Surname Diversity, Social Ties and Innovation*

Max Posch[†] Jonathan Schulz[‡] Joseph Henrich[§]

September 14, 2023

Abstract

We study whether interactions between individuals with different skills, expertise and perspectives influenced innovation in U.S. counties from 1850 to 1940. We introduce and validate a new measure of social interactional diversity based on the distribution of surnames: lower surname diversity indicates more concentrated social interactions among like-minded people. Leveraging quasi-random variation in counties' surname compositions—stemming from the interplay between historical fluctuations in immigration and local factors that attract immigrants—we find that surname diversity increases both the quantity and quality of innovation. The results support the view that social interactions between diverse minds are key drivers of innovation.

Keywords: Innovation, social interactions, surname diversity, immigration
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, 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, Cologne, 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), economists and other scholars have argued that social interactions among diverse minds encourage innovation and creativity (e.g., Jacobs, 1969; Glaeser et al., 1992; Muthukrishna and Henrich, 2016; Akcigit et al., 2018; Galor, 2022; Andrews, 2023). From this perspective, innovations emerge primarily through the recombination of ideas created as individuals with diverse expertise, skills and ways of thinking interact and share their ideas (Weitzman, 1998; Jones, forthcoming). Recent research has revealed remarkable variation in the degree to which interactions among diverse individuals occur across different countries and even among communities within the same country (Alesina and Giuliano, 2014; Moscona et al., 2017; Schulz et al., 2019; Enke, 2019; Henrich, 2020; Ghosh et al., 2023). Indeed, some communities exhibit concentrated social networks, often structured around strong family ties, which can cultivate an inward-looking psychology marked by mistrust of outsiders and substantial reluctance to engage beyond one's family or in-group. If these mindsets and social ties hamper the free exchange of knowledge and ideas beyond one's family or in-group, then less concentrated, or more diverse, social interactions should foster innovation. However, quantifying the impact of this diversity on innovation has proven elusive due to empirical challenges related to measurement—the need for a fine-grained measure of the diversity of social interactions—and causal identification—individuals and their social interactions are not randomly distributed across locations or time.

In this paper, we propose a novel approach to measuring the diversity of social interactions based on surnames within U.S. counties and then use this to assess the existence of a causal relationship with innovation, as measured using data on patents. Conceptually, social interactions among diverse minds should spur innovation, largely because many new ideas arise from the recombination of existing ideas. The potential informational exchanges that occur during these social interactions are assumed to depend on two components: an informational element, in which different individuals possess distinct areas of knowledge (skills, ways of thinking, etc.); and a social-psychological element, in which individuals place trust in and engage with people outside the confines of their familial and in-group circle.

Surnames in the U.S. are well-suited for capturing both components. From an informational perspective, surnames, which are typically patrilineally inherited in the U.S., can serve as markers of different kinds of knowledge, skills and perspectives that originate from learning within familial, ethnic, or professional networks and persist across generations because of intergenerational cultural transmission (Boyd and Richerson, 1985; Bisin and Verdier, 1998). For instance, recent work by Bell et al. (2019) shows that adult sons in the U.S. are much more likely to patent if their father had patented, and sons usually invent in the same technology subclass (e.g., amplifiers, antennas, etc.).

From a social-psychological perspective, since individuals with the same surname are more likely to share the same family or ancestral background, a high prevalence of shared surnames may indicate the presence of strong familial or cultural ties among those individuals. This fosters trust and cooperation primarily within the family or in-group.[1] However, such interactions are usually between informationally similar individuals, making innovative recombinations less likely. Conversely, high surname diversity implies a limited ability to meet needs within one's own relatively small family or surname network, encouraging broader interactions and the development of impersonal trust. This, in turn, fosters innovation through recombination.

To reconcile the described social-psychological mechanism with the established finding that increased diversity is associated with rising mistrust (e.g., Alesina and Ferrara, 2005; Ashraf and Galor, 2013), it is crucial to differentiate between diversity in fine-grained indicators like surnames and broader categories like race, ethnicity, or nationality. As the relative size of extended family and surname groups diminishes within the general population, interactions outside these circles become increasingly necessary and advantageous. Thus, unlike broader measures of diversity, a rise in surname diversity is likely to expand opportunities for beneficial exchanges beyond one's immediate in-group. For example, recent work by Bazzi et al. (2019) finds that a high degree of fractionalization (the presence of many small groups) fosters impersonal trust, whereas polarization (the presence of a few large groups) leads to intergroup antagonism.

In our empirical approach, we utilize all surnames reported in the full-count U.S. Census data from 1850 to 1940. We deploy these data to compute the diversity of surnames

[1] In population genetics, low surname diversity is often used as a marker for cousin marriage and other forms of inbreeding (e.g., Barrai et al., 1996). This approach has recently been adopted by the economics literature (Ghosh et al., 2023).

across U.S. counties, which are presumed to be the primary locations of social interaction during this period. While counties might not encapsulate every social interaction, particularly in today's highly interconnected world, they provide a reasonable approximation in the pre-1950 historical context.

We validate the use of surname diversity as a measure of the diversity of social interactions using two approaches. First, to evaluate the informational aspect of social interactional diversity, we analyze the extent to which surnames cluster within specific occupations and immigrants' ancestral regions. Using the census data, we find that two individuals selected at random, who share the same surname, have a higher than random probability of also sharing an occupation or having origins in the same country or subnational region. This finding aligns with recent work on social mobility (Clark, 2014; Güell et al., 2015; Bell et al., 2019; Barone and Mocetti, 2021), bolstering the assertion that surnames capture distinct facets of knowledge tied to specific occupations and ethnicities.

Second, for the social-psychological aspect of social interactional diversity, we probe the relationship between surname diversity and two key measures of inward-looking psychology: impersonal trust, as measured by surveys, and the strength of family ties, as inferred from census data.[2] We find strong associations between surname diversity and both measures of inward-looking psychology.

Importantly, we do not find these patterns for the more conventional measures of diversity based on country of birth and race, which is consistent with prior evidence (Glaeser et al., 2000; Alesina and La Ferrara, 2000, 2002). This highlights that surname diversity, with its focus on family networks, is not only conceptually but also empirically distinct from these other forms of diversity. Taken together, these results support the view that surname diversity is an effective measure of social interactional diversity that encapsulates both informational and social-psychological components.

To measure innovation, we rely on two patent indicators. First, we calculate the total number of patents per capita for each U.S. county for 5 or 10-year periods from the 1850s to the 1940s, based on the Comprehensive Universe of U.S. Patents (Berkes, 2018). Second, we use the breakthrough patent indicator created by Kelly et al. (2021) to capture highly important patents. Breakthrough patents are identified via textual similarity to previous and subsequent patents; breakthrough patents have low similarity to previous patents but high similarity to subsequent ones.

Using these diversity and innovation measures, our analyses proceed as follows. First, we study the correlation between surname diversity and both patents and breakthrough

[2] Previous work has shown that strong family ties are tightly linked to lower impersonal trust (Alesina and Giuliano, 2014; Schulz et al., 2019; Enke, 2019).

patents per capita across U.S. counties from the 1850s to the 1940s. We find positive and economically meaningful relationships between surname diversity and both innovation measures. A one standard deviation increase in surname diversity within a county is associated with approximately 78% more patents per capita and a slightly larger increase in breakthrough patents per capita. These relationships are remarkably stable over time—our sample spans almost a century of U.S. innovation—and hold when controlling for county and period-state fixed effects, population-scale effects, and the composition of immigrants and race within these counties. We find that the more conventional country-of-birth diversity measure is also a significant predictor of patents in most specifications. However, in line with the distinct ability of surname diversity to capture fine-grained social interactional diversity, we find that surname diversity offers additional explanatory power, even when controlling for country-of-birth diversity and country-of-birth-specific immigrant shares.

Second, we study the causal link between surname diversity and innovation. As noted, such causal evidence has proven elusive because individuals do not allocate randomly across space, possibly creating a spurious correlation between surname diversity and patents. To address this concern, we employ an instrumental variable (IV) strategy, building on the approach developed by Burchardi et al. (2019). This strategy leverages historical immigration patterns as a significant determinant of surname diversity in U.S. counties.

Migration, beyond births, deaths and marriages, is the key driver of counties' surname composition. However, immigration does not monotonically increase surname diversity. Its impact depends on the preexisting surname distribution in a county. In other words, the arrival of the same set of immigrants can increase surname diversity in some counties and decrease it in others. We hypothesize that this relationship between immigration and surname diversity affects both the informational and social-psychological channels (detailed in section 2). When individuals carrying *locally* rare surnames arrive, they enhance surname diversity, and in turn, may create opportunities for diverse social interactions, knowledge acquisition, and the cultivation of trust towards individuals with differing cultural and family backgrounds. On the other hand, an inflow of individuals bearing locally common surnames decreases surname diversity. This movement of individuals, who are culturally and genealogically related to the dominant groups within counties, may limit opportunities for novel knowledge acquisition, strengthen ties within families or culturally homogeneous groups, and nurture a low-trust mentality towards outsiders.

Our IV approach isolates quasi-random variation in counties' surname composition, which stems solely from the historical interplay of two forces: (i) the staggered arrival of migrants with different surnames and (ii) the temporal variation in the relative attractiveness of different destination counties for the average migrant arriving at the time. The interaction of these two historical forces enables us to isolate the variation in surname distributions across counties, which is essentially inherited from plausibly exogenous shocks to historical migrations dating back to the 19th century.

Using data across counties from 1900 to 1940, our IV-estimates provide evidence that a one standard deviation increase in surname diversity raises patents and breakthrough patents (per 1,000 people) by 76-93% and 144-149%, respectively.

These results hold across key robustness checks. First, to scrutinize the potential for reverse causality, i.e., an increase in a county's innovation leading to increased diversity, we perform a falsification exercise by regressing past patents on future surname diversity. The coefficients from this exercise are near zero, or even negative, and statistically insignificant, providing strong evidence against the concern that reverse causality might confound our results.

Second, to address the potential influence of scale effects, including through immigration, we control for quasi-random variation in population size isolated by the IV procedure. Again, the estimates align with our primary findings.

Third, we confront the possibility that our results are region-specific, especially in light of regional variation in factors such as racial segregation that could simultaneously affect innovation, surname diversity and immigration. Estimating the impact of surname diversity on patents across the four major U.S. census regions (Northeast, Midwest, South, and West), we find consistently positive coefficients, most of which are accurately estimated.

Fourth, recent contributions to the literature on social mobility (e.g., Clark, 2014; Barone and Mocetti, 2021) raise the concern that unobserved characteristics embedded in specific (rare) surnames, such as abilities, interests, or knowledge, drive the results rather than diversity per se. To explore this, we change the unit of observation from county-period to surname-county-period and include surname-fixed effects in our specifications to absorb any surname-specific traits. We find that our estimates remain highly significant across all specifications and hardly change by including these fixed effects.

Fifth, we address the concern that a direct effect of immigration, which is not channeled through diversity, confounds our estimates. Replicating our analysis within a subsample of U.S.-born individuals, we find very similar effects of surname diversity on patents and breakthrough patents, reinforcing the interpretation that it is diversity rather than immigration per se driving our results.

We conclude our analysis by substantiating our proposed mechanism: diverse social interactions catalyzing idea recombination. Firstly, to shed more light on the informational channel of social interaction diversity, we utilize the extensive technology code system implemented by the U.S. Patent and Trademark Office (USPTO), which comprises over 140,000 unique codes, to categorize patents. We find that surname diversity increases the number of technologies per patent and the proportion of patents characterized as novel combinations. This suggests that surname diversity has helped to propel the recombination of existing technologies and enhance the complexity of innovation. Secondly, to provide further evidence for the social-psychological channel, we investigate the causal relationship between surname diversity and the strength of family ties. Our findings indicate that greater surname diversity not only correlates with weaker family ties, as previously discussed, but actually drives this correlation. Additionally, we find that reductions in surname population shares within counties lead to weaker family ties within these surname groups. These results suggest that as family networks shrink, the benefits of interacting with non-family members increase. This likely has downstream effects on impersonal trust, the exchange of ideas among diverse individuals, and the rate of innovation.

Taken together, our results indicate that the diversity of social interactions is causally linked to the rate, quality and type of patenting over much of U.S. history. These findings point to an important role of social interactions between diverse individuals in driving recombinant innovation.

1.1 Contributions and Related Literature

Understanding the drivers of innovation is central to many lines of research in economics, from endogenous growth (Romer, 1990; Galor and Weil, 2000) to the origins of the industrial revolution (Mokyr, 2002). Here, we focus narrowly on those recent lines of research that connect most closely with our efforts.

Our paper picks up on ideas related to the impact of cities on innovation and the role of agglomeration, population density and geographic connections (Carlino and Kerr, 2015; Akcigit et al., 2017; Glaeser, 2011). Research in this area emphasizes the importance of skill complementarities, localized knowledge spillovers and other information transfers. Consistent with our approach, several studies have linked innovation to the formation of immigrant clusters and to a greater diversity of social interactions (Kerr, 2010, 2008). This paper extends these observations and insights both more broadly—across the entire U.S. and back to the mid-19th century—and offers a viable approach to measuring the diversity of social interactions across many contexts.

Our findings directly add to the emerging empirical literature on social interaction and innovation, which explores how social institutions and organizations spur innovation.

For example, the closure of saloons during Prohibition reduced patenting rates (Andrews, 2023), demonstrating a crucial role of social establishments for innovation. Similar mechanisms operate today, as illustrated by evidence suggesting that the spread of coffee shops has spurred innovation (Andrews and Lensing, 2020). By potentially tapping the same mechanism, the historical rise of economic societies in Germany reduced information access costs, thereby fostering innovation (Cinnirella et al., 2022). Furthermore, de la Croix et al. (2018) emphasize the role of pre-industrial apprenticeship institutions in Western Europe, including journeymanship, which facilitated the exchange of knowledge and ultimately contributed to Europe's growth. These studies, among others (Atkin et al., 2022), underscore the premise that social interactions stimulate knowledge diffusion, thereby contributing to human capital and innovation-based growth (Akcigit et al., 2018).

Our work intersects with studies exploring how various forms of diversity shape economic prosperity. For example, the seminal work by Ashraf and Galor (2013) shows that genetic diversity fosters innovation while reducing trust, resulting in an inverse U-shaped relationship between genetic diversity and economic prosperity across countries. Subsequent work corroborates these findings across ethnic groups and among second-generation immigrants (Arbatli et al., 2020; Ashraf et al., 2021). Our work complements this line of research. Conceptually, our measure of surname diversity is related to genetic diversity because, similar to genes, surnames are typically transmitted vertically from parents to offspring, and research in population genetics has shown that under certain conditions, genetic heterogeneity can be approximated using surname diversity (Barrai et al., 1996). Diverging from Ashraf and Galor (2013)'s global perspective, we focus on a specific historical episode within the context of a single country: the United States. This allows us to investigate regionally fine-grained changes in surname diversity over time in a panel setting. By using surname-fixed effects, we empirically establish that the results are not confounded by specific surnames or any genes associated with such surnames. That is, when comparing people with the same surnames, those located in counties with greater social interactional diversity are more innovative. Similarly, in a context paralleling our own, Fiszbein (2022) links economic development across counties from 1860 to 1940 with agricultural diversity. Here, a greater diversity in agricultural products resulted in greater economic prosperity, including more patents per capita, more technology classes per patent and more new manufacturing skills.[3]

[3]Beyond innovation, previous studies have also highlighted the positive effects of birthplace or country-of-ancestry diversity on local economic growth or wages, both within the U.S. (Ottaviano and Peri, 2006; Ager and Brückner, 2013; Docquier et al., 2020; Fulford et al., 2020) and across countries (Alesina et al., 2016). Our use of surname diversity complements Buonanno and Vanin (2017) who, using it as a measure of social closure, focus on crime.

Our paper also enriches the literature connecting migration to innovation and economic prosperity (Abramitzky and Boustan, 2017). Drawing on historical data from 1850 to 1920, Sequeira et al. (2020) show how rising flows of immigrants into U.S. counties resulted in faster rates of patenting. Based on an analysis of foreign patents and consistent with the social interactional diversity hypothesis, the authors argue that much of this effect occurred through making native-born Americans more creative—or at least more likely to patent. Similarly, focusing on the period from the mid-1920s to the mid-1960s in the U.S., Moser and San (2020) show how anti-immigration policies in the form of quotas seeking to preserve ethnic homogeneity reduced the inflow of migrants from Eastern and Southern Europe, which in turn stifled the production of innovations in the scientific fields favored by such immigrants prior to the quotas. Revealing the importance of social interactional diversity, their work finds a 62% decline in patenting in these particular fields by native-born scientists. The authors argue that resident scientists lost the mentorship and fresh approaches that inevitably flow in with those trained elsewhere. Similarly, Abramitzky et al. (2023) show that quotas did not benefit US-born workers. On the flip side, after the U.S.'s broad immigration quotas were lifted in 1964, Burchardi et al. (2021) show that by the mid-1970s, American innovation was again powerfully fueled by immigrants, now coming from places such as Mexico, China, India, the Philippines, and Vietnam. Exploiting America's relative openness to immigrants fleeing Germany and Austria prior to World War II, Moser et al. (2014) also demonstrate the impact of Jewish immigrants on U.S. patents. Their analysis reveals not only how refugee chemists stimulated innovation and interest among native-born individuals, but also how their impact reverberated through social networks to impact the patenting of collaborators of the immigrants' collaborators. Our work supports these findings by highlighting an important channel through which immigration affects innovation, via increasing the diversity of social interactions.

2 Concepts and Measurement

In this section, we first describe the recombinative process that arguably underlies much innovation and then highlight supporting lines of evidence. Next, we detail our measure of surname diversity, explain how and why it proxies for the diversity of social interactions, and then empirically demonstrate the key conceptual linkages using census information on occupations and ancestral regions along with measures of psychological openness and family ties. Finally, we discuss how we use U.S. patents to measure innovation.

2.1 Diversity of Social Interactions and Innovation

In 1933, marking an important step on the road to modern radar, three men in Washington D.C.—Tylor, Young and Hyland—filed a patent for a “system for detecting objects by radio” (US patent number 1981884). This initial step toward radar began three years earlier when the Canadian-born immigrant, Lawrence “Pat” Hyland, was testing a directional radio receiver in an aircraft. While tuning the receiver to a transmitter two miles away, he was frustrated by the fact that his signal seemed to randomly grow louder and quieter as he was testing it. He noticed that this occurred whenever a plane flew overhead. Puzzled by the phenomena, he asked a fellow radio engineer, Leo Young, about it. Young, an avid ham radio hobbyist from a farming family in Ohio, recalled an experience from eight years earlier when he worked for the Aircraft Radio Laboratory. For fun, he and a physicist named Albert “Hoyt” Taylor had set up a high-frequency transmitter and receiver on opposite sides of the Potomac River at the mouth of the Anacostia River. Young, following an article he had found in an engineering magazine, had managed to jack up the frequency of his transmitter by a factor of 20. After some tuning, he had a crystal-clear tone from across the Potomac. Then, unexpectedly, the tone doubled in volume. Young looked up and saw a ship, the Dorchester, passing between himself and their receiver across the river. After discussing the event, the duo realized what had happened: their signal had bounced off the Dorchester’s hull and, just for a millisecond, synchronized. They wrote a report about the possibility of using radio signals to detect passing ships, which the U.S. Navy promptly ignored. Hyland had bumbled over what appeared to be the same phenomenon, but now with aircraft. Using these insights, the trio developed a means of using continuous wave radio signals to detect passing ships and planes. Despite now having a working prototype, the U.S. Navy rejected their request for $5,000 to continue their research, explaining that this was “a wild dream with practically no chance of real success” (Bahcall, 2019; DeGering, 2018; Page, 1962).

This patent represents a conceptual recombination—putting existing radio technologies to use in a fresh application, detecting ships and planes at a distance. Of course, people have been trying to extend the reach of our detection abilities for a long time, often using tools like towers or spyglasses. Here, both serendipity and social interactions were central while top-down problem-solving and forward-looking insight were limited. In particular, at the time, many engineers and physicists understood the Doppler effect, but no one had used that understanding to create radar. Instead, these inventors encountered a phenomenon—they accidentally detected ships or planes—and then applied the science of the era to interpret it. After explaining the potential value of their discovery, and later their invention, to the U.S. Navy, the true potential of their insights went unrecognized until the attack on Pearl Harbor in 1941. Interestingly, the patent office assigned two previously used technology codes ('342/27' and '367/128') and two novel codes to this conceptual recombination ('340/991' and '342/453').

The idea illustrated by this patent, that innovation emerges from the recombination of ideas propelled by social interaction, has venerable lineages in both economics (Schumpeter, 1983) and history (Usher, 2013), and has received persistent attention ever since (Jacobs, 1969; Glaeser et al., 1992; Henrich, 2009; Ridley, 2020; Johnson, 2011; Mokyr, 2015; Olsson and Frey, 2002; Lucas Jr and Moll, 2014; Akcigit et al., 2018; Jones, forthcoming). At the population level, the meeting and merging of people and ideas involve both an informational component—different people possess distinct ideas, approaches, skills, and perspectives—and a psychological component—individuals have to interact and be open to sharing their thoughts. Both elements are required since a population of diverse minds that never interact will not generate any recombinations, and a group of cognitive clones who freely interact but all share the same mentality will also fail to generate recombinations. Thus, conceptually, social interactional diversity – the extent to which the free flow of ideas among diverse individuals occurs – should capture the capacity of local populations to generate novel recombinations, some of which will turn into inventions (and for us, patents).

The plausibility of this hypothesis is reinforced by three distinct strands of research. Firstly, a significant body of work posits a central role of recombination in innovation. Secondly, a broad range of research emphasizes the impact of cultural, genetic, disciplinary, and occupational diversity on innovation. Finally, there is a body of evidence illustrating the influence of social dynamics on innovation, focusing either on the institutions that facilitate social interaction, or the role of trust and other psychological factors that mold social interaction and exchange. We will briefly delve into each of these research areas.

Empirically, the concept that most innovations result from recombinations has been explored in economics and related fields. Using 1.8 million U.S. patents from 1975-2004 and their citations to other patents, Acemoglu et al. (2016) model the connections among patents, showing how the production of new patents in specific technological areas depends on progress in other associated fields. In other words, advancements in linked technological domains provide the crucial elements or insights for new patents, supplying the fuel for recombination. Augmenting this work with the complete U.S. patent database, Youn et al. (2015), Strumsky et al. (2011) and Akcigit et al. (2013) use detailed patent class codes to demonstrate that most patents are, indeed, recombinations, drawing from various technological categories. Pushing this idea further, Clancy (2018a,b) a recombinative model of innovation that accounts for both the 'fishing out' of obvious recombinations and the innovation-generating impact of each new recombinative idea (patent). The model's predictions align with the patterns observed in U.S. Patents.[4]

Alongside evidence for the centrality of recombination for innovation, many researchers have studied the connections between innovation and diversity, including measures of genetic, birthplace, academic discipline, and ethnic diversity (Ashraf and Galor, 2013; Alesina et al., 2016; Page et al., 2019; Docquier et al., 2020; Fulford et al., 2020). In general, greater diversity generates more rapid innovation.[5] Conceptually, our approach suggests that a particular kind of diversity fuels innovation because these factors are associated with individuals possessing different skills, techniques, knowledge (explicit beliefs), tacit know-how, intuitions and perspectives.

Finally, both social institutions and psychological traits that facilitate the exchange of ideas have been linked to innovation. As noted above, saloons, cafés and knowledge societies have all been linked to innovation (Mokyr, 1995; Andrews, 2023; Andrews and Lensing, 2020; Cinnirella et al., 2022; Henrich, 2020). Similarly, psychological traits that motivate people to (1) tolerate, trust and cooperate with strangers and (2) express non-conforming ideas, views and perspectives have been linked to innovation. For example, focusing on trust at the levels of countries and U.S. states, Algan and Cahuc (2014) reveal positive correlations between impersonal trust, based on the impersonal trust question, and three measures of innovation. Similarly, using U.S. firm-level data, Nguyen (2021) shows that more trusting CEOs generate an uptick in innovation upon their arrival. Conceptually, these social institutions and aspects of psychology foster the flow of ideas among diverse minds, increasing the likelihood of useful recombinations.

The economics literature often describes the flow of information or exchange of skills among minds as 'knowledge spillovers' or 'skill complementarities.' While our approach here certainly includes these, we think it is important to consider a broader class of cultural and cognitive diversity (Muthukrishna and Henrich, 2016; Page et al., 2019). Across populations, people rely on different languages, thinking styles, decision heuristics,

[4]Work on patents converges with efforts in other domains. Consider three examples. First, using scientific citations to assess recombination, Uzzi et al. (2013) find that the highest-impact scientific papers drew on journals rarely referenced by others in the same journal but were, in the main, otherwise highly conventional in their referencing patterns. Second, using detailed analyses of 21,745,538 lines of computer code based on entries in programming competitions over 14 years, Miu et al. (2018) shows that entries largely copied prior leading entries, which were publicly available, and then added novelty by recombining code drawn from other prior entries. Recombination was, by far, the key element that led to the gradual improvement of these algorithms over time. Finally, Thagard (2012) coded lists of the top 100 most important inventions and scientific discoveries of all time and found them all to involve conceptual recombinations. Based on work in cognitive science, he argues that all creativity arises from recombination based on neuroscientific models of how brains actually form new ideas.

[5]AlShebli et al. (2018), for example, show how both the ethnic and disciplinary diversity of coauthors are linked to scientific impact.

reading preferences, metaphors, attentional biases and ritual practices (Henrich, 2020; Nisbett, 2003). 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). Indeed, we will show that our measure of surname diversity accounts for substantial variation in innovation across U.S. counties even when occupational diversity is held constant (Table D3).

As such, we take a broad perspective on the link between diversity and innovation. A more diverse local population may increase the diversity of the workforce, which then fuels innovation through skill complementarities in production teams. Yet, casual observations of, and interactions with, people in non-professional contexts can likewise fuel recombinative processes and more so in highly diverse local populations. So, while we do not disentangle these different channels, the fact that most patents are attributed to single inventors suggests that at the turn of the 19th century, the latter channel—interactions other than within production teams—was likely important. The average number of inventors per patent is roughly 1.4 and remarkably flat over most of the period of our analysis, as shown in Figure B4.[6]

2.2 Operationalizing the Diversity of Social Interactions with Surnames

To conceptualize the diversity of social interactions, consider a subpopulation consisting of N individuals partitioned into K groups, each of size N_k such that $\sum_{k=1}^{k=K} N_k = N$. Each group carries unique information (e.g., skills, know-how, metaphors), labeled as $s_k \in \{a, b, c, d, ...\}$ and $s_k \neq s_h$ for all $k \neq h$. In practice, these groups could be extended families, occupations, castes, clans, ethnicities, birth country, or other geographical units, though below we will explain why surnames are a particularly potent partition. When individuals from different groups meet, the likelihood of recombination and innovation increases. Information theory (Shannon, 1948) tells us that the average informational content (or the innovation potential) of such a population is

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

[6]At the same time, our data do permit us to test whether patenting teams (those with multiple authors) are more likely to create breakthrough innovations when the surname diversity of the team is greater. Indeed, as Table B2 shows, if we focus only on patents with multiple inventors (2+), those with only inventors who carry the same surname are less likely to generate breakthroughs. This holds with both year and technology category fixed effects. While this cannot explain our overall effects, because most patents are solos, it does suggest that surname diversity even operates at the level of the individual patent.

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). This is a version of Shannon entropy.

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 not surprising (the outcome can be predicted perfectly), the entropy is 0, and thus, no recombinations can arise through social interaction. On the other hand, entropy for a population with a fixed number of groups is maximized if all groups are equal in size. An individual who goes out for a random social interaction in such a diverse population 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.[7]

Conceptually, we expect surname diversity to proxy for both the informational and social-psychological aspects of social interactional diversity. Informationally, we hypothesize—and later provide evidence—that some of the important informational diversity occurs among groups identified by surnames. The idea here is that people who hold the same surnames share much information among themselves, which can occur through vertical as well as both horizontal and oblique forms of cultural transmission (Cavalli-Sforza and Feldman, 1981; Bell et al., 2019). At the same time, people who hold different surnames differ in their socialization and their social networks as well as their cultural heritage; ultimately, this results in the clustering of many kinds of information—skills, know-how, customs, languages and thinking styles—within groups defined by surnames.

Socially and psychologically, we hypothesize—and later test empirically—that surname diversity is associated with weaker family ties and lower trust in strangers. Consistent with the literature on the impact of kinship and family ties on trust (Enke, 2019; Alesina and Giuliano, 2014; Schulz et al., 2019), we suspect that increasing surname diversity not only constrains individuals' ability to meet their needs within their own (shrinking) group,

[7] In economics, a Herfindahl measure, which population geneticists call Isonomy, is frequently used to capture diversity. For our purposes, however, Shannon's entropy has several advantages to conceptualize informational diversity and has favorable mathematical properties (Carcassi et al., 2021). In particular, a Herfindahl approach 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 ($\rho = 0.80$, Table C1).

but also creates more opportunities for exchanges outside of one's group. These increased interactions outside one's group weaken family ties and foster interpersonal trust.

Notably, this proposal may run counter to the intuitions of many readers, who might expect rising diversity to create social miscoordinations or trigger in-group biases that inhibit social interactions, thereby thwarting increases in the opportunities for recombinative innovation created by informational diversity. First, if these inhibiting factors played a significant role, we would expect to see a negative or concave relationship between our surname entropy measure and innovation. As we will demonstrate, we do not observe this. Second, conceptually, it is important to distinguish between fine-grained partitions like surnames and larger-scale divisions such as tribes, ethnicities, and nationalities. As the size of extended families and surname groups shrink relative to the entire population, interactions outside of one's own group—even if riskier—become both more necessary and more profitable. Thus, unlike larger-scale measures of diversity, greater surname diversity is more likely to create opportunities for beneficial exchange with individuals outside one's group. In line with this view, Bazzi et al. (2019) find that a high degree of fractionalization (the presence of many small groups) fosters trust, whereas polarization (the presence of a few large groups) creates intergroup antagonism. Relatedly, a substantial body of anthropological and other evidence indicates that preferential interactions often occur at the boundaries of languages, ethnicities, social norms, and religions (Handley and Mathew, 2020; White et al., 2021; Desmet et al., 2017). We find a similar pattern in our context. Greater surname diversity is associated with higher levels of sociality, as measured by impersonal trust, while race and country-of-birth diversity tend to show the opposite relationship. This finding underlines the importance of distinguishing between different forms of diversity. Consequently, we will see that diversity measures based on national origins and race do not establish the same convex relationship with innovation, and in the case of racial diversity, the relationship is inversely U-shaped.

2.3 U.S. Surname Diversity 1850-1940

To calculate the Shannon measure of surname diversity, the data source is 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 how we harmonize county boundaries. We implement the Philips (1990) phonetic algorithm, *metaphone*, to deal with misspellings in the name string. Following Burchardi et al. (2021), we also obtain the variables age and `yrimmig` (the year of immigration) to estimate surname diversity for the mid-decades 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the mid-decade.

Figure B1 maps our measure of surname diversity for U.S. counties in the year 1940, partialing out population size.[8] Clear geographical patterns emerge. While counties in California and most of the Northeast score high on surname diversity, Utah and the Southern states score substantially lower, i.e., they are more homogeneous with regard to surnames and hence less culturally diverse. Surname diversity correlates moderately with more common diversity measures based on country of birth and occupation (Table B1). This moderate correlation provides support for the notion that surname diversity not only encompasses variation within these broader categories but also captures a so far unobserved slice of diversity at the level of families.

2.4 Surnames Capture Relevant Social Interactional Diversity

We now empirically establish that our measure of surname diversity captures both the informational dimension—i.e., surnames are indicative of occupation and ancestral origins—and the social-psychological aspect— i.e., surname diversity correlates with impersonal trust and the strength of family ties. Our endeavor here is not to establish causal linkages, but merely to reveal the kinds of empirical relationships one would expect if surname diversity indeed captures a population's diversity of social interactions, i.e., the degree that exchange of ideas among diverse people occurs.

Recent work in the social mobility literature demonstrates that surnames capture unique skills, socialization, and know-how (Clark, 2014; Güell et al., 2015; Barone and Mocetti, 2021). This is not surprising given that surnames are indicative of ancestral geographic origins, professions, and (family) lineages. Even though very frequent surnames will capture family- or profession-specific traditions to a lesser degree, these surnames nevertheless still encapsulate unique knowledge. For example, "Smith" and "Johnson" are the most frequent surnames in the 1880 census. The Smiths outnumber the Johnsons by a factor of about 1.65 times. However, among individuals who reported blacksmith as their occupation, there are 2.46 times more Smiths than Johnsons. This makes sense given that the surname "Smith" has a long history of association with metalworking and blacksmithing, and the data show that it is still a relatively more common surname among metalworkers in 1880.

To gain a systematic sense of the degree to which surnames reflect unique knowledge

[8]Figure B2 displays the variation in surname diversity conditional on log county population from 1850 to 1930. Figure B3 shows the variation in changes in surname diversity from 1850 to 1940, i.e., first-differenced surname diversity conditional on first-differenced log county population.

Table 1: Surnames cluster in occupations, birthplaces, and patent fields

	Occupation				Country of origin		Region of origin	USPO tech category	
Sample:	All		Immigrants				Germans	Inventors	
Year:	1880	1940	1880	1940	1880	1940	1880	1880-9	1940-9
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Surname	0.117	0.045	0.097	0.065	0.393	0.227	0.189	0.092	0.068
U.S. county of residence	0.171	0.071	0.153	0.080	0.288	0.130	0.159	0.014	0.017
Country of origin	0.120	0.043	0.096	0.060					
Age	0.154	0.049	0.101	0.048					

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 1 (2) includes all individuals in the 1880 (1940) census. Columns 3 and 5 (4 and 6) include all immigrants in the 1880 (1940) census. Column 7 restricts the sample to German immigrants in 1880. Column 8 (9) includes all inventors of patents issued from 1880 to 1889 (1940 to 1949). In column 7, we use the 31 subnational regional origins for German immigrants recorded by the Census (the variable `bp1d` with codes 45301 to 45361).

in our dataset, we calculated Herfindahl concentration measures that capture how strongly surnames cluster in several domains, including occupations, country or region of origin, and technology categories of patents. The construction of the concentration measure for each domain proceeds in two steps. For example, in the case of occupation, we first calculate a normalized Herfindahl index for each surname across all occupations. This gives us a measure of how strongly a specific surname clusters in occupations. We normalize this measure such that it is zero in the case of a uniform surname distribution and one if the surname is only found within a single occupation. Second, we average the surname-specific Herfindahl indices across all surnames, weighted by the number of people with a given surname. This averaged index reveals the overall surname concentration in occupations based on the U.S. population. Similarly, we construct the concentration measures for the other domains.

Table 1 reports the surname concentration indices for the different domains, samples, and years in the first row. All surname concentration indices are well above zero, indicating that surnames are concentrated in occupations (columns 1 to 4), originating countries (columns 5 and 6), originating regions within Germany (column 7)[9], and patent technology categories (columns 8 and 9).[10] For example, in 1880, two people with the same surname

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

[10]The patent data set does not allow us to uniquely identify inventors. Hence, we are unable to detect

have a roughly 12% (above chance) probability of holding the same occupation out of 249 possible occupations (column 1), or two same-surname immigrants have about a 39% probability of being from the same country of origin (column 5). Moreover, column 7 reports that surnames even indicate the sub-national origin region of immigrants. Same-surname immigrants from Germany have a 19% probability of being from the same inner-German region (out of 31 regions).

Having established that certain surnames concentrate in occupations, originating countries, regions, and patent categories, we can put the concentration indices into context by comparing them to measures of residence-county (row 2), country of origin (row 3), and age (row 4) concentration. In the year 1880, occupations are relatively more concentrated in counties compared to surnames, though this difference markedly narrows in the year 1940 (row 2). Surnames are substantially more indicative of originating countries and regions compared to immigrants' residence counties. Compared to country of origin and age, surnames are about equally indicative of occupation (rows 3 and 4).[11]

Moving now to focus on the social-psychological side of surname diversity, we analyze the correlation between surname diversity and (i) responses to the impersonal trust question (in the General Social Survey using waves from 1972 to 2016) as well as (ii) the strength of family ties, because many studies provide evidence that strong family ties hamper impersonal trust (Alesina and Giuliano, 2014; Schulz et al., 2019). First, surname diversity correlates negatively with the strength of family ties (Figure C2).[12] The strength of this relationship increases from roughly −0.2 during 1860-80 to −0.60 by 1940 (see

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.

[11] 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 cultural heritage. To address this, we construct a more finely-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 C1 shows that the main surname diversity indicator in 1940 is almost perfectly correlated with those more finely-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, and (iii) surnames of whites only.

[12] Here, following Raz (2023), the strength of family ties is captured 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 person who is not their relative, and (iv) the mean size of families.

Section 6.2 for causal evidence on this relationship). Second, consistent with the family ties literature, we find that surname diversity in 1940 correlates positively with impersonal trust (Figure C1). The strength of this relationship over time is remarkably stable as we travel back in time from 1940 to 1870.

Notably, none of these patterns exist for the more conventional measures of diversity based on country of birth and race (Figure C3 to C6). Thus, these findings underline the distinctive nature of surname diversity compared to these other types of diversity. They are also consistent with previous work showing that high fragmentation fosters trust, while polarization creates intergroup antagonism (Bazzi et al., 2019).

Overall, the analyses presented in this section provide *prima facie* evidence that our measure of surname diversity captures both the informational and social components of recombinative innovation.

2.5 Measuring Innovation

To measure innovation, we rely on patent data. Our first measure is the total number of patents per 1,000 individuals. We calculate this measure for each U.S. county for 5 and 10-year periods from 1850 to 1940, based on the Comprehensive Universe of U.S. Patents (CUSP) data set compiled by Berkes (2018). The primary source of this data set is Google Patents supplemented with information from other sources.

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 be 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 (Burchardi et al., 2021; 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; Burchardi et al., 2021). Overall, while far from perfect, current evidence suggests that patents offer a valuable proxy for innovative activity.

Nevertheless, we also deploy a second measure of innovation based on breakthrough patents (per 1,000 people). Developed by Kelly et al. (2021), this approach analyzes the text accompanying each patent by comparing it with the text from both past and future patents. Assessments of breakthroughs are based on each patent's (1) Novelty: how distinct is it from prior patents? and (2) Impact: how similar is it to future patents? This measure aims to capture patents that meaningfully advance knowledge by filtering out minor patents, often filed for strategic reasons, that add noise to the creative signal we seek. Details on both our patent-based measures can be found in Appendix A.

3 Least Squares Estimates

Our analysis is structured into four sections. Here, in Section 3, we begin by reporting the estimates from least squares regressions to establish a robust positive relationship between surname diversity and innovation, using the full decadal dataset from 1850 to 1940. Section 4 provides causal evidence on this relationship based on our instrumental variable strategy, which exploits the pseudo-random nature of immigration flows into U.S. counties. Section 5 offers a variety of robustness checks, including the use of surname-fixed effects. Finally, Section 6 provides evidence supporting the hypothesized mechanism.

Table 2 reports least squares estimates of the relationship between surname diversity and innovation of both patents (Panel A) and breakthrough patents (Panel B) per 1,000 people. Column 1 reports the bivariate relationships between the two innovation outcomes and surname diversity (including period-year fixed effects to control for trends affecting surname diversity and innovation across all counties). We find positive and significant relationships: a one standard deviation higher surname diversity is associated with approximately 1.76 more patents per capita (Panel A) and 0.15 breakthrough patents per capita (Panel B). Relating these coefficients to the mean of the dependent variable suggests that a one standard deviation increase in surname diversity is associated with roughly 80% more patents and 85% more breakthrough patents.

In column 2 of Table 2, we report estimates of the relationship between innovation and standardized country of origin diversity, which is a more conventional measure of cultural diversity (e.g., Ottaviano and Peri, 2006; Ager and Brückner, 2013; Alesina et al., 2016). We find a qualitatively similar relationship with innovation, suggesting that country-of-birth diversity also fuels the recombinative processes.

Figure D1 displays the relationship between surname diversity and innovation visually. It reveals tight, strong relationships and offers no hint of a "hump" shape. Instead, the plots suggest a convex relationship—that is, greater surname diversity is associated with relatively more (breakthrough) patents per capita. By contrast, the relationship between country of birth diversity and innovation is more linear (Figure D2), and the relationship between race diversity and innovation is concave and hump-shaped (Figure D3).

Table 2: Least squares estimates: surname diversity and innovation from 1850s to 1940s

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	Patents per 1,000 people (mean = 2.26, sd = 2.58)					
Surname diversity	1.76***		0.735***	0.705***	0.856***	0.502***
	(0.175)		(0.156)	(0.160)	(0.176)	(0.168)
Country of origin diversity		1.64***	1.10***	1.18***	1.07***	0.165
		(0.093)	(0.181)	(0.175)	(0.232)	(0.176)
R^2	0.503	0.550	0.574	0.607	0.691	0.864
Panel B:	Breakthrough patents per 1,000 people (mean = 0.18, sd = 0.24)					
Surname diversity	0.154***		0.064***	0.059***	0.064***	0.044**
	(0.021)		(0.012)	(0.012)	(0.015)	(0.021)
Country of origin diversity		0.144***	0.098***	0.106***	0.113***	0.004
		(0.013)	(0.019)	(0.018)	(0.024)	(0.015)
R^2	0.416	0.459	0.480	0.524	0.619	0.787
Immigrant shares by country of origin (59 shares)				✓	✓	✓
Period fixed effects	✓	✓	✓	✓		
Period-State fixed effects					✓	✓
County fixed effects						✓
Observations	22,222	22,222	22,222	22,222	22,222	22,222

Notes: The table reports estimates of least squares regressions of innovation outcomes on surname diversity and immigrant diversities. In Panel A (Panel B), the outcome is the number of (breakthrough) patents issued in a given period per 1,000 people. The unit of observation is a county-period from 1850 to 1940 (excluding the midyears). Observations are weighted by county population in 1850. Standard errors are clustered on states and reported in parentheses. 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.

For ease of exposition, Table 2 reports linear specifications which do not take into account the non-linearities which are visible in Figure D1. However, the size of the estimated linear approximations (expressed in percent) is very similar in case we apply the inverse hyperbolic sine (IHS) transformation to the outcome variable (Table D1). A visual check confirms that such a transformation of the dependent variable linearizes the relationship to some degree, which is displayed in Figure D4. Throughout the paper we report on the untransformed dependent variable because the IHS transformation is not scale invariant (Aihounton and Henningsen, 2020). In any case, all results are qualitatively and quantitatively similar when using IHS transformed dependent variables.

In column 3, we include both diversity measures in the regression. The coefficient of surname diversity is significant, suggesting that surname diversity captures an important slice of diversity that is not captured by the coarser country of origin diversity measure.

In column 4, we consider the share of immigrants from each of the 59 originating countries consistently recorded in the census data between 1850 and 1940 to account for the role of immigration from specific countries. We find that the inclusion of these variables does not significantly alter the coefficients on surname diversity. This implies that recent immigration does not confound the relationship between innovation and surname diversity.

In column 5, we add period-state fixed effects. The coefficients on surname diversity remain virtually unchanged, suggesting that the relationship between surname diversity and innovation is not driven by persistent or time-varying differences across states, including North-South differences (e.g., Cook, 2014).

In the final specification in column 6, we include county fixed effects to examine the relationship between *changes* in surname diversity and innovation. This specification utilizes only the variation within each county over time that is distinct from other counties within the same state and is not due to immigration. As reported in column 6, the estimates for surname diversity decrease but remain substantial: a one standard deviation increase in surname diversity is associated with an increase of roughly 23% in both patents and breakthrough patents per 1,000 people. Interestingly, while the coefficient on surname diversity remains large and well estimated, the coefficients on country-of-origin diversity are no longer significant when surname diversity is held constant.

Although our specification focuses on innovations per capita, there remains a concern that this specification is restrictive and does not adequately capture population-scale effects. To address this, Table D2 controls for counties' population size. The coefficients on surname diversity remain large, well-estimated, and robust across specifications for both of our measures of innovation.

To gain further insights into the relationship between surname diversity and other diversity measures, Table D3 reports regressions that also include diversity measures based on race and occupations. In all specifications, we find that surname diversity is consistently significantly associated with patents.

We also explore the role of education and show that surname diversity is a significant predictor of innovation, controlling for the average education of a county's population in the year 1940 (Table D4).

Lastly, to check the stability of the relationship between surname diversity and innovation over time, we estimate a coefficient for surname diversity for each decade separately (in a specification that otherwise parallels column 1 of Table 2). Figure D5 shows that all the estimated coefficients for surnames are positive, significant, and sizable.

4 Instrumental Variable Estimates

The previous section established that surname diversity strongly correlates with the quantity and quality of patenting across counties from 1850 to 1940. However, these estimates could potentially be biased due to reverse causality or unobserved factors. For instance, migrants who possess rare surnames might be more likely to move to innovative counties, subsequently increasing surname diversity. Similarly, highly-skilled individuals, who are more prone to innovate, might have a preference for more diverse counties. In both instances, we could witness a correlation between surname diversity and innovation even if there is no causal relationship.

To address these potential biases, we employ an instrumental variable (IV) approach to isolate quasi-random variation in surname composition, and consequently, in surname diversity across counties. We will use this instrument to estimate the causal impact of changes in surname diversity on innovation at both the county level and surname-county level, from 1900 to 1940. As we detail in the following sections, our instrumental variable is based on predicted surname stocks, which are influenced by factors we can reasonably argue are exogenous to local innovation dynamics. This approach has the advantage of mitigating concerns about reverse causality and local unobservable factors that might simultaneously affect both innovation and surname diversity.

4.1 Construction of the Instrument and Estimating Equations

The central idea underlying our IV strategy stems from the observation that immigration, a major determinant of the change in counties' surname diversity, provides an avenue to isolate variation in surname diversity that is independent of any unobserved determinants of innovation. Thus, we can build on recent advances in the immigration literature to remove potential bias from our estimates. By leveraging this approach, we estimate the local average treatment effect (LATE) of changes in surname diversity on innovation, as induced specifically by immigration to the U.S., not those changes stemming from births, deaths, marriage, or domestic migration.

While we rely on changes in surname composition induced by immigration for our IV strategy, it is crucial to understand that immigration inflow does not necessarily lead to a straightforward, monotonically increasing relationship with surname diversity. Immigration can both decrease and increase surname diversity, depending on the preexisting local distribution of surnames. For instance, if individuals with the surname 'A' immigrate to a county where few or no other individuals named 'A' reside, then surname diversity increases. Conversely, if they migrate to a county where many individuals with the name 'A' already reside, then surname diversity decreases. To reinforce this point, we demonstrate that even when controlling for the changes in population induced by migration, our IV results on surname diversity remain consistent (see section 4.2 for a detailed discussion on identification). The essence of our strategy is that immigration affects surname composition in complex ways, and we exploit quasi-random variation in immigration to create an instrument for surname diversity.

Increased surname diversity due to immigration implies a more heterogeneous knowledge pool. Diverse surnames bring with them varied experiences, skills, and knowledge from different cultures and regions, facilitating the innovative recombination of ideas. This variety contributes to the informational component of our hypothesis. Moreover, when surname diversity is high, the relative sizes of extended families or surname groups shrink. This scenario promotes engagement with outsiders beyond immediate kin members and cultivates impersonal trust, constituting our social-psychological component. We will provide evidence for both channels in Section 6, when we discuss mechanisms.

Our methodology adapts the IV strategy developed by Burchardi et al. (2019) to our context. The construction of this instrument requires two steps. First, we isolate quasi-random variation in the stock $N^t k, i$ of each surname k residing in county i in census year t based on historical migration patterns. Second, we compute the instrument for surname diversity by calculating diversity based on these (predicted) quasi-random stocks of surnames $\hat{N}^t k, i$. We will now delve into the details of these two steps.

Step 1: Isolating Quasi-random Variation in Counties' Surname Stocks We adopt Burchardi et al. (2019)'s historical push-pull approach to isolate quasi-random variation in the composition of surnames in U.S. counties. The approach assumes that a combination of push and pull factors jointly determines the allocation of immigrants with specific surnames to counties and that the historical interactions of these two factors generate quasi-random variation in surname stocks that persists over time.

Empirically, the push factor is summarized by the total number of immigrants with a given surname entering the U.S. during a specific period; the pull factor is the attractiveness of a county in this period, operationalized by the share of immigrants (out of all immigrants entering the US) who settle in this county during the same period. These two factors vary over time, and their interactions, which we can trace back to 1880, generate quasi-random variation in a county's distribution of surnames.

Formally, we predict the stock of people $N_{k,i}^t$ (in thousands) with surname k residing in county i in year t by estimating the following zero-stage equation:

$$N_{k,i}^{t} = \delta_i + \delta_{k,r(i)} + \sum_{\tau=1880}^{t-1} b^{\tau} \underbrace{I_{k,-r(i)}^{\tau}}_{\text{Push}} \underbrace{\frac{I_{i,-k}^{\tau}}{I_{-k}^{\tau}}}_{\text{Pull}} + \sum_{\tau=1880}^{t-1} d^{\tau} \frac{I_{i,-k}^{\tau}}{I_{-k}^{\tau}} + u_{i,k}, \tag{2}$$

where i indexes counties, k denotes surnames, t indexes census years from 1900 to 1940, including the midyears, and $r(i)$ denotes the census region containing county i. The variable $I_{k,-r(i)}^{\tau}$ is the push factor in the period ending in year τ (1880, 1895, 1900, 1905, 1910, 1915, 1920, 1925, 1930). It is given by the total number of migrants (in thousands) with surname k who arrive in the U.S. during this period and settle *outside* the region containing county i. The pull factor captures the relative attractiveness of a specific county i during the period ending in τ. It is given by the share of migrants a county attracts $\frac{I_{i,-k}^{\tau}}{I_{-k}^{\tau}}$, where $I_{i,-k}^{\tau}$ is the total number of migrants who settle in county i during this period and who do not have surname k, and $I_{-k}^{\tau} = \sum_i I_{i,-k}^{\tau}$ is the total number of migrants who settled in the U.S. during the same period and who do not have surname k.[13]

Core to the identification strategy are the historical interactions between the push and pull factors in each period τ (up to period $t-1$). We estimate a coefficient for this interaction, b^{τ}, for each period stretching back to the year 1880 (the earliest period for which we have data on immigrants or their parents). That is, equation (2) attributes the stock of each name in a county (in a given year t) to the past inflow of migrants who are allocated according to the push-pull factors over the course of several decades.

In addition to the push-pull factors, equation (2) also includes the term $\sum_{\tau=1880}^{t-1} d^{\tau} \frac{I_{i,-k}^{\tau}}{I_{-k}^{\tau}}$, 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. Moreover, δ_i, denotes county fixed effects, removing any time-invariant factors that make specific counties more attractive to all migrants. $\delta_{k,r(i)}$ are name-region fixed effects. They remove time-invariant unobserved factors that may make specific census regions more attractive to migrants with certain surnames.

Based on equation (2) we estimate the coefficients $\hat{b}^{\tau}$ for each period τ and then

[13] We follow Burchardi et al. (2019) and estimate equation (2) using a leave-out approach. That is, we exclude migrants with surname n from the pull factor (denoted by $-k$), and we exclude the census regions r that county i is located in from the push factor (denoted by $-r(i)$). This leave-out approach ensures that our estimates are not driven by the settlement outcomes of migrants with surname k who settled in region $r(i)$. We note, though, that at the level of surnames, this is likely less of a concern because the fractions of surnames relative to all migrants are small.

calculate the predicted stocks of name k in county i at time t as

$$\hat{N}^t_{k,i} = \sum_{\tau=1880}^{t-1} \hat{b}^\tau \left(I^\tau_{k,-r(i)} \frac{I^\tau_{i,-k}}{I^\tau_{-k}} \right)^\perp$$

where $\hat{b}^\tau$ is the estimate of b^τ from equation (2), and $\left(I^\tau_{k,-r(i)} \frac{I^\tau_{i,-k}}{I^\tau_{-k}} \right)^\perp$ are residuals of a regression of the push-pull interaction, $I^\tau_{k,-r(i)} \frac{I^\tau_{i,-k}}{I^\tau_{-k}}$, on δ_i, $\delta_{k,r(i)}$ and $\frac{I^\tau_{i,-k}}{I^\tau_{-k}}$. This residualization ensures that the predicted stock of each name $\hat{N}^t_{k,i}$ relies on the component of the push-pull factors that is orthogonal to the control variables included in equation (2). This orthogonalization is particularly useful with regard to $\frac{I^\tau_{i,-k}}{I^\tau_{-k}}$, because it ensures that the instrument is orthogonal to the past attractiveness of a county, which could be driven by an underlying factor that also determines innovation decades later.

Step 2: Calculating the Instrument for Surname Diversity In step 2, we compute the instrument for surname diversity by applying the entropy formula on the predicted stock of each surname $\hat{N}^t_{k,i}$:

$$\widehat{\text{Surname diversity}}^t_i = -\sum_k \left(\frac{\hat{N}^t_{k,i}}{\sum_k \hat{N}^t_{k,i}} \log \left(\frac{\hat{N}^t_{k,i}}{\sum_k \hat{N}^t_{k,i}} \right) \right)$$

We repeat steps 1 and 2 eight times to obtain an instrument for diversity in each of the eight periods (ranging from $t = 1900$ to $t = 1940$) that form part of our panel analysis.

Step 3: IV Estimating Equations We implement our IV procedure using 2SLS. The equations are given by equations (3) and (4), where equation (3) is the first stage and equation (4) is the second stage.

$$\text{Surname diversity}^t_i = \gamma \, \widehat{\text{Surname diversity}}^t_i + \mu_{t,s(i)} + \mu_i + v^t_i \tag{3}$$

$$Y^t_i = \beta \, \text{Surname diversity}^t_i + \alpha_{t,s(i)} + \alpha_i + \varepsilon^t_i \tag{4}$$

where i indexes counties, s states, and t census years (including the midyears). Y^t_i is county i's number of (breakthrough) patents per 1,000 people in the five-year period starting in t. Surname diversity^t_i is county i's surname diversity in t; and $\widehat{\text{Surname diversity}}^t_i$ is county i's predicted surname diversity in t described above. The coefficient β is our main interest.

Equations (3) and (4) also include state-period fixed effects, $\mu_{t,s(i)}$ and $\alpha_{t,s(i)}$, and county fixed effects, μ_i and α_i. By including these fixed effects, β is estimated from changes in surname diversity within the same county over time while controlling for persistent and time-varying differences across states.

In addition, several specifications include county-specific linear time trends, such that β captures the relationship between deviations in the changes in diversity and innovation within counties over time relative to their overall trend. Comparing the estimates of these specifications to the baseline estimates of equation (4) provides another exogeneity check of the instrument. If the estimates remain similar, this suggests that the instrument is orthogonal to persistent or gradually growing county-level confounding factors.

4.2 Identification

Our identification strategy is valid if $\widehat{\text{Surname diversity}}_i^t$ is truly exogenous in the specification of equation (4). A sufficient condition for this to hold is

$$\left(I_{k,-r(i)}^{\tau} \frac{I_{i,-k}^{\tau}}{I_{-k}^{\tau}} \right)^{\perp} \perp \varepsilon_i^t.$$

It requires that any factor affecting a county's innovation in t is independent of the interaction between the orthogonalized historical push-pull factors. If this condition holds, the predicted stocks of surnames are exogenous to innovation (Step 1), and so is the instrument for diversity (Step 2).

An important question regarding the validity of this empirical strategy is whether past push-pull factors are independent of a county's future innovative capacity. It is possible, for example, that migrants preferred to settle in counties that were more innovative in the past, likely increasing their diversity, and those same counties are subsequently still more innovative. This possibility would give rise to reverse causality. More generally, (persistent) unobserved factors may determine both the past pull factors and future innovation, which may create a correlation between the push-pull instrument and the error term.

In their paper, Burchardi et al. (2019) detail why this possibility is unlikely given the substantial variation in the push-pull factors over time and space. Empirically, we address this concern in three ways: First, we orthogonalize our push-pull instrument with regard to the historical attractiveness of a county as captured by the fraction of immigrants who settled there over time (see our zero-stage equation (2)). Consequently, our IV estimates do not reflect unobserved persistent factors that had already manifested themselves in immigrants' past settlement decisions. Second, in several specifications, we control for county-specific linear time trends. To the degree that these linear time-trends capture persistent unobserved factors, they will mitigate concerns of estimation bias. Lastly, and most importantly, we conduct a falsification exercise and regress previous-period innovation on subsequent surname diversity. We do not find any evidence for reverse causality, i.e., a shock to surname diversity is statistically unrelated to previous-period innovation (see Section 5.1). Therefore, it is unlikely that our estimates are driven by reverse causality or persistent unobserved confounders.

Another concern is that unobserved individual characteristics co-determine settlement patterns and innovation. For example, people with a high (unobserved) propensity to innovate may prefer to settle in relatively more diverse counties. In this case, the observed relationship between diversity and innovation would be due to the settlement preferences of individuals with high innovative capacity and not due to diversity per se. The IV approach addresses this concern because the predicted surname stocks in a county are solely determined by the interaction of the historical push and pull factors, i.e., the allocation of migrants to counties does not rest on individual preferences.[14] This push-pull instrument is orthogonal to county fixed effects and surname-region fixed effects. Thus, our estimates cannot be biased by the unobserved stable settlement preferences of people with a certain surname. In addition, in Section 5.4, we further address this concern by devising a specification with surname-fixed-effects. This specification absorbs any genetic, environmental, or acquired characteristics embodied in surnames and, thus, it captures the pure' diversity effect, which is independent of the type of information embedded in surnames. Taken together, it is unlikely that our results are biased due to individual characteristics that co-determine settlement patterns and innovation.

A final concern is the possibility that there is a direct effect of immigration (other than through diversity) that confounds the estimates. Although we establish that immigration impacts counties' surname composition, conceptually, there is not a simple monotonically increasing relationship between immigration and surname diversity. The extent to which immigrants impact surname diversity depends on the surname composition of the immigrants in comparison to the local (county) population. For example, if migrants predominantly hold the same names as the dominant local groups, then immigration will decrease surname diversity. The fact that the least squares estimates in Table 2 hardly change when we add controls for the share of immigrants from different countries likely attests to these considerations and provides evidence that the relationship between surname

[14]The exclusion restrictions could be violated if the push-pull factors primarily reflect the migration decisions (= preferences) of people with a specific surname. Yet, this is unlikely because any specific surname makes up only a tiny fraction of all people entering the U.S. in a given period (the push factor) and a small fraction of migrants settling in a county (the pull factor). Nevertheless, we follow Burchardi et al. (2019) and report leave-out estimators such that the push factor does not contain individuals with surname n and the pull factor does not contain regions in which a county is located in $r(i)$.

diversity and innovation is not confounded by a direct effect of immigration, operating via alternative channels.

In the IV specifications, we avoid controlling for endogenous variables such as the number of immigrants, and our push-pull IV strategy (which is based on surname stocks, not flows) does not allow for a straightforward way to instrument for it. Rather, we rely on alternative strategies. First, as explained above, our instrument is orthogonal to each county's immigration history, which mitigates endogeneity concerns. Second, we conduct a robustness check by instrumenting for each county's population size. This instrument is likewise based on the push-pull approach (calculated by taking the sum over all predicted stocks of surnames). As such, the IV specification controls for the *migration-induced* changes in county populations, which proxies for immigration inflows. We show that our IV results hold when controlling for the instrumented population. Finally, when estimating the IV specifications in a subsample that excludes immigrant innovators, we find positive and sizeable effects of surname diversity on patenting (Section 5.5).

4.3 Zero-stage Estimates

We report the zero-stage estimates of equation (2) in Table D5. These estimates allow us to obtain predicted stocks for each surname in each county for each time period, which we will use to compute the diversity instrument. In total, we estimate equation (2) eight times, once for each period from 1900 to 1940.

The results indicate that we identify variation in the stock of surnames based on the push-pull factors stretching across the full range of periods in our sample. For example, the estimates reported in column 8 suggest that push-pull factors as far back as 1880 and all the way up to 1930 are significant predictors of the stock of surnames in 1940. [15]

Using the estimated models shown in Table D5, we calculate the predicted (and orthogonalized) stock of each surname in each county for each of the eight periods. Finally, we compute the instrument for surname diversity by applying the entropy formula to the predicted surname stocks.

The predicted values of surname diversity for each period from 1900 to 1940 are depicted in Figure 1. To ensure that the variation we are highlighting is truly due to differences in exogenous surname diversity, and not to other county or state-period specific factors, we remove county and state-period fixed effects. The maps demonstrate that our instrument picks up substantial variation both over time and across counties, validating

[15] Qualitatively, our results parallel those of Burchardi et al. (2019), who estimate the push-pull factor at the level of originating countries (not surnames). They, for example, likewise obtain a negative coefficient for the interaction for the period ending in 1930, a period with a high degree of out-migration.

its utility in our analysis.

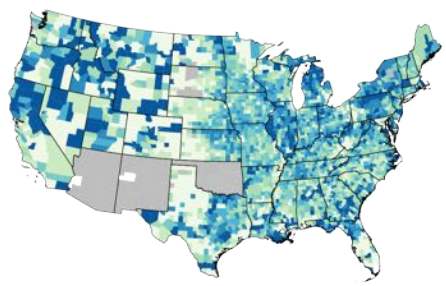
4.4 IV Estimates

Turning to the IV analysis, Table 3 reports first-stage, reduced-form, and second-stage estimates. Starting with the first-stage estimates reported in Panel D, we find that the instrument is strongly correlated with actual surname diversity, with a Kleibergen-Paap F-statistic of around 51 in our baseline specification in columns 2 and 5. The F-statistic shrinks to roughly 28 if we add controls for county-specific linear time trends (columns 3 and 6). The point estimates imply that a one standard deviation increase in the instrument is associated with 0.43 standard deviation greater surname diversity in the baseline (columns 2 and 5) and with a 0.39 standard deviation greater surname diversity when conditioning on county-specific linear time trends (columns 3 and 6). Taken together, the first-stage relationship of the instrument with surname diversity is highly significant, and the F-statistics of the excluded instrument in all specifications exceed the conventional thresholds commonly used in the literature.

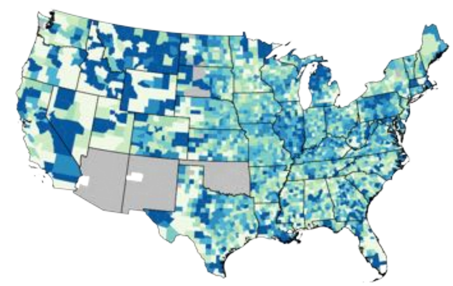
Figure D6 shows binscatter plots that show the first-stage relationship between the instrument and actual surname diversity, both with and without controls for county-specific time trends. They demonstrate that the relationship is strong, linear and not driven by a small set of observations.

We next turn to the estimates relating surname diversity to innovation. Table 3 presents the estimates for both main outcome variables—patents per 1,000 people (columns 1 to 3) and breakthrough patents per 1,000 people (columns 4 to 6). For comparison, Panel A reports least squares estimates, Panel B reports reduced-form estimates and Panel C reports the IV estimates. All specifications control for county fixed effects and period fixed effects (column 1 and 4) or state-period fixed effects (columns 2 to 3 and 5 to 6). In addition, the specifications reported in columns 3 and 6 control for county-specific linear time trends. We report estimates for weighted regressions, with the weights determined by county populations in 1900. The reported standard errors are clustered at the state level.

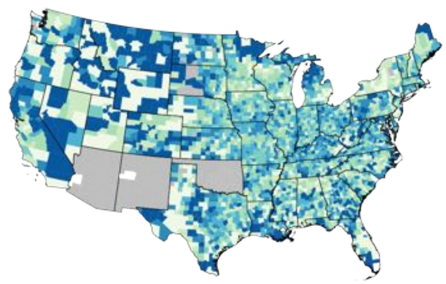
The least squares estimates reveal a significantly positive relationship between surname diversity and both patents and breakthrough patents. In columns 2 and 4, a one standard deviation increase in a county's diversity is associated with 1.5 more patents per 1,000 people, roughly a 74% increase relative to the sample mean, and 0.15 more breakthrough patents per 1,000 people, a 104% increase in breakthrough patents. These estimates are similar to those reported in Table 2 in the previous section, even though they cover a different time span.



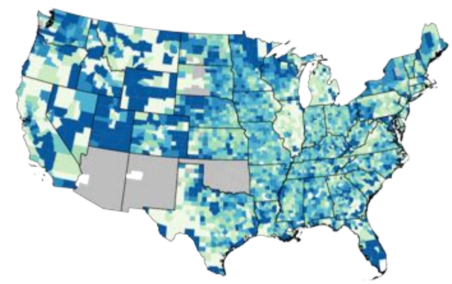
(a) 1900



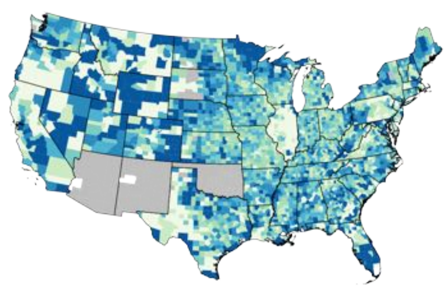
(b) 1905



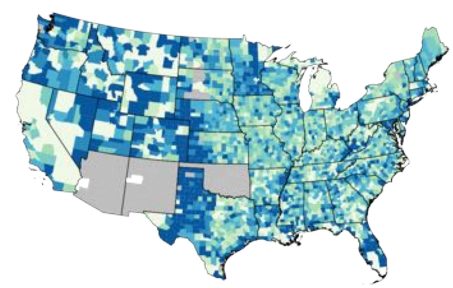
(c) 1910



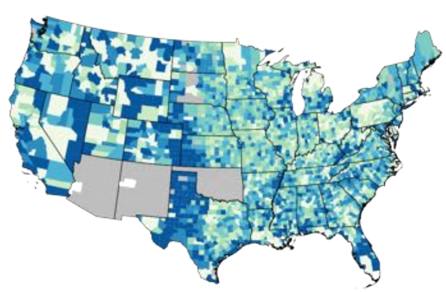
(d) 1915



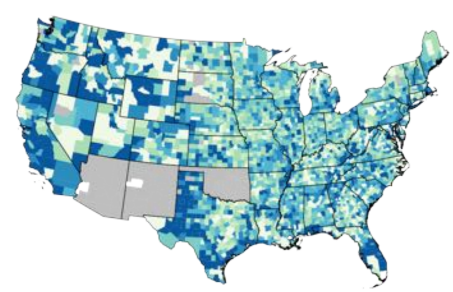
(e) 1920



(f) 1925



(g) 1930



(h) 1940

Figure 1: Predicted surname diversity conditional on county and state-period fixed effects.

Notes: This figure maps residualized instrumented surname diversity for each of the eight periods. We regress the instrument for surname diversity on county and state-year fixed effects and calculate the residuals. This visualization depicts the instrument used in the regression in Table 3. The color coding depicts 7 intervals across counties and within census periods, with darker colors indicating higher values. Grey indicates a lack of data in 1900.

Table 3: Panel estimates of the effect of surname diversity on innovation

	Patents per 1,000 people (mean = 2.04, sd = 2.60)			Breakthrough patents per 1,000 people (mean = 0.14, sd = 0.24)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.528***	1.511***	1.319***	0.183***	0.146***	0.131***
	(0.332)	(0.361)	(0.320)	(0.042)	(0.043)	(0.046)
Panel B: Reduced-form estimates						
Surname diversity (push-pull IV)	0.686***	0.773***	0.734***	0.090***	0.085***	0.080***
	(0.204)	(0.165)	(0.173)	(0.018)	(0.018)	(0.024)
Panel C: Instrumental-variable estimates						
Surname diversity	1.542***	1.794***	1.902***	0.202***	0.197***	0.208**
	(0.382)	(0.378)	(0.543)	(0.043)	(0.055)	(0.090)
Kleibergen-Paap *F*-statistic	63.280	51.050	28.341	63.280	51.050	28.341
Panel D: First-stage estimates	Surname diversity					
Surname diversity (push-pull IV)	0.445***	0.431***	0.386***	0.445***	0.431***	0.386***
	(0.056)	(0.060)	(0.073)	(0.056)	(0.060)	(0.073)
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	23,660	23,660	23,660	23,660	23,660	23,660

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (4) and first-stage estimates for equation (3). An observation is a county in a period from 1900 to 1940. Observations are weighted by county population in 1900. The endogenous variable is county-level surname diversity 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, the dependent variable 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.

Focusing on the IV specifications, Panel B shows statistically significant reduced-form relationships between the dependent variables and our historical push-pull instrument (predicted surname diversity based on the zero stage). To visualize these relationships, Figure D7 shows partial correlation plots. Finally, Panel C presents the IV estimates. The coefficients for surname diversity are all positive and highly significant for both innovation outcomes. The estimates in the baseline specifications (columns 2 and 5) suggest that a one standard deviation increase in a county's surname diversity increases patents (per 1,000 inhabitants) by about 88% relative to the sample mean and breakthrough patents by about 141%. When controlling for county-specific linear trends, the effect of surname diversity remains remarkably stable at around 93% patents per 1,000 inhabitants (column 3) and around 149% for breakthrough patents (column 6), bolstering our confidence that the instrument for diversity is orthogonal to persistent or gradually growing county-level confounding factors. Overall, the estimates suggest that surname diversity has large positive effects on both the quantity and quality of innovation. A comparison of the least-square estimates (Panel A) with the IV estimates (Panel C) reveals that the latter are slightly larger in magnitude, but simple t-tests suggest that the differences are not significant.

5 Robustness and Sensitivity Checks

We now check for the robustness of our estimates in five ways: we examine (1) a placebo test that regresses past innovation on surname diversity, (2) specifications that control or instrument for population size, (3) separate specifications for each of our four major census regions, (4) an approach that uses surname fixed effects to assess whether surname-specific traits affect the results, and (5) analyses that remove immigrant innovators from our analyses, so that we consider only how greater surname diversity impacts innovation among native-born Americans. We also examine the robustness of our results to the use of alternative procedures to construct an instrument for surname diversity (see Section D.4).

5.1 Placebo Tests and Reverse Causality

A potential concern with our results is a form of reverse causality, i.e., that innovative counties attract relatively more immigrants which then may increase diversity. This possibility is unlikely, because our instrument is orthogonal to a county's past attractiveness as captured by the (time-varying) shares of immigrants who settled in a county over the course of several decades (see Section 4.1 for details). Moreover, we have examined

Table 4: Placebo test and persistence: IV estimates of the effect of surname diversity on past and future innovation

	$t-2$ (1)	$t-1$ (2)	t (3)	$t+1$ (4)	$t+2$ (5)	$t+3$ (6)
Panel A:	Patents per 1,000 people					
Surname diversity	0.132	-0.047	1.902***	1.348***	1.000**	0.097
	(0.323)	(0.218)	(0.543)	(0.451)	(0.429)	(0.173)
Panel B:	Breakthrough patents per 1,000 people					
Surname diversity	-0.019	-0.041	0.208**	0.252**	0.179**	-0.098
	(0.067)	(0.083)	(0.090)	(0.116)	(0.073)	(0.073)
Kleibergen-Paap *F*-statistic	24.092	23.922	28.341	20.842	19.700	16.600
County fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
County-specific linear time trends	✓	✓	✓	✓	✓	✓
Observations	17,743	17,746	23,660	17,746	17,743	14,785

Notes: The table reports IV estimates of the leads and lags of innovation outcomes on surname diversity for the specifications described in equation (4). Columns 1 and 2 use the two-period and one-period lag of the dependent variables, respectively. Column 3 repeats the baseline specification (contemporaneous values of the dependent variables). Columns 4 to 6 use the one-period, two-period and three-period lead of the dependent variables, respectively. Observations are county-periods and weighted by county population in 1900. 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.

specifications that include county-specific linear time trends, which absorb the effects of trending unobserved factors associated with innovation and migration.

Nevertheless, to further challenge the validity of our instrument, we conduct a placebo exercise to determine whether contemporaneous diversity affects past innovation activity. A significant estimate in such a regression would provide evidence of reverse causality—i.e., innovative counties attracting immigrants, thereby increasing surname diversity. In columns 1 and 2 of Table 4, we regress our measures of innovation for one (in $t-1$) and two (in $t-2$) periods prior to our measure of surname diversity in the current period (in t). Column 3 replicates our IV specification in which we regress innovation on same-period diversity (reported in Table 3, column 5). The estimates suggest no significant positive relationship between innovation in previous periods and subsequent diversity. This applies to both patents (Panel A) and breakthrough patents (Panel B). When we regress patenting in t (i.e., that occurs between t and $t+1$) on diversity in period t, the coefficients increase in size and become significantly different from zero (column 3). In summary, we find no evidence for reverse causality affecting our identification strategy.

Turning from the past to the future, we investigate the persistence of the impact of diversity on innovation by regressing the one-period lead (innovation in $t+1$, column 4), the two-period lead ($t+2$, column 5), and the three-period lead ($t+3$, column 6) on surname diversity (in t). The estimates in columns 4 to 6 suggest that the impact of diversity on patenting in the two following periods is significantly positive and decreasing in magnitude over time (panel A). The effect on breakthrough patents shows similar persistence (panel B). We fail to detect an effect in period $t+3$, suggesting that the effect of surname diversity on innovation persists for about 15 years.

5.2 Sensitivity of Estimates to Population Size

In our IV specifications, population enters through the dependent variables, which is per capita (per the population in 1900), and each county is weighted based on the population size in 1900. Here, we follow the literature (e.g., Burchardi et al., 2021) and choose a time-invariant base year because population growth is likely endogenous—innovative regions may attract more people. A concern with this specification is that it may not be able to adequately capture scale effects due to an increasing population. The least squares estimates reported in Table D2 suggest that this is unlikely to be the case—the coefficients on surname diversity are not sensitive to controlling for counties' population size. Here, we assess the robustness of our IV estimates to the inclusion of population size in the specification.

The results of this robustness check are reported in Table D6. We use actual population in Panel A, predicted population in Panel B, and we instrument population with predicted population in Panel C. The construction of predicted population parallels the construction of the surname-diversity instrument: based on the historical push-pull interaction of the zero stage (see Section 4), we obtain the predicted stocks of each surname in a county in a given period. By aggregating these stocks, we obtain quasi-random estimates of county population at a given point in time.

Table D6 shows that our IV results are robust to controlling for population. All estimates show little fluctuation compared to our baseline estimates.

5.3 Estimates for Major U.S. Regions

To ascertain if the relationship between surname diversity and innovation holds in different U.S. regions, Table D7 shows the results of interacting surname diversity with each region, the Midwest, Northeast, South and West. Across the full battery of specifications used in our IV analysis, all region-specific estimates are positive and nearly all are statistically significant. The coefficients for the Midwest tend to be larger than for the other regions. However, the estimates are not precise enough to draw strong conclusions about regional differences. We also note that some IV estimates may suffer from a weak first stage.

5.4 Surname Fixed Effects

Another potential concern with the interpretation of our findings is that surname-specific traits, such as abilities, interests, or knowledge, drive innovation rather than the diversity of these traits. For example, Clark (2014) and Barone and Mocetti (2021) find that rare surnames are proxies for the vertical transmission of traits, and these traits might affect innovation. We assess this concern by estimating specifications that include surname fixed effects which absorb any surname-specific traits. This requires us to change the unit of observation from county-period to surname-county-period. The estimating equations are given by (5) and (6), where (5) is the first stage and (6) is the second stage.

$$\text{Surname diversity}_i^t = \gamma\ \widehat{\text{Surname diversity}}_i^t + \mu_{t,s(i)} + \mu_i + \mu_{t,k} + \nu_{i,k}^t \tag{5}$$

$$Y_{i,k}^t = \beta\ \text{Surname diversity}_i^t + \alpha_{t,s(i)} + \alpha_i + \alpha_{t,k} + \varepsilon_{i,k}^t \tag{6}$$

where i indexes counties, s states, t census years (including the midyears), and k surnames. As before, Surname diversity$_i^t$ is county i's surname diversity in t, and $\widehat{\text{Surname diversity}}_i^t$ is county i's predicted surname diversity in t. $Y_{i,k}^t$ is now the number of (breakthrough) patents filed by people with surname k residing in county i per 1,000 of these residents in the five-year period starting in t. For example, 36,984 individuals with the surname Johnson resided in Cook County (IL) in 1940 and filed about 105 patents and 11 breakthrough patents between 1940 and 1944. Therefore, while the surname diversity remains defined at the county-period level, the innovation outcomes vary at the surname-county-period level.[16] Crucially, we can now include surname-period fixed effects, denoted by the param-

[16]Consequently, the number of observations increases because they are now determined by the total number of unique surnames in a given county. Consistent with our baseline specification, we normalize the number of patents and breakthrough patents by the surname population in the year 1900. If a surname does not exist in a given county in 1900, we drop it from the sample. See Appendix A for all the details on how we construct the sample.

Table 5: Surname fixed effects

	Patents per 1,000 people (mean = 2.06, sd = 44.86)				Breakthrough patents per 1,000 people (mean = 0.16, sd = 9.29)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.694***	1.736***	1.667***	1.111***	0.262***	0.267***	0.195**	0.132***
	(0.454)	(0.466)	(0.504)	(0.297)	(0.079)	(0.081)	(0.074)	(0.039)
Panel B: Reduced-form estimates								
Surname diversity (push-pull IV)	0.639***	0.657***	0.655***	0.815**	0.137***	0.140***	0.126***	0.117**
	(0.236)	(0.230)	(0.223)	(0.337)	(0.039)	(0.040)	(0.028)	(0.053)
Panel C: Instrumental-variable estimates								
Surname diversity	1.477***	1.524***	1.571***	2.205**	0.316***	0.325***	0.301***	0.317**
	(0.458)	(0.451)	(0.483)	(0.941)	(0.091)	(0.094)	(0.091)	(0.156)
Kleibergen-Paap *F*-statistic	61.575	61.793	49.541	31.909	61.575	61.793	49.541	31.909
Panel D: First-stage estimates	Surname diversity							
Surname diversity (push-pull IV)	0.433***	0.431***	0.417***	0.370***	0.433***	0.431***	0.417***	0.370***
	(0.055)	(0.055)	(0.059)	(0.065)	(0.055)	(0.055)	(0.059)	(0.065)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓				✓			
Surname-Period fixed effects		✓	✓	✓		✓	✓	✓
State-Period fixed effects			✓	✓			✓	✓
County-specific linear time trends				✓				✓
Observations	30,416,997	30,416,997	30,416,997	30,416,997	30,416,997	30,416,997	30,416,997	30,416,997

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation 6 and first-stage estimates for equation 5. An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted by the surname population in a given county in the year 1900. 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 county i in 1900 (multiplied by 1,000). The dependent variable in columns 4 to 6 is the corresponding number of breakthrough patents. Standard errors are two-way clustered on states and surnames and reported in parentheses. 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.

eter $\alpha_{t,k}$, which implies we non-parametrically control for surname-specific traits across periods (i.e., traits specific to all individuals named Johnson in 1940). The remaining parameters and variables are as in equations (5) and (6). As before, the coefficient of interest is β. The observations are weighted by the number of people in a county carrying the surname in the year 1900. Standard errors are clustered in two ways, on states and surnames.

The results reported in Table 5 show that the estimates are large and highly significant across all specifications and the inclusion of surname fixed effects (in columns 2 to 4 and 6 to 8) hardly changes the estimates. The IV specification suggests that a one standard deviation increase in a county's surname diversity increases patent filings per 1,000 people with the same surname by around 1.6 patents (as seen in column 3) and 0.3 breakthrough patents (as seen in column 7). When expressed as a percentage (by relating it to the mean values of the dependent variable), this is an increase of 76% in patents (column 3) and 188% in breakthrough patents (column 7). Despite changing the unit of observation, we obtain estimates that are comparable to the county-level estimates for patents reported in Table 3, though the estimates for breakthrough patents are larger. Crucially, in this specification, the surname fixed effects ensure that the estimates are independent of any unobserved surname-specific characteristics. [17]

5.5 Innovation among native-born Americans only

A final concern is that immigration per se, rather than diversity, biases our estimates. We tackle this in the county-level least squares specifications reported in Table 2 (Section 3) by regressing innovation on diversity while controlling for immigrant shares (by each originating country). The coefficients on surname diversity are remarkably robust to the inclusion of these control variables. This finding is evidence against the possibility that immigration per se biases our estimates because, for example, immigrants may be more highly skilled, entrepreneurial, or possess novel patentable knowledge. Furthermore, the concern that our IV estimates are confounded by immigration is mitigated by the fact that our instrument is orthogonal to counties' past immigration history (see Section 4.1 for details on the construction of the instrument). The estimates are also robust to controlling for migration-induced population changes (see Section 5.2).

Nevertheless, we conduct an additional robustness check. In the surname fixed effects specification, the unit of observation is surname-county-period. This method allows us to drop all names for which we know that at least one immigrant also holds this name.[18] We then estimate equations (5) and (6) with this non-immigrant sample. We acknowledge that this approach is coarse because we drop many native innovators who happen to share their surname with an immigrant living in the same county at the same time. Consequently, the sample size decreases substantially by roughly 44% from 30.4 to 17.1 million observations. This will skew the estimates towards zero because the average number of patents per 1,000 inhabitants (with the same surname) in the full sample is 2.06, while it is only 0.64 (or a third) in the non-immigrant subsample. This difference is even more pronounced for breakthrough patents (mean of 0.16 in the full vs. 0.04, or a fourth, in the subsample). Therefore, in this robustness check, we are less concerned about the effect size—the approach is too coarse—but rather we are interested in whether we can still detect significant effects in this subsample.

[17] Using a similar approach (for details see Appendix Section D.3), we estimate specifications that include patent technology class fixed effects which absorb any patent class-specific factors. These specifications address the concern that systematic variation in patenting practices across technologies may bias our results. The estimates, reported in Table D9, show that the results hold with this patent class fixed effects specification. This suggests that patent class-specific factors do not bias our estimates.

[18] We are not able to drop immigrants directly—this individual-level data (in contrast to the surname-level data) would require us to match the patent data to the Census at the individual level. This is currently not feasible at a reasonable level of accuracy.

Table D8 shows that in this subsample, the point estimates are all positive and significant; the only exception is the weakly significant estimate in column 8. As expected, the estimates are smaller than the ones reported in Table 5. Yet, they remain sizable: Using our measure of surname diversity, instrumented via our push-pull IV, we estimate a 59% increase in patents per 1,000 natives in column 3 and a 155% increase in breakthrough patents per 1,000 natives in column 7. These sizable estimates—in a sample that rules out the possibility of a patent being filed by an immigrant—provide further evidence for the causal impact of social interactional diversity on innovation.

6 Mechanisms

Conceptually, following much prior work, we view innovation as arising from the recombination of ideas due to the social interactions that occur among diverse minds. To explore this more deeply, we consider (1) if surname diversity spurs the recombination of existing technologies, (2) the impact of surname diversity on the strength of family ties, and (3) the extent of spatial spillovers across counties (in Appendix Section D.5).

6.1 Informational Channel: Recombinations of Technologies

A fundamental aspect of our investigation pertains to the informational channel, whereby a diverse pool of surnames—indicative of a varied range of experiences, skills, perspectives and knowledge—facilitates the recombination of ideas, potentially sparking innovation. As we delve further into the impact of increased surname diversity on patent generation, one key question emerges: what types of patents are stimulated by greater surname diversity? Specifically, we aim to discern whether increased surname diversity triggers the patenting of novel technology types, perhaps introduced from other counties, or whether it stimulates the creation of novel recombinations of existing technologies or their further refinement. Understanding the answers to these questions will provide a deeper insight into the mechanism through which diversity fosters innovation.

To address this, we rely on the methodology developed by Strumsky et al. (2011) and Akcigit et al. (2013), utilizing the comprehensive system of over 140,000 technology codes provided by the United States Patent and Trademark Office (USPTO). We classify patents into three distinct categories: (1) novel technologies, (2) novel combinations, and (3) reuse/refinement combinations. A patent is deemed a 'novel technology' if any of its technology codes emerge for the first time in that patent's grant year (e.g., the radar patent discussed earlier). However, if a patent does not incorporate novel codes but

Table 6: Effects of surname diversity on recombinations

	Technologies per patent (mean = 2.34, sd = 0.67)		Share novel technology patents (mean = 0.02, sd = 2.32)		Share novel combination patents (mean = 0.30, sd = 0.33)		Share reuse/ refinement patents (mean = 0.68, sd = 0.42)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.132***	0.073**	-0.005	0.002	0.024***	0.010	-0.018**	-0.012
	(0.035)	(0.034)	(0.004)	(0.003)	(0.008)	(0.009)	(0.008)	(0.009)
Panel B: Reduced-form estimates								
Surname diversity (push-pull IV)	0.091***	0.072**	-0.006**	-0.004	0.020***	0.017**	-0.014**	-0.013**
	(0.029)	(0.035)	(0.003)	(0.003)	(0.007)	(0.008)	(0.006)	(0.006)
Panel C: Instrumental-variable estimates								
Surname diversity	0.214**	0.177*	-0.015*	-0.009	0.048**	0.041*	-0.033**	-0.032*
	(0.085)	(0.104)	(0.008)	(0.007)	(0.021)	(0.025)	(0.016)	(0.019)
Kleibergen-Paap *F*-statistic	57.177	45.949	57.177	45.949	57.177	45.949	57.177	45.949
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓		✓		✓		✓	
State-Period fixed effects		✓		✓		✓		✓
Observations	20,246	20,246	20,246	20,246	20,246	20,246	20,246	20,246

Notes: The table reports least squares, reduced-form, and IV estimates for the specifications described in equation (4) of the effect surname diversity on the average number of technologies per patent and three different types of patents per 1,000 people. A patent is classified as a novel technology if any of the technology codes listed on the patents appear in the grant year of the patent for the first time. A patent is deemed a novel combination if a unique set of three technologies (a 'triplewise' combination) is observed in the county for the first time. Any remaining patents are considered reuse/refinement patents, where all possible triplewise combinations of technologies have been observed previously. An observation is a county in a period from 1900 to 1940. Observations are weighted by county population 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.

contains a unique triple combination of technologies for the first time in the grant year, we classify it as a 'novel combination'. An example would be a patent combining the previously observed technologies A, B, and C for the first time. All other patents, not meeting these criteria, are deemed 'reuse/refinement' combinations. We also consider the average number of technologies recorded per patent as an additional measure for recombinant innovation.

Our findings, presented in Table 6, display a positive effect of surname diversity on the number of technologies per patent (columns 1 and 2). The IV estimates indicate that a one standard deviation increase in surname diversity leads to a 0.17-0.21 increase in the average number of technologies per patent, signifying a 7.3-9% enhancement relative to the sample mean. Additionally, we find a positive impact on the proportion of patents classified as novel combinations (columns 5 and 6). The IV estimates suggest that a one standard deviation increase in surname diversity causes a 0.041-0.048 percentage point rise in the share of combinations, translating to a 13.6-16% increase relative to the sample mean. This effect occurs at the cost of the share of patents deemed as novel technologies (columns 3 and 4) and reuse or refinements (columns 7 and 8). This evidence suggests that diversity bolsters innovation predominantly through the recombination of existing technologies, which is consistent with the informational channel of our hypothesis.

6.2 Social-Psychological Channel: Strength of Family Ties

We now turn to the social-psychological channel, focusing on a key aspect of our hypothesis: as surname diversity increases and the prevalence of large extended families or surname groups shrinks, the strength of family ties erodes which enhances engagement with outsiders. As mentioned before, previous research has connected the strength of family ties and kinship to impersonal trust (Alesina and Giuliano, 2011; Moscona et al., 2017; Schulz et al., 2019; Enke, 2020).

In Section 2, we have established a robust correlation between surname diversity and family ties. Here, we investigate the causal relationship between surname diversity and the strength of family ties. To do this, we estimate our baseline IV-specification presented in equation (4), but replace the dependent variable with the strength of family ties.

The results from this analysis are reported in columns 1-3 of Table 7. Least-squares (panel A), reduced-form (panel B), and IV estimates (panel C) reveal negative coefficients for surname diversity, which become statistically significant when we control for state-period fixed effects (column 5 and 6). The finding is robust even after controlling for county-specific linear time trends (column 6). Thus, our findings suggest that greater surname diversity leads to weaker family ties in a county.

The structure of the data allows us to complement the county-level analysis with a specification that directly focuses on the impact of the relative size of surname (or family) groups within a county on the strength of family ties. To do so, we refine Raz (2023)'s county-level family ties measure to now capture the strength of family ties among individuals in a county who share the same surname. We then analyze this "within-surname" strength of family ties by regressing it on the instrumented shares of individual surnames within a county. Formally, we estimate the following equations, where equation (7) is the first stage and equation (8) is the second stage:

$$\text{Surname population share}_{i,k}^{t} = \gamma \; \text{Surname} \; \widehat{\text{population}} \; \text{share}_{i,k}^{t} + \mu_{t,s(i)} + \mu_{i,k} + v_{i,k}^{t} \tag{7}$$

$$\text{Strength of family ties}_{i,k}^{t} = \beta \; \text{Surname population share}_{i,k}^{t} + \alpha_{t,s(i)} + \alpha_{i,k} + \varepsilon_{i,k}^{t} \tag{8}$$

where i indexes counties, k names, s states, and t census years (including the midyears). Surname population share$_{i,k}^{t}$ is surname group k's population share in county i in t; and Surname $\widehat{\text{population}}$ share$_{i,k}^{t}$ surname group k's predicted population share in i and t. Equations (7) and (8) also include state-period fixed effects, $\mu_{t,s(i)}$ and $\alpha_{t,s(i)}$, and surname-county fixed effects, $\mu_{i,k}$ and $\alpha_{i,k}$. These fixed effects allow us to focus on changes in

Table 7: Effects of surname shares and surname diversity on strength of family ties

	Strength of family ties, County-level (mean = -0.14, sd = 0.80)			Strength of family ties, Surname-County-level (mean = -0.02, sd = 0.94)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	-0.038	-0.183***	-0.470***			
	(0.075)	(0.037)	(0.087)			
Surname population share				1.669***	1.671***	1.669***
				(0.152)	(0.153)	(0.186)
Panel B: Reduced-form estimates						
Surname diversity (push-pull IV)	-0.162*	-0.264***	-0.328***			
	(0.091)	(0.093)	(0.063)			
Surname population share (push-pull IV)				0.009**	0.008***	0.004
				(0.004)	(0.003)	(0.004)
Panel C: Instrumental-variable estimates						
Surname diversity	-0.365	-0.613**	-0.851***			
	(0.229)	(0.277)	(0.266)			
Surname population share				0.653**	0.635***	0.619
				(0.249)	(0.175)	(0.402)
F-statistic, Surname diversity	63.349	51.127	28.360			
F-statistic, Surname population share				38.896	35.231	6.883
County fixed effects	✓	✓	✓			
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			
Surname-County fixed effects				✓	✓	✓
Surname-County-specific linear time trends						✓
Observations	23,639	23,639	23,639	9,692,638	9,692,638	9,692,638

Notes: The table reports least squares, reduced-form, and IV estimates. Columns 1-3 report estimates for the specifications described in equation (4) with the strength of family ties as the outcome variable. An observation is a county in a period from 1900 to 1940. Observations are weighted by county population in 1900. Standard errors are clustered at the state level. Surname Diversity is standardized to mean zero and unit variance. Columns 4-6 report estimates for the specifications described in equation (8). An observation is a surname within a county in a period from 1900 to 1940. Observations are unweighted. Standard errors are two-way clustered on states and surnames and reported in parentheses. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

counties' surname shares over time while controlling for persistent surname-county-specific and time-varying state-specific factors.

Columns 4-6 of Table 7 report the results. The OLS estimates (panel A) reveal a positive and highly significant relationship between the surname share of the county population and the strength of family ties. As the relative sizes of extended families and surname groups shrink, so does the strength of family ties. The reduced-form (panel B) and IV estimates (panel C) corroborate is finding. The IV coefficients imply that a shift in a surname share from 0 to 1—essentially a change from no presence to complete dominance of a surname in a county—strengthens family ties by around 0.6 standard deviations. Although the first stage of the IV regression in column 3 is relatively weak (with an *F*-statistic of 6.9) and the coefficient is not significant, the direction and magnitude of the relationship remains intact in this demanding specification that includes surname-county-specific linear time trends.[19]

In sum, these results indicate that as surname diversity increases and surname shares within a county shrink, kinship ties weaken. This is in line with our hypothesis that smaller group sizes limit people's ability to meet their needs within their family, which creates more opportunities for exchanges with unrelated individuals. As a result, greater diversity fosters engagement with non-kin individuals, arguably cultivating a culture of impersonal trust, promoting recombinant innovation.

7 Conclusion

Focusing on the United States, during the period when it rose to dominate global innovation (1850-1940), we study the impact of diverse social interactions on innovation. The core idea is that many, if not most, innovations arise from the recombinations of existing ideas, approaches and techniques that come together through the connections among diverse minds. To measure this diversity, we introduce and benchmark an entropic diversity measure that exploits a widely available data source, surnames, obtained from the complete U.S. Census. To measure innovation, we use patents per capita at the county level and a text-based measure of breakthrough patents per capita. In our analysis, we first use least squares regressions across U.S. counties. These analyses show that surname diversity is robustly correlated with both patent-based innovation measures across the entire period and that this holds even when controlling for more common diversity mea-

[19] In previous analyses throughout this paper, observations were weighted based on their population share in 1900. However, for this particular analysis (columns 1-3), we opted not to employ these weights. This choice is driven by a noticeable reduction in the strength of the *F*-statistic when weights are applied, which weakens the instrument's relevance in our IV regressions.

sures such as those based on birth country. Next, we employ an instrumental variable approach, using immigrant flows to extract an exogenous component of surname diversity to examine the effect of surname diversity on our innovation outcomes. These analyses suggest that greater surname diversity causes faster innovation. Third, we subject these results to a battery of robustness and sensitivity checks including a placebo test for reverse causality, explorations of the role of population size, and surname fixed effects, which shows that people with the same surname get more innovative when they live in a more diverse county. Our analysis closes by showing that surname diversity increases novel combinations of existing technologies, weakens family ties and that the impact of surname diversity degrades rapidly with spatial distance.

Conceptually, our efforts raise two important issues for future research. First, we show that surname diversity captures a form of cultural diversity that is distinct from more commonly used measures of diversity, including those based on differences in race, ethnicity, genes, national origins, birthplace and occupation. This suggests that understanding the impact of diversity may require a more nuanced set of conceptual or theoretical distinctions. Indeed, the power of surname diversity to explain innovation over and above the impact of occupational diversity suggests that economists may want to generalize beyond such descriptors as "skill complementarities" and "knowledge spillovers" to consider the role of cultural differences in decision-making heuristics, attentional biases, thinking styles and approaches to problem-solving (Heine, 2010; Nisbett, 2003). Second, if innovation is propelled by recombination, as many have suggested (Weitzman, 1998; Jones, forthcoming), then we might have expected a concave relationship between cultural diversity and innovation because a tradeoff could exist between the informational and social-psychological effects of diversity (Schimmelpfennig et al., 2022). That is, at a certain point, people might have very different skills, perspectives and expertise but refuse to interact or share what they know. Populations that get ’too’ diverse might end up with less innovation due to declining social interactions and restricted informational flows. We, however, do not observe any concavity in the relationship between surname diversity and innovation. Indeed, surname diversity is positively associated with both informational diversity and impersonal trust (and negatively with the strength of family ties). This suggests that if such a tradeoff creates an optimal amount of cultural diversity (as captured by surnames), the United States has remained on the upward-sloping (left) side of the hypothesized hump-shared curve for much of its history.

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.

Aihounton, Ghislain B D & Arne Henningsen (2020) "Units of measurement and the inverse hyperbolic sine transformation", *The Econometrics Journal*, 24 (2), pp. 334–351.

Akcigit, Ufuk, Santiago Caicedo, Ernest Miguelez, Stefanie Stantcheva & Valerio Sterzi (2018) "Dancing with the Stars: Innovation Through Interactions",Technical Report w24466, National Bureau of Economic Research, Cambridge, MA.

Akcigit, Ufuk, John Grigsby & Tom Nicholas (2017) "The Rise of American Ingenuity: Innovation and Inventors of the Golden Age",Technical Report w23047, National Bureau of Economic Research, Cambridge, MA.

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

Alesina, Alberto & Paola Giuliano (2011) "Family Ties and Political Participation", *Journal of the European Economic Association*, 9 (5), pp. 817–839.

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 (2000) “Participation in Heterogeneous Communities”, *The Quarterly journal of economics*, 115 (3), pp. 847–904.

Alesina, Alberto & Eliana La Ferrara (2002) “Who trusts others?”, *Journal of Public Economics*, 85 (2), pp. 207–234.

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.

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

Altonji, Joseph G & David Card (1991) “The effects of immigration on the labor market outcomes of less skilled natives”, *Immigration, Trade and the Labor Market*, Chicago, University of Chicago Press, pp. 201–234.

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

Andrews, Michael J. & Chelsea Lensing (2020) “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.

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”.

Bahcall, S. (2019) *Loonshots: How to Nurture the Crazy Ideas that Win Wars, Cure Diseases, and Transform Industries*, St. Martin’s Press.

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.

Barrai, I., C. Scapoli, M. Beretta, C. Nesti, E. Mamolini & A. Rodriguez-Larralde (1996) "Isonymy and the genetic structure of Switzerland I. The distributions of surnames", *Annals of Human Biology*, 23 (6), pp. 431–455.

Bazzi, Samuel, Arya Gaduh, Alexander D. Rothenberg & Maisy Wong (2019) "Unity in Diversity? How Intergroup Contact Can Foster Nation Building", *American Economic Review*, 109 (11), pp. 3978–4025.

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*.

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, A & T Verdier (1998) "On the cultural transmission of preferences for social status", *Journal of Public Economics*, 70 (1), p. 75–97.

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.

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

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.

Burchardi, Konrad B, Thomas Chaney, Tarek A Hassan, Lisa Tarquinio & Stephen J Terry (2021) "Immigration, Innovation, and Growth", *Working paper*.

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

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

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

Cattaneo, Matias D, Richard K Crump, Max H Farrell & Yingjie Feng (2019) "On binscatter", *arXiv preprint arXiv:1902.09608*.

Cavalli-Sforza, Luigi Luca & Marc W Feldman (1981) *Cultural Transmission and Evolution: A Quantitative Approach*, Princeton University Press.

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

Clancy, Matthew S (2018a) "Combinations of Technology in US Patents, 1926-2009: A Weakening Base for Future Innovation?", *Economics of Innovation and New Technology*, 27 (8), pp. 770–785.

Clancy, Matthew S. (2018b) "Inventing by combining pre-existing technologies: Patent evidence on learning and fishing out", *Research Policy*, 47 (1), pp. 252–265.

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

Cook, Lisa D. (2014) "Violence and economic activity: Evidence from African American patents, 1870–1940", *Journal of Economic Growth*, 19 (2), pp. 221–257.

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.

DeGering, Randall (2018) ""radar contact!": The beginnings of army air forces radar and fighter control",Technical report, Air University Press.

Desmet, Klaus, Avner Greif & Stephen L. Parente (2017) "Spatial Competition, Innovation and Institutions: The Industrial Revolution and the Great Divergence", *SSRN Electronic Journal*.

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 (2020) "Moral Values and Voting", *Journal of Political Economy*, 128 (10), pp. 3679–3729.

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, O & D N Weil (2000) "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and beyond", *The American Economic Review*, 90 (4), pp. 806–828.

Galor, Oded (2022) *The Journey of Humanity: The Origins of Wealth and Inequality*, New York, Dutton.

Ghosh, Arkadev, Sam Il Myoung Hwang & Munir Squires (2023) "Economic Consequences of Kinship: Evidence from US Bans on Cousin Marriage", *Working paper, University of British Columbia*.

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.

Glaeser, Edward L., David I. Laibson, José A. Scheinkman & Christine L. Soutter (2000) "Measuring Trust", *The Quarterly Journal of Economics*, 115 (3), pp. 811–846.

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.

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.

Handley, Carla & Sarah Mathew (2020) "Human large-scale cooperation as a product of competition between cultural groups", *Nature Communications*, 11 (1), p. 702.

Heine, Steven J (2010) *Cultural Psychology.*, John Wiley & Sons, Inc.

Henrich, J (2009) "The evolution of innovation-enhancing institutions", Stephen J Shennan & Michael J O'Brien eds. *Innovation in Cultural Systems: Contributions in Evolutionary Anthropology*, Cambridge, MIT, pp. 99–120.

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

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

Johnson, S. (2011) *Where Good Ideas Come From: The Natural History of Innovation*, Penguin Publishing Group.

Jones, Charles I (forthcoming) "Recipes and economic growth: A combinatorial march down an exponential tail", *Journal of Political Economy*.

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 (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.

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.

Lucas Jr, Robert E & Benjamin Moll (2014) "Knowledge growth and the allocation of time", *Journal of Political Economy*, 122 (1), pp. 1–51.

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.

Miu, Elena, Ned Gulley, Kevin N. Laland & Luke Rendell (2018) "Innovation and cumulative culture through tweaks and leaps in online programming contests", *Nature Communications*, 9 (2321).

Mokyr, Joel (1995) "Urbanization, technological progress, and economic history", H. Giersch ed. *Urban Agglomeration and Economic Growth*, Berlin; Heidelberg, Springer, pp. 51–54.

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

Mokyr, Joel (2015) "ECONOMICS. Intellectuals and the rise of the modern economy.", *Science (New York, N.Y.)*, 349 (6244), pp. 141–2.

Moscona, Jacob, Nathan Nunn & James A Robinson (2017) "Keeping it in the family: Lineage organization and the scope of trust in sub-Saharan Africa", *American Economic Review*, 107 (5), pp. 565–571.

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.

Olsson, Ola & Bruno Frey (2002) "Entrepreneurship as recombinant growth", *Small Business Economics*, 19 (2), p. 69–80.

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.

Page, R.M. (1962) *The Origin of Radar*, Anchor Books, Anchor Books.

Page, S.E., N. Cantor & K. Phillips (2019) *The Diversity Bonus: How Great Teams Pay Off in the Knowledge Economy*, Our Compelling Interests, Princeton University Press.

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

Raz, Itzchak T. (2023) "Soil Heterogeneity and the Formation of Close-knit Communities", *Working paper*.

Ridley, Matt (2020) *How Innovation Works: And Why It Flourishes in Freedom*, London, Harper.

Romer, Paul M (1990) "Endogenous Technological Change", *Journal of Political Economy*, 98 (5, Part 2), pp. S71–S102.

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.

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.

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

Schumpeter, J.A. (1983) *The Theory of Economic Development: An Inquiry Into Profits, Capital, Credit, Interest, and the Business Cycle*, Economics Third World studies, Transaction Books.

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.

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.

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*.

Thagard, Paul (2012) "Creative combination of representations: Scientific discovery and technological invention.", *Psychology of Science: Implicit and Explicit Processes.*, New York, NY, US, Oxford University Press, pp. 389–404.

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

Uzzi, B., S. Mukherjee, M. Stringer & B. Jones (2013) "Atypical Combinations and Scientific Impact", *Science*, 342 (6157), pp. 468–472.

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

White, Cindel J. M., Michael Muthukrishna & Ara Norenzayan (2021) "Cultural similarity among coreligionists within and between countries", *Proceedings of the National Academy of Sciences*, 118 (37), p. e2109650118.

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 Surname Diversity, Social Ties and Innovation

Max Posch, Jonathan Schulz, and Joseph Henrich

A Data Sources and Construction

Surname-based surname diversity

To construct county-level surname diversity 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, we apply the Philips (1990) phonetic algorithm *metaphone* to deal with misspellings.

We harmonize all historical Census data to the 2000 boundaries of U.S. 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 2000 as the reference year because the patent dataset is geocoded to 2000 county boundaries. The harmonization procedure sometimes results in counties with very few people, predominantly in the Midwest and West, and for Census years before 1900. As a remedy, we winsorize all harmonized variables from the lower tail at the 1% level.

Following Burchardi et al. (2021), we also obtain individuals' age and year of immigration, the variables `age` and `yrimmig`, to estimate surname diversity 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 also compute alternative measures of surname diversity by interacting surnames with a male indicator (`sex`) or the main categories of race (`race`) or birthplace (`bpl`). We recode U.S. states and territories (`bpl` codes <10000) to a single code.

Construction of the instrumental variable

We build on the Burchardi et al. (2019) approach to construct an instrumental variable for surname-based surname diversity. We identify the number of individuals in a given U.S. county i at the time of each census who immigrated to the U.S. 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 U.S. 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- and second-generation immigrants, regardless of the date of arrival in the U.S.

When we predict the stock of people $N_{i,k}^{t}$ in equation (2), we obtain negative values for some observations. The logarithmic transformation of a negative value is undefined. To obtain Shannon entropy for counties containing $N_{i,k}^{t}$ with negative values, we truncate those negative values at the smallest positive value we observe in the data in a given year. The resulting variable is highly correlated with the original variable ($\rho = 0.965$).

Construction of other demographic measures

We collect county-level data on population size, racial, origin country and occupational diversity, and immigrant shares for each census year from 1850 to 1940. All data are taken from the full-count IPUMS available on the NBER servers. To compute racial diversity, we obtain the variable `race` and use the nine main categories. To compute origin country diversity and immigrant shares, we draw on the variable `bpl` with 188 main categories. To compute occupational diversity, we draw on the variable `occ1950`. As before, we recode U.S. states and territories (codes <100) to a single category, transform the data from each period to 2000 U.S. counties using the `M4` weights from the Ferrara et al. (2021) cross-walks, and winsorize all demographic variables from the lower tail at the 1% level.

We follow the instructions in Raz (2023) to 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 captured 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 person who is not their relative, and (iv) the mean size of families. We use the variables age and `yrimmig` to estimate the strength of family ties for the midyears 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the midyear.

We also construct the strength of family ties within surname groups across counties. This necessitates having a positive number of married and old individuals within the surname groups of each county. However, many surname groups within counties do not meet this requirement, which prevents the construction of the principal component for these groups and subsequently reduces the number of observations.

Finally, we obtain the variable `higrade` and compute the average years of schooling for each county in 1940 and harmonize the data to 2000 county boundaries.

Construction of the innovation measures

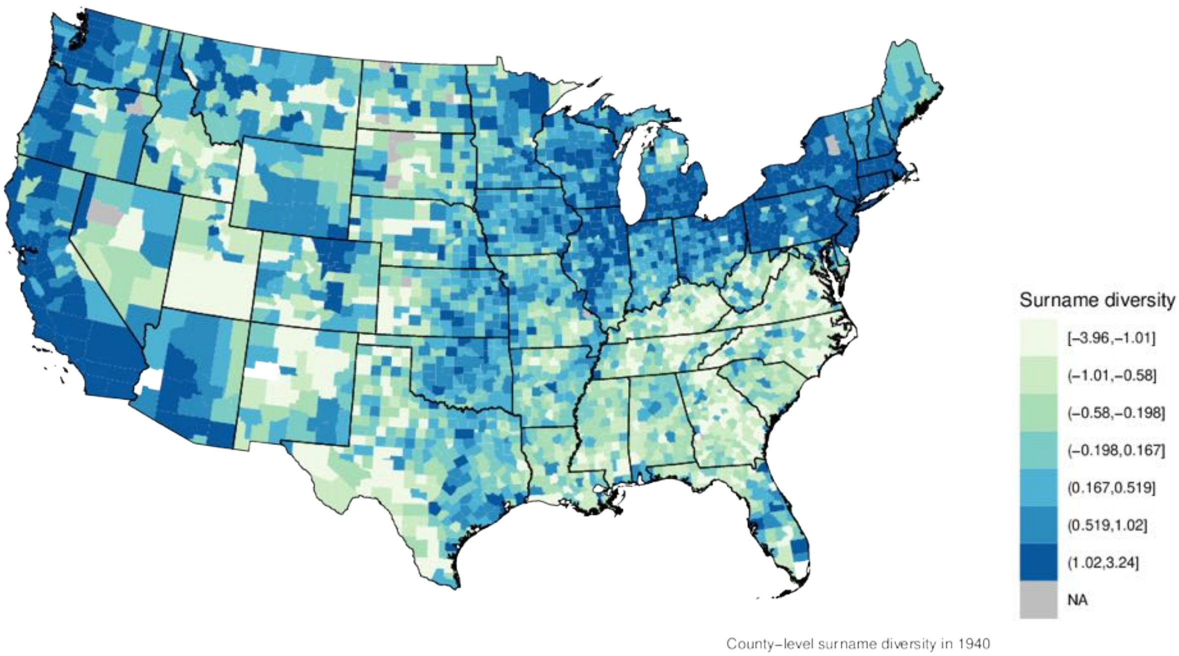
We use the *Comprehensive Universe of U.S. Patents* (CUSP) compiled by Berkes (2018). The data set contains U.S. patents from 1836-2015 and is primarily constructed from Google Patents with supplementary information from other sources. For each patent, the data set provides inventor names and location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

We also draw on the breakthrough patent indicator created by Kelly et al. (2021). The authors use the text in patent documents to estimate patent quality. 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 use this measure of patent quality 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 the innovation outcome variables at the county-period and surname-county-period levels. The county-period-level outcomes are per capita number of (breakthrough) patents filed by inventors residing in county i during the period starting t. If a patent is filed by more than one inventor, possibly residing in different counties, we divide the patent count by the number of inventors. We use county population sizes at the beginning of t, computed using the full-count IPUMS and the Ferrara et al. (2021) border harmonization procedure (winsorized from the lower tail at the 1% level), to get per capita rates of (breakthrough) patents filed. We use patent issuing years rather than filing years in the least-squares analysis in Section 3, because filing years are not consistently recorded in the CUSP data set before 1870. When the unit of observation is county-period, we winsorize the innovation outcome variables from the upper tail at the 99% level to reduce the influence of outlier counties with a very large number of patents. We do not winsorize the surname-county-period-level outcomes because the number of breakthrough patents filed by inventors with a given surname in a given county during a given period is typically small. We also report results using non-winsorized, inverse hyperbolic sine transformed patent counts.

The surname-county-period-level outcomes are per capita number of (breakthrough) patents filed by inventors with surname k residing in county i during the period starting t. The construction of these outcome variables requires inventor surnames. The CUSP data includes inventor names. This string variable contains the surname, first name, and sometimes middle names or initials. Identifying surnames from this string variable is not straightforward because the order of first names and surnames is inconsistent: surnames follow first names in some entries but not in others. When a semicolon, colon, or comma delineates surnames from first and middle names, we use these characters to discern the 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 distinguish the surnames accordingly. For the remaining entries, we tokenize the string variable based on whitespace and keep the first token and the last token, which are the first name and surname in most cases. To determine which of the two tokens is the surname, we compute the frequencies of all names (first name and surname) from the pooled census years 1900, 1910, 1920, 1930, and 1940 and compare which constellation is more common. 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 steps described above.

B Additional Descriptive Results



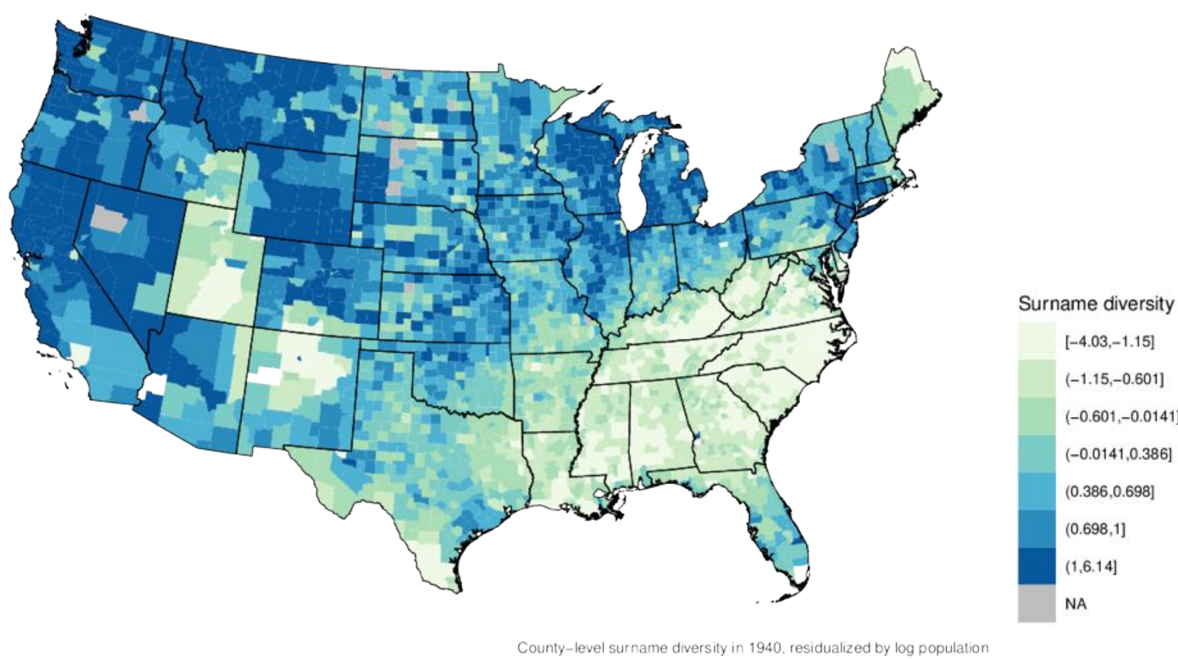
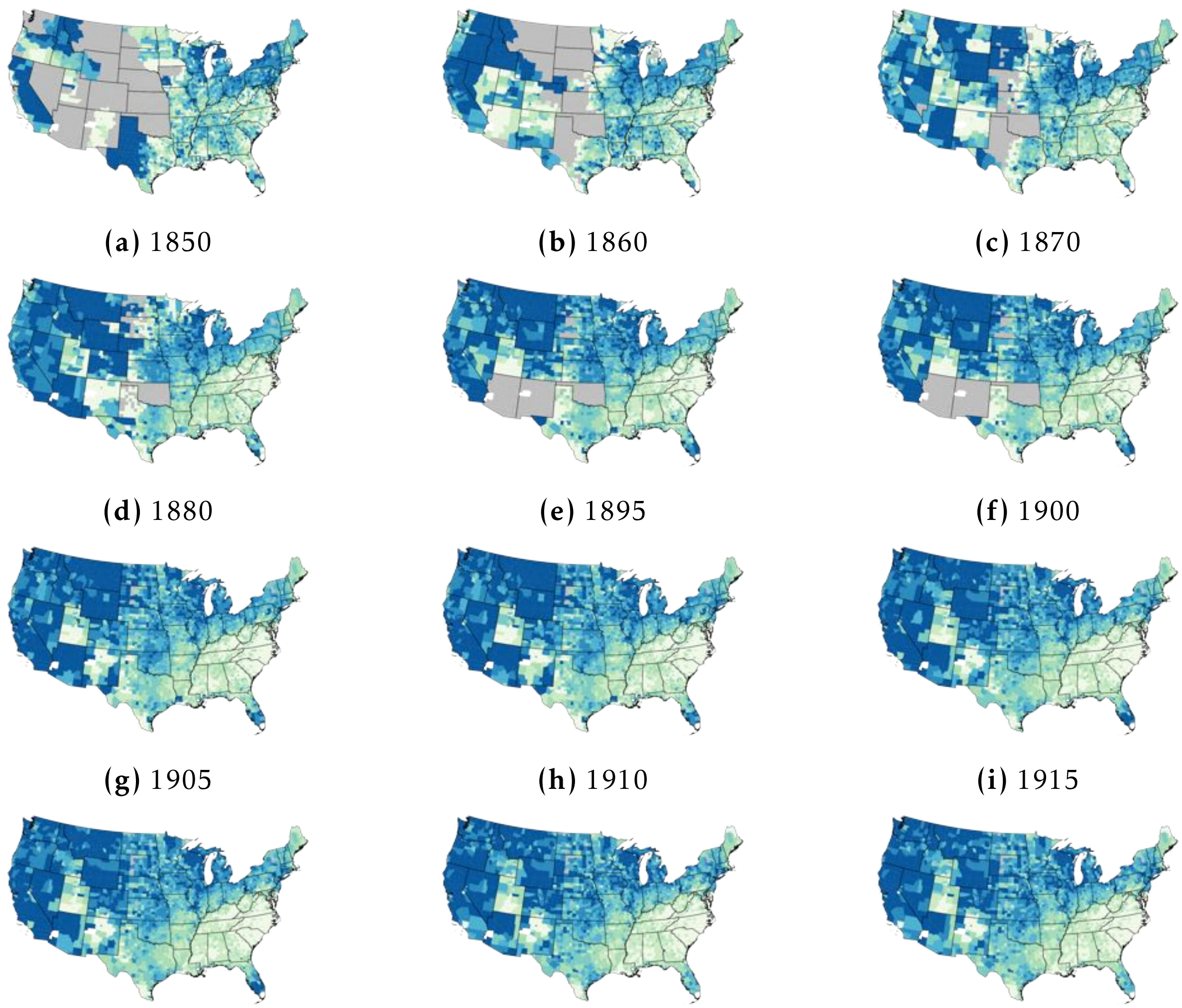


Figure B1: The figures show the geographic variation in surname diversity. Top: Raw data; bottom: diversity residualized by log county population



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

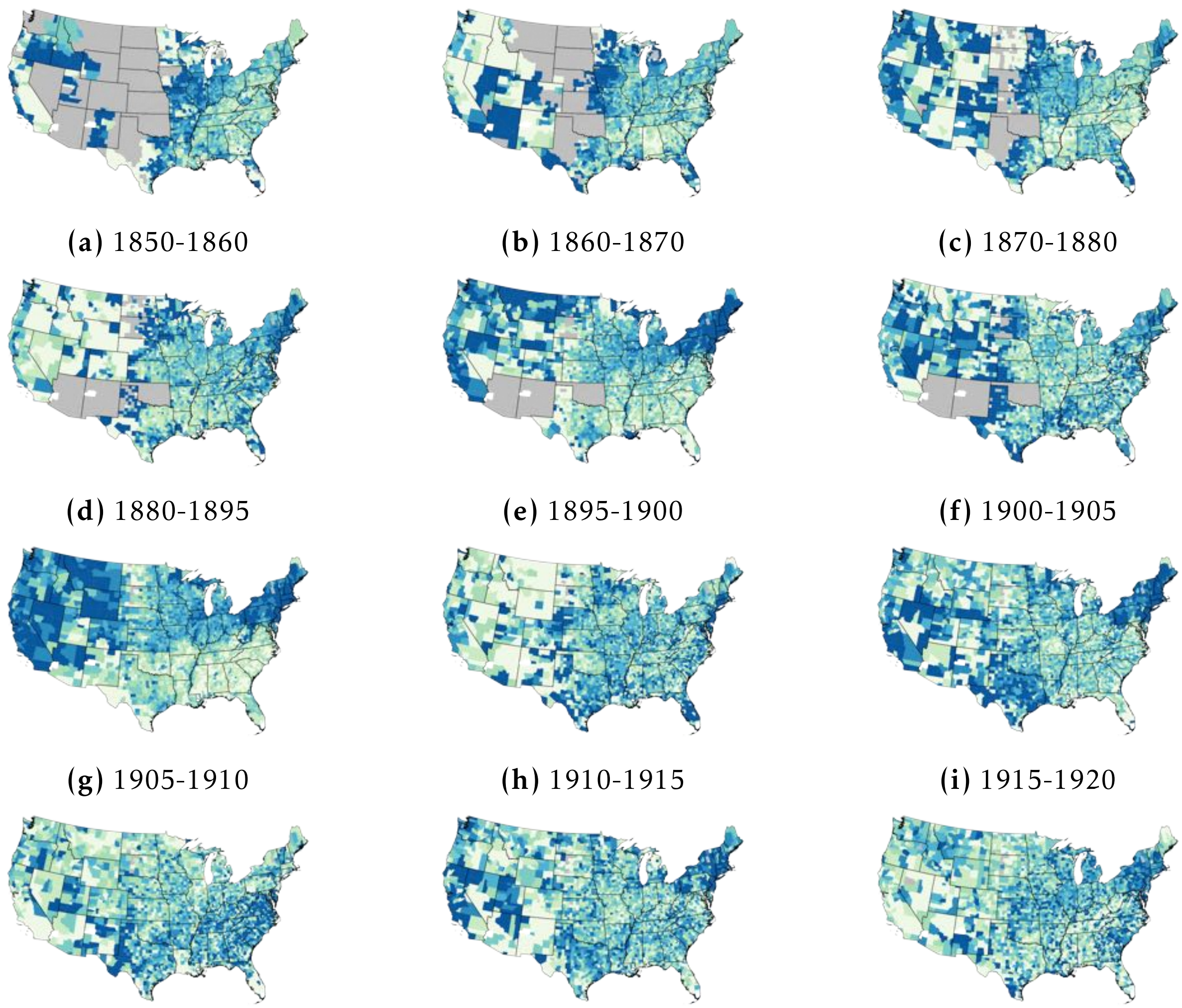
(**d**) 1880 (**e**) 1895 (**f**) 1900

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

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

Figure B2: Surname diversity from 1850 to 1930

Notes: The figure shows standardized county-level surname diversity residualized by log county population in the respective year.



(a) 1850-1860 **(b)** 1860-1870 **(c)** 1870-1880

(d) 1880-1895 **(e)** 1895-1900 **(f)** 1900-1905

(g) 1905-1910 **(h)** 1910-1915 **(i)** 1915-1920

(j) 1920-1925 **(k)** 1925-1930 **(l)** 1930-1940

Figure B3: Changes in surname diversity from 1850 to 1940

Notes: The figure shows first-differenced standardized county-level surname diversity residualized by first-differenced log county population in the respective years.

Table B1: Correlations between surname diversity and other diversities

	Country of origin diversity	Share immigrants	Race diversity	Occupational diversity
Raw Corr.	0.39	0.27	-0.24	0.60
Partial Corr. (Log Population)	0.60	0.50	-0.29	0.57
Partial Corr. (State FE, Log Population)	0.40	0.27	0.08	0.48

Notes: This table reports standardized coefficients of regressions of county-level surname diversity on other dimensions of sociocultural diversity from 1850 to 1940. The first row reports the relationship conditional on year fixed effects. The second row reports the coefficients of regressions additionally controlling for log county population. The third row reports the correlations additionally controlling for state fixed effects. An observation is a county from 1850 to 1940 (exluding the midyears). The sources and construction of all variables are explained in Appendix Section A.

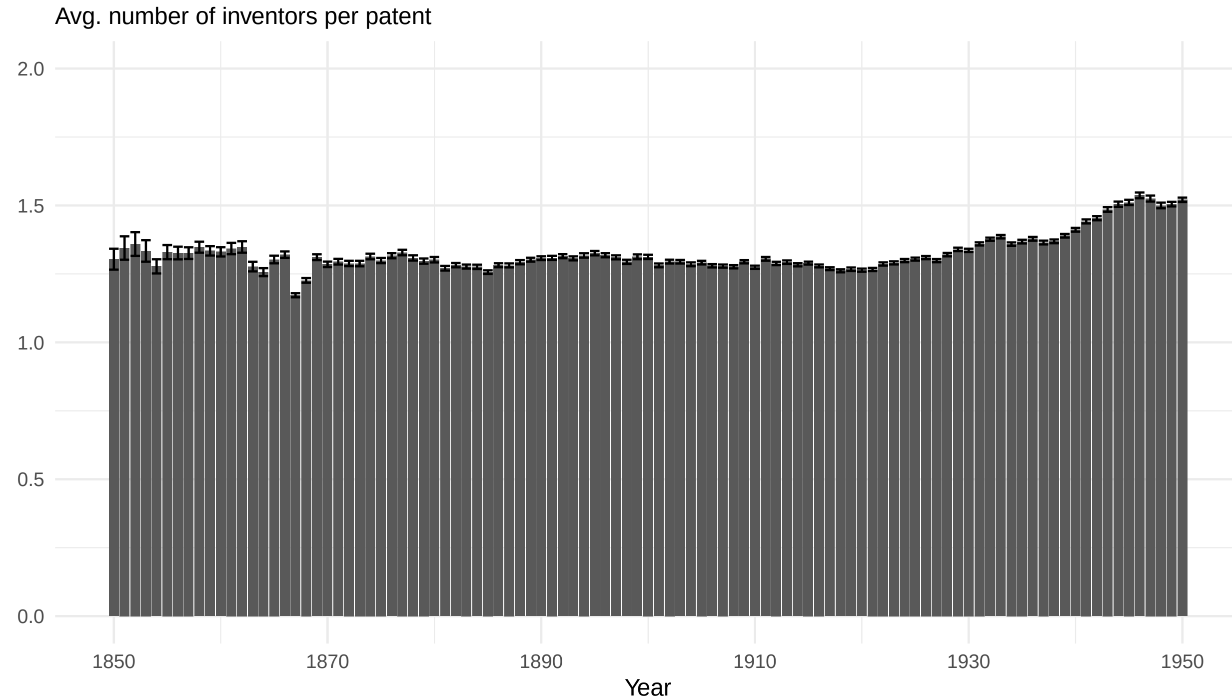


Figure B4: Average number of inventors per patent over time

Notes: The figures show the average number of distinct inventors per patent from 1850 to 1950. Error bars indicate standard errors.

Table B2: Do same-surname inventors produce lower-quality patents?

	Breakthrough patent indicator		
	(1)	(2)	(3)
Constant	0.107***		
	(0.001)		
Same-surname indicator	-0.048***	-0.030***	-0.011***
	(0.002)	(0.002)	(0.002)
R^2	0.002	0.033	0.184
Observations	200,818	200,818	200,818
Year fixed effects		✓	✓
Patent technology class fixed effects			✓

Notes: An observation is a patent from 1850 to 1949 with at least two or more distinct inventors. The same-surname indicator takes value one if all inventors on the patent have the same surname. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

C Additional Validation Results

Table C1: Correlations between baseline surname diversity and alternative surname diversity measures

Herfindahl surname	Surname, uncorrected	Surname, men	Surname, whites	Surname-race	Surname-country of birth
0.80	0.96	1.00	0.99	0.98	0.98

Notes: This table reports the correlations between county-level surname diversity (based on Shannon entropy) and (i) a surname-based Herfindahl index, (ii) diversity of surnames that are not phonetically corrected, (iii) surname diversity among men, (iv) surname diversity among white individuals, and (v) alternative diversity measures that interact surnames with race or birthplace. An observation is a county from 1850 to 1940 (exluding the midyears). The sources and construction of all variables are explained in Appendix Section A.

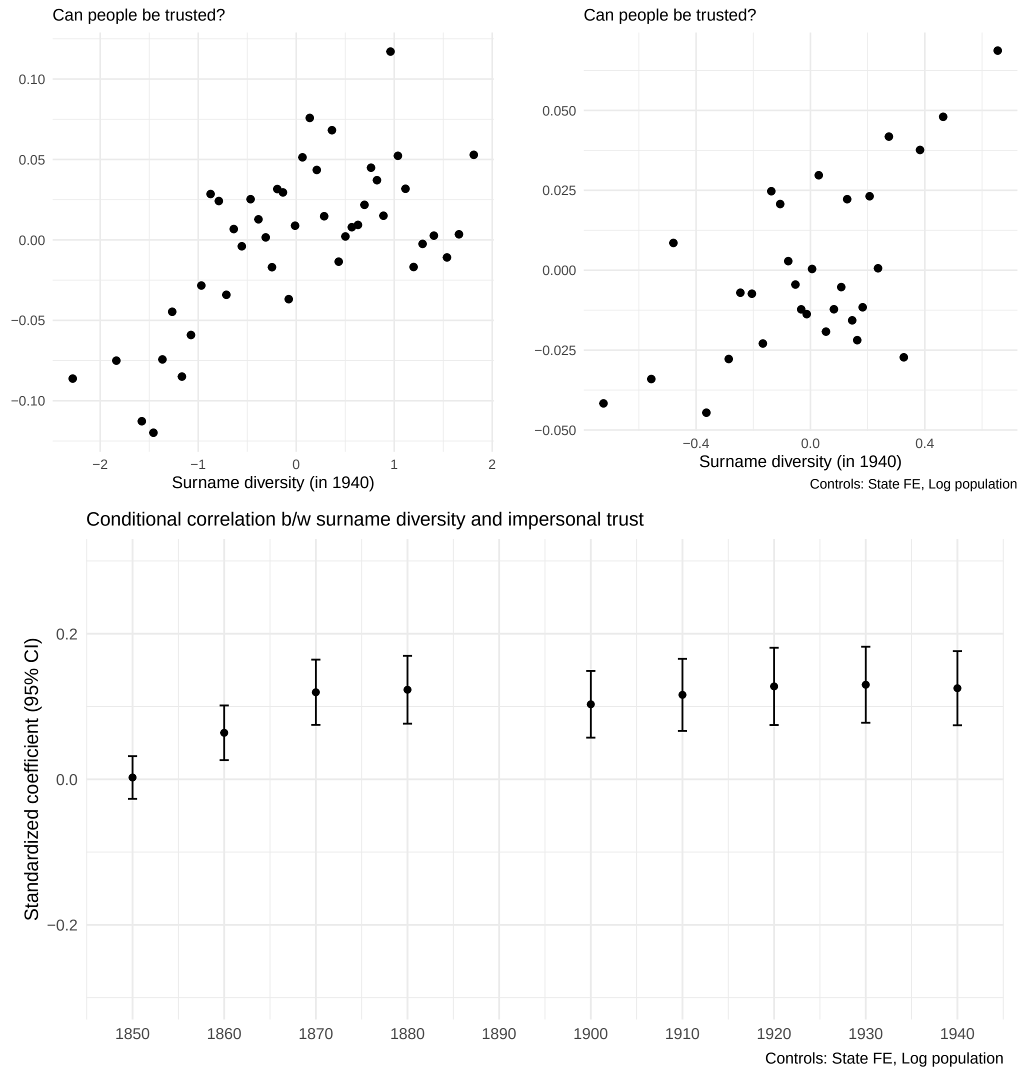


Figure C1: Relationship between surname diversity and impersonal trust

Notes: An observation is an individual. The top left graph show the relationship between impersonal trust and surname diversity in 1940 conditional on state fixed effects and fixed effects for survey year, sex, age, and race. The top right graph additionally controls for log county population in 1940. Bottom: Coefficients of regressions of impersonal trust on surname diversity conditional on state fixed effects and log county population by census year (1850-1940) and survey year, sex, age, and race fixed effects. The trust question is taken from the General Social Survey, waves 1972 to 2016.

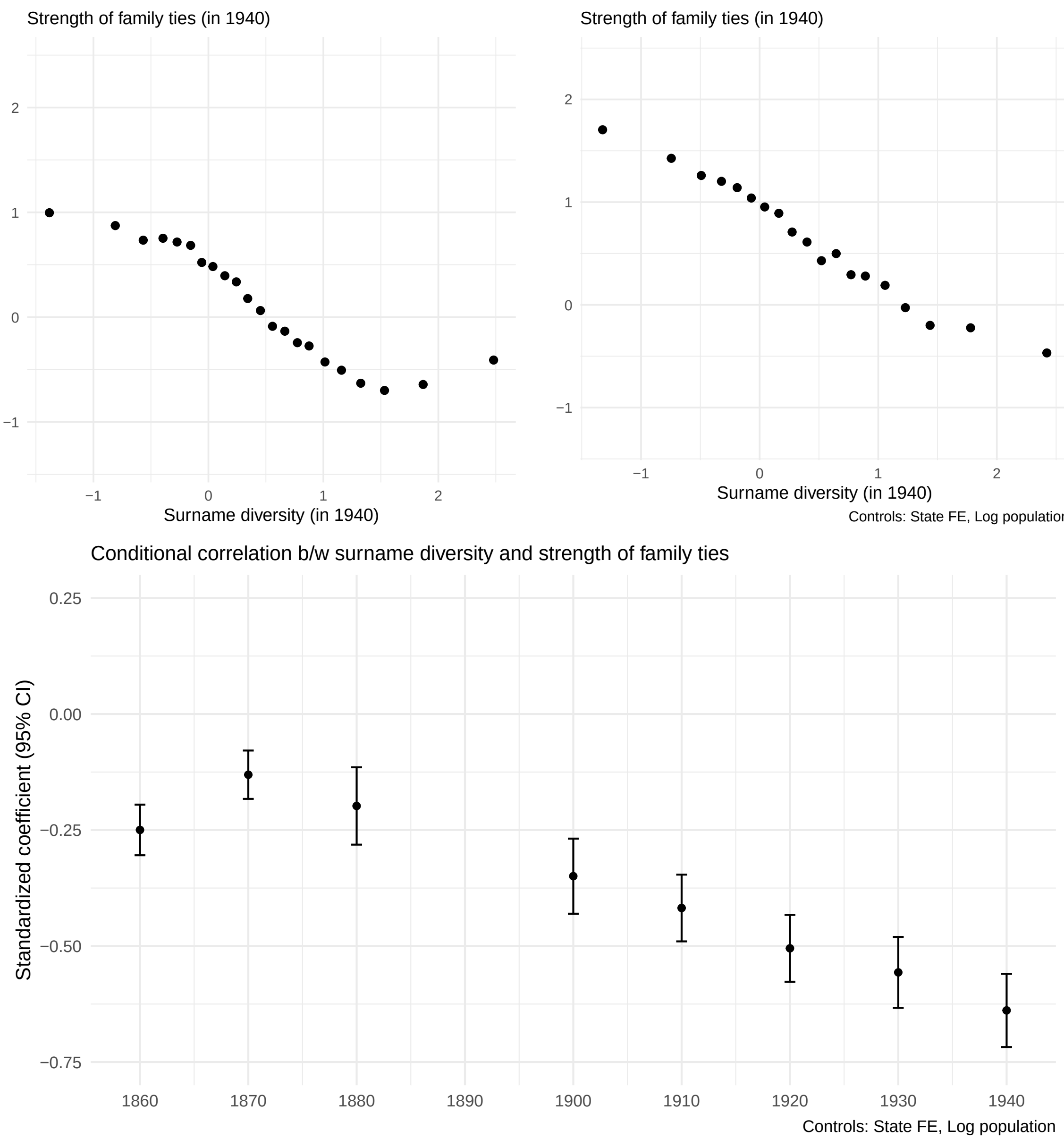


Figure C2: Relationship between surname diversity and strength of family ties

Notes: An observation is a county. Top left: Bivariate relationship in 1940. Top right: Variables residualized by state fixed effects and log county population in 1940. Bottom: Coefficients of regressions of strength of family ties on surname diversity conditional on state fixed effects and log county population by census year (1860-1940). The strength of family ties data is constructed following Raz (2023).

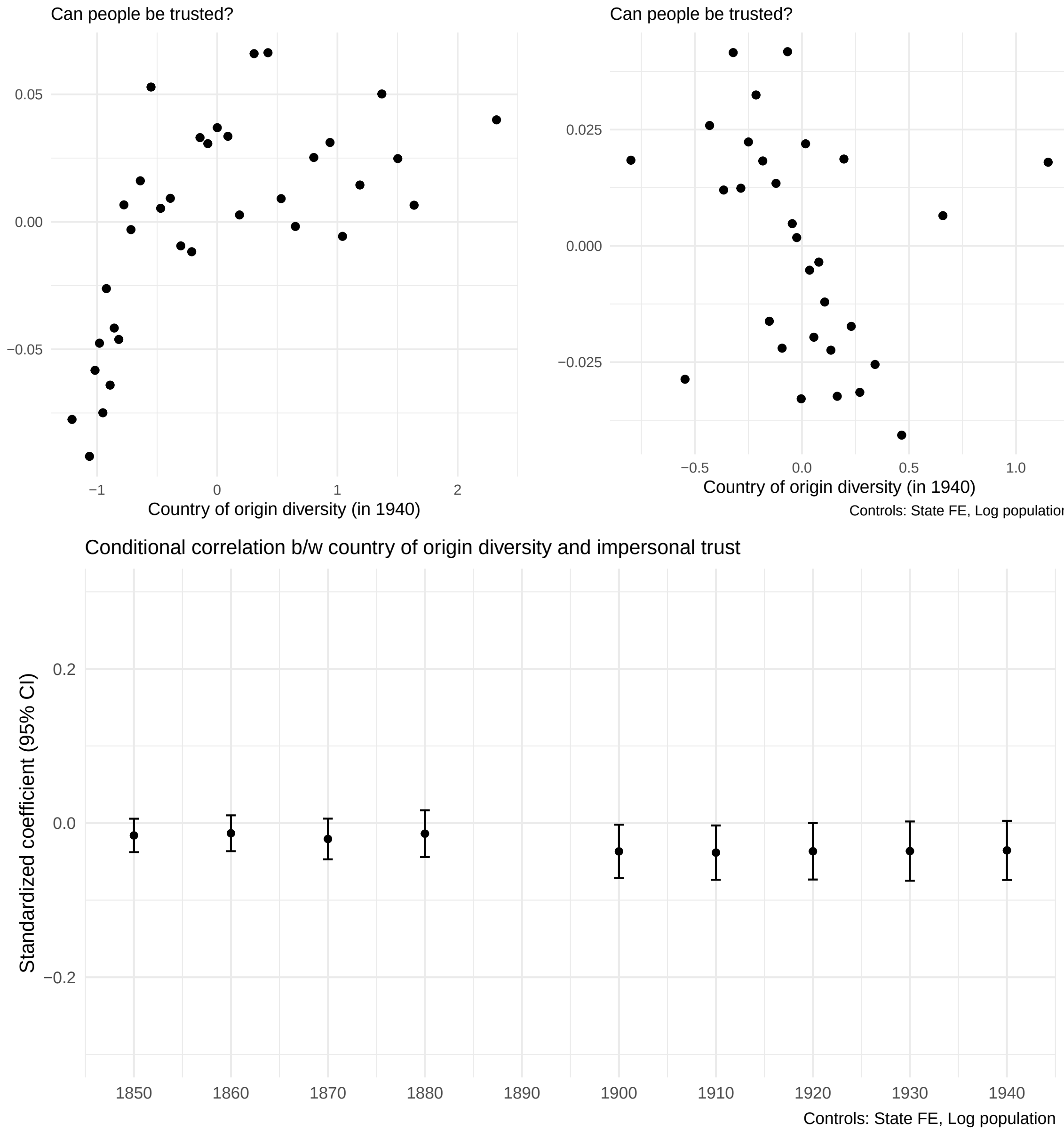


Figure C3: Relationship between country of origin diversity and impersonal trust

Notes: An observation is an individual. Top left: Bivariate relationship in 1940. Top right: Variables residualized by state fixed effects and log county population in 1940. Bottom: Coefficients of regressions of impersonal trust today on country of origin diversity conditional on state fixed effects and log county population by census year (1850-1940) and survey year, sex, age, and race fixed effects. The trust question is taken from the General Social Survey, waves 1972 to 2016.

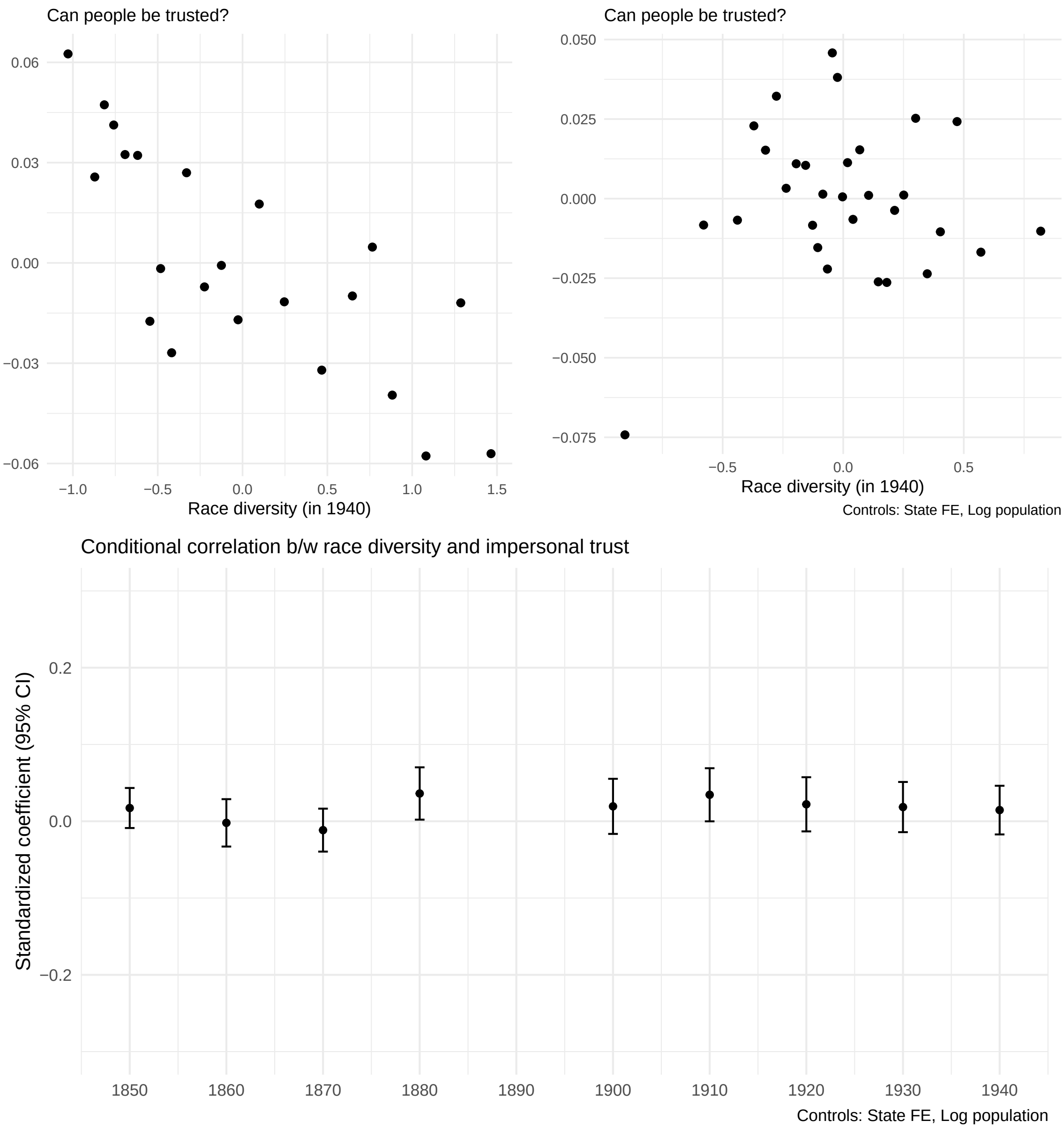


Figure C4: Relationship between race diversity and impersonal trust

Notes: An observation is an individual. Top left: Bivariate relationship in 1940. Top right: Variables residualized by state fixed effects and log county population in 1940. Bottom: Coefficients of regressions of impersonal trust today on race diversity conditional on state fixed effects and log county population by census year (1850-1940) and survey year, sex, age, and race fixed effects. The trust question is taken from the General Social Survey, waves 1972 to 2016.

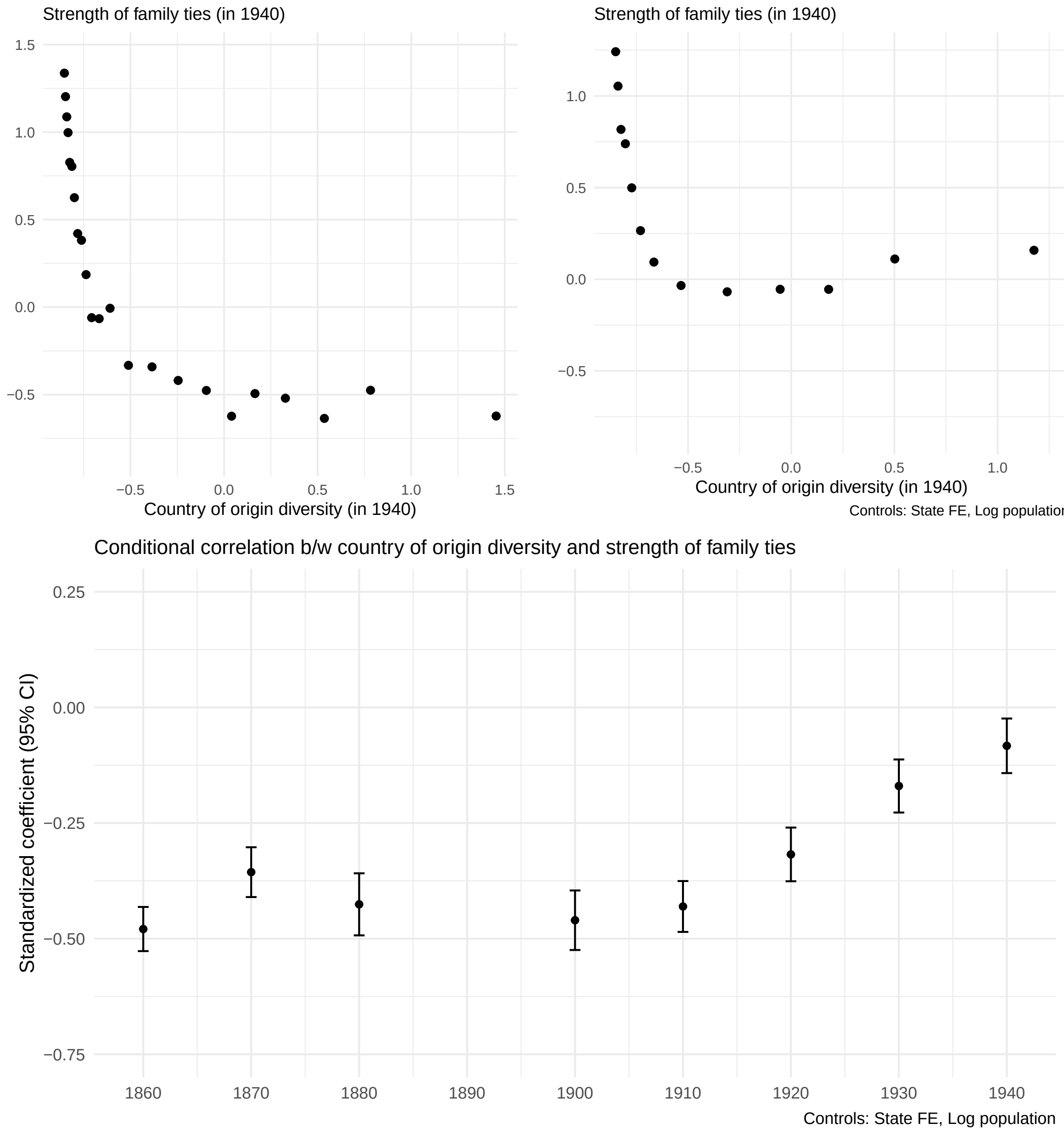


Figure C5: Relationship between country of origin diversity and strength of family ties

Notes: An observation is a county. Top left: Bivariate relationship in 1940. Top right: Variables residualized by state fixed effects and log county population in 1940. Bottom: Coefficients of regressions of strength of family ties on country of origin diversity conditional on state fixed effects and log county population by census year (1860-1940). The strength of family ties data is constructed following Raz (2023).

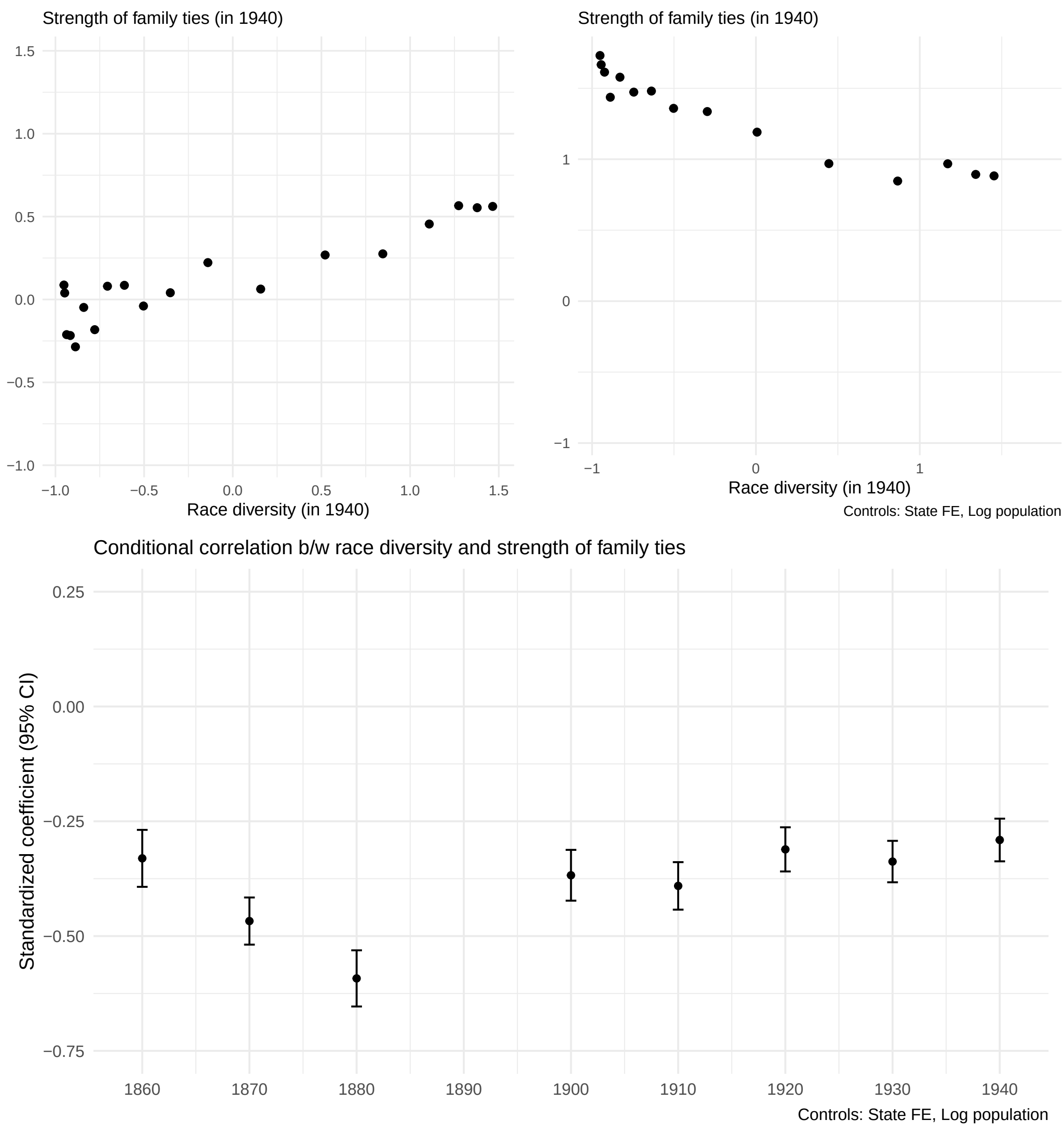


Figure C6: Relationship between race diversity and strength of family ties

Notes: An observation is a county. Top left: Bivariate relationship in 1940. Top right: Variables residualized by state fixed effects and log county population in 1940. Bottom: Coefficients of regressions of strength of family ties on race diversity conditional on state fixed effects and log county population by census year (1860-1940). The strength of family ties data is constructed following Raz (2023).

D Robustness of Main Results

D.1 Least-Squares Results

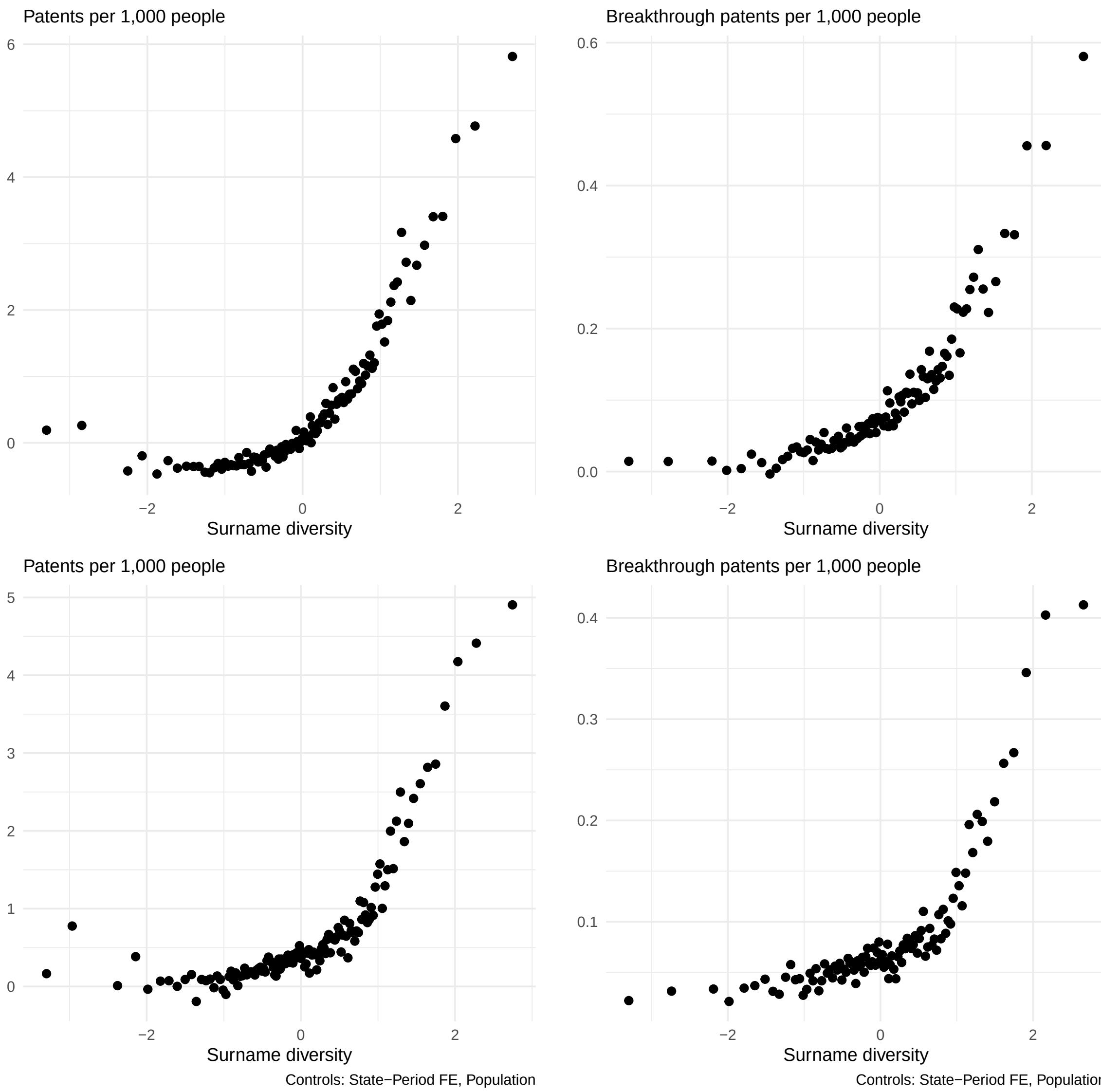


Figure D1: Relationships between surname diversity and (breakthrough) patents

Notes: County-level data from 1850 to 1940 (excluding the midyears). Observations are weighted by county population in 1850 and residualized by census year fixed effects. Bottom graphs: observations are additionally residualized by state-period fixed effects and county population. Binscatter plot created using the R package written by Cattaneo et al. (2019).

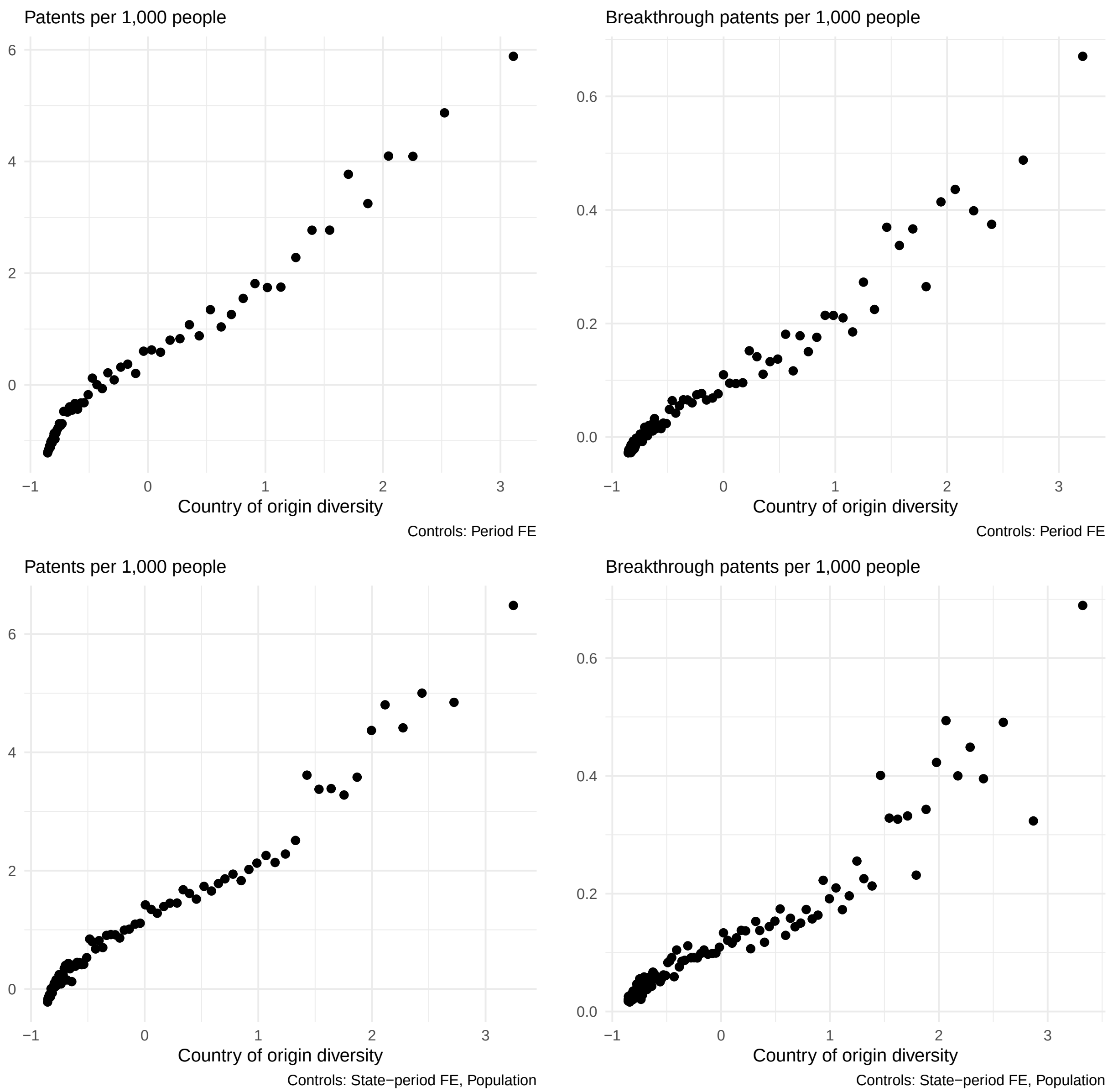


Figure D2: Relationships between country of origin diversity and (breakthrough) patents
Notes: County-level data from 1850 to 1940 (excluding the midyears). Observations are weighted by county population in 1850 and residualized by census year fixed effects. Bottom graphs: observations are additionally residualized by state-period fixed effects and county population size. Binscatter plot created using the R package written by Cattaneo et al. (2019).

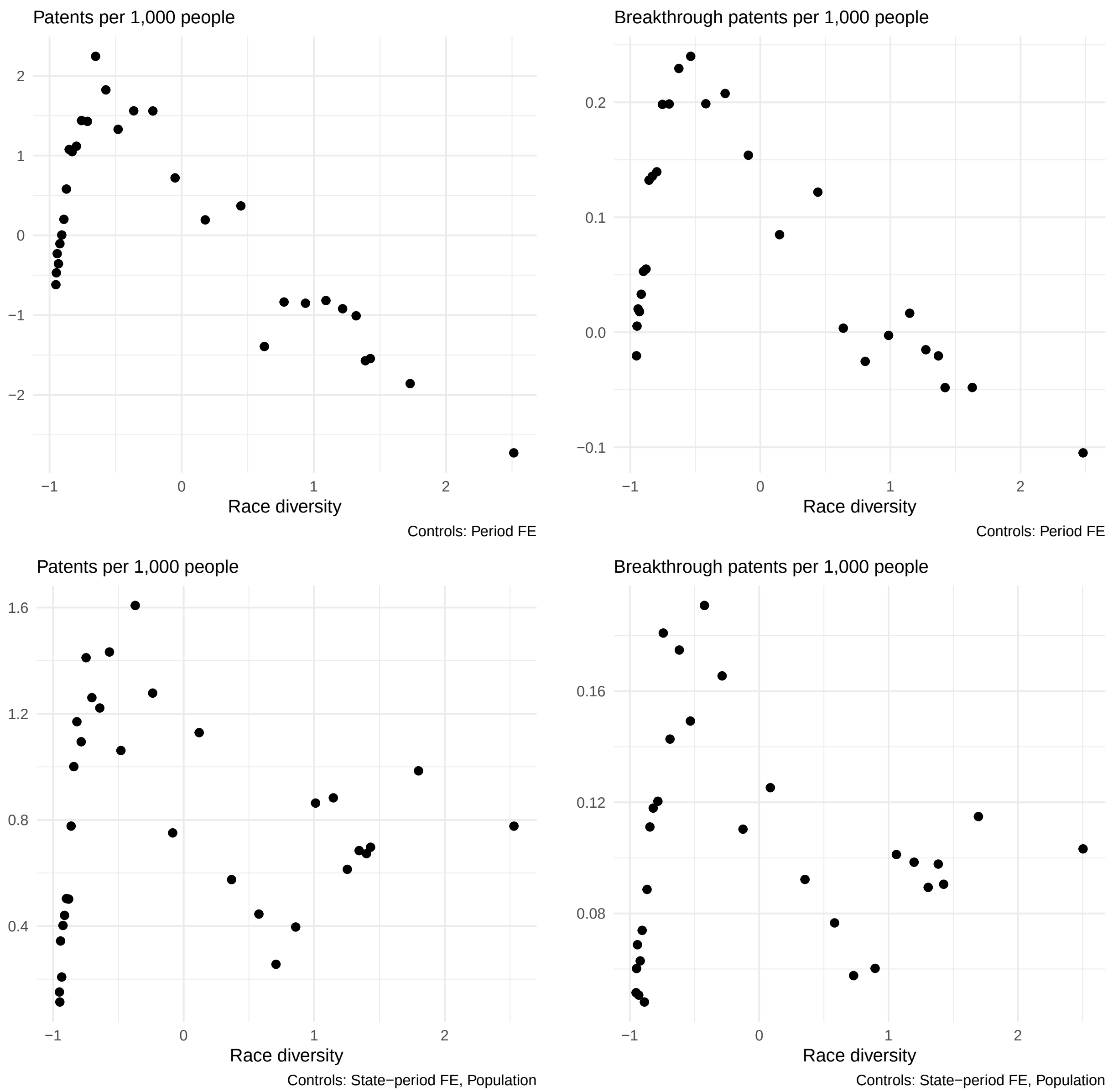


Figure D3: Relationships between race diversity and (breakthrough) patents

Notes: County-level data from 1850 to 1940 (excluding the midyears). Observations are weighted by county population in 1850 and residualized by census year fixed effects. Bottom graphs: observations are additionally residualized by state-period fixed effects and county population size. Binscatter plot created using the R package written by Cattaneo et al. (2019).

Table D1: Least-squares estimates: inverse hyperbolic sine transformed outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	IHS Patents per 1,000 people (mean = 1.20, sd = 0.90)					
Surname diversity	0.688***		0.435***	0.435***	0.441***	0.123***
	(0.038)		(0.041)	(0.045)	(0.050)	(0.038)
Country of origin diversity		0.590***	0.273***	0.284***	0.215***	0.033
		(0.027)	(0.045)	(0.046)	(0.054)	(0.060)
R^2	0.635	0.600	0.671	0.697	0.768	0.901
Panel B:	IHS Breakthrough patents per 1,000 people (mean = 0.18, sd = 0.27)					
Surname diversity	0.164***		0.063***	0.053***	0.059***	0.063*
	(0.027)		(0.014)	(0.014)	(0.017)	(0.033)
Country of origin diversity		0.155***	0.109***	0.120***	0.131***	0.002
		(0.020)	(0.027)	(0.026)	(0.036)	(0.019)
R^2	0.366	0.412	0.428	0.467	0.573	0.760
Immigrant shares by country of origin (59 shares)				✓	✓	✓
Period fixed effects	✓	✓	✓	✓		
Period-State fixed effects					✓	✓
County fixed effects						✓
Observations	22,222	22,222	22,222	22,222	22,222	22,222

Notes: The table reports estimates of least-squares regressions of innovation outcomes on surname diversity and other dimensions of sociocultural diversity. In Panel A (Panel B), the outcome is inverse hyperbolic sine transformed number of (breakthrough) patents issued in a given period per 1,000 people. The unit of observation is a county-period from 1850 to 1940 (excluding the midyears). Observations are weighted by county population in 1850. Standard errors are clustered on states and reported in parentheses. 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.

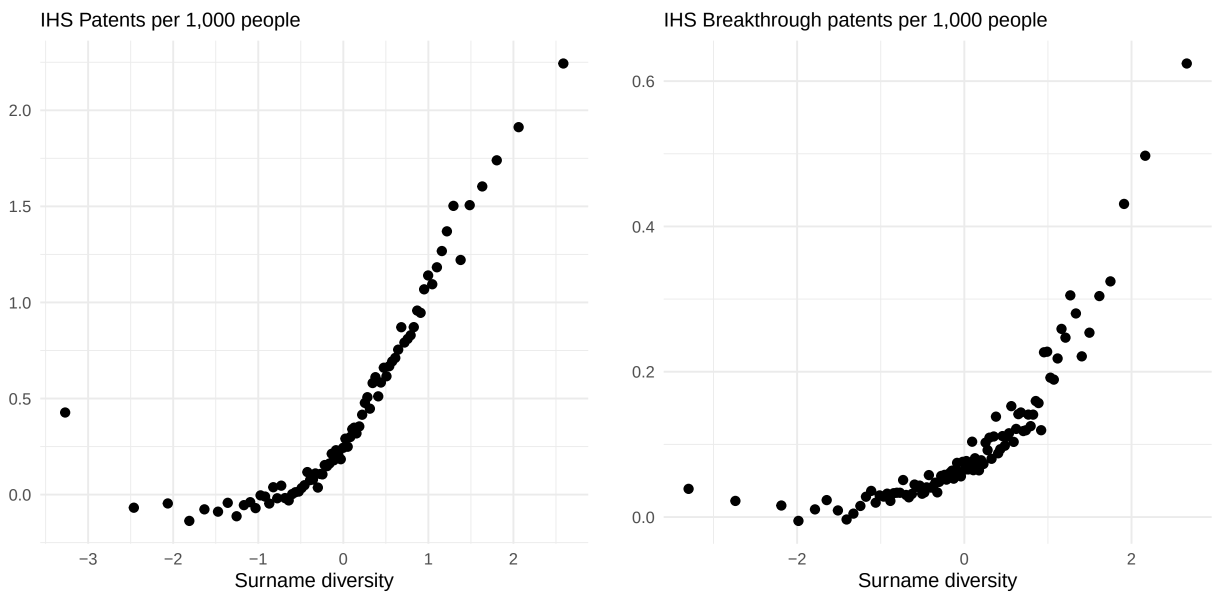


Figure D4: Relationships between surname diversity and inverse hyperbolic sine transformed innovation outcomes

Notes: County-level data from 1850 to 1940 (excluding the midyears). Observations are weighted by county population in 1850 and residualized by census year fixed effects. Binscatter plot created using the R package written by Cattaneo et al. (2019).

Table D2: Least-squares estimates: population size

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	Patents per 1,000 people (mean = 2.26, sd = 2.58)					
Surname diversity		1.51***	0.677***	0.665***	0.802***	0.602***
		(0.134)	(0.149)	(0.154)	(0.168)	(0.187)
Population	0.345***	0.115***	0.058***	0.043**	0.070***	-0.070***
	(0.050)	(0.033)	(0.021)	(0.021)	(0.021)	(0.015)
Country of origin diversity			1.03***	1.13***	0.930***	0.147
			(0.177)	(0.177)	(0.239)	(0.196)
R^2	0.305	0.521	0.579	0.609	0.695	0.865
Panel B:	Breakthrough patents per 1,000 people (mean = 0.18, sd = 0.24)					
Surname diversity		0.123***	0.054***	0.052***	0.057***	0.057**
		(0.014)	(0.011)	(0.011)	(0.014)	(0.024)
Population	0.033***	0.015***	0.010***	0.007***	0.009***	-0.009***
	(0.005)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
Country of origin diversity			0.085***	0.098***	0.095***	0.002
			(0.016)	(0.016)	(0.024)	(0.018)
R^2	0.283	0.448	0.494	0.530	0.626	0.790
Immigrant shares by country of origin (59 shares)				✓	✓	✓
Period fixed effects	✓	✓	✓	✓		
Period-State fixed effects					✓	✓
County fixed effects						✓
Observations	22,299	22,299	22,299	22,299	22,299	22,299

Notes: The table reports least-squares estimates of regressions of innovation outcomes on surname diversity, immigrant diversities and population size. In Panel A (Panel B), the outcome is number of (breakthrough) patents per 1,000 people. The unit of observation is a county-period from 1850 to 1940 (excluding the midyears). Standard errors are clustered on states and reported in parentheses. 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 D3: Least-squares estimates: race and occupational diversity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A:	Patents per 1,000 people (mean = 1.20, sd = 0.90)						
Surname diversity	1.76***		0.734***	0.692***	0.342**	0.577***	0.488**
	(0.175)		(0.157)	(0.161)	(0.128)	(0.130)	(0.203)
Country of origin diversity		1.62***	1.10***	1.16***	1.08***	1.03***	0.137
		(0.103)	(0.184)	(0.178)	(0.179)	(0.242)	(0.194)
Race diversity		-0.086	-0.010	-0.135	-0.146	-0.086	-0.478***
		(0.107)	(0.113)	(0.094)	(0.092)	(0.082)	(0.148)
Occupational diversity					0.535***	0.422***	0.180
					(0.095)	(0.133)	(0.134)
R^2	0.503	0.550	0.574	0.608	0.620	0.696	0.866
Panel B:	Breakthrough patents per 1,000 people (mean = 0.18, sd = 0.27)						
Surname diversity	0.154***		0.067***	0.060***	0.036***	0.046***	0.044*
	(0.021)		(0.012)	(0.012)	(0.009)	(0.011)	(0.023)
Country of origin diversity		0.148***	0.100***	0.108***	0.103***	0.110***	0.002
		(0.015)	(0.019)	(0.018)	(0.019)	(0.025)	(0.016)
Race diversity		0.018	0.025**	0.015	0.015	0.003	-0.047***
		(0.011)	(0.012)	(0.010)	(0.011)	(0.009)	(0.017)
Occupational diversity					0.037***	0.027**	0.013
					(0.009)	(0.011)	(0.008)
R^2	0.416	0.461	0.485	0.525	0.531	0.622	0.789
Immigrant shares by country of origin (59 shares)				✓	✓	✓	✓
Period fixed effects	✓	✓	✓	✓	✓		
Period-State fixed effects						✓	✓
County fixed effects							✓
Observations	22,206	22,206	22,206	22,206	22,206	22,206	22,206

Notes: The table reports least-squares estimates of regressions of innovation outcomes on surname diversity and other dimensions of sociocultural diversity, including race and occupational diversity. In Panel A (Panel B), the outcome is number of (breakthrough) patents per 1,000 people. The unit of observation is a county-period from 1850 to 1940 (excluding the midyears). Standard errors are clustered on states and reported in parentheses. 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 D4: Least-squares estimates: education

	Patents per 1,000 people (mean = 1.78, sd = 1.88)			Breakthrough patents per 1,000 people (mean = 0.18, sd = 0.21)		
	(1)	(2)	(3)	(4)	(5)	(6)
Average years of schooling	1.10***	0.072	0.292**	0.114***	0.009	0.037**
	(0.125)	(0.114)	(0.140)	(0.015)	(0.009)	(0.014)
Surname diversity		0.589***	0.360**		0.056***	0.028*
		(0.145)	(0.146)		(0.013)	(0.014)
Surname diversity × Average years of schooling		0.573***	0.653***		0.064***	0.069***
		(0.105)	(0.121)		(0.013)	(0.016)
Constant	1.28***	0.147**		0.129***	0.012*	
	(0.096)	(0.061)		(0.011)	(0.007)	
R^2	0.282	0.563	0.650	0.247	0.495	0.592
State fixed effects			✓			✓
Observations	3,078	3,078	3,078	3,078	3,078	3,078

Notes: The table reports least-squares estimates of regressions of innovation outcomes on surname diversity and individuals' average years of schooling. The unit of observation is a county-period in 1940. Observations are weighted by county population in 1940. Standard errors are clustered on states and reported in parentheses. 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.

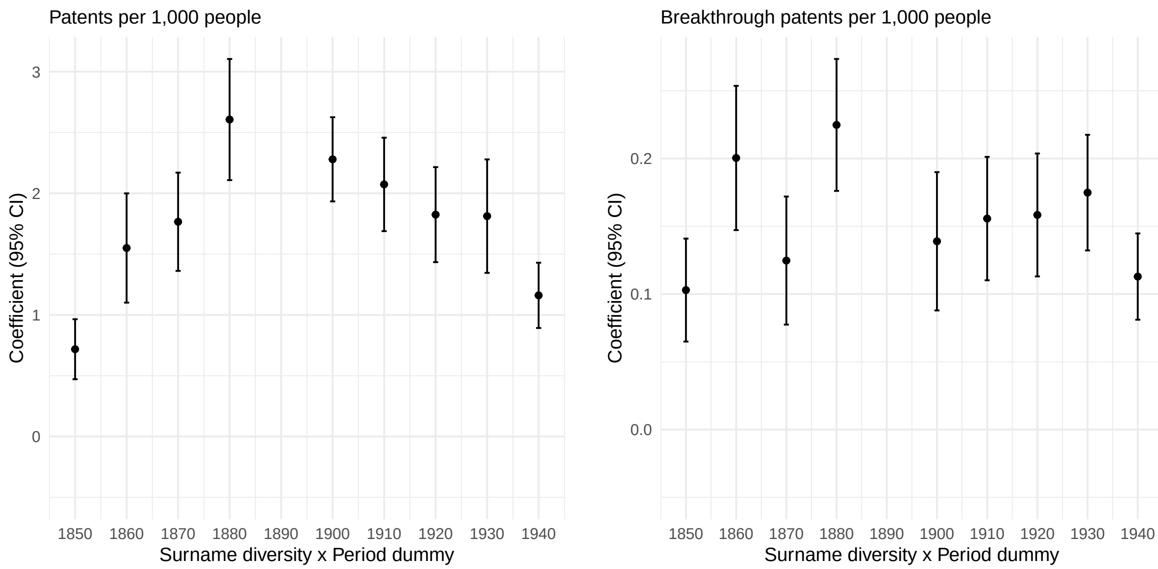


Figure D5: Correlations between surname diversity in years 1850-1940 and innovation

Notes: Each figure shows coefficients of a regression of an innovation outcome on surname diversity interacted with year dummies conditional on year fixed effects.

D.2 IV Results

Table D5: Zero-stage panel estimates

	No. of people in county i with surname n in year:							
	1900	1905	1910	1915	1920	1925	1930	1940
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$I^{1880}_{n,-r(i)} \times \frac{I^{1880}_{-n,i}}{I^{1880}_{-n}}$	3.181***	3.192***	3.683***	3.982***	4.386***	4.516***	4.945***	5.030***
	(0.113)	(0.121)	(0.130)	(0.095)	(0.113)	(0.136)	(0.108)	(0.101)
$I^{1895}_{n,-r(i)} \times \frac{I^{1895}_{-n,i}}{I^{1895}_{-n}}$	1.962***	2.336***	2.902***	2.276***	2.588***	4.071***	4.471***	4.551***
	(0.274)	(0.336)	(0.311)	(0.125)	(0.120)	(0.199)	(0.243)	(0.290)
$I^{1900}_{n,-r(i)} \times \frac{I^{1900}_{-n,i}}{I^{1900}_{-n}}$		-2.405	-0.262	-6.766**	-5.988**	-9.868***	-10.297***	-8.711***
		(3.101)	(3.497)	(2.632)	(2.742)	(3.227)	(3.513)	(3.312)
$I^{1905}_{n,-r(i)} \times \frac{I^{1905}_{-n,i}}{I^{1905}_{-n}}$			12.702***	17.984***	20.613***	27.031***	30.116***	32.922***
			(0.885)	(0.911)	(1.110)	(1.401)	(1.219)	(1.250)
$I^{1910}_{n,-r(i)} \times \frac{I^{1910}_{-n,i}}{I^{1910}_{-n}}$				14.937***	16.972***	24.672***	26.990***	28.736***
				(2.540)	(2.918)	(3.346)	(3.213)	(2.902)
$I^{1915}_{n,-r(i)} \times \frac{I^{1915}_{-n,i}}{I^{1915}_{-n}}$					8.268***	8.023***	9.419***	10.294***
					(0.803)	(0.830)	(0.546)	(0.602)
$I^{1920}_{n,-r(i)} \times \frac{I^{1920}_{-n,i}}{I^{1920}_{-n}}$						3.659*	5.087***	9.319***
						(2.198)	(1.181)	(1.584)
$I^{1925}_{n,-r(i)} \times \frac{I^{1925}_{-n,i}}{I^{1925}_{-n}}$							25.957***	31.662***
							(1.293)	(1.490)
$I^{1930}_{n,-r(i)} \times \frac{I^{1930}_{-n,i}}{I^{1930}_{-n}}$								-29.396***
								(2.731)
Observations	5,933,320	7,336,530	7,806,098	7,365,155	7,448,013	7,977,906	8,053,917	8,811,918
R^2	0.710	0.687	0.698	0.701	0.711	0.660	0.706	0.690
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Surname-Region fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
$I^{\tau}_{-n,i}/I^{\tau}_{-n}$ controls	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table reports OLS estimates for the specification described in equation (2), corresponding to step 1 of the instrument construction. An observation is a surname-county in a period from 1900 to 1940. Standard errors clustered at the surname level. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

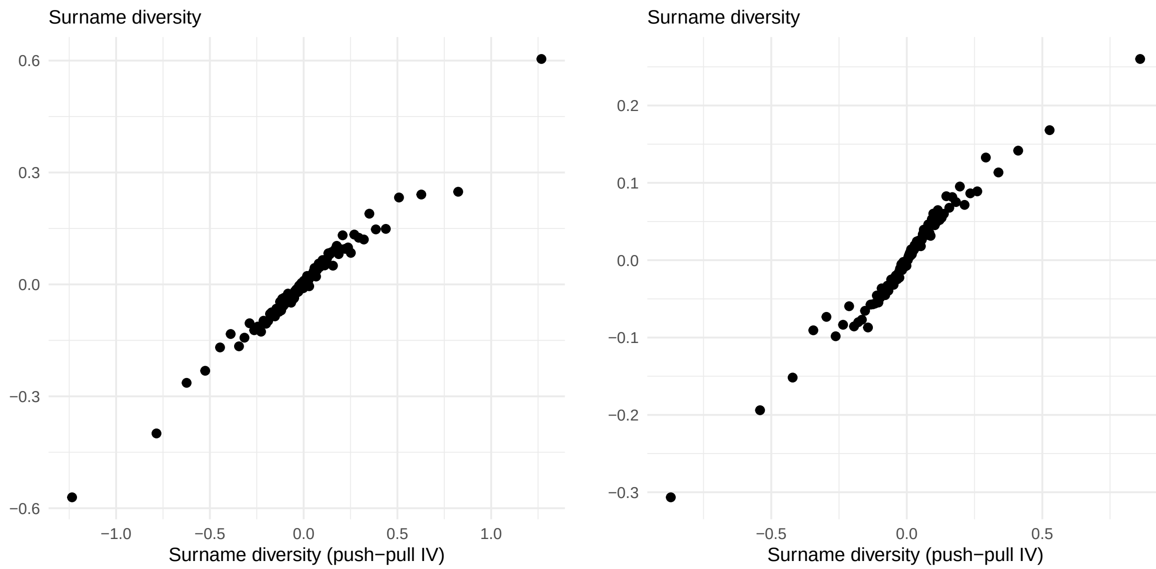


Figure D6: Binned scatter plots of surname diversity (pull-push IV) and actual surname diversity from 1900 to 1940

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

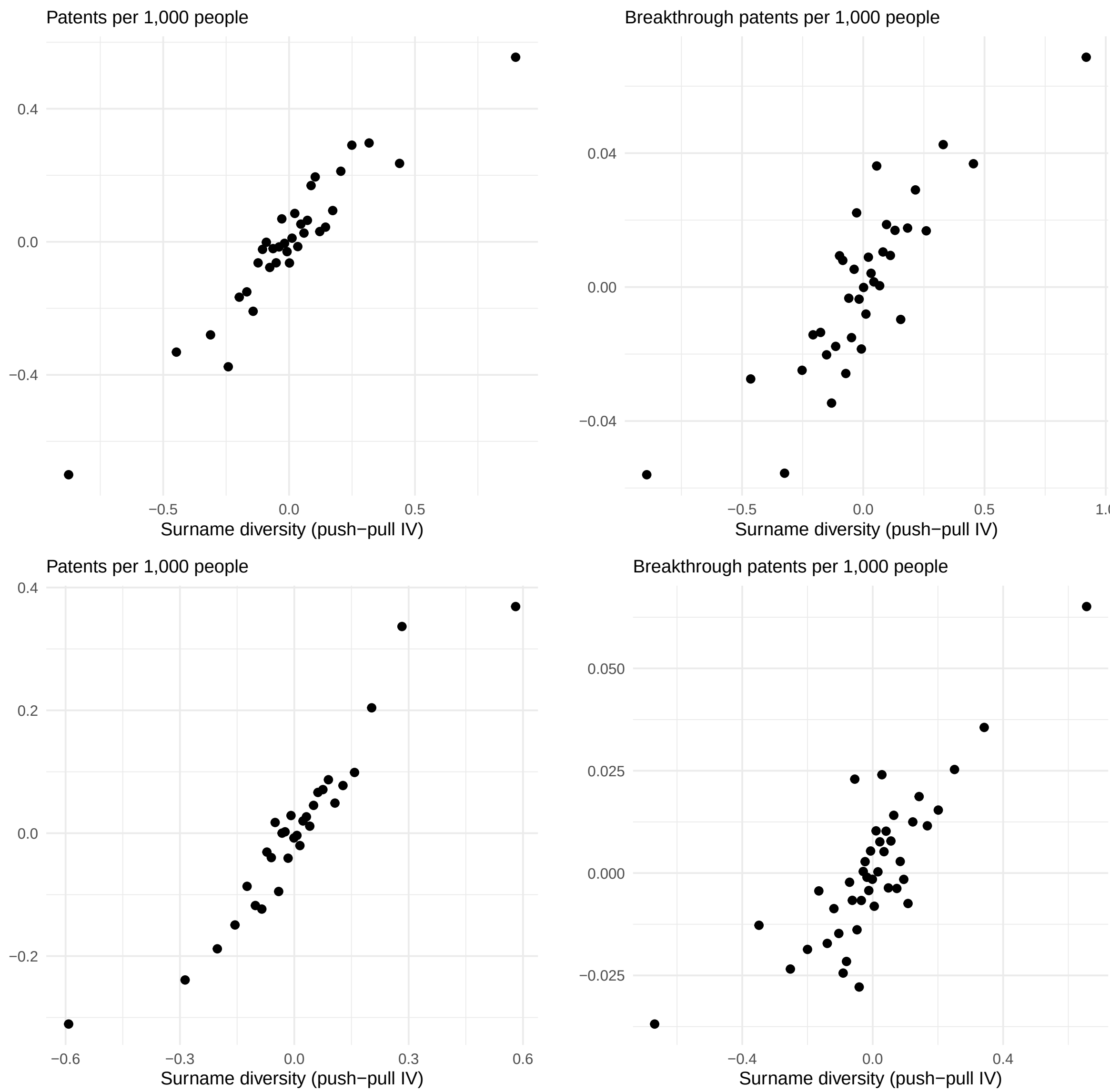


Figure D7: Binned scatter plots of surname diversity (pull-push IV) and innovation outcomes from 1900 to 1940

Notes: County-level data from 1900 to 1940 (including midyears). Observations are weighted by county population in 1900 and residualized by county fixed effects and state-period fixed effects (top plots), and county-specific time trends (bottom plots).

Table D6: Controlling and instrumenting for population size

	Patents per 1,000 people (mean = 2.04, sd = 2.6)			Breakthrough patents per 1,000 people (mean = 0.14, sd = 0.24)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity	1.428***	1.425***	1.206***	0.170***	0.136***	0.111***
	(0.322)	(0.339)	(0.292)	(0.041)	(0.041)	(0.038)
Population	0.069**	0.079**	0.118***	0.009***	0.010***	0.021***
	(0.030)	(0.036)	(0.042)	(0.003)	(0.002)	(0.007)
Panel B: Reduced-form estimates						
Surname diversity (push-pull IV)	0.651***	0.753***	0.638***	0.084***	0.081***	0.060*
	(0.184)	(0.131)	(0.164)	(0.018)	(0.021)	(0.034)
Population (push-pull IV)	0.023	0.013	0.082	0.004	0.003	0.017
	(0.043)	(0.062)	(0.121)	(0.004)	(0.005)	(0.018)
Panel C: Instrumental-variable estimates						
Surname diversity	1.460***	1.707***	1.523***	0.189***	0.184***	0.137
	(0.363)	(0.352)	(0.522)	(0.044)	(0.056)	(0.096)
Population	0.032	0.031	0.163	0.005	0.005	0.031
	(0.043)	(0.058)	(0.178)	(0.003)	(0.004)	(0.027)
KP *F*-statistic, Surname diversity	44.298	39.749	17.979	44.298	39.749	17.979
KP *F*-statistic, Population	76.245	93.742	295.820	76.245	93.742	295.820
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	23,660	23,660	23,660	23,660	23,660	23,660

Notes: The table reports the estimates of the least-squares, reduced-form, and IV estimates for the specification described in equation (4) but additionally controlling or instrumenting for (the instrument for) county population size, as predicted by our estimates for equation (2). An observation is a county-period from 1900 to 1940. Observations are weighted by county population in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table D7: Regional heterogeneity in the effect of surname diversity on innovation

	Patents per 1,000 people (mean = 2.04, sd = 2.60)			Breakthrough patents per 1,000 people (mean = 0.14, sd = 0.24)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname diversity × Region = Midwest	2.927***	2.807***	3.069***	0.251***	0.241***	0.286***
	(0.433)	(0.446)	(0.509)	(0.038)	(0.038)	(0.068)
Surname diversity × Region = Northeast	1.596**	3.112**	1.983***	0.264***	0.378**	0.358*
	(0.697)	(1.171)	(0.721)	(0.084)	(0.156)	(0.178)
Surname diversity × Region = South	0.826***	0.433***	0.367**	0.071***	0.037***	0.015**
	(0.150)	(0.140)	(0.153)	(0.017)	(0.008)	(0.008)
Surname diversity × Region = West	1.697***	0.941**	1.325***	0.201**	0.066***	0.071***
	(0.564)	(0.380)	(0.380)	(0.090)	(0.023)	(0.021)
Panel B: Reduced-form estimates						
Surname diversity (push-pull IV) × Region = Midwest	1.685***	1.919***	1.648***	0.151***	0.175***	0.168***
	(0.362)	(0.328)	(0.419)	(0.035)	(0.039)	(0.062)
Surname diversity (push-pull IV) × Region = Northeast	0.430	0.646***	0.547***	0.099***	0.114***	0.098***
	(0.305)	(0.170)	(0.120)	(0.036)	(0.018)	(0.026)
Surname diversity (push-pull IV) × Region = South	0.540***	0.357**	0.287***	0.056***	0.035***	0.012**
	(0.155)	(0.170)	(0.106)	(0.014)	(0.011)	(0.005)
Surname diversity (push-pull IV) × Region = West	0.729***	0.369***	0.660***	0.061***	0.002	0.031*
	(0.197)	(0.085)	(0.128)	(0.022)	(0.010)	(0.019)
Panel C: Instrumental-variable estimates						
Surname diversity × Region = Midwest	3.300***	3.721***	3.978***	0.313***	0.339***	0.405**
	(0.653)	(0.670)	(1.093)	(0.065)	(0.079)	(0.171)
Surname diversity × Region = Northeast	1.237*	2.875***	2.889***	0.254***	0.509***	0.517**
	(0.631)	(0.640)	(0.893)	(0.077)	(0.139)	(0.246)
Surname diversity × Region = South	1.098***	0.606**	0.464***	0.122***	0.060***	0.019*
	(0.244)	(0.265)	(0.169)	(0.025)	(0.017)	(0.010)
Surname diversity × Region = West	1.637**	0.928*	1.753*	0.142*	0.005	0.083
	(0.697)	(0.461)	(0.910)	(0.073)	(0.023)	(0.071)
Kleibergen-Paap *F*-statistic 1st coefficient	200.951	105.238	27.024	200.951	105.238	27.024
Kleibergen-Paap *F*-statistic 2nd coefficient	21.441	3.995	2.791	21.441	3.995	2.791
Kleibergen-Paap *F*-statistic 3rd coefficient	106.312	29.253	21.172	106.312	29.253	21.172
Kleibergen-Paap *F*-statistic 4th coefficient	30.594	2.363	1.981	30.594	2.363	1.981
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	23,660	23,660	23,660	23,660	23,660	23,660

Notes: The table reports regional heterogeneity in the least-squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (4). An observation is a county in a period from 1900 to 1940. Observations are weighted by county population in 1900. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Table D8: The effect of surname diversity on innovation among native-born Americans only

	Patents per 1,000 people (mean = 0.64, sd = 43.76)				Breakthrough patents per 1,000 people (mean = 0.04, sd = 9.61)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.317***	0.380***	0.331***	0.301***	0.062***	0.063***	0.047***	0.019*
	(0.072)	(0.077)	(0.081)	(0.086)	(0.021)	(0.020)	(0.015)	(0.010)
Panel B: Reduced-form estimates								
Surname diversity (push-pull IV)	0.213***	0.260***	0.237***	0.246***	0.044***	0.045***	0.039***	0.022*
	(0.064)	(0.065)	(0.065)	(0.066)	(0.015)	(0.015)	(0.013)	(0.012)
Panel C: Instrumental-variable estimates								
Surname diversity	0.333***	0.406***	0.375***	0.390***	0.069***	0.070***	0.062***	0.035*
	(0.095)	(0.095)	(0.100)	(0.114)	(0.023)	(0.023)	(0.020)	(0.019)
Kleibergen-Paap *F*-statistic	460.250	457.336	343.932	217.417	460.250	457.336	343.932	217.417
Panel D: First-stage estimates	Surname diversity							
Surname diversity (push-pull IV)	0.638***	0.639***	0.631***	0.631***	0.638***	0.639***	0.631***	0.631***
	(0.030)	(0.030)	(0.034)	(0.043)	(0.030)	(0.030)	(0.034)	(0.043)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓				✓			
Surname-Period fixed effects		✓	✓	✓		✓	✓	✓
State-Period fixed effects			✓	✓			✓	✓
County-specific linear time trends				✓				✓
Observations	17,061,661	17,061,661	17,061,661	17,061,661	17,061,661	17,061,661	17,061,661	17,061,661

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation 6 and first-stage estimates for equation 5. An observation is a surname in a given county in a period from 1900 to 1940. The sample is restricted to observations with no immigrants in a given surname-county-period cell. Observations are weighted by the surname population in a given county in year 1900. Standard errors are two-way clustered on states and surnames 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.

D.3 Patent Technology Class Fixed Effects

Another potential concern with the interpretation of our findings is that patenting practices vary across industries and technologies (Moser, 2013), and these differences might affect our results.

Using the fact that the USPTO assigns a technology class to each granted patent, we assess this concern by estimating specifications that include patent class fixed effects to absorb any technology-specific traits. Similar to the surname fixed effects specifications in our main analysis, this requires us to change the unit of observation from county-period to patent class-county-period. The estimating equations are given by equations (9) and (10), where equation (9) is the first stage and equation (10) is the second stage.

$$\text{Surname diversity}_i^t = \gamma\ \widehat{\text{Surname diversity}}_i^t + \mu_{t,s(i)} + \mu_i + \mu_{t,c} + v_{i,c}^t \tag{9}$$

$$Y_{i,c}^t = \beta\ \text{Surname diversity}_i^t + \alpha_{t,s(i)} + \alpha_i + \alpha_{t,c} + \varepsilon_{i,c}^t \tag{10}$$

where i indexes counties, s states, t census years (including the midyears), and c patent class. There are 408 patent classes in our sample from 1900 to 1944. Examples of the patent class level are "Geometrical Instruments", "Stoves and Furnace", and "Chemistry: Electrical and Wave Energy". As before, Surname diversity$_i^t$ is county i's surname diversity in t, and $\widehat{\text{Surname diversity}}_i^t$ is county i's predicted surname diversity in t. $Y_{i,c}^t$ now is the number of (breakthrough) patents (per 1,000 residents) in patent class c, filed in county i in the five-year period starting in t. Therefore, the innovation outcomes vary at the patent class-county-period level, while surname diversity remains defined at the county-period level. Importantly, we can now include patent class-period fixed effects, denoted by the parameter $\alpha_{t,c}$, which implies we non-parametrically control for patent class-specific confounders across periods, including differences in patenting practices across industries. The coefficient of interest is β. Observations are weighted by the number of people in a county in the year 1900. Standard errors are clustered in two ways, on states and patent class.

The results are reported in Table D9 and show that estimates are virtually unaffected by the inclusion of patent class fixed effects (in columns 2 to 4 and 6 to 8). All the estimates are highly significant in all specifications. Thus, we conclude that differences across technological categories do not affect our results.

Table D9: Patent technology class fixed effects

	Patents per 1,000 people (mean = 0.01, sd = 0.03)				Breakthrough patents per 1,000 people (mean = <0.01, sd = 0.01)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.006***	0.006***	0.006***	0.004***	0.001***	0.001***	0.001***	0.000***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Panel B: Reduced-form estimates								
Surname diversity (push-pull IV)	0.002***	0.002***	0.002***	0.002**	0.000***	0.000***	0.000***	0.000**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Panel C: Instrumental-variable estimates								
Surname diversity	0.005***	0.005***	0.005***	0.007**	0.001***	0.001***	0.001***	0.001**
	(0.001)	(0.001)	(0.001)	(0.003)	(0.000)	(0.000)	(0.000)	(0.000)
Kleibergen-Paap F-statistic	57.355	57.355	46.053	30.024	57.355	57.355	46.053	30.024
Panel D: First-stage estimates	Surname diversity							
Surname diversity (push-pull IV)	0.423***	0.423***	0.407***	0.356***	0.423***	0.423***	0.407***	0.356***
	(0.056)	(0.056)	(0.060)	(0.065)	(0.056)	(0.056)	(0.060)	(0.065)
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓				✓			
Patent class-Period fixed effects		✓	✓	✓		✓	✓	✓
State-Period fixed effects			✓	✓			✓	✓
County-specific linear time trends				✓				✓
Observations	8,264,856	8,264,856	8,264,856	8,264,856	8,264,856	8,264,856	8,264,856	8,264,856

Notes: The table reports least-squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation 10 and first-stage estimates for equation 9. An observation is a patent class in a given county in a period from 1900 to 1940. Observations are weighted by the population in a given county in the year 1900. In columns 1 to 3, the dependent variable is number of patents with c as the main technological category and filed by individuals in county i in the five-year period starting in t divided by population size in county i in 1900. The dependent variable in columns 4 to 6 is the corresponding number of breakthrough patents. Standard errors are two-way clustered on states and technological category 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.

D.4 Alternative Shift-share Instruments

The conventional shift-share approach in the immigration literature rests on the observations that migrants locate near people from the same country of origin (Altonji and Card, 1991; Card, 2001). We can adapt this approach to our context because migrants also settle near family members. That is, other than in the push-pull approach, we allocate newly arriving migrants (i.e., between t and $t-1$) according to the preexisting share of people in a county (in year $t-1$) with the same surname. This procedure allows us to calculate the predicted inflow of migrants by surnames in this period.

The construction of the instrument for surname diversity based on this method requires counties' previous-period stocks (and not just inflow) of each surname. To get the current stocks (i.e., in t), we add the last-period inflow (predicted via shift-share) to counties' previous-period stocks of each surname (i.e., in $t-1$). Then, we apply the entropy formula to obtain the instrument for diversity.[20]

A concern with this calculation is that previous-period surname stocks are endogenous. The inclusion of county fixed effects in the estimation mitigates this concern somewhat because it shifts the focus from levels to changes in diversity; the previous period surname stocks are hence less important as a source of bias. Nevertheless, we follow the approach in Burchardi et al. (2021) and construct an additional shift-share instrument of surname diversity, which relies on the *predicted* stock of surnames in the previous period. These predicted previous-period stocks are calculated based on the historical push-pull approach. We add the predicted stock of surname in $t-1$ (calculated based on the push-pull approach) to the predicted inflow between t and $t-1$ (calculated based on the shift-share approach) to arrive at the predicted surname stocks in t.

Appendix Table D10 reports the estimates for both the IV specification that rests on the shift-share instrument alone (Panel A) and on the one that combines the shift-share and historical push-pull approach (Panel B). Consistent with our baseline results, the estimates are positive and highly significant for both patents and breakthrough patents per capita. Their point estimates are larger, though, they also estimated with more noise. In summary, our results are robust to the use of more conventional shift-share IV strategies.

[20]Since the 1940 census does not provide information on the immigration year, we cannot calculate a shift-share instrument for this period. Therefore, the sample in this robustness check is restricted to 1900 to 1930.

Table D10: Robustness: Alternative shift-share instruments

	Patents per 1,000 people (mean = 2.10, sd = 2.59)			Breakthrough patents per 1,000 people (mean = 0.14, sd = 0.25)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Shift-share IV using realized surname shares						
Surname diversity	2.216***	2.197***	3.624***	0.301***	0.222***	0.474***
	(0.449)	(0.488)	(0.706)	(0.061)	(0.063)	(0.140)
Kleibergen-Paap *F*-statistic	1,697	1,845	87	1,697	1,845	87
Panel B: Shift-share IV using push-pull predicted surname shares						
Surname diversity	2.539***	3.223***	4.917***	0.394***	0.379***	0.816**
	(0.670)	(0.697)	(1.142)	(0.075)	(0.090)	(0.373)
Kleibergen-Paap *F*-statistic	85	60	16	85	60	16
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	20,704	20,704	20,704	20,704	20,704	20,704

Notes: The table reports IV estimates for the specifications described in equation (4), but based on alternative shift-share procedures to construct the instrument for surname diversity. Panel A reports estimates for a shift-share instrument using realized surname shares, akin to Card (2001). Panel B reports estimates for a shift-share instrument using predicted surname shares based on the push-pull approach described in equation (2), akin to Burchardi et al. (2021). An observation is a county-period from 1900 to 1940. Observations are weighted by county population in 1900. 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.

D.5 Spatial Spillovers

People acquire inspiration, knowledge and ideas from others they frequently observe and interact with in their daily activities. Mostly, these will be people living in proximity and, consequently, we expect that local diversity drives innovation. (Carlino and Kerr, 2015).

Here, we investigate how local the relationship between diversity and innovation is and whether there are spillovers from nearby counties. To do so, we compute surname diversity among individuals residing in surrounding regions successively further away from the county. Specifically, for each county i at time t, we pool the individuals and compute surname diversity and construct a separate instrument for individuals living within 100 miles, excluding i itself, individuals living between 100 miles and 200 miles, and between 200 miles and 300 miles.[21]

Table D11 reports the results for patents (columns 1 to 4) and breakthrough patents (columns 5 to 8). According to columns 1 to 4 (panel C), an increase in surname diversity just outside and within 100 miles of county i increases its number of patents. A similar relation is observed for breakthrough patents, though these estimates are less accurate. Moving to regions still further away from county i (i.e., between 100 and 200 miles or 200 and 300 miles), we do not find evidence for spillover effects. Our findings suggest that the causal link between diversity and innovation tends to be local, including the neighboring areas that fall within a 100-mile radius of the county.

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

Table D11: Spillover analysis

	Patents per 1,000 people (mean = 2.04, sd = 2.59)				Breakthrough patents per 1,000 people (mean = 0.14, sd = 2.21)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	1.474***	1.471***	1.469***	1.180***	0.145***	0.144***	0.144***	0.114**
	(0.382)	(0.382)	(0.382)	(0.297)	(0.046)	(0.046)	(0.046)	(0.046)
Surname diversity (< 100 miles)	0.872	0.879	0.855	1.505***	0.066	0.067	0.065	0.187**
	(0.573)	(0.579)	(0.604)	(0.487)	(0.064)	(0.064)	(0.067)	(0.076)
Surname diversity (100 < 200 miles)		-0.522	-0.524	-1.278*		-0.053	-0.053	-0.221**
		(0.446)	(0.452)	(0.661)		(0.042)	(0.043)	(0.093)
Surname diversity (200 < 300 miles)			-0.205	0.121			-0.017	-0.024
			(0.383)	(0.458)			(0.042)	(0.084)
Panel B: Reduced-form estimates								
Surname diversity (push-pull IV)	0.780***	0.780***	0.783***	0.751***	0.086***	0.086***	0.087***	0.084***
	(0.161)	(0.161)	(0.162)	(0.176)	(0.019)	(0.019)	(0.019)	(0.026)
Surname diversity (push-pull IV, < 100 miles)	0.815***	0.814***	0.837***	0.653***	0.076**	0.075**	0.081**	0.059
	(0.280)	(0.284)	(0.304)	(0.179)	(0.031)	(0.031)	(0.034)	(0.036)
Surname diversity (push-pull IV, 100 < 200 miles)		-0.016	0.020	0.303		-0.007	0.002	0.023
		(0.174)	(0.188)	(0.301)		(0.017)	(0.021)	(0.037)
Surname diversity (push-pull IV, 200 < 300 miles)			0.106	0.235			0.026	0.046
			(0.166)	(0.194)			(0.026)	(0.040)
Panel C: Instrumental-variable estimates								
Surname diversity	1.674***	1.668***	1.668***	1.855***	0.188***	0.187***	0.188***	0.209**
	(0.357)	(0.358)	(0.357)	(0.573)	(0.058)	(0.058)	(0.057)	(0.099)
Surname diversity (< 100 miles)	2.533**	2.558**	2.552**	1.458**	0.217	0.222	0.227	0.115
	(1.202)	(1.204)	(1.190)	(0.660)	(0.143)	(0.146)	(0.151)	(0.157)
Surname diversity (100 < 200 miles)		-0.279	-0.286	0.698		-0.054	-0.047	0.050
		(0.552)	(0.552)	(1.029)		(0.068)	(0.072)	(0.132)
Surname diversity (200 < 300 miles)			-0.055	0.628			0.053	0.162
			(0.679)	(0.872)			(0.115)	(0.168)
F-statistic: Surname diversity	88.191	59.925	44.633	26.361	88.191	59.925	44.633	26.361
F-statistic: Surname diversity (< 100 miles)	17.373	14.392	12.706	37.700	17.373	14.392	12.706	37.700
F-statistic: Surname diversity (100 < 200 miles)		13.568	13.652	25.706		13.568	13.652	25.706
F-statistic: Surname diversity (200 < 300 miles)			6.150	10.814			6.150	10.814
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear time trends				✓				✓
Observations	23,093	23,093	23,093	23,093	23,093	23,093	23,093	23,093

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates of regressions of innovation outcomes on surname diversity. The unit of observation is a county-period from 1900 to 1940 (including the midyears). The table sequentially adds surname diversity in areas within 100 miles (excluding *i*), 100 miles to 200 miles, and 200 miles to 300 miles of county *i*. Observations are weighted by county population in 1900. 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.