, when and why diversity and immigration foster innovation. Our comparisons of surname diversity versus ancestral-country diversity suggest that the former, which captures more granular cultural differences, may be the more powerful stimulant for innovation. Indeed, our findings indicate that the latter influences innovation by capturing part of what surname diversity includes. Our analysis of mechanisms suggests that this difference is most likely due to how the two forms of diversity influence social interaction. Greater surname diversity results in weaker family ties, less residential segregation and more diversity at the patent level. By contrast, ancestral-country diversity does not stimulate more social interaction; indeed, there are hints that it may result in less interaction. The differences here likely arise from the parochialism and distrust that appear among large ethnic groups, which are not usually matched at similar levels among different families or smaller groupings. Such findings suggest there may be optimal forms of diversity that jointly maximize the richness of information and the potential for social interactions.

Our findings offer insights for designing policies aimed at enhancing innovation, suggesting that any factors that increase the likelihood of open-minded, face-to-face social interactions among people with diverse backgrounds can foster faster innovation. Policies related to immigration, housing, education, urban design (zoning, density) and transportation (among others), as well as local norms and beliefs that influence tolerance, cooperation and trust towards strangers, outsiders and foreigners, can play a role. Crucially, the most powerful forms of diversity come from having smaller and more uniformly distributed kinship groups or networks rather than large country or nationality-based groups. Specifically, our results suggest that policies that encourage family-based migration, resulting in clan-like clusters of relatives, may suppress innovation and that educational investments may have larger innovation payoffs when applied to local populations with more fine-grained cultural diversity.

References

Abramitzky, Ran, Philipp Ager, Leah Boustan, Elior Cohen & Casper W. Hansen (2023) "The effect of immigration restrictions on local labor markets: Lessons from the 1920s border closure", *American Economic Journal: Applied Economics*, 15 (1), pp. 164–191.

Abramitzky, Ran & Leah Boustan (2017) "Immigration in American economic history", *Journal of Economic Literature*, 55 (4), pp. 1311–45.

Acemoglu, Daron, Ufuk Akcigit & William R. Kerr (2016) "Innovation network", *Proceedings of the National Academy of Sciences*, 113 (41), p. 11483–11488.

Ager, Philipp & Markus Brückner (2013) "Cultural diversity and economic growth: Evidence from the US during the age of mass migration", *European Economic Review*, 64, pp. 76–97.

Agneman, Gustav & Esther Chevrot-Bianco (2023) "Market participation and moral decision-making: Experimental evidence from greenland", *The Economic Journal*, 133 (650), p. 537–581.

Agrawal, Ajay, Devesh Kapur & John McHale (2008) "How do spatial and social proximity influence knowledge flows? Evidence from patent data", *Journal of Urban Economics*, 64 (2), pp. 258–269.

Akcigit, Ufuk, John Grigsby & Tom Nicholas (2017) "Immigration and the rise of American ingenuity", *American Economic Review*, 107 (5), pp. 327–331.

Akcigit, Ufuk, William R Kerr & Tom Nicholas (2013) "The mechanics of endogenous innovation and growth: Evidence from historical U.S. patents", *Working paper*.

Alesina, Alberto F. & Paola Giuliano (2015) "Culture and institutions", *Journal of Economic Literature*, 53 (4), p. 898–944, Citation Key: 134896.

Alesina, Alberto & Paola Giuliano (2010) "The power of the family", *Journal of Economic Growth*, 15 (2), p. 93–125.

Alesina, Alberto & Paola Giuliano (2014) "Chapter 4 - Family Ties", Philippe Aghion & Steven N. Durlauf eds. *Handbook of Economic Growth*, 2, Elsevier, pp. 177–215.

Alesina, Alberto, Johann Harnoss & Hillel Rapoport (2016) "Birthplace diversity and economic prosperity", *Journal of Economic Growth*, 21 (2), pp. 101–138.

Alesina, Alberto & Eliana La Ferrara (2005) "Ethnic Diversity and Economic Performance", *Journal of Economic Literature*, 43 (3), pp. 762–800.

Algan, Yann & Pierre Cahuc (2014) "Chapter 2 - trust, growth, and well-being: New evidence and policy implications", Philippe Aghion & Steven N. Durlauf eds. *Handbook of Economic Growth*, 2 of Handbook of Economic Growth, Elsevier, pp. 49–120.

Allport, Gordon W. (1954) *The Nature of Prejudice*, Oxford, England, Addison-Wesley.

AlShebli, Bedoor K., Talal Rahwan & Wei Lee Woon (2018) "The preeminence of ethnic diversity in scientific collaboration", *Nature Communications*, 9 (1), p. 5163.

Andersson, David, Thor Berger & Erik Prawitz (2023) "Making a market: Infrastructure, integration, and the rise of innovation", *The Review of Economics and Statistics*, 105 (2), p. 258–274.

Andrews, Michael (2023) "Bar Talk: Informal Social Interactions, Alcohol Prohibition, and Invention", *Working paper*.

Andrews, Michael J. & Chelsea Lensing (2024) "Cup of joe and knowledge flow: Coffee shops and invention", *Working paper*.

Arbatli, Cemal Eren, Quamrul H Ashraf, Oded Galor & Marc Klemp (2020) "Diversity and Conflict", *Econometrica*, 88 (2), pp. 727–797, Publisher: Wiley Online Library.

Arkolakis, Costas, Sun Kyoung Lee & Michael Peters (2020) "European Immigrants and the United States' Rise to the Technological Frontier", p. 69.

Ashraf, Quamrul & Oded Galor (2013) "The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development", *American Economic Review*, 103 (1), pp. 1–46.

Ashraf, Quamrul H., Oded Galor & Marc Klemp (2021) "Chapter 22 - The Ancient Origins of the Wealth of Nations", Alberto Bisin & Giovanni Federico eds. *The Handbook of Historical Economics*, Academic Press, pp. 675–717.

Atkin, David, Keith Chen & Anton Popov (2022) "The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley".

Bahrami-Rad, Duman, Jonathan Beauchamp, Joseph Henrich & Jonathan F. Schulz (2022) "Kin-based institutions and Economic Development".

Barone, Guglielmo & Sauro Mocetti (2021) "Intergenerational mobility in the very long run: Florence 1427–2011", *The Review of Economic Studies*, 88 (4), pp. 1863–1891.

Bell, Alexander, Raj Chetty, Xavier Jaravel, Neviana Petkova & John Van Reenen (2019) "Who Becomes an Inventor in America? The Importance of Exposure to Innovation", *Quarterly Journal of Economics*, 134 (2), pp. 647–713.

Berkes, Enrico (2018) "Comprehensive universe of US patents (CUSP): Data and facts", *Working paper*.

Berkes, Enrico & Ruben Gaetani (2021) "The Geography of Unconventional Innovation", *The Economic Journal*, 131 (636), pp. 1466–1514.

Bernstein, Shai, Rebecca Diamond, Abhisit Jiranaphawiboon, Timothy McQuade & Beatriz Pousada (2022) "The contribution of high-skilled immigrants to innovation in the united states",Technical report, National Bureau of Economic Research.

Bettencourt, L M A, J Lobo & D Strumsky (2007) "Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size", *Research Policy*, 36 (1), pp. 107–120.

Bisin, Alberto & Thierry Verdier (2001) "The Economics of Cultural Transmission and the Dynamics of Preferences", *Journal of Economic Theory*, 97 (2), pp. 298–319.

Blasi, Damián E., Joseph Henrich, Evangelia Adamou, David Kemmerer & Asifa Majid (2022) "Over-reliance on english hinders cognitive science", *Trends in Cognitive Sciences*, 26 (12), p. 1153–1170.

Borusyak, Kirill, Peter Hull & Xavier Jaravel (2022) "Quasi-Experimental Shift-Share Research Designs", *The Review of Economic Studies*, 89 (1), pp. 181–213.

Boyd, Robert & Peter J Richerson (1985) *Culture and the Evolutionary Process*, Chicago, IL, University of Chicago Press.

Buggle, Johannes C (2020) "Growing collectivism: Irrigation, group conformity and technological divergence", *Journal of Economic Growth*, 25 (2), pp. 147–193.

Buonanno, Paolo & Paolo Vanin (2017) "Social Closure, Surnames and Crime", *Journal of Economic Behavior & Organization*, 137, pp. 160–175.

Burchardi, Konrad B, Thomas Chaney & Tarek A Hassan (2019) "Migrants, Ancestors, and Foreign Investments", *The Review of Economic Studies*, 86 (4), pp. 1448–1486.

Bursztyn, Leonardo, Thomas Chaney, Tarek A. Hassan & Aakaash Rao (2024) "The Immigrant Next Door", *American Economic Review*, 114 (2), pp. 348–384.

Carcassi, Gabriele, Christine A Aidala & Julian Barbour (2021) "Variability as a better characterization of Shannon entropy", *European Journal of Physics*, 42 (4), p. 045102.

Carlino, G A, S Chatterjee & R M Hunt (2007) "Urban density and the rate of invention", *Journal of Urban Economics*, 61 (3), pp. 389–419.

Carlino, Gerald & William R. Kerr (2015) "Agglomeration and Innovation", *Handbook of Regional and Urban Economics*, 5, Elsevier, pp. 349–404.

Chen, Jiafeng & Jonathan Roth (2023) "Logs with Zeros? Some Problems and Solutions", *The Quarterly Journal of Economics*, p. qjad054.

Chiopris, Caterina (2024) "Spatial networks and the diffusion of ideas".

Christakis, N.A. & J.H. Fowler (2009) *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*, Little, Brown.

Cinnirella, Francesco, Erik Hornung & Julius Koschnick (2022) "Flow of ideas: Economic societies and the rise of useful knowledge", *CESifo Working Paper*, 9836.

Clark, Gregory (2014) *The Son also Rises: a Surprising Look how Ancestry still Determines Social Outcomes*, Princeton, Princeton University Press.

Cook, Lisa D., John Parman & Trevon Logan (2022) "The antebellum roots of distinctively black names", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55 (1), pp. 1–11.

de la Croix, David, Matthias Doepke & Joel Mokyr (2018) "Clans, guilds, and markets: Apprenticeship institutions and growth in the pre-industrial economy", *The Quarterly Journal of Economics*, 133 (1), pp. 1–70.

Docquier, Frédéric, Riccardo Turati, Jérôme Valette & Chrysovalantis Vasilakis (2020) "Birthplace diversity and economic growth: Evidence from the US states in the Post-World War II period", *Journal of Economic Geography*, 20 (2), pp. 321–354.

Dowey, James (2017) "Mind over matter: Access to knowledge and the British industrial revolution", Ph.D. dissertation, Dissertation, The London School of Economics and Political Science.

Enke, Benjamin (2019) "Kinship, Cooperation, and the Evolution of Moral Systems*", *The Quarterly Journal of Economics*, 134 (2), pp. 953–1019.

Enke, Benjamin (2022) "Market exposure and human morality", *Nature Human Behaviour*.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman & Uwe Sunde (2018) "Global evidence on economic preferences*", *The Quarterly Journal of Economics*, 133 (4), p. 1645–1692.

Feldman, Maryann P & David B Audretsch (1999) "Innovation in cities : Science-based diversity , specialization and localized competition", *European Economic Review*, 43 (2), pp. 409–429.

Ferrara, Andreas, Patrick Testa & Liyang Zhou (2021) "New Area-and Population-based Geographic Crosswalks for US Counties and Congressional Districts, 1790-2020", *SSRN 4019521*.

Fiszbein, Martin (2022) "Agricultural Diversity, Structural Change, and Long-Run Development: Evidence from the United States", *American Economic Journal: Macroeconomics*, 14 (2), pp. 1–43.

Fulford, Scott L., Ivan Petkov & Fabio Schiantarelli (2020) "Does it matter where you came from? Ancestry composition and economic performance of US counties, 1850–2010", *Journal of Economic Growth*, 25 (3), pp. 341–380.

Galor, Oded, Marc Klemp & Daniel C. Wainstock (2024) "Roots of cultural diversity", *SSRN Electronic Journal*.

Ghosh, Arkadev, Sam Il Myoung Hwang & Munir Squires (2023) "Economic Consequences of Kinship: Evidence From U.S. Bans on Cousin Marriage", *The Quarterly Journal of Economics*, 138 (4), pp. 2559–2606.

Giuliano, Paola & Nathan Nunn (2021) "Understanding cultural persistence and change", *The Review of Economic Studies*, 88 (4), p. 1541–1581.

Glaeser, Edward (2011) "Engines of Innovation", *Scientific American*, p. 7.

Glaeser, Edward L., Hedi D. Kallal, José A. Scheinkman & Andrei Shleifer (1992) "Growth in cities", *Journal of Political Economy*, 100 (6), p. 1126–1152.

Goldsmith-Pinkham, Paul, Isaac Sorkin & Henry Swift (2020) "Bartik Instruments: What, When, Why, and How", *American Economic Review*, 110 (8), pp. 2586–2624.

Gomez-Lievano, Andres, Oscar Patterson-Lomba & Ricardo Hausmann (2017) "Explaining the prevalence, scaling and variance of urban phenomena", *Nature Human Behaviour*, 1 (1), p. No. 0012.

Gorodnichenko, Yuriy & Gerard Roland (2016) "Culture, institutions, and the wealth of nations", *The Review of Economics and Statistics*, 99 (3), p. 402–416.

Griliches, Zvi (1990) "Patent Statistics as Economic Indicators: A Survey", *Journal of Economic Literature*, 28 (4), pp. 1661–1707.

Güell, Maia, José V. Rodríguez Mora & Christopher I. Telmer (2015) "The Informational Content of Surnames, the Evolution of Intergenerational Mobility, and Assortative Mating", *The Review of Economic Studies*, 82 (2), pp. 693–735.

Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. P. Tracer & J. Ziker (2010) "Market, religion, community size and the evolution of fairness and punishment", *Science*, 327, p. 1480–1484.

Henrich, Joseph (2016) *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*, Princeton, Princeton University Press.

Henrich, Joseph (2020) *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*, New York, Farrar, Straus and Giroux.

Henrich, Joseph & James Broesch (2011) "On the nature of cultural transmission networks: evidence from fijian villages for adaptive learning biases", *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366 (1567), p. 1139–1148.

Henrich, Joseph & Michael Muthukrishna (2024) "What makes us smart?", *Topics in Cognitive Science*, 16 (2), pp. 322–342.

Hewlett, Barry S. (2016) *Social Learning and Innovation in Hunter-Gatherers*, p. 1–15, Tokyo, Springer Japan.

Jacobs, J. (1985) *Cities and the Wealth of Nations: Principles of Economic Life*, Vintage Books.

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

Jaffe, Adam B., Manuel Trajtenberg & Rebecca Henderson (1993) "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations*", *The Quarterly Journal of Economics*, 108 (3), pp. 577–598.

Jones, Charles I (2023) "Recipes and economic growth: A combinatorial march down an exponential tail", *Journal of Political Economy*, 131 (8), pp. 1994–2031.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru & Matt Taddy (2021) "Measuring Technological Innovation over the Long Run", *American Economic Review: Insights*, 3 (3), pp. 303–320.

Kerr, William R. (2008) "The Ethnic Composition of US Inventors", *HBS Finance Working Paper No. 08-006*.

Kerr, William R. (2010) "Breakthrough inventions and migrating clusters of innovation", *Journal of Urban Economics*, 67 (1), pp. 46–60.

Knudsen, Anne Sofie Beck (2024) "Those who stayed: Selection and cultural change in the age of mass migration".

Lerner, Josh & Amit Seru (2022) "The use and misuse of patent data: Issues for finance and beyond", *The Review of Financial Studies*, 35 (6), p. 2667–2704.

Lerner, Josh & Julie Wulf (2007) "Innovation and Incentives: Evidence from Corporate R&D", *the Review of Economics and Statistics*, 89 (4), pp. 634–644.

Lindquist, Matthew J, Joeri Sol & Mirjam Van Praag (2015) "Why do entrepreneurial parents have entrepreneurial children?", *Journal of Labor Economics*, 33 (2), pp. 269–296.

Lissoni, Francesco & Ernest Miguelez (2024) "Migration and innovation: Learning from patent and inventor data", *Journal of Economic Perspectives*, 38 (1), p. 27–54.

Logan, Trevon D. & John M. Parman (2017) "The National Rise in Residential Segregation", *The Journal of Economic History*, 77 (1), pp. 127–170.

Maddux, William W, Hajo Adam & Adam D Galinsky (2010) "When in rome... learn why the romans do what they do: How multicultural learning experiences facilitate creativity", *Personality and Social Psychology Bulletin*, 36 (6), p. 731–741.

Mill, J.S. (1871) *Principles of Political Economy: With Some of Their Applications to Social Philosophy*, Principles of Political Economy: With Some of Their Applications to Social Philosophy, Longmans, Green, Reader, and Dyer.

Mokyr, Joel (2002) *The Gifts of Athena: Historical Origins of the Knowledge Economy*, Princeton, NJ, Princeton University Press.

Moser, Petra (2013) "Patents and Innovation: Evidence from Economic History", *The Journal of Economic Perspectives*, 27 (1), pp. 23–44.

Moser, Petra & Shmuel San (2020) "Immigration, Science, and Invention. Lessons from the Quota Acts", *SSRN Electronic Journal*.

Moser, Petra, Alessandra Voena & Fabian Waldinger (2014) "German Jewish Émigrés and US Invention", *American Economic Review*, 104 (10), pp. 3222–3255.

Muthukrishna, Michael & Joseph Henrich (2016) "Innovation in the collective brain", *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371 (1690), pp. 1–14.

Nguyen, Kieu-Trang (2021) "Trust and innovation within the firm: Evidence from matched CEO-Firm data", *Working paper*.

Nisbett, Richard E (2003) *The Geography of Thought: How Asians and Westerners Think Differently– and Why*, New York, Free Press.

Nunn, Nathan (2021) "History as evolution", *The handbook of historical economics*, Elsevier, pp. 41–91.

Ottaviano, Gianmarco IP & Giovanni Peri (2006) "The economic value of cultural diversity: Evidence from US cities", *Journal of Economic Geography*, 6 (1), pp. 9–44.

Packalen, Mikko & Jay Bhattacharya (2015) "Cities and ideas".

Philips, Lawrence (1990) "Hanging on the metaphone", *Computer Language*, 7 (12), pp. 39–43.

Raz, Itzchak T. (2024) "Soil Heterogeneity and the Formation of Close-knit Communities", *Working paper*.

Ruggles, Steven, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Evan Roberts, Megan Schouweiler & Matthew Sobek (2021) *IPUMS Ancestry Full Count Data: Version 3.0 [Dataset]*, Minneapolis, MN, IPUMS.

Rustagi, Devesh (2024) "Market exposure, civic values, and rules".

Schimmelpfennig, Robin, Layla Razek, Eric Schnell & Michael Muthukrishna (2022) "Paradox of diversity in the collective brain", *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377 (1843), p. 20200316.

Schniter, Eric, Hillard S. Kaplan & Michael Gurven (2023) "Cultural transmission vectors of essential knowledge and skills among tsimane forager-farmers", *Evolution and Human Behavior*, 44 (6), p. 530–540.

Schulz, Jonathan F (2022) "Kin Networks and Institutional Development", *The Economic Journal*, 132 (647), pp. 2578–2613.

Schulz, Jonathan F., Duman Bahrami-Rad, Jonathan P. Beauchamp & Joseph Henrich (2019) "The Church, Intensive Kinship, and Global Psychological Variation", *Science*, 366 (6466).

Sequeira, Sandra, Nathan Nunn & Nancy Qian (2020) "Immigrants and the Making of America", *Review of Economic Studies*, 87, pp. 382–419.

Shannon, Claude Elwood (1948) "A mathematical theory of communication", *The Bell system technical journal*, 27 (3), pp. 379–423.

Spolaore, Enrico & Romain Wacziarg (2009) "The Diffusion of Development", *The Quarterly Journal of Economics*, 124 (2), pp. 469–529.

Squicciarini, Mara P & Nico Voigtländer (2015) "Human capital and industrialization: Evidence from the age of the Enlightenment", *The Quarterly Journal of Economics*, 130 (4), pp. 1825–1883.

Stirling, Andy (2007) "A general framework for analysing diversity in science, technology and society", *Journal of The Royal Society Interface*, 4 (15), pp. 707–719, Publisher: Royal Society.

Strumsky, Deborah, Jose Lobo & Sander Van der Leeuw (2011) "Measuring the relative importance of reusing, recombining and creating technologies in the process of invention", *SFI Working Paper 2011-02–003: 23*.

Tabellini, Marco (2020) "Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration", *The Review of Economic Studies*, 87 (1), pp. 454–486.

Terry, Stephen J, Thomas Chaney, Konrad B Burchardi, Lisa Tarquinio & Tarek A Hassan (2024) "Immigration, Innovation, and Growth", *Working paper*.

Usher, A.P. (2013) *A History of Mechanical Inventions: Revised Edition*, Dover Publications.

Weitzman, Martin L. (1998) "Recombinant Growth", *The Quarterly Journal of Economics*, 63 (2).

Youn, H J, D Strumsky, L M A Bettencourt & J Lobo (2015) "Invention as a combinatorial process: Evidence from US patents", *Journal of the Royal Society Interface*, 12 (106).

For Online Publication

Appendix to "How Cultural Diversity Drives Innovation: Surnames and Patents in U.S. History"

Max Posch, Jonathan Schulz, and Joseph Henrich

A Data Sources and Construction

Surnames and baseline surname diversity measure

To construct county-level surname diversities up until the year 1940, we use the 1850, 1860, 1870, 1880, 1900, 1910, 1920, 1930, and 1940 waves of the full-count Integrated Public Use Microdata Series (IPUMS) compiled by Ruggles et al. (2021) and available on the NBER servers. For each wave, we obtain county identifiers and the variable `namelast` of all individuals. We perform the following steps to clean the surname variable. First, we transform non-ASCII characters into ASCII characters—e.g., we convert characters with accents or umlauts to the closest letter in English. Second, we convert all characters to upper case. Third, we remove all non-alphabetic characters, including all spaces (e.g., 'MAC ARTHUR' becomes 'MACARTHUR'). Fourth, we drop entries with one or fewer letters. Last, to address potential biases in surnames from misspellings and Anglicization, we apply the Philips (1990) phonetic algorithm *metaphone* to standardize name strings. This method reduces the risk that our analysis is biased by such variations. For instance, the algorithm treats phonetically similar names—like "Heinrich" and its Anglicized form "Henrich"—as identical ("HNRX"). This mitigates concerns that discrimination-induced name changes in less innovative regions could spuriously suggest a link between low surname entropy and lack of innovation. There are 10,633,834 non-standardized unique surnames in the data set, which are reduced to 1,084,159 unique surnames after standardization.

Following Terry et al. (2024), we also obtain individuals' age and year of immigration, the variables `age` and `yrimmig`, to estimate surname diversities for the midyears 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the midyear. Ideally, we would also remove all individuals who moved to the county after the midyear, but this information is unavailable.

We harmonize all historical Census data to the 1900 boundaries of US counties using the Ferrara et al. (2021) crosswalks. Specifically, we use the `M4` weights, which account for urban and rural areas and topographic suitability. We use 1900 as the reference year

because this is the first year of our panel data set in our main analysis.

Counties harmonized with very few people in a given census year may exhibit very low surname diversity (notably counties in Texas in 1900), potentially due to small sample bias. To mitigate this, we winsorize all entropic surname diversity variables at the 1% level when we run the regressions. Surname fractionalization indices are winsorized at the 2% level. We also standardize all diversity measures to have a mean of zero and a standard deviation of one in the regression analyses, and we compute county-level population size by summing over all surname stocks within a country. We winsorize this variable at the 1% level and standardize it to have a mean of zero and a standard deviation of one when we run the regressions.

Alternative surname diversity measures

We also compute alternative measures of surname diversity using two approaches. First, we interact surnames with a male indicator (`sex`) or the main categories of race (`race`) or birthplace (`bp1`). For the latter, we recode US states and territories (`bp1` codes <10000) to a single code.

Second, we create a distance-weighted measure of surname diversity by integrating the genetic distances between a surname's most common ancestral country to the US into the entropy measure. We identify the most frequent ancestral country for each phonetically grouped surname using the universe of immigrants to the US from 1850 to 1940—e.g. Germany for surnames like SXLS (Schulz), HNRX (Henrich) and PSX (Posch). This approach accurately identifies the ancestral country for approximately 47% of immigrants, for whom we know the country of birth.

We then link each surname to the genetic distance data from Spolaore and Wacziarg (2009). If no immigrants with a surname are found during this period (which occurs for about 30% of surnames), we categorize it a US surname and assign a distance of 0. We choose the genetic distance data because it serves as a common proxy for cultural distance, is well-defined at the country level, easy to link, and available for most countries. Spolaore and Wacziarg (2009) demonstrates its effectiveness in predicting diffusion of development.

Additionally, given the potentially moderate accuracy of this approach, we also compute a weighted average genetic distance for each surname instead of relying solely on the most frequent ancestral country. For example, 87% of immigrants named SXLS originated from Germany, 3% from Poland, 3% from Russia, 1.5% from Austria, etc. We weight their genetic distances to the US accordingly to get the distance measure for SXLS.

We then integrate the distances into our entropy measure by computing $-\sum dp \log p$, where d is the genetic distance. The two approaches (most frequent ancestral country and

weighted average) yield almost perfectly correlated measures ($\rho \approx 0.99$). The correlation with our baseline entropy measure is $\rho \approx 0.88$.

Ancestral-country diversity measures

To construct a measure of ancestral-country diversity, we use the most frequent ancestral country approach described above, but with phonetically ungrouped surnames instead of phonetically grouped ones. This improves the accuracy rate of identifying the ancestral country among immigrants to 68%. We then calculate the entropy of the distribution of the most frequent ancestral countries of surnames in each county. We winsorize this measure at the 1% level and standardize it to have a mean of zero and a standard deviation of one when we run the regressions. The correlation with our baseline surname entropy measure is $\rho \approx 0.41$ (see Table B6).

Construction of the instrumental variables

We build on the Burchardi et al. (2019) approach to construct an instrumental variable for surname diversity. We identify the number of individuals in a given US county i at the time of each census who immigrated to the US since the prior census and have the surname k. For the 1900 to 1930 census waves, we separate this immigration into five-year periods based on the year each migrant arrived in the US We obtain immigration flows for the following bins: 1881-1895, 1896-1900, 1901-1905, 1906-1910, 1911-1915, 1916-1920, 1921-1925, and 1926-1930. From the 1880 census wave, we count all first-generation immigrants as well as second-generation immigrants who had an immigrant father, regardless of the date of arrival in the US.

When we predict the stock of people $N^t_{k,i}$ in equation (3), we obtain negative values for 6.3% of observations. The logarithmic transformation of a negative value is undefined. To obtain Shannon entropy for counties containing $N^t_{i,k}$ with negative values, we truncate those negative values at the smallest positive value we observe in the data in a given year. The resulting predicted stock variable is highly correlated with the original variable ($\rho = 0.962$).

We also construct an instrument for population size by summing up the predicted stocks in each county for each period.

Construction of other demographic measures

We construct county-level measures for other types of diversity (race, country of birth and occupation) for each census year from 1850 to 1940. All data are taken from the full-count

IPUMS available on the NBER servers. To calculate population, we sum over all surname stocks within a country. To compute race entropy, we obtain the variable `race` and use the nine main categories. To compute origin country entropy and immigrant shares, we draw on the variable `bpl` with 188 main categories. We recode US states and territories (codes <100) to a single category. To compute occupational entropy, we draw on the variable `occ1950`, dropping all observations with value greater or equal than 979. As before, we transform the data from each period to 1900 US counties using the `M4` weights from the Ferrara et al. (2021) cross-walks and use the variables `age` and `yrimmig` to estimate these metrics for the midyears 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the midyear. We winsorize all entropy variables at the 1% level and standardize them to have a mean of zero and a standard deviation when we run the regressions.

Following the methodology in Raz (2024), we construct the strength of family ties measure from the full-count census data for all census waves from 1860 to 1940. The strength of family ties is determined by the first principal component of four underlying variables: (i) the divorce-to-marriage ratio, (ii) the share of elderly people living without a relative, (iii) the share of people living with at least one non-relative, and (iv) the mean size of families. The variables `age` and `yrimmig` are also used to estimate the strength of family ties for the midyears by removing all individuals who arrived after the midyear. We winsorize and standardize this variable when running regressions as above.

We closely follow the methodology of Logan and Parman (2017) to construct a measure of residential segregation, focusing on surnames rather than race. We identify household heads using the variable `relate` and sort the dataset using `serial`. An indicator is set to one if a neighbor (above or below on the same page, as indicated by `pageno`) has a different surname. If the line above (or below) is missing, we only consider the available line. We then aggregate these indicators to the county-surname level to determine the number of households with at least one different-surname neighbor for each surname group in each county. We also tally households for which we observe both, one, or no neighbors at the county-surname level, as well as the county population with different surname neighbors. Data transformation to 1900 US counties uses the `M4` weights from the Ferrara et al. (2021) cross-walks. Following Logan and Parman (2017), we compute segregation estimates under random assignment and complete segregation, calculating the final county-surname level segregation measure. County-level segregation estimates are then averaged across all surname groups, weighted by their population size.

We also apply this approach to origin countries to compute residential segregation of origin-country groups.

The segregation measures are constructed for all census years from 1880 to 1940, excluding midyears due to the impact of excluding post-midyear arrivals on household order.

Counties with very few people in a given census year may show anomalously high segregation values, likely due to small sample bias. To address this, we winsorize all segregation variables at the 1% level. We standardize these variables in the analyses as described above.

Finally, we compute the average years of schooling for each county in 1940 using the variable `higrade` and harmonize the data to 1900 county boundaries.

Construction of the innovation measures

We use the *Comprehensive Universe of US Patents* (CUSP) compiled by Berkes (2018). The data set contains US patents from 1836-2015 and is primarily constructed from Google Patents with supplementary information from other sources. For each patent, the data set provides inventor names and location of residence (geocoded and the 2000 county FIPS identifiers), filing and issuing years of patents, and the US Patent and Trademark Office technology classifications. We harmonize the data to 1900 US counties.

We also employ the breakthrough patent indicator created by Kelly et al. (2021). The authors use the text in patent documents to estimate patent novelty and impact. They assign a higher quality to patents that are novel in terms of cosine similarities. Patents are considered novel if they have low similarity with the existing stock of patents and are impactful in that they have high similarity with subsequent patents.

We construct innovation outcome variables at four levels: county, surname-county, inventor and patent. The county-level outcomes measure the number of patents and breakthrough patents filed by inventors residing in county i during period t, normalized by the county population in 1900. For patents filed by multiple inventors possibly from different counties, we divide the patent count by the number of inventors. We calculate county population sizes in 1900 using the full-count IPUMS and the Ferrara et al. (2021) border harmonization procedure.

In our least-squares analysis, depicted in Figure B13, we use patent issuing years rather than filing years and normalize patents by the population size in 1850 because filing years are inconsistently recorded in the CUSP dataset before 1870.

The surname-county-level outcomes count the number of patents and breakthroughs filed by inventors with surname k residing in county i during period t, normalized by the 1900 surname population within county i. Constructing these variables requires inventor surnames. The CUSP data includes inventor names as a string variable containing the

surname, first name, and sometimes middle names or initials. Identifying surnames from this variable can be challenging due to inconsistent name order. We use punctuation marks such as semicolons, colons, or commas to identify surnames. When the string variable starts with initials followed by a token of two or more characters, or when it ends with a whitespace followed by "DE", "DU", "DE LA", "DI", "DEL", "DELLA", "VAN", "VON", "LE", "LA", or "ST", we identify the surnames accordingly. For the remaining entries, we tokenize the string based on whitespace, keeping the first and last tokens, typically representing the first name and surname. To determine the surname, we compare the frequencies of all name combinations from pooled census years 1900, 1910, 1920, 1930, and 1940, identifying the surname based on the most common constellation. For example, for the tokens "JOHN" and "PETER", we identify the surname based on whether there were more individuals named "JOHN PETER" or "PETER JOHN". Finally, we clean the surname variable following the described steps.

The inventor-level outcomes count the number of patents and breakthroughs filed by inventor j during period t. The patent-level outcomes are the breakthrough indicator and the number of technology classes assigned to the patent.

For county-level observations, we winsorize innovation outcome variables at the 1% level to lessen the impact of outlier counties with a large number of patents. For surname-county observations, we winsorize the variables at 0.1% level, because few observations have a positive number of breakthrough patents filed by inventors with a specific surname in a given county during a specified period. For inventor-level observations, we winsorize the variables at 10% level, because the distribution is highly skewed, likely because of inaccuracies assigning unique inventor IDs (see below). We also present results using non-winsorized, log-transformed patent counts (see Tables B16, B18, B17, B19 and B20). We do not use the log-transformed variables in the main analysis to avoid the potential issues around logs with zero (Chen and Roth, 2023).

Construction of the inventor-level dataset

We assign unique inventor IDs to observations with the same first and last name and residing in the same county if they filed patents in either the preceding or following period (or both). These periods are grouped as follows: 1900-1904, 1905-1909, 1910-1914, 1915-1919, 1920-1924, 1925-1929, 1930-1934 and 1940-1944. For robustness, we also use the first letter of the first name instead of the full first name to assign inventor IDs. Using the full first name approach, we assign unique IDs for approximately 47% of the patents filed between 1900 and 1944, while the first-letter approach yields about 49%. Table B53 reports the estimates when we replicate the analysis using for the first-letter

approach. We acknowledge that this method is imperfect, as it lumps together individuals with common names like John Smith in big counties. To mitigate this skew, we winsorize the patent outcomes at the 10% level (see above).

B Additional Tables and Figures

Table B1: County-level summary statistics

	N	Mean	SD	Min	Max
Surname Diversity and Population					
Surname diversity	21984	9.552	0.902	7.130	11.767
Surname diversity (IV)	21984	10.635	1.032	7.827	13.301
Population	21984	30.721	50.040	1.174	367.820
Population (IV)	21984	36.098	41.520	3.415	283.595
Patents and Breakthroughs					
Patents per 1,000 people	21984	1.053	1.733	0.000	11.196
Breakthrough patents per 1,000 people	21984	0.126	0.282	0.000	1.918
Other diversities, Segregation and Family Ties					
Birth-country diversity	21873	0.487	0.512	0.003	1.954
Ancestral-country diversity	21984	2.864	0.253	2.189	3.552
Ancestral-country diversity (IV)	21984	4.768	0.338	3.219	5.283
Racial diversity	21873	0.334	0.361	0.000	1.003
Occupational diversity	21982	3.825	0.951	1.782	5.710
USPTO tech codes diversity	10353	4.323	2.611	0.000	12.264
Segregation of surname groups	13727	0.026	0.018	0.004	0.096
Segregation of ancestral-country groups	13737	1.008	0.008	1.000	1.051
Strength of family ties	21968	0.023	1.567	-5.346	2.807
Spillovers					
Surname diversity (< 100 miles)	21942	11.205	0.641	9.922	12.565
Surname diversity (IV, < 100 miles)	21942	9.177	0.523	7.746	10.606
Surname diversity (100 < 200 miles)	21942	11.472	0.590	10.398	12.566
Surname diversity (IV, 100 < 200 miles)	21942	9.492	0.461	8.554	10.639
Surname diversity (200 < 300 miles)	21942	11.602	0.528	10.513	12.542
Surname diversity (IV, 200 < 300 miles)	21942	9.624	0.433	8.735	10.628
Population (< 100 miles)	21942	1594.204	1722.360	26.672	11537.515
Population (IV, < 100 miles)	21942	1706.251	1650.026	65.373	11214.863
Population (100 < 200 miles)	21942	4320.107	3496.394	173.435	17387.047
Population (IV, 100 < 200 miles)	21942	4633.992	3442.929	343.034	17602.293
Population (200 < 300 miles)	21942	6433.896	4661.430	234.400	21431.629
Population (IV, 200 < 300 miles)	21942	6894.486	4525.089	518.537	21003.459

Notes: The table reports the number of observations, mean, standard deviation, minimum and maximum for all variables considered in the main county-level tables. The first section of the table lists summary statistics for surname diversity and population (in 1,000s of people). The second section provides summary statistics for patents and breakthrough patents filed in the subsequent 5-year period per 1,000 people. The third section contains summary statistics for birth-country diversity, ancestral-country diversity, racial diversity, occupational diversity, USPTO tech codes diversity, segregation of surname and ancestral-country groups, and strength of family ties. The fourth section reports the summary statistics for the surname diversity and population variables used in the spillover analysis.

Table B2: Surname-county, inventor and patent-level summary statistics

	N	Mean	SD	Min	Max
Panel A: Surname-county level					
Surname Diversity and Population					
Surname diversity	19,482,024	10.334	1.012	8.118	12.512
Surname diversity (IV)	19,482,024	11.516	1.111	9.139	14.017
Population	19,482,024	174.544	440.804	4.510	2,764.837
Population (IV)	19,482,024	152.471	368.978	8.212	2,623.951
Patents and Breakthroughs					
Patents per 1,000 people	19,482,024	2.092	23.858	0.000	500.000
Breakthrough patents per 1,000 people	19,482,024	0.193	3.664	0.000	95.238
Panel B: Inventor level					
Surname Diversity and Population					
Surname diversity	209,944	11.474	0.815	9.122	12.652
Surname diversity (IV)	209,944	12.643	0.859	10.189	14.564
Population	209,944	701.607	916.742	10.722	3,984.478
Population (IV)	209,944	517.423	676.921	16.939	3,047.960
Patents and Breakthroughs					
Patents	209,944	2.240	1.692	1.000	6.000
Breakthrough patents	209,944	0.259	0.438	0.000	1.000
Panel C: Patent level					
Surname Diversity and Team Size					
Surname diversity of patent	1,451,459	0.102	0.302	0.000	1.000
Surname diversity (IV)	1,451,459	12.672	0.991	9.911	14.604
Team size	1,451,459	1.122	0.358	1.000	7.000
Patents and Breakthroughs					
Technologies per patent	1,451,459	2.486	1.822	1.000	50.000
Breakthrough patent indicator	1,451,459	0.170	0.376	0.000	1.000

Notes: The table reports the number of observations, mean, standard deviation, minimum and maximum for all variables considered in the main surname-county (Panel A), inventor (Panel B) and patent-level tables (Panel C). The first section of each panel lists summary statistics for county or patent-level surname diversity and population (in 1,000s of people) or team size. The second section of each panel provides summary statistics for patents and breakthrough patents filed in the subsequent 5-year period per 1,000 people (Panel A), not normalized patents and breakthrough patents (Panel B), the breakthrough patent indicator and the number of technologies per patent (Panel C).

Table B3: Surname diversity over time of the counties with the highest and lowest surname diversity in 1940

		Population in year:		Surname entropy in year:				
		1900	1940	1900	1910	1920	1930	1940
Highest surname diverse counties in 1940								
Cook County, IL	Chicago	1,842,480	4,066,566	12.21	12.52	12.65	12.59	12.61
Kings County, NY	Broolyn	1,167,062	2,697,634	11.91	12.28	12.43	12.43	12.56
New York County, NY	Manhattan	2,056,956	3,283,844	12.16	12.44	12.49	12.44	12.55
Milwaukee County, MN	Milwaukee	331,519	766,385	12.09	12.37	12.45	12.47	12.48
Queens County, NY	Queens	153,329	1,298,365	11.64	12.05	12.19	12.34	12.46
Wayne County, MI	Detroit	347,974	2,017,092	11.95	12.28	12.51	12.44	12.46
Cuyahoga County, OH	Cleveland	437,737	1,217,402	12.01	12.35	12.45	12.44	12.44
Hudson County, NJ	Jersey City	386,474	651,688	11.74	12.12	12.34	12.33	12.43
Erie County, NY	Buffalo	432,269	797,945	11.98	12.23	12.31	12.28	12.39
Passaic County, NJ	Paterson	155,265	309,293	11.59	12.02	12.18	12.25	12.35
Sample mean		27,123	46,580	9.28	9.51	9.53	9.65	9.71
Lowest surname diverse counties in 1940								
Loving County, TX		43	289	4.09	5.69	4.28	5.89	6.01
Alpine County, CA		379	323	6.56	6.01	5.69	5.88	6.34
Hinsdale County, CO		1,608	361	8.58	7.53	6.87	6.95	6.64
Greene County, VA		6,210	5,220	7.03	6.94	6.68	7.03	6.79
Wayne County, UT		1,899	2,390	7.08	7.03	6.89	7.42	6.86
Charles City, VA		5,088	4,278	7.30	7.27	6.92	7.08	7.03
Kane County, UT		1,744	2,563	6.68	6.65	6.65	6.87	7.03
Rich County, UT		1,946	2,069	7.02	7.01	6.88	7.12	7.13
Hyde County, NC		9,187	7,864	7.26	7.17	7.08	7.16	7.15
Craig County, VA		4,286	3,769	7.20	7.44	7.14	7.09	7.18

Notes: The table lists the surname diversity (entropy) from 1900 to 1940 and population in 1900 and 1940 of the counties with the highest and lowest surname diversity in 1940 as well as the sample means. Column 2 reports the largest city within the county.

Table B4: Surname diversity over time of the counties with the highest and lowest immigrant shares in 1900

	Population in year:	Immigrant share in year:	Surname entropy in year:				
	1900	1900	1900	1910	1920	1930	1940
Counties with largest immigrant share in 1900							
Logan, ND	1,577	52.9%	7.50	9.08	9.05	8.86	8.74
Stark, ND	7,614	51.4%	9.57	10.8	10.7	10.8	10.7
Webb, TX	21,798	51.3%	8.56	8.63	8.85	8.96	9.32
Pembina, ND	17,850	50.9%	9.98	9.85	9.84	9.86	9.93
Kittson, MN	7,864	49.9%	8.63	8.77	8.76	8.81	8.81
Gogebic, MI	16,723	49.5%	10.1	10.6	10.8	10.6	10.9
Chippewa, MI	21,071	49.5%	10.1	10.4	10.3	10.4	10.5
Cavalier, ND	12,271	49.3%	9.88	10.1	9.96	9.92	9.73
Luce, MI	2,982	49.0%	8.85	9.25	9.64	9.78	9.83
Lake, MN	4,639	48.6%	8.99	9.28	9.13	8.86	8.86
Counties with lowest immigrant share in 1900							
Harris, GA	18,014	0.006%	8.55	8.61	8.61	8.48	8.86
Jackson, TN	15,010	0.007%	8.37	8.29	8.15	8.24	8.05
Johnson, KY	13,625	0.007%	7.77	8.03	8.17	8.16	8.19
Heard, GA	11,134	0.009%	8.67	8.63	8.64	8.43	8.57
Lee, GA	10,310	0.010%	8.56	8.61	8.59	8.71	8.86
Franklin, GA	17,671	0.011%	8.90	9.20	9.09	9.22	9.05
Perry, KY	8,275	0.012%	6.79	6.83	8.53	8.93	8.73
Alleghany, NC	7,731	0.013%	7.64	7.68	7.47	7.69	7.90
Owsley, KY	6,845	0.015%	7.70	7.45	7.41	7.48	7.63
Baker, GA	6,743	0.015%	8.44	8.59	8.41	8.52	8.47

Notes: The table lists the surname diversity (entropy) from 1900 to 1940, population and immigrant shares in 1900 of the U.S. counties with the highest and lowest immigrant shares in 1900.

Table B5: Correlations between baseline surname diversity and alternative surname diversity measures

Suranme fract. index	Surname, uncorrected	Surname, men	Surname, household heads	Surname, whites	Surname-race	Surname-country of birth	Genetic distance weighted
0.87	0.99	1.00	1.00	0.99	0.98	0.98	0.88

Notes: This table reports the correlations between our baseline county-level surname diversity measure (surname entropy) and (i) a surname fractionalization index, (ii) entropy of surnames that are not phonetically corrected, surname entropy among (iii) men, (iv) household heads, (v) white individuals, (vi) alternative entropy measures that interact surnames with race or country of birth, and (vii) genetic distance weighted surname entropy. An observation is a county from 1900 to 1940 (including the midyears). The sources and construction of all variables are explained in Appendix A.

Table B6: Correlations between surname diversity and other dimensions of cultural diversity

	Ancestral-country diversity	Birth-country diversity	Racial diversity	Occupational diversity
Partial Corr. (Period FE)	0.41	0.56	-0.27	0.71
Partial Corr. (State-Period FE)	0.33	0.40	0.23	0.64
Partial Corr. (County FE, State-Period FE)	0.40	0.24	0.11	0.27
Partial Corr. (Period FE, Population)	0.36	0.49	-0.33	0.61
Partial Corr. (State-Period FE, Population)	0.26	0.22	0.19	0.55
Partial Corr. (County FE, State-Period FE, Population)	0.38	0.25	0.10	0.29

Notes: This table reports conditional correlations between surname diversity and other dimensions of cultural diversity. An observation is a county from 1900 to 1940 (including the midyears). The sources and construction of all variables are explained in Appendix A.

Table B7: Controlling for immigration (county level)

	Patents per 1,000 people (mean = 1.09, sd = 1.68)				Breakthrough patents per 1,000 people (mean = 0.12, sd = 0.27)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.079***	1.068***	1.028***	1.217***	0.087***	0.085***	0.076***	0.122***
	(0.152)	(0.153)	(0.204)	(0.196)	(0.014)	(0.014)	(0.015)	(0.018)
Immigration		0.115***	0.125***	0.033		0.023**	0.028***	0.014
		(0.037)	(0.034)	(0.039)		(0.009)	(0.007)	(0.011)
Population	0.779***	0.766***	0.730***	0.338*	0.139***	0.137***	0.117***	0.042
	(0.099)	(0.091)	(0.092)	(0.187)	(0.020)	(0.019)	(0.017)	(0.031)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.641***	0.644***	0.648***	0.891***	0.032**	0.033**	0.036**	0.087***
	(0.194)	(0.193)	(0.230)	(0.230)	(0.016)	(0.016)	(0.017)	(0.020)
Immigration (IV)		0.281***	0.308***	0.255***		0.051***	0.049***	0.049***
		(0.082)	(0.065)	(0.088)		(0.016)	(0.013)	(0.017)
Population (IV)	0.628***	0.394***	0.326**	-0.205	0.135***	0.093***	0.069**	-0.025
	(0.136)	(0.140)	(0.149)	(0.175)	(0.026)	(0.029)	(0.026)	(0.027)
Panel C: Instrumental-variable estimates								
Surname diversity	1.087***	1.025***	1.030***	1.211***	0.083***	0.072***	0.074***	0.114***
	(0.164)	(0.165)	(0.205)	(0.203)	(0.016)	(0.014)	(0.014)	(0.019)
Immigration		0.351*	0.475**	0.407**		0.067*	0.085**	0.085**
		(0.188)	(0.193)	(0.165)		(0.035)	(0.035)	(0.033)
Population	0.800***	0.773***	0.644***	-0.303	0.168***	0.163***	0.111***	-0.076
	(0.120)	(0.099)	(0.105)	(0.238)	(0.026)	(0.026)	(0.021)	(0.069)
Sanderson-Windmeijer *F*-stat	183; 287	151; 47; 143	151; 35; 26	138; 25; 21	183; 287	151; 47; 143	151; 35; 26	138; 25; 21
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.705***	0.706***	0.735***	0.717***	-0.157***	-0.156***	-0.211***	-0.091**
	(0.071)	(0.071)	(0.069)	(0.078)	(0.033)	(0.033)	(0.042)	(0.045)
Immigration (IV)		0.030**	0.055***	0.029***		0.095***	0.094**	0.021
		(0.012)	(0.012)	(0.009)		(0.034)	(0.042)	(0.021)
Population (IV)	-0.025	-0.050	-0.130***	-0.174**	0.818***	0.739***	0.831***	0.481***
	(0.042)	(0.044)	(0.043)	(0.069)	(0.062)	(0.077)	(0.109)	(0.124)
	Immigration							
Surname diversity (IV)		0.118**	0.056	-0.011				
		(0.053)	(0.053)	(0.067)				
Immigration (IV)		-0.362**	-0.160	0.373**				
		(0.161)	(0.135)	(0.178)				
Population (IV)		0.504***	0.401***	0.556***				
		(0.101)	(0.078)	(0.123)				
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	19,236	19,236	19,236	19,236	19,236	19,236	19,236	19,236

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1930. The endogenous variables are county-level surname diversity and population (both in t) and the number of recent immigrants (between $t-1$ and t). Predicted number of immigrants is based on the shift-share approach described in Section 4.4. In columns 1 to 4, the dependent variable is the number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, the dependent variable is the number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B8: Controlling for immigration (county-surname level)

	Patents per 1,000 people (mean = 1.09, sd = 1.68)				Breakthrough patents per 1,000 people (mean = 0.12, sd = 0.27)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.929***	0.924***	0.940***	0.941***	0.072***	0.071***	0.071***	0.058***
	(0.156)	(0.156)	(0.160)	(0.236)	(0.022)	(0.022)	(0.022)	(0.015)
Immigration		0.583**	0.602**	0.185		0.119*	0.122*	0.092
		(0.269)	(0.276)	(0.179)		(0.062)	(0.064)	(0.079)
Population	4.173***	4.040***	4.054***	2.853***	0.628***	0.601***	0.608***	0.436***
	(0.812)	(0.726)	(0.740)	(0.362)	(0.104)	(0.097)	(0.099)	(0.083)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.482***	0.469**	0.477**	0.519***	0.036*	0.032	0.031	0.055
	(0.174)	(0.175)	(0.179)	(0.147)	(0.021)	(0.021)	(0.022)	(0.038)
Immigration (IV)		0.525***	0.527***	0.243		0.155***	0.159***	0.065
		(0.177)	(0.182)	(0.157)		(0.050)	(0.052)	(0.052)
Population (IV)	3.522***	2.952***	2.973***	0.620	0.474***	0.306**	0.309**	-0.199
	(1.126)	(1.018)	(1.043)	(0.668)	(0.142)	(0.121)	(0.123)	(0.370)
Panel C: Instrumental-variable estimates								
Surname diversity	1.150***	1.234***	1.253***	1.216***	0.101***	0.126***	0.126***	0.116***
	(0.216)	(0.230)	(0.236)	(0.268)	(0.030)	(0.041)	(0.043)	(0.038)
Immigration		-3.864	-3.904	-1.116		-1.116	-1.153	-0.172
		(5.854)	(5.983)	(1.468)		(1.774)	(1.859)	(0.285)
Population	3.569***	3.896**	3.942**	0.043	0.482***	0.577*	0.595	-0.442
	(0.679)	(1.562)	(1.624)	(1.224)	(0.090)	(0.341)	(0.364)	(0.450)
Sanderson-Windmeijer *F*-stat	212; 40	199; 1; 15	197; 1; 12	111; 1; 16	212; 40	199; 1; 15	197; 1; 12	111; 1; 16
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.544***	0.543***	0.542***	0.497***	-0.040**	-0.040**	-0.040**	-0.024***
	(0.041)	(0.041)	(0.041)	(0.044)	(0.016)	(0.017)	(0.017)	(0.005)
Immigration (IV)		0.027	0.028	0.019		-0.017	-0.018	-0.064**
		(0.019)	(0.019)	(0.015)		(0.040)	(0.040)	(0.025)
Population (IV)	0.044	0.015	0.016	-0.065	0.972***	0.991***	0.992***	0.667***
	(0.032)	(0.030)	(0.031)	(0.077)	(0.154)	(0.135)	(0.137)	(0.082)
	Immigration							
Surname diversity (IV)		0.012	0.012	0.076***				
		(0.017)	(0.018)	(0.019)				
Immigration (IV)		-0.145	-0.144	-0.200				
		(0.223)	(0.225)	(0.259)				
Population (IV)		0.240	0.246	-0.601**				
		(0.230)	(0.231)	(0.234)				
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	18,110,491	18,110,491	18,110,491	18,110,491	18,110,491	18,110,491	18,110,491	18,110,491

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1930. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname diversity and population (both in t) and the number of recent immigrants (between $t-1$ and t). Predicted number of immigrants is based on the shift-share approach described in Section 4.4. In columns 1 to 4, the dependent variable is number of patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. In columns 5 to 8, the dependent variable is number of breakthrough patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B9: Controlling for immigration (inventor level)

	Patents (mean = 2.22, sd = 1.68)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity	0.227***	0.229***	0.045*	0.045*
	(0.070)	(0.070)	(0.024)	(0.024)
Immigration		0.006		-0.005
		(0.012)		(0.004)
Population	-0.019	-0.021	-0.016	-0.015
	(0.040)	(0.039)	(0.010)	(0.010)
Panel B: Reduced-form estimates				
Surname diversity (IV)	0.080***	0.081***	0.019***	0.018***
	(0.023)	(0.024)	(0.006)	(0.006)
Immigration (IV)		0.004		-0.003
		(0.023)		(0.006)
Population (IV)	-0.008	-0.011	-0.014	-0.012
	(0.038)	(0.042)	(0.009)	(0.011)
Panel C: Instrumental-variable estimates				
Surname diversity	0.456***	0.482**	0.109***	0.109**
	(0.149)	(0.202)	(0.042)	(0.045)
Immigration		-0.087		0.002
		(0.281)		(0.046)
Population	-0.060	-0.093	-0.028***	-0.027
	(0.044)	(0.126)	(0.010)	(0.023)
Sanderson-Windmeijer *F*-stat	44; 1147	20; 1; 17	44; 1147	20; 1; 17
Panel D: First-stage estimates				
	Surname diversity		Population	
Surname diversity (IV)	0.179***	0.177***	0.026	0.035
	(0.025)	(0.025)	(0.028)	(0.028)
Immigration (IV)		-0.013*		0.044**
		(0.007)		(0.019)
Population (IV)	0.097***	0.107***	0.869***	0.834***
	(0.017)	(0.016)	(0.051)	(0.057)
	Immigration			
Surname diversity (IV)		0.013		
		(0.076)		
Immigration (IV)		-0.158		
		(0.185)		
Population (IV)		-0.171		
		(0.221)		
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	198,625	198,625	198,625	198,625

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population (both in t) and the number of recent immigrants (between $t-1$ and t). Predicted number of immigrants is based on the shift-share approach described in Section 4.4. The dependent variables are number of patents filed by the inventor in the five-year period starting in t (column 1–2) and number of breakthrough patents filed by the inventor in the five-year period starting in t (column 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B10: Placebo and dynamic effects (county level)

	Patents per 1,000 people				Breakthrough patents per 1,000 people			
	$t-5$ (1)	t (2)	$t+5$ (3)	$t+10$ (4)	$t-5$ (5)	t (6)	$t+5$ (7)	$t+10$ (8)
Panel A: Least-squares estimates								
Surname diversity	0.119	1.338***	0.797***	0.112	-0.015	0.099***	0.142***	-0.009
	(0.129)	(0.174)	(0.120)	(0.205)	(0.023)	(0.018)	(0.023)	(0.032)
Population	0.222	0.736***	-0.081	-0.164	0.049	0.099***	-0.110*	0.011
	(0.186)	(0.189)	(0.169)	(0.233)	(0.030)	(0.029)	(0.062)	(0.034)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.077	0.939***	0.566***	0.049	-0.025	0.061***	0.107***	0.000
	(0.127)	(0.236)	(0.127)	(0.143)	(0.022)	(0.019)	(0.025)	(0.017)
Population (IV)	0.140	0.205	-0.125	0.006	0.019	0.051	-0.019	0.011
	(0.131)	(0.205)	(0.127)	(0.155)	(0.024)	(0.035)	(0.041)	(0.027)
Panel C: Instrumental-variable estimates								
Surname diversity	0.154	1.392***	0.784***	0.073	-0.035	0.101***	0.149***	0.003
	(0.205)	(0.191)	(0.138)	(0.188)	(0.032)	(0.019)	(0.024)	(0.021)
Population	0.358	0.731***	0.080	0.039	0.038	0.118**	0.025	0.023
	(0.297)	(0.236)	(0.248)	(0.294)	(0.058)	(0.054)	(0.081)	(0.056)
Sanderson-Windmeijer F-stat	91; 22	149; 29	132; 12	135; 15	91; 22	149; 29	132; 12	135; 15
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.639***	0.729***	0.730***	0.719***	-0.061	-0.103**	-0.074	-0.088*
	(0.089)	(0.074)	(0.082)	(0.077)	(0.038)	(0.041)	(0.047)	(0.045)
Population (IV)	-0.079	-0.142**	-0.205**	-0.164**	0.426***	0.550***	0.447***	0.476***
	(0.067)	(0.063)	(0.079)	(0.072)	(0.104)	(0.107)	(0.131)	(0.127)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓	✓	✓
Observations	16,488	21,984	16,488	16,488	16,488	21,984	16,488	16,488

Notes: The table reports OLS, reduced-form, and IV estimates of the leads and lags of innovation outcomes on surname diversity and population for the specifications described in equation (2). Columns 1 and 5 use the five-year lag of the dependent variables. Columns 2 and 6 repeat the baseline specification (contemporaneous values of the dependent variables). Columns 3–4 and 7–8 use the five-year and ten-year lead of the dependent variables, respectively. An observation is a county in a period. Standard errors are clustered on states and reported in parentheses. Independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B11: Placebo and dynamic effects (county-surname level)

	Patents per 1,000 people				Breakthrough patents per 1,000 people			
	$t-5$ (1)	t (2)	$t+5$ (3)	$t+10$ (4)	$t-5$ (5)	t (6)	$t+5$ (7)	$t+10$ (8)
Panel A: Least-squares estimates								
Surname diversity	-0.016	1.143***	1.113***	0.629***	-0.069**	0.061***	0.149***	0.053*
	(0.274)	(0.182)	(0.219)	(0.130)	(0.032)	(0.015)	(0.033)	(0.029)
Population	2.631***	3.679***	-2.454	-2.484	0.216*	0.728***	-1.543**	0.327***
	(0.703)	(0.597)	(2.224)	(1.622)	(0.128)	(0.100)	(0.688)	(0.076)
Panel B: Reduced-form estimates								
Surname diversity (IV)	-0.196	0.675***	0.538***	0.404***	-0.064*	0.044	0.092**	0.038*
	(0.185)	(0.161)	(0.158)	(0.116)	(0.034)	(0.030)	(0.042)	(0.020)
Population (IV)	2.834*	0.958	-0.543	-2.185	0.312	0.112	-0.428	0.106
	(1.480)	(0.971)	(1.779)	(1.590)	(0.333)	(0.280)	(0.670)	(0.095)
Panel C: Instrumental-variable estimates								
Surname diversity	-0.238	1.358***	1.041***	0.627***	-0.118**	0.091**	0.148***	0.086**
	(0.362)	(0.248)	(0.203)	(0.141)	(0.058)	(0.043)	(0.036)	(0.040)
Population	4.038**	1.478	-0.346	-2.894	0.450	0.170	-0.481	0.151
	(1.859)	(1.509)	(2.048)	(1.933)	(0.445)	(0.424)	(0.736)	(0.134)
Sanderson-Windmeijer F-stat	250; 188	215; 300	156; 26	146; 70	250; 188	215; 300	156; 26	146; 70
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.462***	0.520***	0.505***	0.499***	-0.021***	-0.031***	-0.021***	-0.036***
	(0.032)	(0.046)	(0.045)	(0.046)	(0.005)	(0.007)	(0.004)	(0.010)
Population (IV)	0.045	-0.028	-0.251*	-0.072	0.705***	0.739***	0.674***	0.812***
	(0.052)	(0.059)	(0.135)	(0.078)	(0.079)	(0.090)	(0.049)	(0.162)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓	✓	✓
Observations	14,611,518	19,482,024	14,611,518	14,611,518	14,611,518	19,482,024	14,611,518	14,611,518

The table reports OLS, reduced-form, and IV estimates of the leads and lags of innovation outcomes on surname diversity and population for the specifications described in equation (8). Columns 1 and 5 use the five-year lag of the dependent variables. Columns 2 and 6 repeat the baseline specification (contemporaneous values of the dependent variables). Columns 3–4 and 7–8 use the five-year and ten-year lead of the dependent variables, respectively. An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. Standard errors are clustered on states and reported in parentheses. Independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B12: Placebo and dynamic effects (inventor level)

	Patents				Breakthrough patents			
	$t-5$	t	$t+5$	$t+10$	$t-5$	t	$t+5$	$t+10$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.247***	0.219***	0.143*	0.186***	0.035	0.043*	0.035	0.012
	(0.083)	(0.069)	(0.079)	(0.064)	(0.023)	(0.023)	(0.023)	(0.021)
Population	-0.034	-0.016	-0.027	-0.047	-0.017*	-0.011	-0.040***	-0.008
	(0.045)	(0.039)	(0.034)	(0.044)	(0.010)	(0.012)	(0.008)	(0.014)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.010	0.078***	-0.015	0.111***	-0.003	0.016**	0.008	0.014
	(0.028)	(0.022)	(0.030)	(0.031)	(0.008)	(0.006)	(0.009)	(0.009)
Population (IV)	-0.045	0.003	-0.048	-0.044	-0.017	-0.008	-0.028***	-0.010
	(0.043)	(0.038)	(0.042)	(0.040)	(0.012)	(0.010)	(0.009)	(0.011)
Panel C: Instrumental-variable estimates								
Surname diversity	0.078	0.438***	-0.058	0.410***	-0.022	0.094**	0.029	0.052
	(0.192)	(0.147)	(0.119)	(0.111)	(0.059)	(0.040)	(0.033)	(0.033)
Population	-0.062	-0.045	-0.047	-0.117**	-0.018	-0.019	-0.032***	-0.020
	(0.055)	(0.044)	(0.048)	(0.051)	(0.015)	(0.012)	(0.009)	(0.015)
Sanderson-Windmeijer F-stat	37; 973	49; 691	106; 1168	118; 696	37; 973	49; 691	106; 1168	118; 696
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.139***	0.179***	0.258***	0.263***	0.019	0.020	-0.007	-0.025
	(0.023)	(0.025)	(0.025)	(0.025)	(0.027)	(0.025)	(0.026)	(0.020)
Population (IV)	0.091***	0.095***	0.082***	0.120***	0.836***	0.867***	0.936***	0.793***
	(0.016)	(0.018)	(0.022)	(0.020)	(0.052)	(0.056)	(0.057)	(0.033)
Inventor fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Observations	178,201	209,944	176,094	159,120	178,201	209,944	176,094	159,120

Notes: The table reports OLS, reduced-form, and IV estimates of the leads and lags of innovation outcomes on surname diversity and population for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. Columns 1 and 5 use the five-year lag of the dependent variables. Columns 2 and 6 repeat the baseline specification (contemporaneous values of the dependent variables). Columns 3–4 and 7–8 use the five-year and ten-year lead of the dependent variables, respectively. An observation is an inventor residing in a given county in a period from 1900 to 1940. Standard errors are clustered on counties and reported in parentheses. The independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B13: Regional heterogeneity in the effect of surname diversity on innovation (county level)

	Patents per 1,000 people		Breakthrough patents per 1,000 people	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity × Region = Midwest	0.821***	1.352***	0.058**	0.087*
	(0.148)	(0.230)	(0.025)	(0.050)
Surname diversity × Region = Northeast	0.681	0.423***	0.173	0.180*
	(0.496)	(0.124)	(0.106)	(0.097)
Surname diversity × Region = South	1.158***	1.083***	0.099***	0.065***
	(0.282)	(0.325)	(0.017)	(0.014)
Surname diversity × Region = West	1.481***	2.267***	0.173***	0.200***
	(0.243)	(0.423)	(0.034)	(0.035)
Population × Region = Midwest	0.974***	1.299***	0.149***	0.139***
	(0.181)	(0.201)	(0.037)	(0.033)
Population × Region = Northeast	0.968***	0.519***	0.173***	0.139***
	(0.202)	(0.049)	(0.034)	(0.042)
Population × Region = South	0.349***	0.147	0.095***	0.049
	(0.126)	(0.279)	(0.021)	(0.064)
Population × Region = West	0.548**	1.780**	0.136***	0.150**
	(0.233)	(0.733)	(0.040)	(0.059)
Panel B: Reduced-form estimates				
Surname diversity (IV) × Region = Midwest	0.496***	1.119***	0.028	0.070
	(0.169)	(0.270)	(0.023)	(0.048)
Surname diversity (IV) × Region = Northeast	0.045	0.160	-0.013	-0.051*
	(0.337)	(0.144)	(0.069)	(0.029)
Surname diversity (IV) × Region = South	0.852**	0.803*	0.063***	0.049**
	(0.331)	(0.411)	(0.024)	(0.021)
Surname diversity (IV) × Region = West	1.106***	1.586***	0.093***	0.135***
	(0.255)	(0.257)	(0.030)	(0.025)
Population (IV) × Region = Midwest	0.927***	0.675*	0.147***	0.113*
	(0.200)	(0.354)	(0.036)	(0.064)
Population (IV) × Region = Northeast	0.905***	0.352**	0.191***	0.147***
	(0.208)	(0.133)	(0.039)	(0.050)
Population (IV) × Region = South	0.077	-0.396	0.076***	-0.006
	(0.108)	(0.292)	(0.023)	(0.036)
Population (IV) × Region = West	0.176	0.705**	0.108**	0.013
	(0.334)	(0.345)	(0.042)	(0.061)
Panel C: Instrumental-variable estimates				
Surname diversity × Region = Midwest	0.831***	1.464***	0.069**	0.101*
	(0.172)	(0.267)	(0.028)	(0.054)
Surname diversity × Region = Northeast	0.791	0.549***	0.136	0.000
	(0.512)	(0.146)	(0.120)	(0.055)
Surname diversity × Region = South	1.155***	1.135***	0.099***	0.073***
	(0.280)	(0.352)	(0.021)	(0.022)
Surname diversity × Region = West	1.738***	2.439***	0.190***	0.195***
	(0.300)	(0.454)	(0.034)	(0.029)
Population × Region = Midwest	0.989***	0.901**	0.156***	0.165*
	(0.225)	(0.420)	(0.042)	(0.089)
Population × Region = Northeast	0.930***	0.510**	0.198***	0.238**
	(0.189)	(0.208)	(0.028)	(0.097)
Population × Region = South	0.334***	-0.388	0.113***	0.039
	(0.108)	(0.342)	(0.019)	(0.123)
Population × Region = West	0.217	1.692**	0.133***	0.079
	(0.389)	(0.736)	(0.045)	(0.062)
Sanderson-Windmeijer *F*-stat: Surname diversity	547; 448; 569; 213	390; 101; 210; 237	547; 448; 569; 213	390; 101; 210; 237
Sanderson-Windmeijer *F*-stat: Population:	616; 109; 133; 668	183; 122; 23; 47	616; 109; 133; 668	183; 122; 23; 47
County fixed effects	✓	✓	✓	✓
Period fixed effects	✓	✓	✓	✓
County-specific linear trends		✓		✓
Observations	21,984	21,984	21,984	21,984

Notes: The table reports regional heterogeneity in the estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B14: Regional heterogeneity in the effect of surname diversity on innovation (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)	Breakthrough patents per 1,000 people (mean = 0.1, sd = 2.23)
	(1)	(2)
Panel A: Least-squares estimates		
Surname diversity × Region = Midwest	1.140***	0.125***
	(0.239)	(0.042)
Surname diversity × Region = Northeast	2.021	0.392
	(1.247)	(0.281)
Surname diversity × Region = South	0.521**	0.025
	(0.200)	(0.016)
Surname diversity × Region = West	1.655***	0.144***
	(0.513)	(0.032)
Population × Region = Midwest	3.197***	0.375***
	(0.439)	(0.088)
Population × Region = Northeast	2.530***	0.586***
	(0.485)	(0.125)
Population × Region = South	3.721	0.632***
	(2.283)	(0.233)
Population × Region = West	5.092***	0.843***
	(0.127)	(0.020)
Panel B: Reduced-form estimates		
Surname diversity (IV) × Region = Midwest	1.023***	0.115***
	(0.249)	(0.041)
Surname diversity (IV) × Region = Northeast	1.549*	0.343
	(0.805)	(0.212)
Surname diversity (IV) × Region = South	0.174	-0.013
	(0.159)	(0.015)
Surname diversity (IV) × Region = West	0.801***	0.043*
	(0.258)	(0.025)
Population (IV) × Region = Midwest	2.335**	0.256**
	(0.892)	(0.095)
Population (IV) × Region = Northeast	1.863***	0.407***
	(0.308)	(0.101)
Population (IV) × Region = South	3.698	0.670*
	(3.025)	(0.362)
Population (IV) × Region = West	7.058***	1.100***
	(0.471)	(0.061)
Panel C: Instrumental-variable estimates		
Surname diversity × Region = Midwest	1.433***	0.161***
	(0.310)	(0.054)
Surname diversity × Region = Northeast	2.882*	0.638
	(1.434)	(0.381)
Surname diversity × Region = South	0.564**	0.017
	(0.267)	(0.019)
Surname diversity × Region = West	2.383***	0.207**
	(0.748)	(0.085)
Population × Region = Midwest	2.719***	0.298***
	(0.585)	(0.090)
Population × Region = Northeast	1.863***	0.407***
	(0.295)	(0.100)
Population × Region = South	4.782*	0.799***
	(2.558)	(0.245)
Population × Region = West	4.496***	0.714***
	(0.190)	(0.022)
Sanderson-Windmeijer *F*-stat: Surname diversity	473; 144; 159; 220	473; 144; 159; 220
Sanderson-Windmeijer *F*-stat: Population:	311; 726; 94; 10035	311; 726; 94; 10035
County-Surname fixed effects	✓	✓
Surname-Period fixed effects	✓	✓
State-Period fixed effects	✓	✓
Observations	19,482,024	19,482,024

Notes: The table reports regional heterogeneity in the estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B15: Regional heterogeneity in the effect of surname diversity on innovation (inventor level)

	Patents		Breakthrough patents	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity × Region = Midwest	0.330***	0.324***	0.090***	0.086***
	(0.094)	(0.095)	(0.029)	(0.028)
Surname diversity × Region = Northeast	0.115	0.115	-0.013	-0.012
	(0.142)	(0.141)	(0.049)	(0.050)
Surname diversity × Region = South	0.069	0.052	0.053	0.041
	(0.168)	(0.170)	(0.042)	(0.040)
Surname diversity × Region = West	0.273*	0.273*	0.029	0.029
	(0.141)	(0.140)	(0.051)	(0.051)
Population × Region = Midwest		-0.052		-0.038***
		(0.059)		(0.013)
Population × Region = Northeast		0.003		0.014
		(0.098)		(0.026)
Population × Region = South		0.242		0.169
		(0.388)		(0.121)
Population × Region = West		0.036		0.021
		(0.045)		(0.017)
Panel B: Reduced-form estimates				
Surname diversity (IV) × Region = Midwest	0.097*	0.101*	0.029**	0.023*
	(0.055)	(0.060)	(0.012)	(0.012)
Surname diversity (IV) × Region = Northeast	0.059	0.059	0.005	0.005
	(0.038)	(0.037)	(0.014)	(0.014)
Surname diversity (IV) × Region = South	0.059	0.109	0.019	-0.001
	(0.091)	(0.104)	(0.019)	(0.022)
Surname diversity (IV) × Region = West	0.085***	0.096***	0.021***	0.028***
	(0.018)	(0.019)	(0.007)	(0.007)
Population (IV) × Region = Midwest		0.014		-0.022**
		(0.056)		(0.011)
Population (IV) × Region = Northeast		-0.036		0.019
		(0.067)		(0.014)
Population (IV) × Region = South		-0.995		0.394
		(1.306)		(0.303)
Population (IV) × Region = West		0.068		0.043
		(0.073)		(0.030)
Panel C: Instrumental-variable estimates				
Surname diversity × Region = Midwest	0.388*	0.348*	0.118***	0.071*
	(0.214)	(0.200)	(0.045)	(0.037)
Surname diversity × Region = Northeast	0.322	0.328	0.025	0.024
	(0.218)	(0.217)	(0.078)	(0.078)
Surname diversity × Region = South	0.167	0.238	0.053	0.025
	(0.261)	(0.264)	(0.053)	(0.054)
Surname diversity × Region = West	1.927**	1.895**	0.485**	0.491**
	(0.947)	(0.932)	(0.229)	(0.226)
Population × Region = Midwest		-0.032		-0.038***
		(0.062)		(0.013)
Population × Region = Northeast		-0.091		0.026
		(0.114)		(0.022)
Population × Region = South		-0.885		0.350
		(1.239)		(0.265)
Population × Region = West		-0.034		0.006
		(0.053)		(0.021)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,944	209,944	209,944	209,944

Notes: The table reports regional heterogeneity in the estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. Standard errors are clustered at the county level. Independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B16: Log-transformed population-normalized innovation outcomes (county level)

	Log Patents per 1,000 people (mean = 0.54, sd = 0.56)			Log Breakthrough patents per 1,000 people (mean = 0.1, sd = 0.22)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	0.411***	0.395***	0.409***	0.103***	0.098***	0.069***
	(0.064)	(0.087)	(0.058)	(0.023)	(0.029)	(0.013)
Population	0.151***	0.141***	0.115***	0.105***	0.096***	0.081***
	(0.024)	(0.023)	(0.031)	(0.019)	(0.017)	(0.018)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.279***	0.270***	0.307***	0.051*	0.049	0.045***
	(0.075)	(0.093)	(0.075)	(0.029)	(0.034)	(0.015)
Population (IV)	0.106***	0.091**	-0.030	0.099***	0.088***	0.040*
	(0.039)	(0.043)	(0.049)	(0.025)	(0.025)	(0.024)
Panel C: Instrumental-variable estimates						
Surname diversity	0.433***	0.415***	0.429***	0.102***	0.100***	0.074***
	(0.060)	(0.084)	(0.062)	(0.026)	(0.031)	(0.014)
Population	0.131***	0.129***	0.057	0.119***	0.100***	0.092**
	(0.033)	(0.027)	(0.048)	(0.027)	(0.020)	(0.034)
Sanderson-Windmeijer *F*-stat	162; 426	147; 227	149; 29	162; 426	147; 227	149; 29
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.693***	0.723***	0.729***	-0.160***	-0.231***	-0.103**
	(0.075)	(0.073)	(0.074)	(0.033)	(0.042)	(0.041)
Population (IV)	-0.009	-0.076*	-0.142**	0.837***	0.948***	0.550***
	(0.041)	(0.040)	(0.063)	(0.053)	(0.073)	(0.107)
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear trends			✓			✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in t. In columns 1 to 3, the dependent variable is log number of patents filed in the county in the five-year period starting in t divided by county population size in 1900 (plus one, to avoid dropping observations with zero patents). In columns 4 to 6, the dependent variable is log number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900 (plus one). Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B17: Log-transformed population-normalized innovation outcomes (county-surname level)

	Log Patents per 1,000 people (mean = 0.09, sd = 0.52)			Log Breakthrough patents per 1,000 people (mean = 0.01, sd = 0.20)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	0.049***	0.051***	0.058***	0.009***	0.009***	0.007***
	(0.009)	(0.009)	(0.008)	(0.003)	(0.003)	(0.002)
Population	0.209***	0.208***	0.251***	0.090***	0.091***	0.112***
	(0.022)	(0.023)	(0.074)	(0.012)	(0.012)	(0.030)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.026***	0.027***	0.036***	0.003	0.003	0.004
	(0.009)	(0.009)	(0.009)	(0.003)	(0.003)	(0.005)
Population (IV)	0.180***	0.179***	0.106	0.072***	0.073***	0.039
	(0.042)	(0.043)	(0.077)	(0.020)	(0.020)	(0.056)
Panel C: Instrumental-variable estimates						
Surname diversity	0.062***	0.064***	0.076***	0.011***	0.012***	0.010
	(0.012)	(0.012)	(0.013)	(0.004)	(0.004)	(0.007)
Population	0.180***	0.178***	0.160	0.072***	0.073***	0.058
	(0.020)	(0.020)	(0.122)	(0.011)	(0.011)	(0.086)
Sanderson-Windmeijer *F*-stat	201; 50	201; 49	215; 300	201; 50	201; 49	215; 300
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.532***	0.532***	0.520***	-0.040**	-0.040**	-0.021***
	(0.041)	(0.041)	(0.046)	(0.016)	(0.016)	(0.004)
Population (IV)	0.058*	0.060*	-0.028	0.981***	0.982***	0.674***
	(0.030)	(0.031)	(0.059)	(0.142)	(0.144)	(0.049)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects		✓	✓		✓	✓
County fixed effects			✓			✓
County-specific linear trends			✓			✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname entropy and population in t. In columns 1 to 3, the dependent variable is log number of patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900 (plus one, to avoid dropping observations with zero patents). In columns 4 to 6, the dependent variable is log number of breakthrough patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900 (plus one). Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B18: Log-transformed, not population normalized innovation outcomes (county level)

	Log Patents (mean = 2.13, sd = 1.56)			Log Breakthrough patents (mean = 0.77, sd = 1.14)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	0.640***	0.596***	0.422***	0.254***	0.210***	0.154***
	(0.034)	(0.030)	(0.041)	(0.037)	(0.026)	(0.029)
Population	0.340***	0.302***	0.215***	0.244***	0.218***	0.153***
	(0.039)	(0.036)	(0.060)	(0.041)	(0.038)	(0.044)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.437***	0.431***	0.338***	0.163***	0.119***	0.110***
	(0.051)	(0.063)	(0.044)	(0.026)	(0.027)	(0.025)
Population (IV)	0.279***	0.210***	0.015	0.211***	0.192***	0.045
	(0.050)	(0.049)	(0.046)	(0.049)	(0.049)	(0.053)
Panel C: Instrumental-variable estimates						
Surname diversity	0.710***	0.686***	0.485***	0.295***	0.235***	0.169***
	(0.041)	(0.031)	(0.043)	(0.047)	(0.032)	(0.034)
Population	0.341***	0.277***	0.152**	0.255***	0.222***	0.125
	(0.050)	(0.039)	(0.073)	(0.055)	(0.042)	(0.082)
Sanderson-Windmeijer *F*-stat	162; 426	147; 227	149; 29	162; 426	147; 227	149; 29
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.693***	0.723***	0.729***	-0.160***	-0.231***	-0.103**
	(0.075)	(0.073)	(0.074)	(0.033)	(0.042)	(0.041)
Population (IV)	-0.009	-0.076*	-0.142**	0.837***	0.948***	0.550***
	(0.041)	(0.040)	(0.063)	(0.053)	(0.073)	(0.107)
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear trends			✓			✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in year t. In columns 1 to 3, the dependent variable is log number of patents filed in county i in the five-year period starting in t (plus one, to avoid dropping observations with zero patents). In columns 4 to 6, the dependent variable is log number of breakthrough patents filed in county i in the five-year period starting in t (plus one). Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B19: Log-transformed, not population normalized innovation outcomes (county-surname level)

	Log Patents (mean = 0.03, sd = 0.21)			Log Breakthrough patents (mean = 0.01, sd = 0.09)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	0.013***	0.013***	0.013***	0.002	0.002	0.001
	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
Population	0.105***	0.105***	0.175***	0.049***	0.050***	0.100***
	(0.015)	(0.015)	(0.050)	(0.006)	(0.006)	(0.032)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.004	0.004	0.006	0.000	-0.001	-0.001
	(0.003)	(0.003)	(0.005)	(0.001)	(0.001)	(0.003)
Population (IV)	0.093***	0.093***	0.091*	0.041***	0.041***	0.057
	(0.021)	(0.022)	(0.050)	(0.009)	(0.009)	(0.042)
Panel C: Instrumental-variable estimates						
Surname diversity	0.015***	0.015***	0.018**	0.002	0.002	0.001
	(0.004)	(0.004)	(0.007)	(0.002)	(0.002)	(0.004)
Population	0.094***	0.094***	0.136*	0.041***	0.042***	0.085
	(0.013)	(0.013)	(0.081)	(0.005)	(0.005)	(0.067)
Sanderson-Windmeijer *F*-stat	201; 50	201; 49	215; 300	201; 50	201; 49	215; 300
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.532***	0.532***	0.520***	-0.040**	-0.040**	-0.021***
	(0.041)	(0.041)	(0.046)	(0.016)	(0.016)	(0.004)
Population (IV)	0.058*	0.060*	-0.028	0.981***	0.982***	0.674***
	(0.030)	(0.031)	(0.059)	(0.142)	(0.144)	(0.049)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects		✓	✓		✓	✓
County fixed effects			✓			✓
County-specific linear trends			✓			✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname entropy and population in t. In columns 1 to 3, the dependent variable is log number of patents filed by individuals with surname n residing in county i in the five-year period starting in t (plus one, to avoid dropping observations with zero patents). In columns 4 to 6, the dependent variable is log number of breakthrough patents filed by individuals with surname n residing in county i in the five-year period starting in t (plus one). Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B20: Log-transformed, not population normalized innovation outcomes (inventor level)

	Log Patents (mean = 1.11, sd = 0.58)		Log Breakthrough patents (mean = 0.25, sd = 0.49)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity	0.081***	0.081***	0.028	0.026
	(0.025)	(0.025)	(0.037)	(0.037)
Population		-0.005		-0.026*
		(0.015)		(0.014)
Panel B: Reduced-form estimates				
Surname diversity (IV)	0.028***	0.028***	0.020**	0.017*
	(0.008)	(0.008)	(0.010)	(0.010)
Population (IV)		0.000		-0.022
		(0.015)		(0.014)
Panel C: Instrumental-variable estimates				
Surname diversity	0.170***	0.160***	0.118*	0.098*
	(0.055)	(0.052)	(0.062)	(0.059)
Population		-0.017		-0.036**
		(0.017)		(0.016)
Sanderson-Windmeijer *F*-stat	44	49; 691	44	49; 691
Panel D: First-stage estimates				
	Surname diversity		Population	
Surname diversity (IV)	0.167***	0.179***		0.020
	(0.025)	(0.025)		(0.025)
Population (IV)		0.095***		0.867***
		(0.018)		(0.056)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,944	209,944	209,944	209,944

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in year t, and the dependent variables are log one plus number of patents filed by the inventor in the five-year period starting in t (column 1–2) and log one plus number of breakthrough patents filed by the inventor in the five-year period starting in t (column 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B21: Baseline vs. distance-weighted surname diversity (county level)

	Patents per 1,000 people (mean = 1.05, sd = 1.73)					Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Least-squares estimates										
Genetic distance weighted surname diversity	1.123***	0.917***	0.866***	1.097***	-0.205	0.139***	0.099***	0.090***	0.093***	0.034
	(0.109)	(0.134)	(0.175)	(0.126)	(0.155)	(0.022)	(0.014)	(0.013)	(0.018)	(0.021)
Surname diversity					1.522***					0.068***
					(0.150)					(0.024)
Population		0.706***	0.681***	0.802***	0.743***		0.140***	0.127***	0.101***	0.098***
		(0.100)	(0.106)	(0.201)	(0.190)		(0.020)	(0.018)	(0.030)	(0.029)
Panel B: Reduced-form estimates										
Genetic distance weighted surname diversity (IV)	0.867***	0.609***	0.529**	0.659***	-0.146	0.104***	0.044**	0.037*	0.043**	-0.009
	(0.161)	(0.206)	(0.244)	(0.193)	(0.135)	(0.013)	(0.018)	(0.021)	(0.016)	(0.020)
Surname diversity (IV)					1.081***					0.070***
					(0.214)					(0.023)
Population (IV)		0.602***	0.614***	0.441**	0.211		0.139***	0.126***	0.066*	0.051
		(0.132)	(0.145)	(0.197)	(0.205)		(0.024)	(0.024)	(0.034)	(0.035)
Panel C: Instrumental-variable estimates										
Genetic distance weighted surname diversity	1.324***	1.108***	1.029***	1.190***	-0.167	0.159***	0.106***	0.100***	0.088***	-0.005
	(0.147)	(0.172)	(0.243)	(0.173)	(0.281)	(0.018)	(0.016)	(0.019)	(0.019)	(0.038)
Surname diversity					1.543***					0.106***
					(0.244)					(0.036)
Population		0.677***	0.697***	0.985***	0.743***		0.164***	0.140***	0.135**	0.119**
		(0.119)	(0.110)	(0.239)	(0.241)		(0.026)	(0.021)	(0.053)	(0.054)
Sanderson-Windmeijer *F*-stat	106	176; 450	167; 342	145; 36	246; 58; 54	106	176; 450	167; 342	145; 36	246; 58; 54
Panel D: First-stage estimates										
	Genetic distance weighted surname diversity					Population				
Genetic distance weighted surname diversity (IV)	0.655***	0.639***	0.647***	0.623***	0.494***		-0.146***	-0.197***	-0.084**	-0.025*
	(0.064)	(0.087)	(0.075)	(0.068)	(0.091)		(0.029)	(0.033)	(0.031)	(0.015)
Surname diversity (IV)					0.174***					-0.078**
					(0.058)					(0.037)
Population (IV)		0.038	-0.029	-0.072	-0.109*		0.827***	0.923***	0.535***	0.551***
		(0.058)	(0.050)	(0.062)	(0.062)		(0.053)	(0.069)	(0.098)	(0.106)
	Surname diversity									
Genetic distance weighted surname diversity (IV)					-0.029					
					(0.019)					
Surname diversity (IV)					0.757***					
					(0.087)					
Population (IV)					-0.141**					
					(0.063)					
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓				✓	✓			
State-Period fixed effects			✓	✓	✓			✓	✓	✓
County-specific linear trends				✓	✓				✓	✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level genetic distance-weighted surname diversity, baseline surname diversity and population in t. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B22: Baseline vs. distance-weighted surname diversity (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)					Breakthrough patents per 1,000 people (mean = 0.1, sd = 2.23)				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Least-squares estimates										
Genetic distance weighted surname diversity	0.952***	0.707***	0.718***	0.891***	-0.112	0.113***	0.073***	0.073***	0.054***	0.020
	(0.174)	(0.136)	(0.139)	(0.145)	(0.152)	(0.030)	(0.023)	(0.023)	(0.015)	(0.027)
Surname diversity					1.246***					0.043
					(0.254)					(0.027)
Population		3.723***	3.742***	3.749***	3.685***		0.617***	0.624***	0.729***	0.727***
		(0.597)	(0.609)	(0.626)	(0.597)		(0.107)	(0.108)	(0.099)	(0.099)
Panel B: Reduced-form estimates										
Genetic distance weighted surname diversity (IV)	0.615***	0.429***	0.435***	0.517***	-0.026	0.065***	0.037	0.036	0.033*	-0.004
	(0.132)	(0.149)	(0.152)	(0.115)	(0.129)	(0.018)	(0.022)	(0.022)	(0.018)	(0.032)
Surname diversity (IV)					0.700***					0.048
					(0.215)					(0.058)
Population (IV)		3.140***	3.157***	1.162	0.960		0.479***	0.485***	0.126	0.112
		(0.947)	(0.971)	(0.927)	(0.971)		(0.156)	(0.159)	(0.263)	(0.278)
Panel C: Instrumental-variable estimates										
Genetic distance weighted surname diversity	1.177***	1.052***	1.066***	1.126***	-0.119	0.124***	0.105***	0.105***	0.075***	-0.013
	(0.210)	(0.200)	(0.204)	(0.188)	(0.302)	(0.033)	(0.030)	(0.030)	(0.025)	(0.076)
Surname diversity					1.471***					0.103
					(0.420)					(0.110)
Population		3.077***	3.090***	1.728	1.488		0.476***	0.483***	0.188	0.171
		(0.564)	(0.580)	(1.465)	(1.509)		(0.102)	(0.104)	(0.403)	(0.420)
Sanderson-Windmeijer *F*-stat	168	174; 51	173; 49	216; 251	202; 62; 365	168	174; 51	173; 49	216; 251	202; 62; 365
Panel D: First-stage estimates										
	Genetic distance weighted surname diversity					Population				
Genetic distance weighted surname diversity (IV)	0.522***	0.515***	0.515***	0.485***	0.463***		-0.037**	-0.037**	-0.017***	-0.003
	(0.040)	(0.041)	(0.041)	(0.038)	(0.063)		(0.014)	(0.014)	(0.004)	(0.005)
Surname diversity (IV)					0.029					-0.018***
					(0.059)					(0.005)
Population (IV)		0.124***	0.124***	0.006	-0.003		0.978***	0.979***	0.669***	0.674***
		(0.036)	(0.037)	(0.064)	(0.058)		(0.139)	(0.142)	(0.050)	(0.050)
	Surname diversity									
Genetic distance weighted surname diversity (IV)					0.023					
					(0.020)					
Surname diversity (IV)					0.497***					
					(0.059)					
Population (IV)					-0.030					
					(0.059)					
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects			✓	✓	✓			✓	✓	✓
County-specific linear trends				✓	✓				✓	✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level genetic distance-weighted surname diversity, baseline surname diversity and population in t. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B23: Baseline vs. distance-weighted surname diversity (inventor level)

	Patents (mean = 2.24, sd = 1.69)			Breakthrough patents (mean = 0.26, sd = 0.44)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Genetic distance weighted surname diversity	0.124**	0.127**	-0.238**	0.025	0.026	-0.039
	(0.058)	(0.059)	(0.115)	(0.020)	(0.020)	(0.034)
Surname diversity			0.461***			0.082**
			(0.138)			(0.041)
Population		-0.028	-0.007		-0.013	-0.009
		(0.041)	(0.040)		(0.012)	(0.012)
Panel B: Reduced-form estimates						
Genetic distance weighted surname diversity (IV)	0.056***	0.055**	-0.096*	0.014**	0.013**	-0.010
	(0.021)	(0.022)	(0.054)	(0.006)	(0.006)	(0.015)
Surname diversity (IV)			0.167***			0.025
			(0.056)			(0.016)
Population (IV)		-0.013	0.000		-0.010	-0.008
		(0.039)	(0.040)		(0.011)	(0.011)
Panel C: Instrumental-variable estimates						
Genetic distance weighted surname diversity	0.289**	0.262**	-0.678**	0.107**	0.064**	-0.097
	(0.123)	(0.115)	(0.283)	(0.044)	(0.031)	(0.078)
Surname diversity			1.119***			0.192*
			(0.357)			(0.102)
Population		-0.060	-0.012		-0.022*	-0.014
		(0.045)	(0.049)		(0.012)	(0.013)
Sanderson-Windmeijer *F*-stat	58	64; 723	323; 200; 807	58	64; 723	323; 200; 807
Panel D: First-stage estimates						
	Genetic distance weighted surname diversity			Population		
Genetic distance weighted surname diversity (IV)	0.194***	0.210***	0.304***		0.005	-0.105***
	(0.026)	(0.025)	(0.023)		(0.020)	(0.037)
Surname diversity (IV)			-0.104***			0.122**
			(0.034)			(0.055)
Population (IV)		0.146***	0.138***		0.863***	0.873***
		(0.020)	(0.020)		(0.055)	(0.057)
	Surname diversity					
Genetic distance weighted surname diversity (IV)			0.097***			
			(0.022)			
Surname diversity (IV)			0.087**			
			(0.036)			
Population (IV)			0.093***			
			(0.019)			
Inventor fixed effects	✓	✓	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓	✓	✓
Observations	208,450	208,450	208,450	208,450	208,450	208,450

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level genetic distance-weighted surname diversity, baseline surname diversity and population in t. Standard errors are clustered at the county level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B24: Baseline surname diversity and its components (county level)

	Patents per 1,000 people (mean = 1.05, sd = 1.73)			Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)		
	(1)	(2)	(3)	(4)	(5)	(6)
Surname diversity			1.025***			0.161***
			(0.222)			(0.031)
Log distinct surnames	0.995***	1.218***	0.354***	0.051**	0.065*	-0.071**
	(0.122)	(0.119)	(0.114)	(0.022)	(0.036)	(0.030)
Surname dispersion		0.335***	0.052		0.020	-0.025
		(0.091)	(0.063)		(0.024)	(0.016)
Population	0.747***	0.765***	0.708***	0.109***	0.110***	0.101***
	(0.187)	(0.195)	(0.186)	(0.031)	(0.031)	(0.029)
County fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓
Observations	21,984	21,983	21,983	21,984	21,983	21,983

Notes: The table reports OLS estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The independent variables are county-level entropic surname diversity (variety and dispersion), log number of distinct surnames (variety), surname dispersion and population in t. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B25: Baseline surname diversity and its components (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)			Breakthrough patents per 1,000 people (mean = 0.10, sd = 2.23)		
	(1)	(2)	(3)	(4)	(5)	(6)
Surname diversity			0.524			0.071***
			(0.372)			(0.019)
Log number distinct surnames	1.021***	1.310***	0.768	0.054***	0.055***	-0.019
	(0.204)	(0.234)	(0.490)	(0.014)	(0.017)	(0.017)
Surname dispersion		0.276***	0.102		0.001	-0.023**
		(0.088)	(0.140)		(0.010)	(0.009)
Population	3.624***	3.617***	3.593***	0.725***	0.725***	0.722***
	(0.572)	(0.581)	(0.577)	(0.099)	(0.099)	(0.098)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear trend	✓	✓	✓	✓	✓	✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The independent variables are county-level entropic surname diversity (variety and dispersion), log number of distinct surnames (variety), surname dispersion and population in t. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B26: Baseline surname diversity and its components (inventor level)

	Patents (mean = 2.24, sd = 1.69)			Breakthrough patents (mean = 0.26, sd = 0.44)		
	(1)	(2)	(3)	(4)	(5)	(6)
Surname diversity			0.294*			0.021
			(0.165)			(0.065)
Log distinct surnames	0.226***	0.244**	-0.177	0.031	0.068**	0.037
	(0.074)	(0.100)	(0.224)	(0.020)	(0.034)	(0.091)
Surname dispersion		0.016	-0.159		0.034*	0.021
		(0.071)	(0.109)		(0.021)	(0.041)
Population	-0.057	-0.056	-0.034	-0.017	-0.013	-0.012
	(0.041)	(0.041)	(0.043)	(0.013)	(0.013)	(0.013)
Inventor fixed effects	✓	✓	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓	✓	✓
Observations	208,450	208,450	208,450	208,450	208,450	208,450

Notes: The table reports OLS estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The independent variables are county-level entropic surname diversity (variety and dispersion), log number of distinct surnames (variety), surname dispersion and population in t. Standard errors are clustered at the county level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B27: Estimates of the effect of surname fractionalization on innovation (county level)

	Patents per 1,000 people (mean = 1.05, sd = 1.73)				Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname fractionalization index	0.798***	0.745***	0.702***	0.764***	0.065***	0.055***	0.050***	0.039**
	(0.145)	(0.163)	(0.203)	(0.114)	(0.010)	(0.011)	(0.012)	(0.015)
Population		0.878***	0.811***	1.028***		0.160***	0.141***	0.124***
		(0.106)	(0.111)	(0.226)		(0.020)	(0.018)	(0.029)
Panel B: Reduced-form estimates								
Surname fractionalization index (IV)	0.563***	0.464***	0.481**	0.546***	0.049***	0.030**	0.033***	0.035***
	(0.135)	(0.160)	(0.184)	(0.138)	(0.008)	(0.012)	(0.012)	(0.011)
Population (IV)		0.820***	0.757***	0.767***		0.154***	0.134***	0.088***
		(0.116)	(0.120)	(0.174)		(0.022)	(0.021)	(0.029)
Panel C: Instrumental-variable estimates								
Surname fractionalization index	1.050***	0.984***	0.989***	1.025***	0.091***	0.078***	0.082***	0.070***
	(0.108)	(0.133)	(0.173)	(0.104)	(0.015)	(0.012)	(0.011)	(0.023)
Population		1.039***	0.957***	1.708***		0.201***	0.170***	0.199***
		(0.127)	(0.122)	(0.309)		(0.024)	(0.019)	(0.052)
Sanderson-Windmeijer *F*-stat	33	32; 268	38; 200	37; 49	33	32; 268	38; 200	37; 49
Panel D: First-stage estimates								
	Surname fractionalization index				Population			
Surname fractionalization index (IV)	0.536***	0.531***	0.556***	0.570***		-0.056***	-0.071***	-0.023**
	(0.094)	(0.096)	(0.094)	(0.098)		(0.012)	(0.013)	(0.009)
Population (IV)		0.043**	0.004	0.020		0.749***	0.787***	0.437***
		(0.021)	(0.023)	(0.032)		(0.058)	(0.067)	(0.072)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname fractionalization index, defined as $1 - \sum p$, and population in t. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B28: Estimates of the effect of surname fractionalization on innovation (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)				Breakthrough patents per 1,000 people (mean = 0.10, sd = 2.23)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname fractionalization index	0.296***	0.279***	0.285***	0.450***	0.017**	0.014*	0.014*	0.010*
	(0.058)	(0.062)	(0.064)	(0.103)	(0.008)	(0.008)	(0.008)	(0.005)
Population		3.882***	3.902***	4.012***		0.634***	0.641***	0.749***
		(0.564)	(0.575)	(0.654)		(0.102)	(0.103)	(0.104)
Panel B: Reduced-form estimates								
Surname fractionalization index (IV)	-0.003	-0.006	-0.006	0.001	-0.001	-0.001	-0.001	0.000
	(0.006)	(0.006)	(0.006)	(0.006)	(0.002)	(0.002)	(0.002)	(0.001)
Population (IV)		3.355***	3.377***	1.988**		0.497***	0.504***	0.179
		(0.915)	(0.938)	(0.925)		(0.149)	(0.152)	(0.238)
Panel C: Instrumental-variable estimates								
Surname fractionalization index	-7.943	-10.106	-10.483	5.164	-2.489	-2.911	-2.696	-0.174
	(26.519)	(34.296)	(33.852)	(28.512)	(7.933)	(9.900)	(8.995)	(4.179)
Population		5.240	5.382	-1.944		1.021	1.004	0.449
		(5.928)	(6.038)	(27.971)		(1.644)	(1.534)	(4.052)
Sanderson-Windmeijer *F*-stat	0	0; 0	0; 0	0; 0	0	0; 0	0; 0	0; 0
Panel D: First-stage estimates								
	Surname fractionalization index				Population			
Surname fractionalization index (IV)	0.000	0.000	0.000	0.000		-0.001	-0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.002)		(0.000)	(0.000)	(0.000)
Population (IV)		0.166**	0.171**	0.627**		0.960***	0.961***	0.642***
		(0.079)	(0.080)	(0.254)		(0.134)	(0.136)	(0.052)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname fractionalization index, defined as $1 - \sum p$, and population in t. In columns 1 to 3, the dependent variable is number of patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. In columns 4 to 6, the dependent variable is number of breakthrough patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B29: Estimates of the effect of surname fractionalization on innovation (inventor level)

	Patents (mean = 2.24, sd = 1.69)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname fractionalization index	0.121**	0.119**	0.040***	0.039***
	(0.050)	(0.049)	(0.015)	(0.015)
Population		-0.015		-0.009
		(0.040)		(0.012)
Panel B: Reduced-form estimates				
Surname fractionalization index (IV)	0.029	0.033	0.018**	0.019***
	(0.026)	(0.027)	(0.007)	(0.007)
Population (IV)		0.037		0.010
		(0.050)		(0.012)
Panel C: Instrumental-variable estimates				
Surname fractionalization index	0.413	0.471	0.262**	0.249**
	(0.376)	(0.384)	(0.102)	(0.098)
Population		0.015		-0.003
		(0.048)		(0.012)
Sanderson-Windmeijer *F*-stat	26	31; 220	26	31; 220
Panel D: First-stage estimates				
	Surname diversity		Population	
Surname diversity (IV)	0.070***	0.076***		-0.181***
	(0.014)	(0.014)		(0.022)
Population (IV)		0.051**		0.875***
		(0.024)		(0.065)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,944	209,944	209,944	209,944

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname fractionalization index, defined as $1 - \sum p$, and population in t. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the county level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B30: Excluding controls from zero-stage specification (county level)

	Patents per 1,000 people (mean = 1.05, sd = 1.73)				Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Reduced-form estimates								
Surname diversity (IV)	0.945***	0.732***	0.689***	0.877***	0.108***	0.059***	0.057***	0.055***
	(0.178)	(0.214)	(0.249)	(0.204)	(0.014)	(0.018)	(0.020)	(0.015)
Population (IV)		0.714***	0.698***	0.472***		0.162***	0.142***	0.096***
		(0.153)	(0.164)	(0.161)		(0.027)	(0.027)	(0.030)
Panel B: Instrumental-variable estimates								
Surname diversity	1.393***	1.146***	1.098***	1.347***	0.159***	0.101***	0.099***	0.087***
	(0.171)	(0.211)	(0.269)	(0.189)	(0.017)	(0.018)	(0.021)	(0.019)
Population		0.741***	0.732***	0.938***		0.172***	0.147***	0.180***
		(0.145)	(0.150)	(0.248)		(0.026)	(0.025)	(0.053)
Sanderson-Windmeijer *F*-stat	130	127; 355	118; 323	148; 103	130	127; 355	118; 323	148; 103
Panel C: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.678***	0.672***	0.669***	0.661***		-0.052**	-0.062***	-0.015
	(0.059)	(0.067)	(0.066)	(0.068)		(0.021)	(0.021)	(0.020)
Population (IV)		0.020	-0.009	-0.029		0.932***	0.967***	0.545***
		(0.032)	(0.029)	(0.045)		(0.053)	(0.059)	(0.055)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports reduced-form and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in t. The instruments are computed from estimates of a version of equation (3) that includes only the push-pull instruments, not the fixed effects and controls. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B31: Excluding controls from zero-stage specification (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)				Breakthrough patents per 1,000 people (mean = 0.1, sd = 2.23)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Reduced-form estimates								
Surname diversity (IV)	0.714***	0.641***	0.651***	0.695***	0.075***	0.064***	0.065***	0.050***
	(0.157)	(0.162)	(0.166)	(0.135)	(0.019)	(0.020)	(0.020)	(0.019)
Population (IV)		2.594***	2.602***	0.223		0.385***	0.390***	-0.020
		(0.721)	(0.742)	(1.522)		(0.114)	(0.116)	(0.381)
Panel B: Instrumental-variable estimates								
Surname diversity	1.367***	1.221***	1.243***	1.384***	0.143***	0.121***	0.122***	0.100**
	(0.249)	(0.249)	(0.255)	(0.244)	(0.032)	(0.033)	(0.034)	(0.042)
Population		2.761***	2.765***	0.252		0.416***	0.421***	-0.048
		(0.523)	(0.543)	(2.837)		(0.085)	(0.087)	(0.715)
Sanderson-Windmeijer *F*-stat	170	286; 76	288; 74	162; 216	170	286; 76	288; 74	162; 216
Panel C: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.522***	0.520***	0.519***	0.501***		0.002	0.002	0.003
	(0.040)	(0.040)	(0.040)	(0.046)		(0.008)	(0.008)	(0.003)
Population (IV)		0.086***	0.087***	0.063*		0.902***	0.902***	0.539***
		(0.015)	(0.015)	(0.036)		(0.105)	(0.106)	(0.041)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports reduced-form and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname diversity and population in t. The instruments are computed from estimates of a version of equation (3) that includes only the push-pull instruments, not the fixed effects and controls. In columns 1 to 4, the dependent variable is number of patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. In columns 5 to 8, the dependent variable is number of breakthrough patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B32: excluding controls from zero-stage specification (inventor level)

	Patents (mean = 2.24, sd = 1.69)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Reduced-form estimates				
Surname diversity (IV)	0.086***	0.080***	0.023***	0.021**
	(0.030)	(0.031)	(0.009)	(0.009)
Population (IV)		-0.023		-0.009
		(0.035)		(0.010)
Panel B: Instrumental-variable estimates				
Surname diversity	0.313***	0.285***	0.085**	0.076**
	(0.111)	(0.105)	(0.033)	(0.032)
Population		-0.067		-0.022*
		(0.045)		(0.012)
Sanderson-Windmeijer *F*-stat	73	70; 2939	73	70; 2939
Panel C: First-stage estimates				
	Surname diversity		Population	
Surname diversity (IV)	0.276***	0.302***		0.080**
	(0.032)	(0.035)		(0.033)
Population (IV)		0.101***		0.768***
		(0.019)		(0.026)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,944	209,944	209,944	209,944

Notes: The table reports reduced-form and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in year t. The instruments are computed from estimates of a version of equation (3) that includes only the push-pull instruments, not the fixed effects and controls. The dependent variables are number of patents filed by the inventor in the five-year period starting in t (column 1–2) and number of breakthrough patents filed by the inventor in the five-year period starting in t (column 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B33: Dropping counties with highest surname diversity in 1900 (county level)

	Patents per 1,000 people			Breakthrough patents per 1,000 people		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.360***	1.367***	1.370***	0.105***	0.102***	0.100***
	(0.173)	(0.181)	(0.189)	(0.021)	(0.022)	(0.024)
Population	0.493	0.322	0.239	0.007	0.007	-0.003
	(0.341)	(0.453)	(0.604)	(0.067)	(0.089)	(0.126)
Panel B: Reduced-form estimates						
Surname diversity (IV)	1.163***	1.245***	1.234***	0.088***	0.096***	0.090***
	(0.239)	(0.242)	(0.256)	(0.020)	(0.018)	(0.020)
Population (IV)	-0.410*	-0.700***	-0.717**	-0.028	-0.060	-0.056
	(0.224)	(0.255)	(0.292)	(0.039)	(0.037)	(0.041)
Panel C: Instrumental-variable estimates						
Surname diversity	1.454***	1.489***	1.466***	0.110***	0.115***	0.107***
	(0.199)	(0.196)	(0.208)	(0.023)	(0.024)	(0.025)
Population	0.170	-0.313	-0.241	0.024	-0.049	-0.035
	(0.423)	(0.535)	(0.719)	(0.109)	(0.133)	(0.163)
Sanderson-Windmeijer *F*-stat	252; 13	291; 8	243; 5	252; 13	291; 8	243; 5
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.803***	0.836***	0.844***	-0.023	0.001	0.010
	(0.076)	(0.082)	(0.087)	(0.031)	(0.032)	(0.033)
Population (IV)	-0.320***	-0.417***	-0.453**	0.327***	0.254**	0.217*
	(0.103)	(0.145)	(0.173)	(0.097)	(0.103)	(0.109)
County fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓
Observations	20,880	19,784	17,584	20,880	19,784	17,584
Dropping top % of counties	5%	10%	20%	5%	10%	20%

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in t. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Column 1 and 5 drop the 5% of counties with the highest surname diversity in 1900; Column 2 and 6 drop the top 10%; and Column 3 and 7 drop the top 20%. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B34: Dropping counties with highest surname diversity in 1900 (county-surname level)

	Patents per 1,000 people			Breakthrough patents per 1,000 people		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.044***	1.143***	1.127***	0.041***	0.054***	0.046***
	(0.207)	(0.212)	(0.233)	(0.014)	(0.012)	(0.013)
Population	7.630*	3.197	2.889	0.932**	0.304	0.568
	(4.434)	(5.958)	(7.967)	(0.347)	(0.342)	(0.449)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.695***	0.884***	0.853***	0.029**	0.046***	0.036***
	(0.176)	(0.188)	(0.197)	(0.012)	(0.012)	(0.011)
Population (IV)	0.279	-3.201	-2.994	0.190	-0.147	-0.040
	(2.186)	(2.580)	(2.831)	(0.155)	(0.158)	(0.169)
Panel C: Instrumental-variable estimates						
Surname diversity	1.170***	1.369***	1.312***	0.051***	0.071***	0.054***
	(0.260)	(0.297)	(0.344)	(0.017)	(0.019)	(0.019)
Population	6.159	-1.808	-0.081	0.868**	-0.013	0.427
	(6.060)	(9.297)	(12.278)	(0.419)	(0.511)	(0.631)
Sanderson-Windmeijer *F*-stat	186; 10	164; 6	108; 4	186; 10	164; 6	108; 4
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.609***	0.646***	0.650***	-0.003	0.000	0.002
	(0.049)	(0.053)	(0.058)	(0.005)	(0.005)	(0.005)
Population (IV)	-1.323***	-2.022***	-2.270***	0.296***	0.239**	0.193*
	(0.400)	(0.652)	(0.820)	(0.102)	(0.112)	(0.113)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓
Observations	14,590,920	12,536,792	9,720,304	14,590,920	12,536,792	9,720,304
Dropping top % of counties	5%	10%	20%	5%	10%	20%

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname diversity and population in t. In columns 1 to 4, the dependent variable is number of patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed by individuals with surname n residing in county i in the five-year period starting in t divided by surname population size in 1900. Column 1 and 5 drop the 5% of counties with the highest surname diversity in 1900; Column 2 and 6 drop the top 10%; and Column 3 and 7 drop the top 20%. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B35: Dropping counties with highest surname diversity in 1900 (inventor level)

	Patents			Breakthrough patents		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	0.268***	0.207**	0.071	0.082***	0.085***	0.041
	(0.079)	(0.097)	(0.104)	(0.021)	(0.028)	(0.026)
Population	0.036	0.047	1.061	-0.361	-0.019	0.253
	(0.827)	(1.415)	(2.204)	(0.227)	(0.324)	(0.423)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.098	0.176*	0.204**	0.053**	0.058**	0.033
	(0.074)	(0.093)	(0.095)	(0.020)	(0.022)	(0.022)
Population (IV)	0.049	-1.018	-3.724***	-0.219	-0.208	-0.202
	(0.814)	(1.613)	(1.013)	(0.232)	(0.372)	(0.316)
Panel C: Instrumental-variable estimates						
Surname diversity	0.199	0.301*	0.352*	0.101***	0.103***	0.060
	(0.130)	(0.159)	(0.192)	(0.033)	(0.036)	(0.041)
Population	0.399	-1.256	-6.946	-0.226	-0.245	-0.333
	(1.012)	(2.153)	(4.526)	(0.392)	(0.463)	(0.674)
Sanderson-Windmeijer *F*-stat	253; 128	407; 87	246; 30	253; 128	407; 87	246; 30
Panel D: First-stage estimates						
	Surname diversity			Population		
Surname diversity (IV)	0.507***	0.543***	0.541***	-0.008	-0.011*	-0.002
	(0.046)	(0.029)	(0.036)	(0.005)	(0.006)	(0.007)
Population (IV)	-0.898	-0.216	-0.541	0.570***	0.759***	0.509
	(0.666)	(0.752)	(0.641)	(0.208)	(0.203)	(0.313)
Inventor fixed effects	✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	
Period-State fixed effects	✓	✓	✓	✓	✓	✓
✓	✓	✓	✓	✓	✓	
Observations	55,450	35,192	19,910	55,450	35,192	19,910
Dropping top % of counties	5%	10%	20%	5%	10%	20%

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in year t. The instruments are computed from estimates of a version of equation (1) that includes only the push-pull instruments, not the fixed effects and controls. The dependent variables are number of patents filed by the inventor in the five-year period starting in t (column 1–3) and number of breakthrough patents filed by the inventor in the five-year period starting in t (column 4–6). Column 1 and 5 drop the 5% of counties with the highest surname diversity in 1900; Column 2 and 6 drop the top 10%; and Column 3 and 7 drop the top 20%. Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B36: Excluding top and bottom 5% of surnames within each county (county level)

	Patents per 1,000 people (mean = 1.05, sd = 1.73)				Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.322***	1.125***	1.081***	1.338***	0.148***	0.107***	0.099***	0.099***
	(0.169)	(0.202)	(0.263)	(0.174)	(0.016)	(0.015)	(0.017)	(0.018)
Population		0.674***	0.645***	0.736***		0.139***	0.125***	0.099***
		(0.106)	(0.106)	(0.189)		(0.020)	(0.018)	(0.029)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.956***	0.698***	0.667**	0.939***	0.111***	0.046**	0.043*	0.061***
	(0.179)	(0.235)	(0.283)	(0.236)	(0.014)	(0.021)	(0.024)	(0.019)
Population (IV)		0.546***	0.521***	0.205		0.138***	0.122***	0.051
		(0.147)	(0.161)	(0.205)		(0.026)	(0.025)	(0.035)
Panel C: Instrumental-variable estimates								
Surname diversity	1.388***	1.161***	1.128***	1.392***	0.161***	0.104***	0.104***	0.101***
	(0.171)	(0.205)	(0.262)	(0.191)	(0.018)	(0.018)	(0.019)	(0.019)
Population		0.664***	0.640***	0.731***		0.166***	0.137***	0.118**
		(0.137)	(0.120)	(0.236)		(0.028)	(0.021)	(0.054)
Sanderson-Windmeijer *F*-stat	129	162; 426	147; 227	149; 29	129	162; 426	147; 227	149; 29
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.689***	0.693***	0.723***	0.729***		-0.160***	-0.231***	-0.103**
	(0.061)	(0.075)	(0.073)	(0.074)		(0.033)	(0.042)	(0.041)
Population (IV)		-0.009	-0.076*	-0.142**		0.837***	0.948***	0.550***
		(0.041)	(0.040)	(0.063)		(0.053)	(0.073)	(0.107)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	21,981	21,981	21,981	21,981	21,981	21,981	21,981	21,981

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in t. When constructing these variables, we dropped the most and least frequent 5% surnames within each county. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B37: Excluding top and bottom 5% of surnames within each county (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)				Breakthrough patents per 1,000 people (mean = 0.1, sd = 2.23)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.703***	1.036***	1.056***	1.293***	0.208***	0.092***	0.093***	0.072***
	(0.322)	(0.221)	(0.227)	(0.182)	(0.051)	(0.033)	(0.034)	(0.020)
Population		3.376***	3.397***	4.302***		0.586***	0.595***	0.761***
		(0.703)	(0.720)	(0.753)		(0.122)	(0.124)	(0.132)
Panel B: Reduced-form estimates								
Surname diversity (IV)	1.036***	0.355	0.363	0.555***	0.124***	0.000	0.000	-0.004
	(0.217)	(0.239)	(0.246)	(0.200)	(0.032)	(0.038)	(0.039)	(0.034)
Population (IV)		2.592***	2.608***	1.880*		0.470***	0.477***	0.399*
		(0.751)	(0.771)	(0.950)		(0.138)	(0.142)	(0.210)
Panel C: Instrumental-variable estimates								
Surname diversity	1.589***	1.019***	1.037***	1.124***	0.190***	0.087**	0.087**	0.044*
	(0.297)	(0.228)	(0.235)	(0.185)	(0.047)	(0.033)	(0.034)	(0.026)
Population		3.274***	3.300***	5.828**		0.594***	0.604***	1.170**
		(0.699)	(0.716)	(2.172)		(0.143)	(0.147)	(0.532)
Sanderson-Windmeijer *F*-stat	324	420; 96	425; 93	382; 40	324	420; 96	425; 93	382; 40
Panel D: First-stage estimates								
	Surname diversity				Population			
Surname diversity (IV)	0.652***	0.650***	0.649***	0.638***		-0.094***	-0.094***	-0.028**
	(0.036)	(0.039)	(0.039)	(0.043)		(0.025)	(0.025)	(0.010)
Population (IV)		0.006	0.007	-0.120**		0.790***	0.788***	0.346***
		(0.025)	(0.025)	(0.051)		(0.094)	(0.096)	(0.070)
County-Surname fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	19,482,016	19,482,016	19,482,016	19,482,016	19,482,016	19,482,016	19,482,016	19,482,016

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname entropy and population in t. When constructing these variables, we dropped the most and least frequent 5% surnames within each county. In columns 1 to 4, the dependent variable is number of patents filed by individuals with surname n residing in county in i in the five-year period starting in t divided by surname population size in 1900. In columns 5 to 8, the dependent variable is number of breakthrough patents filed by individuals with surname n residing in county in i in the five-year period starting in t divided by surname population size in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B38: Excluding top and bottom 5% of surnames within each county (inventor level)

	Patents (mean = 2.24, sd = 1.69)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity	0.267***	0.282***	0.035	0.043*
	(0.074)	(0.078)	(0.023)	(0.023)
Population		-0.040		-0.021*
		(0.043)		(0.012)
Panel B: Reduced-form estimates				
Surname diversity (IV)	0.170***	0.166***	0.027**	0.034**
	(0.050)	(0.058)	(0.014)	(0.015)
Population (IV)		0.007		-0.010
		(0.033)		(0.008)
Panel C: Instrumental-variable estimates				
Surname diversity	0.333***	0.333***	0.054*	0.059**
	(0.104)	(0.109)	(0.027)	(0.028)
Population		-0.001		-0.017
		(0.053)		(0.013)
Sanderson-Windmeijer *F*-stat	275	240; 525	275	240; 525
Panel D: First-stage estimates				
	Surname diversity		Population	
Surname diversity (IV)	0.511***	0.496***		-0.270***
	(0.031)	(0.036)		(0.045)
Population (IV)		0.023		0.671***
		(0.018)		(0.032)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,943	209,943	209,943	209,943

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9) as well as IV first-stage estimates. An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname entropy and population in year t. When constructing these variables, we dropped the most and least frequent 5% surnames within each county. The dependent variables are number of patents filed by the inventor in the five-year period starting in t (column 1–2) and number of breakthrough patents filed by the inventor in the five-year period starting in t (column 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B39: Surname diversity square (county level)

	Patents per 1,000 people (mean = 1.05, sd = 1.73)				Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.487***	1.132***	1.068***	1.513***	0.202***	0.140***	0.125***	0.147***
	(0.182)	(0.161)	(0.187)	(0.171)	(0.023)	(0.020)	(0.020)	(0.019)
Surname diversity square	0.152**	0.005	-0.009	0.115	0.049***	0.024***	0.019***	0.032***
	(0.068)	(0.068)	(0.076)	(0.083)	(0.011)	(0.009)	(0.007)	(0.008)
Population		0.668***	0.653***	0.635***		0.116***	0.110***	0.072**
		(0.120)	(0.127)	(0.195)		(0.019)	(0.019)	(0.030)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.961***	0.461***	0.396**	0.749***	0.121***	0.028	0.025	0.059**
	(0.169)	(0.146)	(0.179)	(0.171)	(0.019)	(0.020)	(0.022)	(0.026)
Surname diversity (IV) square	0.011	-0.178***	-0.192***	-0.120*	0.022**	-0.013**	-0.013**	-0.002
	(0.062)	(0.059)	(0.063)	(0.064)	(0.008)	(0.006)	(0.005)	(0.008)
Population (IV)		0.869***	0.894***	0.477**		0.162***	0.147***	0.054
		(0.155)	(0.170)	(0.198)		(0.029)	(0.029)	(0.044)
Panel C: Instrumental-variable estimates								
Surname diversity	1.560***	1.118***	1.036***	1.536***	0.215***	0.126***	0.121***	0.138***
	(0.201)	(0.231)	(0.250)	(0.224)	(0.025)	(0.030)	(0.027)	(0.034)
Surname diversity square	0.175**	-0.028	-0.065	0.094	0.055***	0.014	0.012	0.024
	(0.072)	(0.086)	(0.080)	(0.103)	(0.012)	(0.012)	(0.009)	(0.015)
Population		0.711***	0.725***	0.525*		0.142***	0.121***	0.065
		(0.179)	(0.139)	(0.277)		(0.033)	(0.025)	(0.070)
Sanderson-Windmeijer *F*-stat	200; 312	105; 117; 105	135; 132; 165	99; 77; 84	200; 312	105; 117; 105	135; 132; 165	99; 77; 84
Panel D: First-stage estimates								
	Surname diversity				Surname diversity square			
Surname diversity (IV)	0.657***	0.547***	0.581***	0.562***	-0.361***	-0.557***	-0.458***	-0.362***
	(0.035)	(0.034)	(0.038)	(0.039)	(0.071)	(0.089)	(0.088)	(0.103)
Surname diversity (IV) square	-0.068***	-0.110***	-0.100***	-0.106***	0.669***	0.595***	0.613***	0.619***
	(0.014)	(0.013)	(0.012)	(0.011)	(0.043)	(0.067)	(0.067)	(0.072)
Population (IV)		0.191***	0.119***	0.097***		0.340***	0.157	0.022
		(0.015)	(0.016)	(0.016)		(0.123)	(0.129)	(0.167)
	Population							
Surname diversity (IV)		-0.233***	-0.325***	-0.152***				
		(0.036)	(0.046)	(0.053)				
Surname diversity (IV) square		-0.054***	-0.067***	-0.031***				
		(0.012)	(0.013)	(0.012)				
Population (IV)		0.936***	1.077***	0.621***				
		(0.062)	(0.088)	(0.126)				
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear trends				✓				✓
Observations	21,984	21,984	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity, surname diversity square and population in t. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B40: Surname diversity square (county-surname level)

	Patents per 1,000 people (mean = 1.02, sd = 13.31)			Breakthrough patents per 1,000 people (mean = 0.1, sd = 2.23)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.721***	1.749***	1.840***	0.231***	0.235***	0.180***
	(0.334)	(0.340)	(0.339)	(0.060)	(0.061)	(0.045)
Surname diversity square	0.474***	0.482***	0.364***	0.086***	0.088***	0.062***
	(0.124)	(0.126)	(0.119)	(0.026)	(0.027)	(0.019)
Population	3.241***	3.256***	3.324***	0.532***	0.539***	0.668***
	(0.693)	(0.709)	(0.562)	(0.123)	(0.124)	(0.088)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.690**	0.699**	0.753***	0.083**	0.084**	0.064
	(0.258)	(0.262)	(0.275)	(0.038)	(0.039)	(0.061)
Surname diversity (IV) square	0.174*	0.176*	0.061	0.038**	0.039**	0.015
	(0.096)	(0.098)	(0.114)	(0.018)	(0.018)	(0.026)
Population (IV)	2.911***	2.927***	0.851	0.437***	0.443***	0.085
	(0.983)	(1.009)	(1.121)	(0.156)	(0.160)	(0.316)
Panel C: Instrumental-variable estimates						
Surname diversity	2.158***	2.190***	2.255***	0.295***	0.298***	0.210
	(0.501)	(0.512)	(0.824)	(0.082)	(0.084)	(0.183)
Surname diversity square	0.640***	0.651***	0.517	0.117***	0.118***	0.068
	(0.209)	(0.214)	(0.363)	(0.041)	(0.042)	(0.082)
Population	2.366***	2.373***	0.685	0.348***	0.353***	0.065
	(0.643)	(0.663)	(1.895)	(0.105)	(0.107)	(0.522)
Sanderson-Windmeijer *F*-stat	136; 314	134; 312	99; 142	136; 314	134; 312	99; 142
Panel D: First-stage estimates						
	Surname diversity			Surname diversity square		
Surname diversity (IV)	0.470***	0.470***	0.416***	-0.345***	-0.346***	-0.325***
	(0.035)	(0.036)	(0.040)	(0.058)	(0.058)	(0.053)
Surname diversity (IV) square	-0.064***	-0.065***	-0.082***	0.505***	0.504***	0.479***
	(0.013)	(0.013)	(0.012)	(0.034)	(0.034)	(0.037)
Population (IV)	0.118***	0.119***	0.114**	0.508***	0.501***	0.248*
	(0.026)	(0.026)	(0.052)	(0.078)	(0.078)	(0.134)
	Population					
Surname diversity (IV)	-0.044*	-0.044*	-0.025***			
	(0.024)	(0.024)	(0.007)			
Surname diversity (IV) square	-0.004	-0.004	-0.003			
	(0.009)	(0.009)	(0.003)			
Population (IV)	0.985***	0.986***	0.679***			
	(0.148)	(0.150)	(0.045)			
County-Surname fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects		✓	✓		✓	✓
County-specific linear trends			✓			✓
Observations	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024	19,482,024

Notes: The table reports OLS, reduced-form, and IV estimates for the specifications described in equation (8). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. The endogenous variables are county-level surname diversity, surname diversity square and population in t. In columns 1 to 4, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 5 to 8, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B41: Surname diversity square (inventor level)

	Patents (mean = 2.24, sd = 1.69)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity	0.217***	0.221***	0.042*	0.044*
	(0.074)	(0.075)	(0.025)	(0.025)
Surname diversity square	-0.002	0.002	-0.002	0.001
	(0.023)	(0.023)	(0.006)	(0.007)
Population		-0.017		-0.012
		(0.040)		(0.012)
Panel B: Reduced-form estimates				
Surname diversity (IV)	0.082***	0.083***	0.021***	0.020***
	(0.027)	(0.027)	(0.008)	(0.008)
Surname diversity (IV) square	-0.005	-0.005	-0.004	-0.003
	(0.009)	(0.009)	(0.003)	(0.003)
Population (IV)		0.005		-0.006
		(0.039)		(0.011)
Panel C: Instrumental-variable estimates				
Surname diversity	0.584***	0.542***	0.116*	0.106*
	(0.207)	(0.197)	(0.060)	(0.056)
Surname diversity square	0.099	0.123	0.009	0.015
	(0.064)	(0.084)	(0.021)	(0.025)
Population		-0.127*		-0.029
		(0.072)		(0.021)
Sanderson-Windmeijer *F*-stat	26; 31	35; 44; 239	26; 31	35; 44; 239
Panel D: First-stage estimates				
	Surname diversity		Surname diversity square	
Surname diversity (IV)	0.222***	0.242***	-0.479***	-0.412***
	(0.023)	(0.023)	(0.051)	(0.053)
Surname diversity (IV) square	-0.052***	-0.056***	0.259***	0.245***
	(0.007)	(0.007)	(0.030)	(0.031)
Population (IV)		0.118***		0.399***
		(0.018)		(0.055)
	Population			
Surname diversity (IV)		-0.021		
		(0.015)		
Surname diversity (IV) square		0.036***		
		(0.013)		
Population (IV)		0.852***		
		(0.050)		
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	209,944	209,944	209,944	209,944

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9). An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity, surname diversity square and population in year t, and the dependent variables are number of patents filed by the inventor in the five-year period starting in t (column 1–2) and number of breakthrough patents filed by the inventor in the five-year period starting in t (column 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B42: Surname vs. educational diversity and educational attainment

	Patents per 1,000 people (mean = 0.75, sd = 1.28)			Breakthrough patents per 1,000 people (mean = 0.15, sd = 0.27)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Diversity in educational attainment						
Surname diversity	0.122***		0.110***	0.023***		0.020***
	(0.028)		(0.029)	(0.006)		(0.006)
Years of schooling diversity		0.060***	0.028**		0.012***	0.006**
		(0.014)	(0.012)		(0.003)	(0.003)
Population	0.158***	0.204***	0.159***	0.029***	0.037***	0.029***
	(0.018)	(0.021)	(0.018)	(0.003)	(0.004)	(0.003)
R^2	0.421	0.412	0.422	0.332	0.325	0.333
Panel B: Educational attainment (in levels)						
Surname diversity		0.088***	0.090***		0.017***	0.018***
		(0.024)	(0.026)		(0.005)	(0.005)
Average years of schooling	0.136***	0.105***	0.151***	0.024***	0.018***	0.027***
	(0.029)	(0.024)	(0.030)	(0.005)	(0.004)	(0.005)
Surname diversity × Average years of schooling			0.114***			0.022***
			(0.028)			(0.005)
Population	0.199***	0.161***	0.125***	0.037***	0.029***	0.022***
	(0.021)	(0.018)	(0.018)	(0.004)	(0.004)	(0.004)
R^2	0.423	0.430	0.451	0.332	0.338	0.356
State fixed effects	✓	✓	✓	✓	✓	✓
Observations	2,820	2,820	2,820	2,820	2,820	2,820

Notes: The table reports OLS estimates of regressions of number of patents and breakthrough patents filed between 1940 and 1944 per 1940 population on surname diversity, population, individuals' years of schooling diversity and average years of schooling in 1940. The unit of observation is a county in 1940 harmonized to 1900 borders. Standard errors are clustered on states and reported in parentheses. Independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B43: Ancestral country-county level estimates of the effects of push-pull instruments on ancestry stocks

	No. of people in county i with ancestry county o in year:															
	1900		1905		1910		1915		1920		1925		1930		1940	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
$I^{1880}_{o,-r(i)} \times \frac{I^{1880}_{-c(o),i}}{I^{1880}_{-c(o)}}$	3.205***	3.124***	3.267***	3.252***	3.887***	3.739***	3.916***	3.648***	4.364***	4.053***	4.484***	4.143***	4.981***	4.708***	5.125***	4.863***
	(0.121)	(0.120)	(0.095)	(0.161)	(0.188)	(0.262)	(0.277)	(0.329)	(0.272)	(0.333)	(0.381)	(0.440)	(0.273)	(0.355)	(0.286)	(0.364)
$I^{1895}_{o,-r(i)} \times \frac{I^{1895}_{-c(o),i}}{I^{1895}_{-c(o)}}$	2.796***	3.776***	3.541***	4.838***	4.786***	5.765***	5.721***	5.948***	6.536***	6.723***	8.153***	8.580***	9.097***	9.975***	9.533***	10.803***
	(0.434)	(0.481)	(0.510)	(0.786)	(0.635)	(0.721)	(0.413)	(0.605)	(0.436)	(0.619)	(0.406)	(0.488)	(0.287)	(0.284)	(0.366)	(0.288)
$I^{1900}_{o,-r(i)} \times \frac{I^{1900}_{-c(o),i}}{I^{1900}_{-c(o)}}$			-2.993	-1.065	-0.465	-2.120	-4.675	-9.689	-3.842	-9.645	-14.394	-21.973	-14.400	-17.513	-13.957	-15.498
			(12.587)	(15.644)	(13.226)	(17.426)	(14.477)	(18.910)	(15.779)	(20.454)	(16.932)	(21.772)	(17.418)	(22.636)	(17.165)	(21.836)
$I^{1905}_{o,-r(i)} \times \frac{I^{1905}_{-c(o),i}}{I^{1905}_{-c(o)}}$					18.867***	20.976***	22.565***	23.549***	25.879***	26.700***	30.794***	33.231***	34.257***	35.494***	37.721***	38.884***
					(4.780)	(6.034)	(5.461)	(6.842)	(5.653)	(6.829)	(5.487)	(6.817)	(4.125)	(6.147)	(4.871)	(7.368)
$I^{1910}_{o,-r(i)} \times \frac{I^{1910}_{-c(o),i}}{I^{1910}_{-c(o)}}$							22.109**	18.560	25.430**	21.504	37.181***	33.876**	41.072***	35.021**	43.634***	38.027***
							(9.605)	(12.961)	(11.328)	(14.662)	(13.492)	(16.762)	(11.306)	(14.001)	(11.168)	(13.019)
$I^{1915}_{o,-r(i)} \times \frac{I^{1915}_{-c(o),i}}{I^{1915}_{-c(o)}}$									12.814***	12.628***	22.541***	22.884***	25.652***	25.606***	29.307***	29.755***
									(3.395)	(2.871)	(5.319)	(4.395)	(3.907)	(3.325)	(4.657)	(3.954)
$I^{1920}_{o,-r(i)} \times \frac{I^{1920}_{-c(o),i}}{I^{1920}_{-c(o)}}$											9.107	8.782	10.159*	7.603*	17.836**	15.182***
											(11.877)	(9.171)	(5.689)	(4.355)	(7.107)	(5.344)
$I^{1925}_{o,-r(i)} \times \frac{I^{1925}_{-c(o),i}}{I^{1925}_{-c(o)}}$													28.357***	27.532***	34.635***	34.147***
													(2.221)	(3.398)	(3.272)	(4.152)
$I^{1930}_{o,-r(i)} \times \frac{I^{1930}_{-c(o),i}}{I^{1930}_{-c(o)}}$															-32.971***	-28.977***
															(9.814)	(7.604)
Constant	0.392*		0.393		-0.019		-0.693***		-1.128***		-1.763***		-2.414***		-2.383***	
	(0.212)		(0.277)		(0.236)		(0.244)		(0.281)		(0.387)		(0.244)		(0.276)	
	F-stat	360	337	13550	789	16329	1586	6370	2214	6636	1948	4233	1297	15059	1246	59658
1492																
Observations	72,905	72,905	76,156	76,156	77,160	77,160	77,711	77,711	77,978	77,978	79,740	79,740	79,934	79,934	81,602	81,602
R^2	0.775	0.881	0.757	0.867	0.795	0.893	0.778	0.878	0.790	0.887	0.750	0.860	0.793	0.886	0.789	0.885
Neighboring ancestral countries-County fixed effects		✓		✓		✓		✓		✓		✓		✓		✓
Ancestral country-Census division fixed effects		✓		✓		✓		✓		✓		✓		✓		✓
I^{τ}_{i}/I^{τ} controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: This table reports OLS estimates for a version of the specification described in equation (3), corresponding to step 1 of the instrument construction for ancestral-country diversity. The dependent variable is the number of individuals from a given ancestral country residing in a given U.S. county in a period from 1900 to 1940. The table includes the F-statistic for the Wald test for the joined nullity of the push-pull instruments. Standard errors clustered at the surname level. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B44: Surname diversity vs. ancestral-country diversity (county level)

	Patents per 1,000 people			Breakthrough patents per 1,000 people		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.154***	1.134***	1.445***	0.105***	0.105***	0.110***
	(0.214)	(0.262)	(0.183)	(0.016)	(0.019)	(0.019)
Ancestral-country diversity	-0.057	-0.110**	-0.199***	0.005	-0.012	-0.020*
	(0.059)	(0.054)	(0.055)	(0.011)	(0.009)	(0.012)
Population	0.680***	0.652***	0.742***	0.139***	0.126***	0.100***
	(0.105)	(0.107)	(0.190)	(0.020)	(0.019)	(0.029)
Panel B: Reduced-form estimates						
Surname diversity (IV)	0.910***	0.893***	0.996***	0.113***	0.108***	0.086***
	(0.149)	(0.185)	(0.194)	(0.016)	(0.018)	(0.018)
Ancestral-country diversity (IV)	-0.095	-0.077	-0.002	-0.037***	-0.032***	-0.014
	(0.099)	(0.106)	(0.053)	(0.009)	(0.010)	(0.011)
Population (IV)	0.323***	0.253***	0.071	0.073***	0.057***	0.019
	(0.090)	(0.092)	(0.104)	(0.013)	(0.012)	(0.020)
Panel C: Instrumental-variable estimates						
Surname diversity	1.197***	1.175***	1.510***	0.117***	0.126***	0.116***
	(0.202)	(0.215)	(0.283)	(0.023)	(0.025)	(0.042)
Ancestral-country diversity	-0.096	-0.084	-0.182	-0.052*	-0.069	-0.058
	(0.247)	(0.375)	(0.219)	(0.028)	(0.043)	(0.044)
Population	0.686**	0.611**	0.585	0.197***	0.160***	0.214
	(0.291)	(0.273)	(0.735)	(0.036)	(0.034)	(0.155)
Sanderson-Windmeijer *F*-stat	197; 171; 67	150; 100; 70	162; 98; 7	197; 171; 67	150; 100; 70	162; 98; 7
Panel D: First-stage estimates						
	Surname diversity			Ancestral-country diversity		
Surname diversity (IV)	0.640***	0.648***	0.641***	0.130***	0.167***	0.210***
	(0.053)	(0.056)	(0.063)	(0.044)	(0.059)	(0.069)
Ancestral-country diversity (IV)	0.032	0.036*	0.032**	0.118***	0.120***	0.128***
	(0.022)	(0.020)	(0.013)	(0.029)	(0.027)	(0.035)
Population (IV)	0.061***	0.022	0.019	0.240***	0.110***	0.095***
	(0.022)	(0.020)	(0.017)	(0.028)	(0.024)	(0.026)
	Population					
Surname diversity (IV)	0.227***	0.237***	0.113***			
	(0.030)	(0.035)	(0.013)			
Ancestral-country diversity (IV)	-0.178***	-0.179***	-0.046**			
	(0.022)	(0.025)	(0.020)			
Population (IV)	0.397***	0.387***	0.102**			
	(0.033)	(0.046)	(0.043)			
County fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear trends			✓		✓	
Observations	21,984	21,984	21,984	21,984	21,984	21,984

Notes: The table reports OLS estimates for the specifications described in equation (2). An observation is a county in a period from 1900 to 1940. The endogenous variables are county-level ancestral-country diversity, surname diversity and population in t. In columns 1 to 3, the dependent variable is number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 4 to 6, it is number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance 5%, and 10% levels.

Table B45: Share of patents with different inventor surname characteristics

Patents with	Number	Share (%)
1 inventor	1,289,035	88.81
> 1 inventors and 1 distinct surname	14,963	1.03
> 1 inventors and > 1 distinct surnames	147,461	10.16

Notes: The table reports the number of patents with (1) a single patentee, (2) multiple patentees who share a single surname and (3) multiple patentees who do not share a single surname.

Table B46: Alternative measures of surname diversity of patents

	Surname diversity of patent		Log distinct surnames of patent		Surname dispersion of patent		Genetic distance weighted surname diversity of patent	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.020	0.011*	0.021	0.012*	0.020	0.012*	0.020	0.010
	(0.019)	(0.006)	(0.019)	(0.006)	(0.019)	(0.006)	(0.019)	(0.006)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.014*	0.008***	0.014	0.009***	0.014*	0.008***	0.014*	0.008***
	(0.008)	(0.003)	(0.009)	(0.003)	(0.008)	(0.003)	(0.008)	(0.003)
Panel C: Instrumental-variable estimates								
Surname diversity	0.067*	0.040***	0.066*	0.041***	0.067*	0.040***	0.067*	0.040***
	(0.039)	(0.013)	(0.040)	(0.013)	(0.039)	(0.013)	(0.039)	(0.013)
Sanderson-Windmeijer *F*-stat	40	40	40	40	40	40	40	40
Panel D: First-stage estimates								
	Surname diversity							
Surname diversity (IV)	0.210***	0.210***						
	(0.033)	(0.033)						
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Patent technology field-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Team size fixed effects		✓		✓		✓		✓
Observations	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459

Notes: The table reports least squares, reduced-form and IV estimates for the specifications described in equation (10) but with county-level surname diversity as the main independent variable instead of patent-level surname diversity. An observation is a patent with inventors residing in a given county in a five-year period from 1900 to 1940. Observations are weighted by the inverse of the number of authors multiplied by the weight used to assign patents to counties within their 1900 borders. Standard errors are clustered on counties and reported in parentheses. All variables are standardized to have a mean of zero and a unit variance. The sources and construction of all variables are detailed in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B47: Alternative measures of surname diversity of patents with 2+ inventors

	Surname diversity of patent		Log distinct surnames of patent		Surname dispersion of patent		Genetic distance weighted surname diversity of patent	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.111**	0.106*	0.101*	0.101*	0.118*	0.117*	0.078	0.077
	(0.056)	(0.056)	(0.054)	(0.054)	(0.061)	(0.061)	(0.053)	(0.053)
Panel B: Reduced-form estimates								
Surname diversity (IV)	0.068**	0.068***	0.061**	0.065***	0.075***	0.075***	0.057**	0.059**
	(0.027)	(0.026)	(0.026)	(0.024)	(0.028)	(0.028)	(0.025)	(0.025)
Panel C: Instrumental-variable estimates								
Surname diversity	0.332**	0.334***	0.298**	0.316***	0.368***	0.367***	0.277**	0.286**
	(0.134)	(0.124)	(0.128)	(0.117)	(0.138)	(0.138)	(0.120)	(0.116)
Sanderson-Windmeijer *F*-stat	35	35	35	35	35	35	35	35
Panel D: First-stage estimates								
	Surname diversity							
Surname diversity (IV)	0.205***	0.205***						
	(0.035)	(0.035)						
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Patent technology field-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Team size fixed effects		✓		✓		✓		✓
Observations	162,424	162,424	162,424	162,424	162,424	162,424	162,424	162,424

Notes: The table reports least squares, reduced-form and IV estimates for the specifications described in equation (10) but with county-level surname diversity as the main independent variable instead of patent-level surname diversity. An observation is a patent with inventors residing in a given county in a five-year period from 1900 to 1940. Observations are weighted by the inverse of the number of authors multiplied by the weight used to assign patents to counties within their 1900 borders. Standard errors are clustered on counties and reported in parentheses. All variables are standardized to have a mean of zero and a unit variance. The sources and construction of all variables are detailed in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B48: Patent-level estimates of the effect of genetic distance-weighted surname diversity on innovation

	Breakthrough patent indicator (mean = 0.17, sd = 0.38)			Technologies per patent (mean = 2.49, sd = 1.82)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Genetic distance weighted surname diversity of patent	0.014***	0.009***	0.009***	0.031***	0.018***	0.017***
	(0.001)	(0.001)	(0.001)	(0.005)	(0.005)	(0.005)
Panel B: Reduced-form estimates						
Genetic distance weighted surname diversity (IV)	0.010**	0.007*		0.044***	0.034***	
	(0.004)	(0.004)		(0.013)	(0.011)	
Panel C: Instrumental-variable estimates						
Genetic distance weighted surname diversity of patent	1.279*	0.966		5.562**	4.730**	
	(0.674)	(0.594)		(2.544)	(2.407)	
Sanderson-Windmeijer *F*-stat	8	7		8	7	
Panel D: First-stage estimates						
	Genetic distance weighted surname entropy of patent					
Genetic distance weighted surname entropy (push-pull IV)	0.008***	0.007***				
	(0.003)	(0.003)				
Team size fixed effects	✓	✓	✓	✓	✓	✓
County fixed effects	✓	✓		✓	✓	
State-Period fixed effects	✓	✓		✓	✓	
Patent technology field-Period fixed effects		✓	✓		✓	✓
County-Period fixed effects			✓			✓
Observations	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459

Notes: The table reports least squares, reduced-form and IV estimates for the specifications described in equation (10). An observation is a patent with inventors residing in a given county in a period from 1900 to 1940. If there are multiple inventors who reside in different counties, the patent is assigned to all counties and the observation is weighted by 1 / number of inventors. The endogenous variables is genetic distance-weighted surname diversity of patent. The dependent variables are a breakthrough patent indicator in columns 1 to 3 and the number of technology classes per patent in columns 4 to 6. Columns 2–3 and 5–6 add one of six broad patent technology fields (chemical, computers, drugs, electrical, mechanical, other) fixed effects interacted with period fixed effects. Standard errors are clustered clustered at the county level. Columns 3 and 6 additionally include country-period fixed effects. All diversity variables are standardized to have a mean of zero and a unit variance. The sources and construction of all variables are detailed in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B49: Patent-level estimates of the effect of ancestral-country diversity on innovation

	Breakthrough patent indicator (mean = 0.17, sd = 0.38)			Technologies per patent (mean = 2.49, sd = 1.82)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Ancestral-country diversity of patent	0.005***	0.003***	0.003***	0.019***	0.014***	0.013***
	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)	(0.004)
Panel B: Least-squares estimates						
Ancestral-country diversity of patent	0.001	0.001	0.001	0.012**	0.011**	0.011**
	(0.001)	(0.001)	(0.001)	(0.005)	(0.005)	(0.005)
Surname diversity of patent	0.014***	0.009***	0.009***	0.022***	0.010	0.010
	(0.001)	(0.001)	(0.001)	(0.007)	(0.006)	(0.006)
Panel C: Reduced-form estimates						
Ancestral-country diversity (IV)	0.001	0.005		-0.041*	-0.031*	
	(0.007)	(0.006)		(0.022)	(0.018)	
Panel D: First-stage estimates						
	Ancestral-country diversity of patent			Surname diversity of patent		
Ancestral-country diversity (IV)	0.000	0.000		0.000	0.000	
	(0.005)	(0.005)		(0.004)	(0.004)	
Team size fixed effects	✓	✓	✓	✓	✓	✓
County fixed effects	✓	✓		✓	✓	
State-Period fixed effects	✓	✓		✓	✓	
Patent technology field-Period fixed effects		✓	✓		✓	✓
County-Period fixed effects			✓			✓
Observations	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459

Notes: The table reports least squares (Panels A and B), reduced-form (Panel C) and first-stage estimates (Panel D) for versions of the specifications described in equation (10), replacing surname diversity of patent with ancestral-country diversity of patent. An observation is a patent with inventors residing in a given county in a five-year period from 1900 to 1940. If there are multiple inventors who reside in different counties, the patent is assigned to all counties and the observation is weighted by 1 / number of inventors. The dependent variables are breakthrough patent indicator in Columns 1 to 3 and the number of technology classes per patent in Columns 4 to 6. Columns 2 and 5 add six broad patent technology fields (chemical, computers, drugs, electrical, mechanical and 'other') fixed effects interacted with period fixed effects. Columns 3 and 6 additionally include country-period fixed effects. Standard errors are clustered clustered at the county level. All diversity variables are standardized to have a mean of zero and a unit variance. The sources and construction of all variables are detailed in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B50: Surnames clustering (full sample)

	Years of schooling	Occupation				Country of origin		Region of origin	USPTO tech class	
Sample:	All			Immigrants				Germans	Inventors	
Year:	1940	1880	1940	1880	1940	1880	1940	1880	1880-9	1940-9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Surname	0.054	0.117	0.045	0.097	0.065	0.393	0.227	0.268	0.092	0.068
Country of birth	0.055	0.120	0.043	0.096	0.060					
Country of origin	0.051	0.112	0.041	0.087	0.048				0.004	0.003
U.S. county of residence	0.065	0.171	0.071	0.153	0.080	0.288	0.130	0.269	0.014	0.017

Notes: This table reports normalized Herfindahl indices, where larger values indicate greater concentration. The indices are calculated as the average Herfindahl indices of the variable in the header computed for each value of the variables on the left. For example, we calculate the normalized Herfindahl index of occupations for each surname and then average over all surnames using the number of individuals with a given surname as weights. Column 2 (1 and 3) includes all individuals in the 1880 (1940) census. Columns 4 and 6 (5 and 7) include all immigrants in the 1880 (1940) census. Column 8 restricts the sample to German immigrants in 1880. We use the 32 subnational regional origins for German immigrants recorded by the Census (the variable `bp1d` with codes 45301 to 45362). Column 9 (10) includes all inventors of patents issued from 1880 to 1889 (1940 to 1949). We use the main technology class on the patent.

Table B51: Effects on proxy measures of informational and social-behavioral mechanisms (birth-country diversity)

	Occupational diversity		USPTO tech codes diversity		Residential segregation of surname groups		Residential segregation of origin-country groups		Strength of family ties	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Surname diversity	0.311***	0.300***	0.303***	0.317***	-0.387***	-0.363***	-1.079***	-1.146***	-0.202***	-0.139***
	(0.034)	(0.035)	(0.048)	(0.054)	(0.081)	(0.085)	(0.049)	(0.048)	(0.041)	(0.039)
Birth-country diversity		0.068*		-0.063		-0.140*		0.382***		-0.375***
		(0.036)		(0.053)		(0.080)		(0.039)		(0.040)
Population	-0.052**	-0.056***	0.059**	0.061**	0.044	0.051	0.091***	0.070***	-0.088**	-0.070**
	(0.021)	(0.021)	(0.029)	(0.029)	(0.033)	(0.034)	(0.025)	(0.024)	(0.036)	(0.033)
Observations	21,982	21,982	10,353	10,353	13,727	13,727	13,727	13,727	21,968	21,968
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear trends	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The table reports OLS estimates for the the specification described in equation (2) but with occupational diversity, USPTO tech codes diversity, residential segregation of surname groups, residential segregation of ancestral-country group and strength of family ties as the outcome variables. Even columns add birth-country diversity to the specification. An observation is a county in a period from 1900 to 1940. Standard errors are clustered at the state level. All variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B52: First-stage estimates of spillover analysis

	(1)	(2)	(3)	(4)	(5)	(6)
	Surname entropy			Surname entropy (< 100 miles)		
Surname entropy (push-pull IV)	0.712***	0.710***	0.710***	0.033***	0.033***	0.032***
	(0.070)	(0.071)	(0.070)	(0.011)	(0.011)	(0.011)
Surname entropy (push-pull IV, < 100 miles)	0.095***	0.099***	0.100***	0.297***	0.297***	0.299***
	(0.029)	(0.027)	(0.026)	(0.042)	(0.041)	(0.040)
Surname entropy (push-pull IV, 100 < 200 miles)		0.038*	0.039*		0.039	0.044
		(0.021)	(0.023)		(0.028)	(0.030)
Surname entropy (push-pull IV, 200 < 300 miles)			0.013			0.035
			(0.031)			(0.027)
Population (push-pull IV)	-0.077*	-0.076*	-0.077*	-0.014	-0.014	-0.014
	(0.038)	(0.038)	(0.038)	(0.010)	(0.010)	(0.009)
Population (push-pull IV, < 100 miles)	0.047***	0.052***	0.049**	0.116**	0.115**	0.113**
	(0.015)	(0.017)	(0.018)	(0.046)	(0.047)	(0.046)
Population (push-pull IV, 100 < 200 miles)		-0.048***	-0.055***		0.002	-0.004
		(0.018)	(0.019)		(0.024)	(0.027)
Population (push-pull IV, 200 < 300 miles)			-0.047**			-0.045
			(0.019)			(0.035)
	Surname entropy (100 < 200 miles)			Surname entropy (200 < 300 miles)		
Surname entropy (push-pull IV)		-0.004	-0.003			-0.014
		(0.004)	(0.004)			(0.009)
Surname entropy (push-pull IV, < 100 miles)		0.053***	0.053***			0.008
		(0.013)	(0.013)			(0.012)
Surname entropy (push-pull IV, 100 < 200 miles)		0.157***	0.166***			0.063**
		(0.038)	(0.036)			(0.027)
Surname entropy (push-pull IV, 200 < 300 miles)			0.050**			0.152***
			(0.020)			(0.037)
Population (push-pull IV)		0.013**	0.014**			0.003
		(0.005)	(0.005)			(0.004)
Population (push-pull IV, < 100 miles)		-0.005	-0.002			0.000
		(0.013)	(0.013)			(0.012)
Population (push-pull IV, 100 < 200 miles)		0.104**	0.107**			0.025
		(0.043)	(0.044)			(0.019)
Population (push-pull IV, 200 < 300 miles)			0.027			0.159***
			(0.038)			(0.055)
	Population			Population (< 100 miles)		
Surname entropy (push-pull IV)	-0.241***	-0.242***	-0.243***	0.013*	0.011*	0.011*
	(0.040)	(0.040)	(0.040)	(0.007)	(0.007)	(0.007)
Surname entropy (push-pull IV, < 100 miles)	0.078***	0.080***	0.080***	-0.087***	-0.084***	-0.084***
	(0.019)	(0.019)	(0.018)	(0.026)	(0.025)	(0.024)
Surname entropy (push-pull IV, 100 < 200 miles)		0.041	0.035		0.006	0.007
		(0.030)	(0.029)		(0.014)	(0.015)
Surname entropy (push-pull IV, 200 < 300 miles)			-0.029			0.006
			(0.019)			(0.011)
Population (push-pull IV)	0.951***	0.952***	0.951***	0.009	0.009	0.009
	(0.077)	(0.077)	(0.077)	(0.007)	(0.007)	(0.007)
Population (push-pull IV, < 100 miles)	0.008	0.009	0.007	0.961***	0.966***	0.966***
	(0.035)	(0.035)	(0.035)	(0.032)	(0.032)	(0.033)
Population (push-pull IV, 100 < 200 miles)		-0.016	-0.021		-0.047*	-0.048
		(0.019)	(0.020)		(0.025)	(0.029)
Population (push-pull IV, 200 < 300 miles)			-0.036			-0.003
			(0.025)			(0.047)
	Population (100 < 200 miles)			Population (200 < 300 miles)		
Surname entropy (push-pull IV)		0.004	0.004			-0.005*
		(0.005)	(0.005)			(0.003)
Surname entropy (push-pull IV, < 100 miles)		0.017	0.017			-0.002
		(0.011)	(0.011)			(0.006)
Surname entropy (push-pull IV, 100 < 200 miles)		-0.117***	-0.114***			0.022*
		(0.031)	(0.030)			(0.012)
Surname entropy (push-pull IV, 200 < 300 miles)			0.022**			-0.092***
			(0.011)			(0.022)
Population (push-pull IV)		0.001	0.001			0.001
		(0.006)	(0.006)			(0.003)
Population (push-pull IV, < 100 miles)		-0.028*	-0.029*			-0.019***
		(0.014)	(0.015)			(0.007)
Population (push-pull IV, 100 < 200 miles)		0.978***	0.975***			-0.017
		(0.032)	(0.029)			(0.015)
Population (push-pull IV, 200 < 300 miles)			-0.013			0.992***
			(0.025)			(0.025)
County fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear time trends	✓	✓	✓	✓	✓	✓
Observations	21,430	21,430	21,430	21,430	21,430	21,430

Notes: The table reports first-stage estimates of the IV regressions reported in Table 9. Standard errors are clustered on states and reported in parentheses. All variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table B53: Alternative disambiguation of inventors

	Patents (mean = 2.24, sd = 1.69)		Breakthrough patents (mean = 0.26, sd = 0.44)	
	(1)	(2)	(3)	(4)
Panel A: Least-squares estimates				
Surname diversity	0.245***	0.248***	0.047**	0.048**
	(0.070)	(0.070)	(0.023)	(0.023)
Population		0.041		0.008
		(0.041)		(0.011)
Panel B: Reduced-form estimates				
Surname diversity (IV)	0.065***	0.071***	0.010*	0.012*
	(0.022)	(0.023)	(0.006)	(0.006)
Population (IV)		0.049		0.011
		(0.040)		(0.011)
Panel C: Instrumental-variable estimates				
Surname diversity	0.395***	0.403***	0.062*	0.066*
	(0.152)	(0.147)	(0.037)	(0.035)
Population		0.013		0.005
		(0.045)		(0.012)
Sanderson-Windmeijer *F*-stat	44	50; 635	44	50; 635
Panel D: First-stage estimates				
	Surname entropy		Population	
Surname diversity (IV)	0.163***	0.176***		0.018
	(0.025)	(0.025)		(0.025)
Population (IV)		0.094***		0.878***
		(0.018)		(0.059)
Inventor fixed effects	✓	✓	✓	✓
Period-State fixed effects	✓	✓	✓	✓
Observations	218,618	218,618	218,618	218,618

Notes: The table reports OLS, reduced-form and IV estimates for the specifications described in equation (9). Inventors are assigned based on the first letter of the first name and county of residence. An observation is an inventor residing in a given county in a period from 1900 to 1940. The endogenous variables are county-level surname diversity and population in year t, and the dependent variables are number of patents filed by the inventor in the five-year period starting in t (column 1–2) and number of breakthrough patents filed by the inventor in the five-year period starting in t (column 3–4). Standard errors are clustered at the county level. The independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

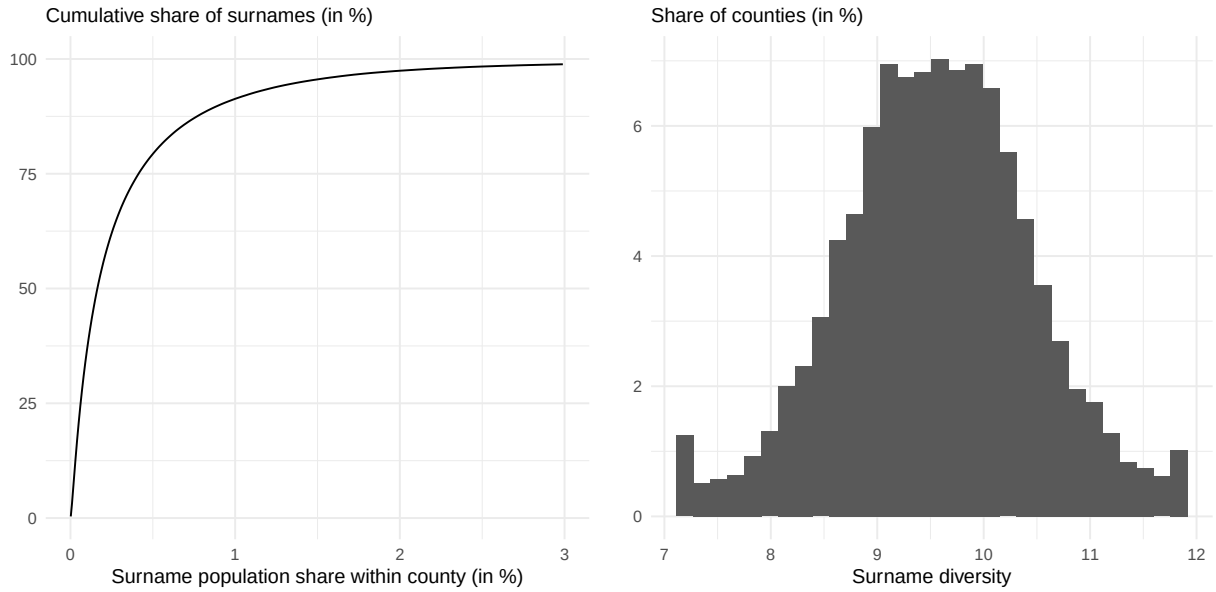
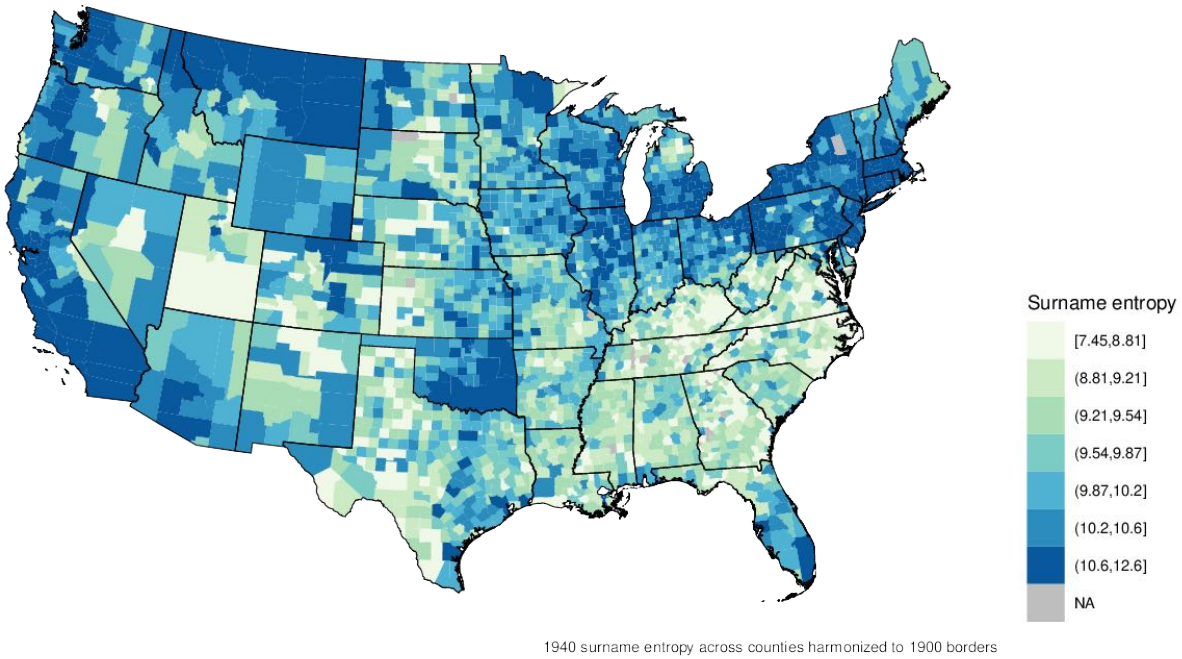


Figure B1: Distribution of surname population shares within counties and surname diversity

Notes: The figures show the distribution of surname population shares within counties (1850–1940, left plot) and surname diversity (1900–1940, right plot).



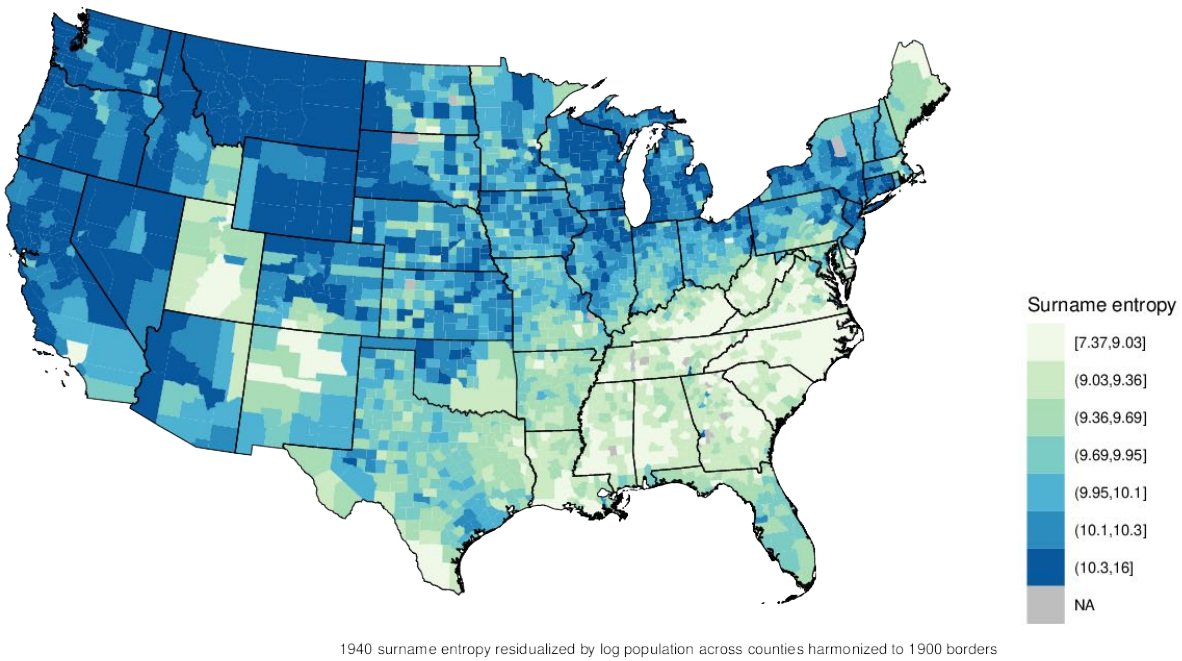
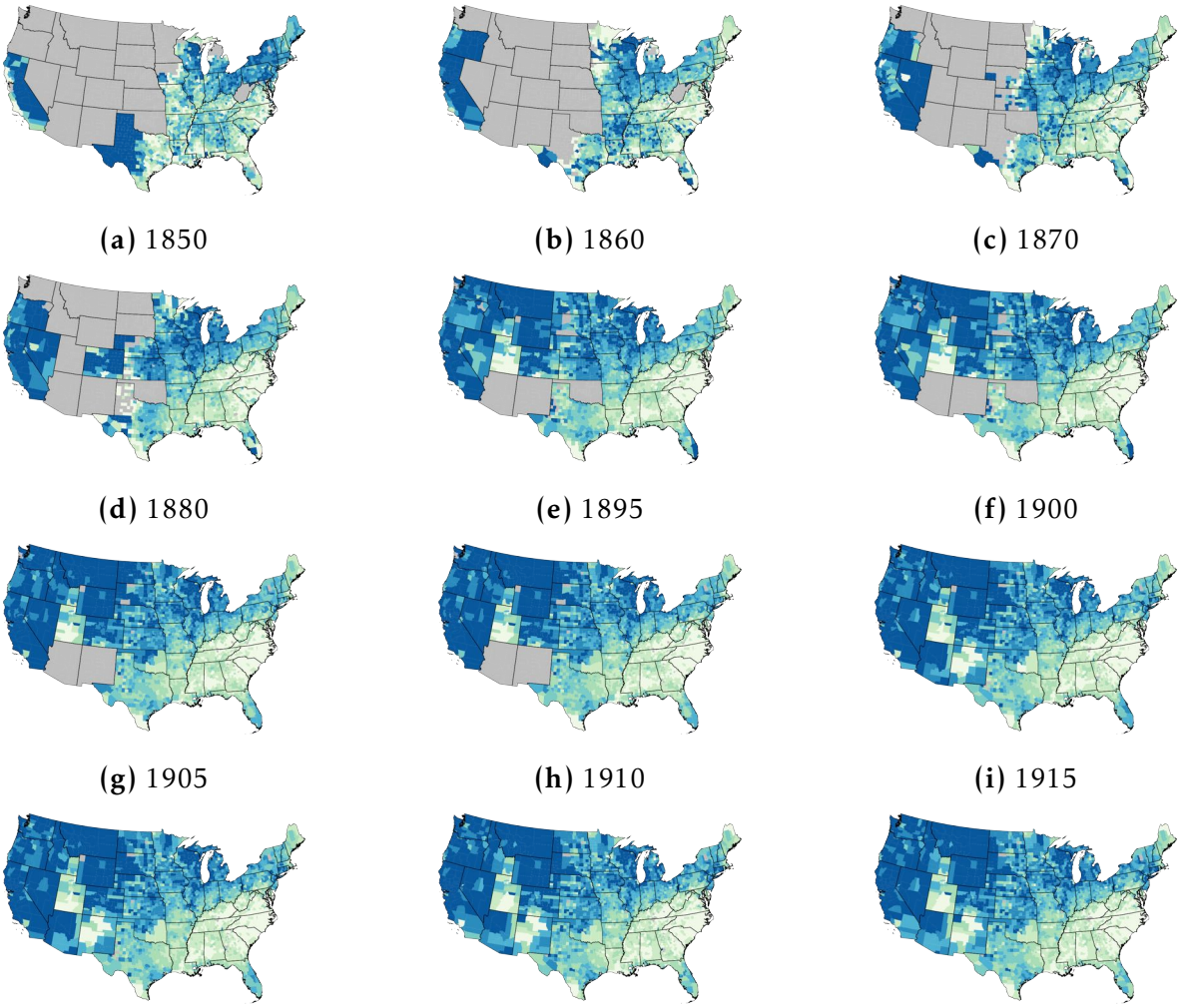


Figure B2: Surname diversity in 1940

Notes: The figures show the geographic variation in surname entropy in 1940 across counties harmonized to 1900 borders. Top: Raw data; bottom: residualized by log county population.



(a) 1850 **(b)** 1860 **(c)** 1870

(d) 1880 **(e)** 1895 **(f)** 1900

(g) 1905 **(h)** 1910 **(i)** 1915

(j) 1920 **(k)** 1925 **(l)** 1930

Figure B3: Surname diversity from 1850 to 1930

Notes: The figure shows surname entropy residualized by log county population in the respective year across counties harmonized to 1900 borders.

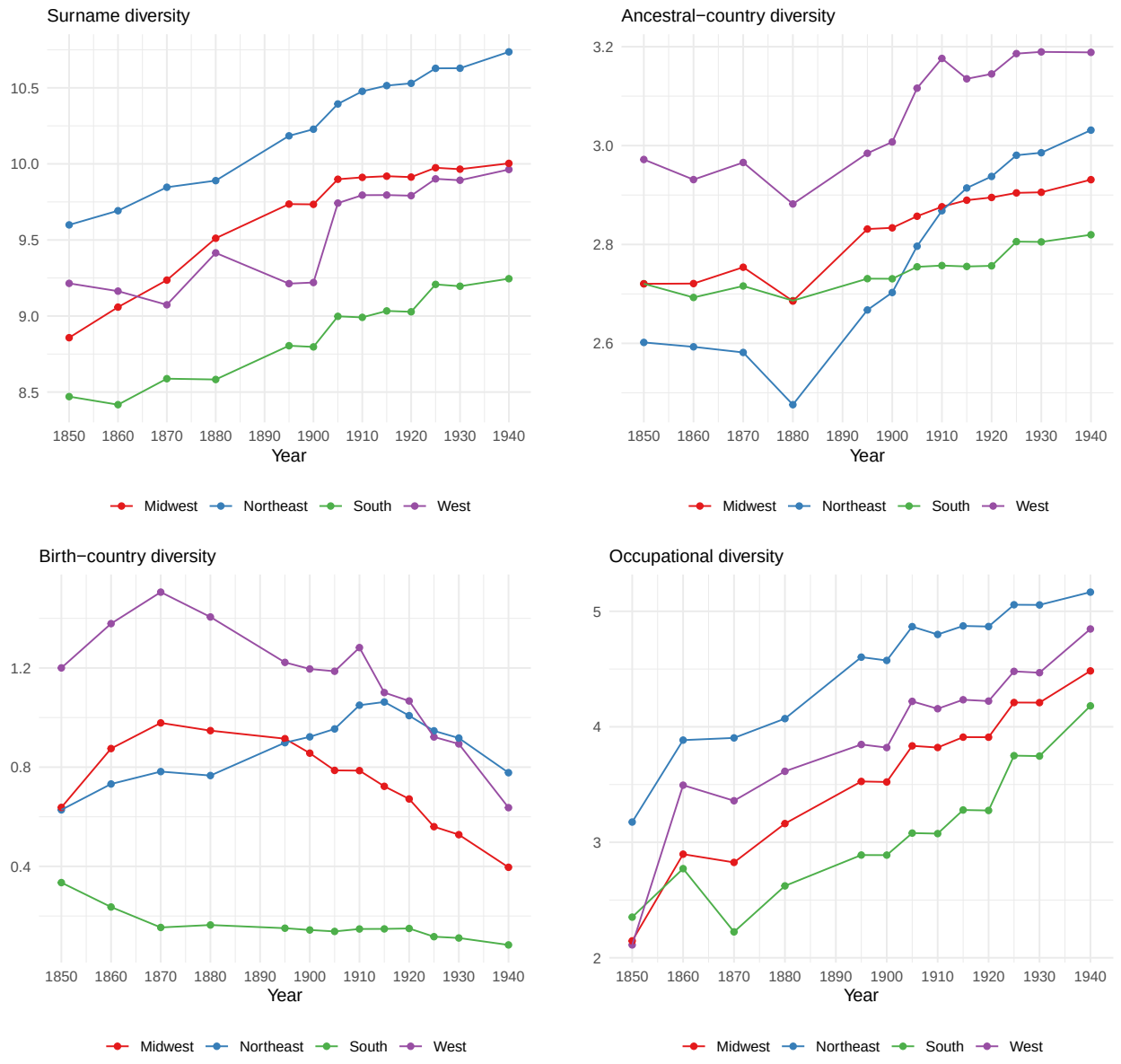


Figure B4: Surname diversity and other diversities from 1850 to 1940 across US census regions

Notes: The figure shows mean surname, ancestral-country, birth-country and occupational entropy in the respective year across census regions.

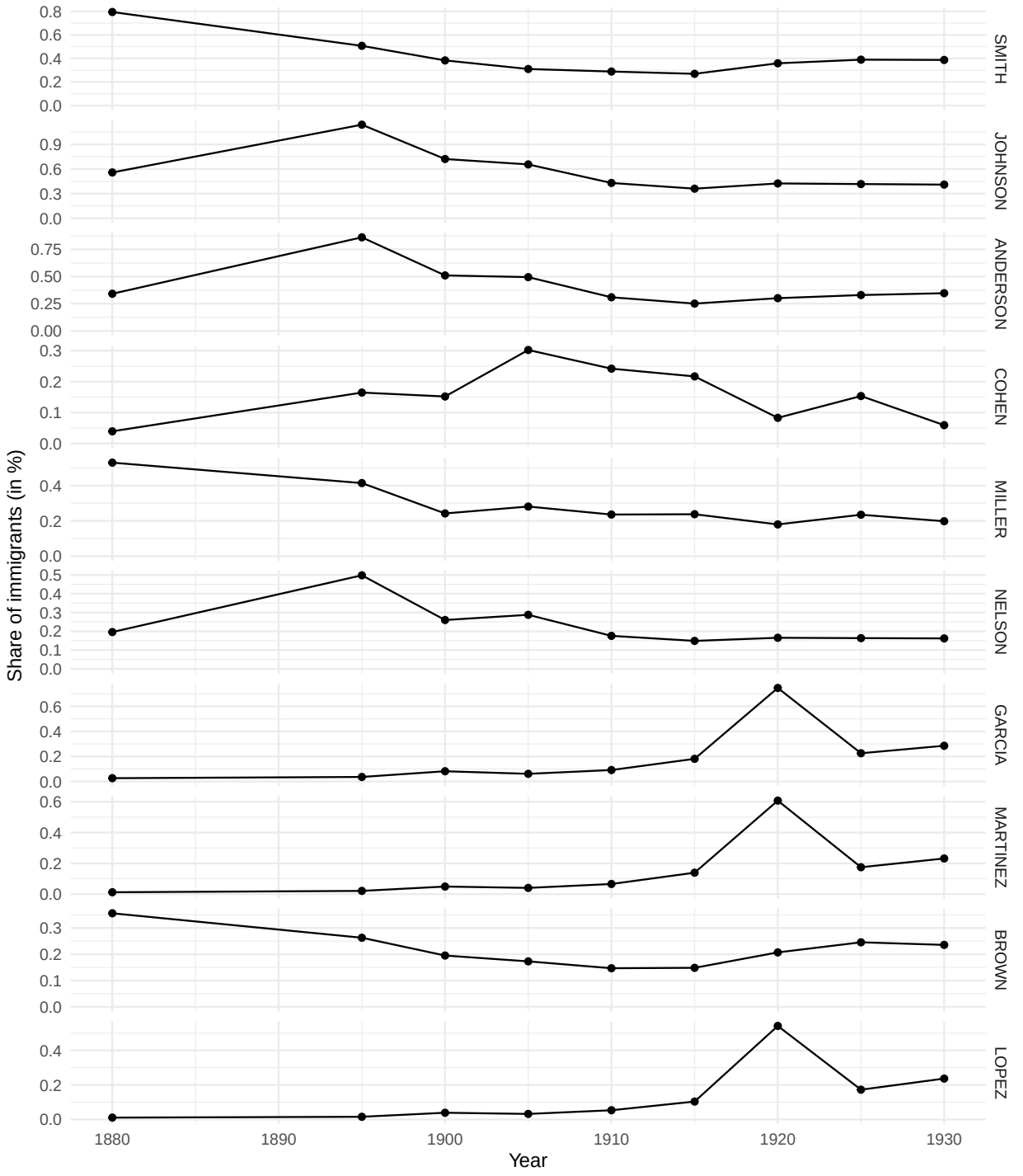


Figure B5: Share immigrants to the US by Surname

Notes: This figure illustrates the share of immigration into the US from the 10 surname groups with the highest immigration rates during specified periods: Smith, Johnson, Anderson, Cohen, Miller, Nelson, Garcia, Martinez, Brown, and Lopez. For each period, the most frequent surname was selected. If a top surname had already been selected for a previous period, the next most frequent surname was chosen, and so forth. The figure highlights variation in the push factors, illustrating how the number of migrants with each given surname changes over time.

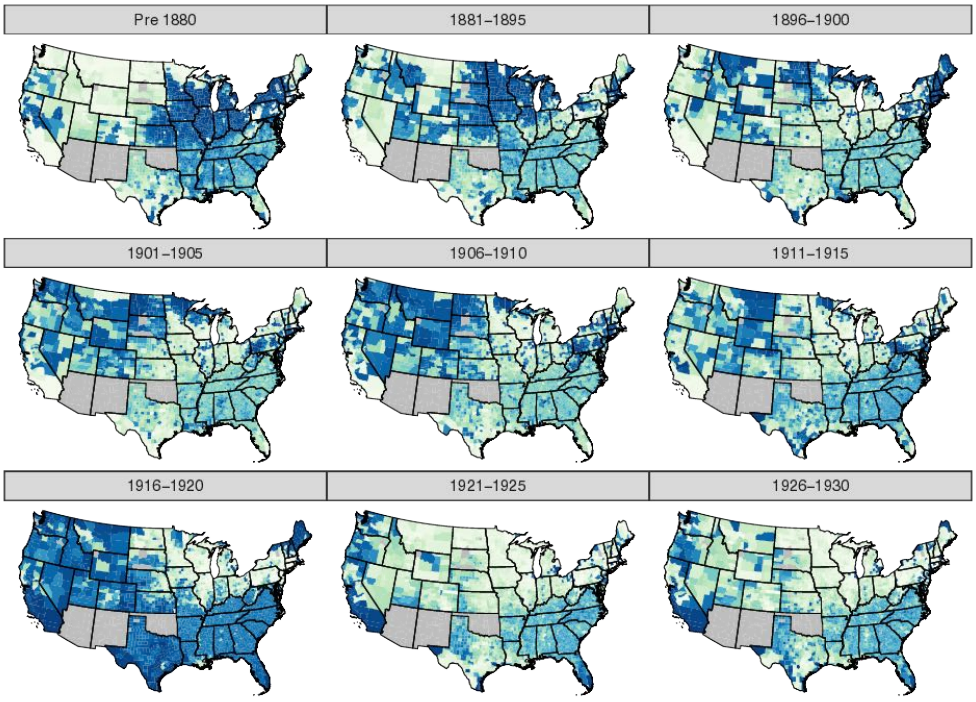
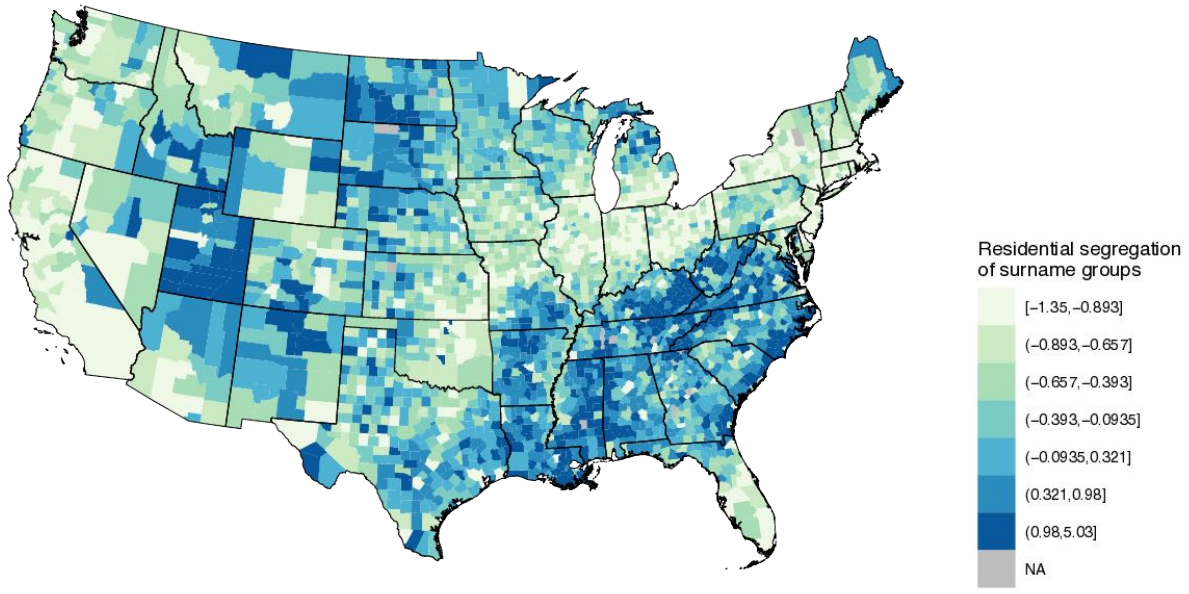


Figure B6: Destinations of immigrants to the US

Notes: This figure maps immigration flows into US counties by 5-year periods (except between 1880 and 1895). We regress the share of immigrants settling in US county i at time t, I_i^t, on destination county i and year t fixed effects, and calculate the residuals. The color coding depicts the 200 quantiles of the residuals across counties and within census periods. Darker colors indicate a higher quantile.



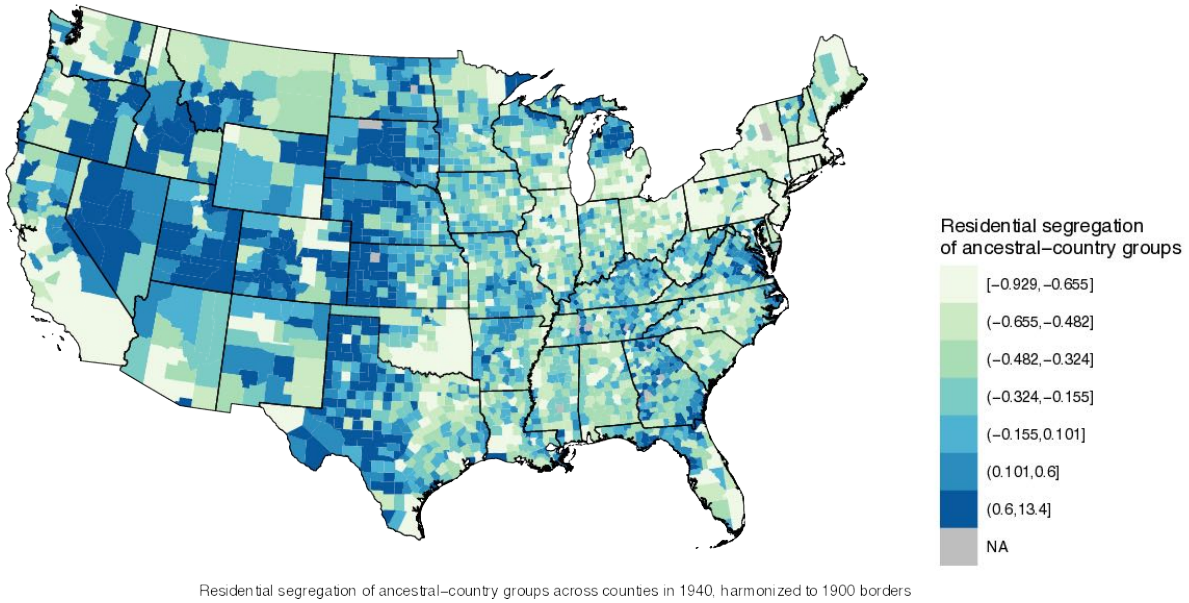


Figure B7: Residential segregation of surname and ancestral-country groups in 1940

Notes: The figures show the geographic variation in residential segregation of surname (top) and ancestral-country (bottom) groups across counties in 1940, harmonized to 1900 borders. The segregation measures are constructed adapting the Logan and Parman (2017) procedure to surnames and country of birth. The sources and construction of all variables are explained in Appendix A.

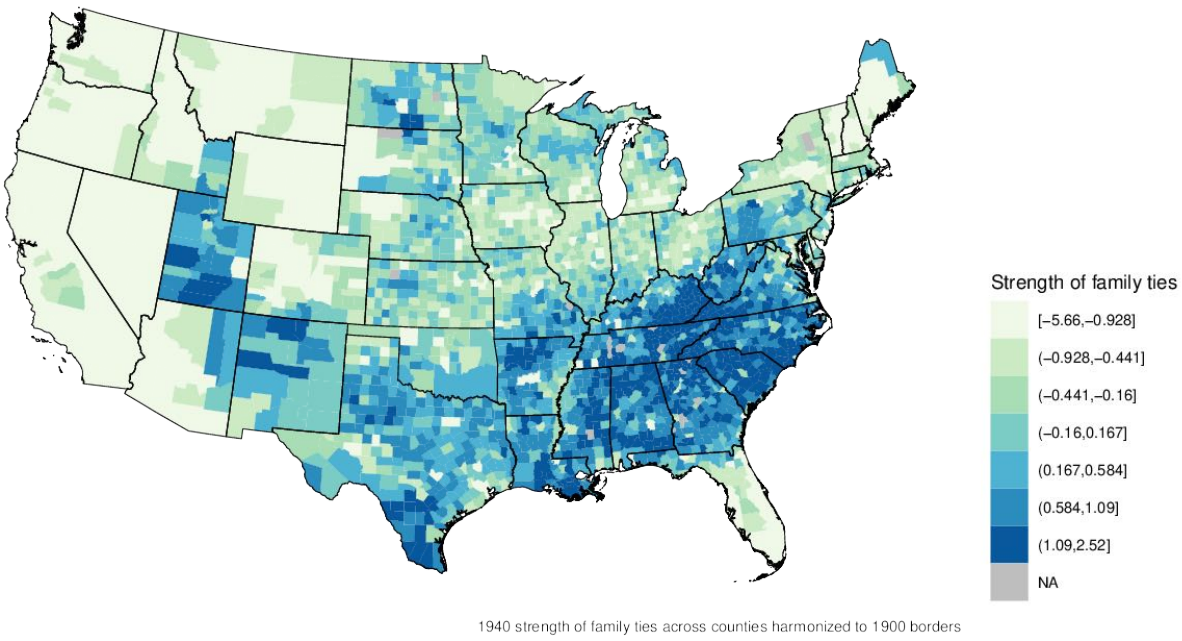


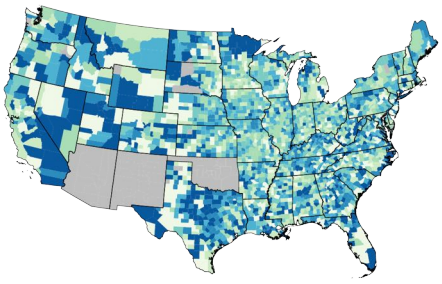
Figure B8: Strength of family ties in 1940

Notes: The figure shows the geographic variation in strength of family ties in 1940 across counties harmonized to 1900 borders. The measure is constructed following Raz (2024). The sources and construction of all variables are explained in Appendix A.

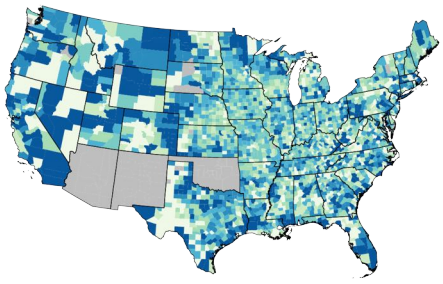


(a) 1900

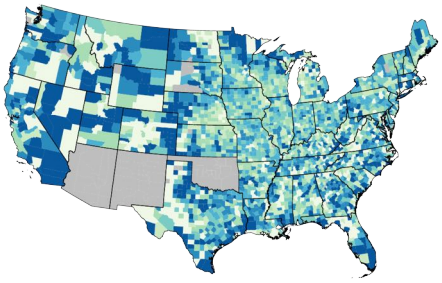
(b) 1905



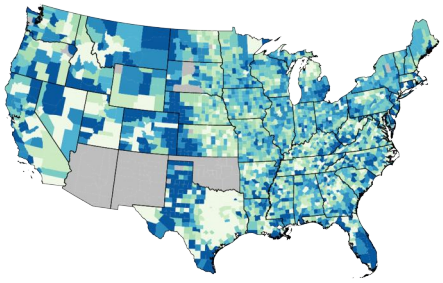
(c) 1910



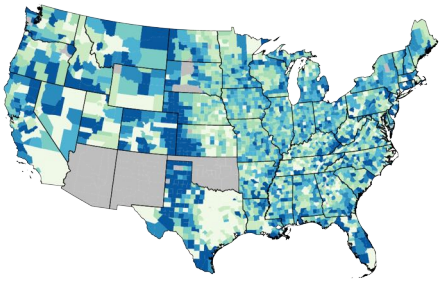
(d) 1915



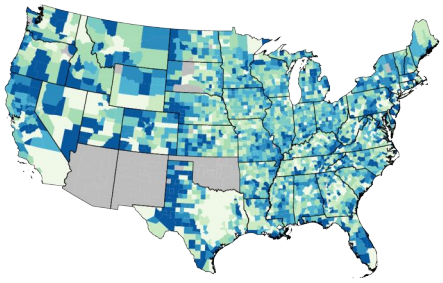
(e) 1920



(f) 1925



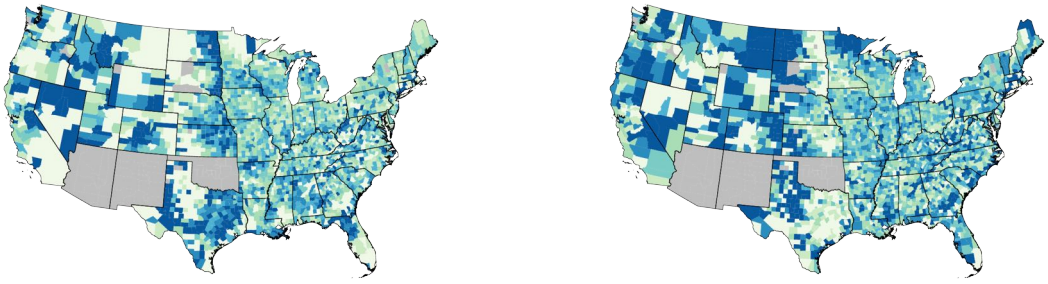
(g) 1930



(h) 1940

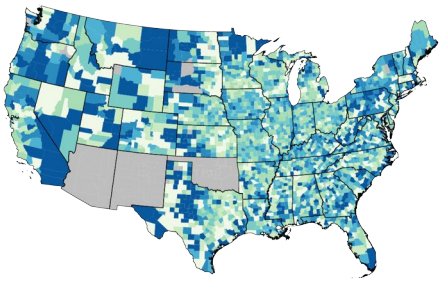
Figure B9: Predicted surname diversity (residuals)

Notes: This figure maps residualized instrumented surname diversity for each of the eight periods. We regress the instrument for surname diversity on the instrument for population and county and state-year fixed effects, and calculate the residuals. The color coding depicts 7 intervals across counties and within census periods, with darker colors indicating higher values. Grey indicates a lack of population data in 1900.

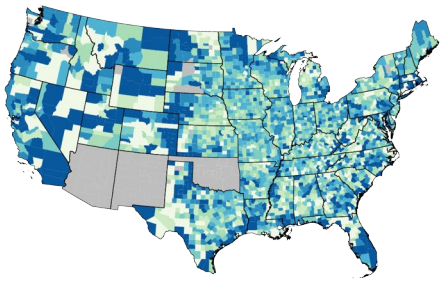


(a) 1900

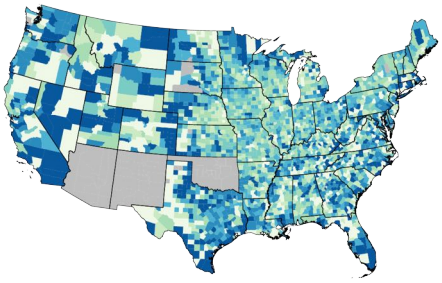
(b) 1905



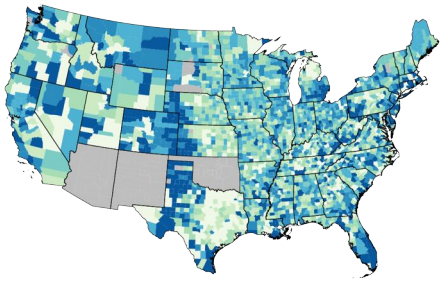
(c) 1910



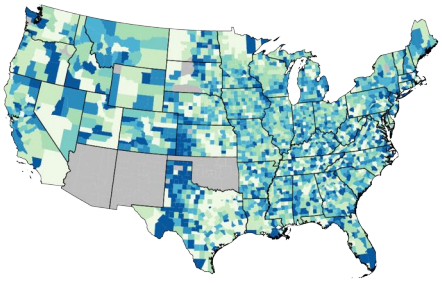
(d) 1915



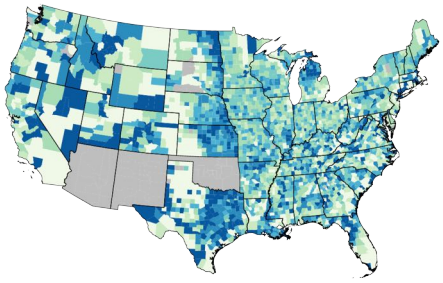
(e) 1920



(f) 1925



(g) 1930



(h) 1940

Figure B10: Predicted surname entropy (residuals)

Notes: This figure maps residualized instrumented surname entropy for each of the eight periods. We regress the instrument for surname entropy on the instrument for population, county and state-year fixed effects, and county specific linear time trends, and calculate the residuals. The color coding depicts 7 intervals across counties and within census periods, with darker colors indicating higher values. Grey indicates a lack of population data in 1900.

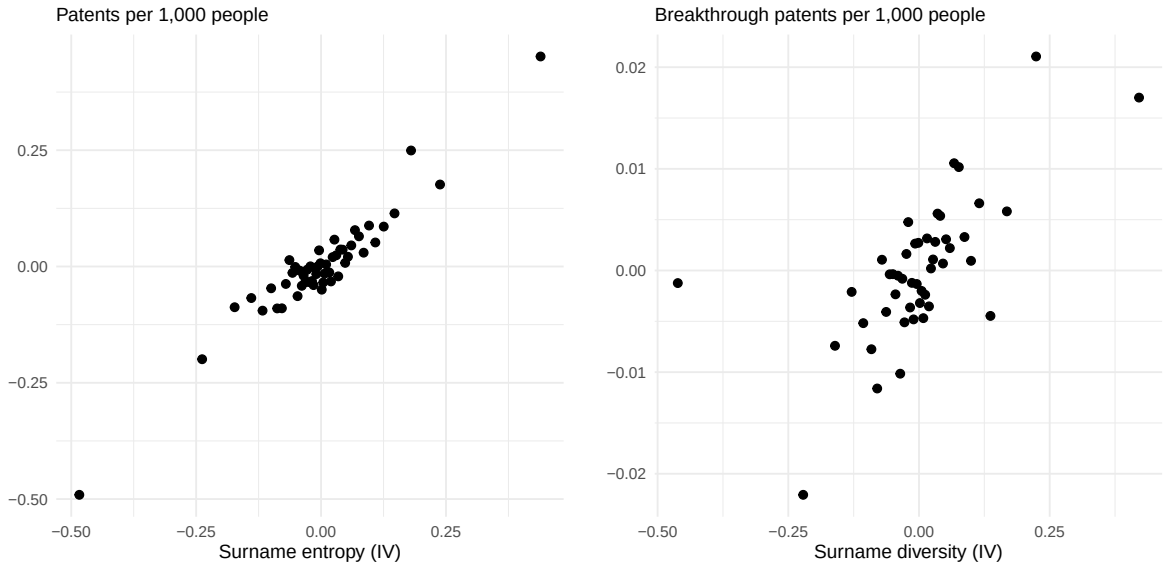


Figure B11: Reduced-from relationships of surname diversity (IV) and innovation outcomes, additionally controlling for county-specific trends

Notes: County-level data from 1900 to 1940. Observations are residualized by population (IV), county fixed effects, state-period fixed effects and county-specific time trends.

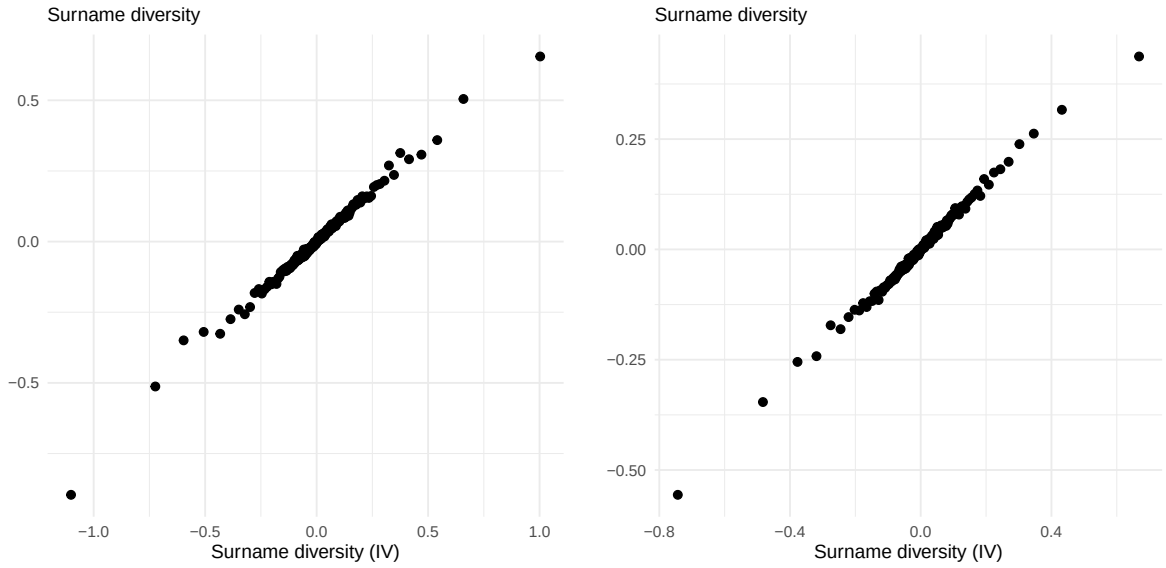


Figure B12: First Stage: Binned scatter plots of surname diversity (IV) and actual surname diversity from 1900 to 1940

Notes: County-level data from 1900 to 1940 (including midyears). Observations are residualized by population (IV), county fixed effects and state-period fixed effects (left plot) and additionally county-specific trends (right plot).

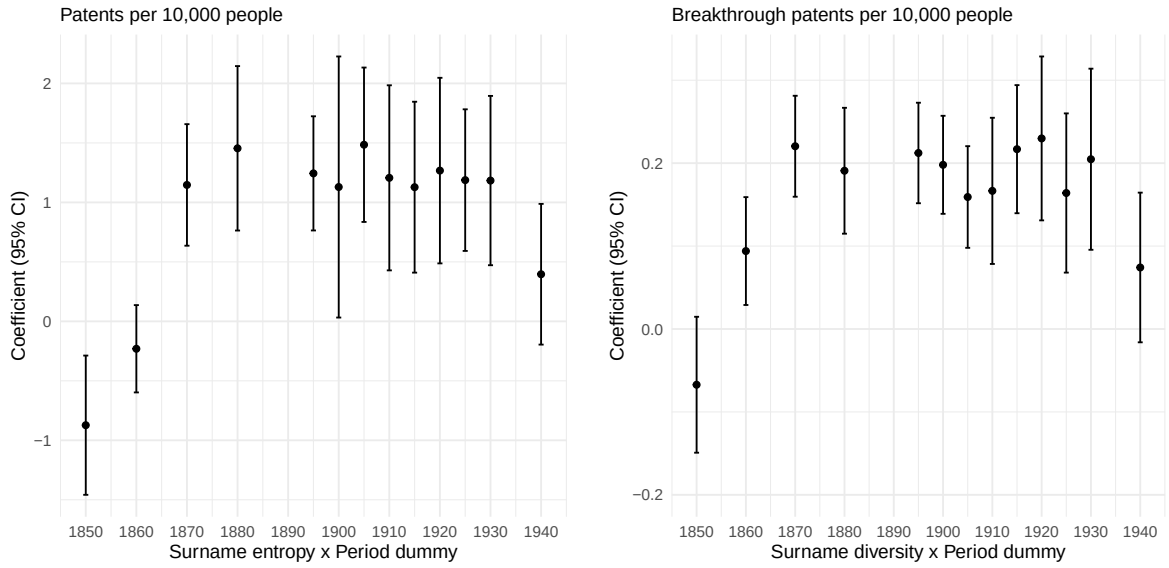


Figure B13: OLS estimates in years 1850-1940

Notes: The figures display coefficients from regressions of the number of patents and breakthrough patents normalized by county population in period t on surname diversity, interacted with period dummies, conditional on population size, county fixed effects and state-period fixed effects. Standard errors are clustered at the state level. Few patents were issued in 1850 compared to the latter years. Surname diversity is standardized to have a mean of zero and unit variance. The sources and construction of all variables are explained in Appendix A. The dependent variables are winsorized at the 5 per cent level. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

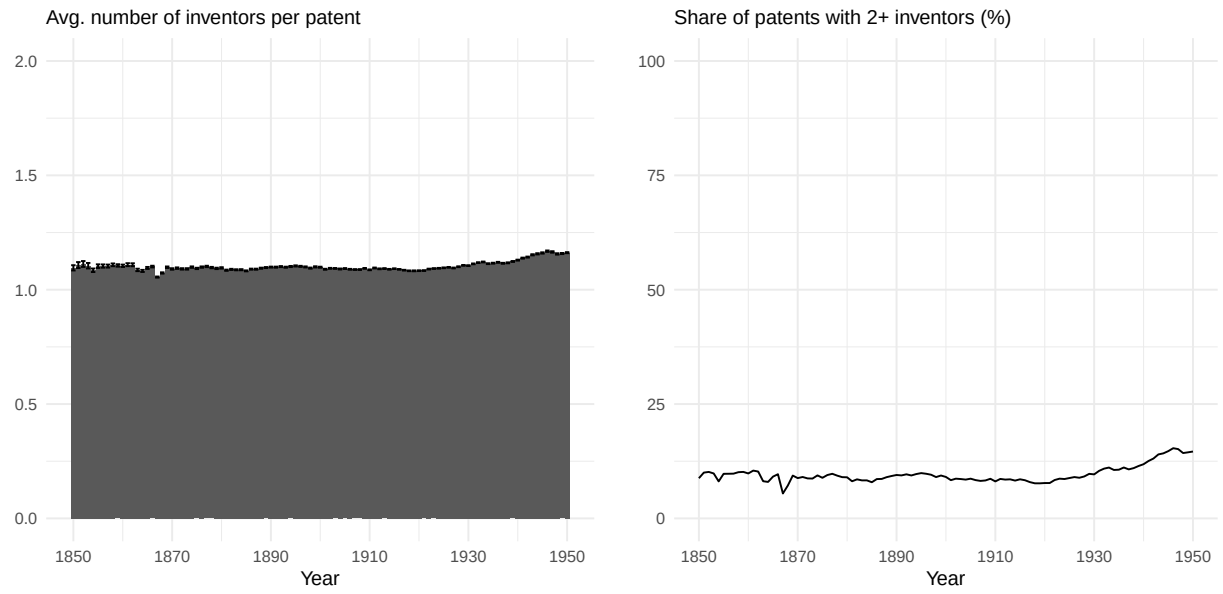


Figure B14: Patentees per patent over time

Notes: The left plot shows the average number of patentees per patent per year; The plot displays the share of patents with two or more patentees per year.