How Social Structure Drives Innovation

Surname Diversity and Patents in U.S. History*

Max Posch† Jonathan Schulz‡ Joseph Henrich§

February 16, 2024

Abstract

We study the impact of social structure, captured by the distribution of surnames, on innovation in U.S. counties from 1850 to 1940. 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 entropy increases both the quantity and quality of innovation. Supporting analyses suggest that the relationship arises from an increase in interactions among individuals with different skills, expertise and perspectives. The results support the view that the free flow of information between diverse minds is a key driver of innovation.

Keywords: Innovation, surnames, immigration, social interactions, diversity
JEL Classification: O33, R11, N92, J15, Z13

*Max Posch and Jonathan Schulz share co-first authorship. For comments and helpful discussions, we thank Alberto Alesina, Mike Andrews, Pablo Balan, Michael Clemens, Tyler Cowen, Klaus Desmet, Dan Fetter, Vicky Fouka, Martin Fiszbein, Oded Galor, Erik Hornung, Chad Jones, Ross Mattheis, Petra Moser, Nathan Nunn, Tzachi Raz, Slava Savitskiy, and seminar and workshop participants at Bonn, 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), scholars from many disciplines have suggested that social interactions among diverse individuals should stimulate more rapid innovation and greater creativity, primarily because such interactions foster the exchange and recombination of ideas, approaches, practices and perspectives (e.g., Jacobs, 1969; Glaeser et al., 1992; Weitzman, 1998; Jones, forthcoming; Muthukrishna and Henrich, 2016). However, a wide range of evidence suggests that people tend to structure their communities in ways that may inhibit recombinative innovation, usually by clustering both geographically and relationally with their kin, co-ethics and others with whom they share cultural affinities (Burchardi et al., 2021; Kerr, 2008; Agrawal et al., 2008; AlShebli et al., 2018). Here, because such clustering or homophily (in network parlance) may influence the frequency of idea sharing among diverse individuals, we explore the connection between social structure and innovation. We tackle this in three steps. First, we develop a measure of social structure based on the distribution of surnames in U.S. counties from the mid-19th century to the mid-20th century. Second, using quasi-random variation in surnames based on immigration flows, we establish a causal link between our surname measure and two measures of innovation based on historical U.S. patents. Third, we conduct a large battery of robustness checks, including analyses at the surname-county level that confirm the relationship between social structure and innovation for those with the same surname. Finally, we confirm that our surname-based measure of social structure at the county-level does indeed lead to (1) lower street-level residential segregation, (2) weaker family ties and (3) more heterogeneous interactions among inventors at the level of individual patents. We also show that greater surname diversity at the patent level is associated with more impactful patents that involve more recombinative elements.

More specifically, hypotheses about recombinative innovation propose that many innovations emerge from the integration of ideas, approaches and techniques that connect during social interactions among diverse minds. At the population level, the meeting

and merging of people and ideas involves both an informational component—different people have to possess distinct ideas, skills, approaches and perspectives—and a social-psychological component—individuals have to be willing to interact and share their thoughts. Both elements are required since neither a population of diverse minds that never interact nor a group of cognitive clones who freely interact but all share the same mentality will generate recombinations.

Our surname-based measure of social structure—surname entropy—influences both the informational and social-psychological components of recombination innovation. To obtain our surname measure, we collect all surnames reported in the full-count U.S. Census data from 1850 to 1940 and compute the diversity of surnames 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, we will show that they provide a reasonable approximation in the pre-1950 historical context.

We connect our measure of social structure to the two components of recombinative innovation in several ways. To start, focusing on the informational component, we demonstrate that surnames cluster with occupations, ancestry, and patent technology codes. This offers *prima facie* evidence that surnames capture an important slice of the informational component necessary for the kind of heterogeneous social interactions thought to foster innovation, as counties with a greater diversity of surnames will also have greater heterogeneity in occupations, expertise, cultural knowledge and perspectives. As we explain below, there's good reason to suspect that surnames capture additional, unobservable (in this context) dimensions of informational diversity that can further fuel innovation.

Of course, even if people vary in their skills, expertise, cultural backgrounds and perspectives, they may still refuse to interact and offer their ideas, thereby inhibiting such informational diversity from cashing out in inventive recombinations (Bursztyn et al., 2024). To examine this, we show that our measure of social structure is highly negatively correlated with both a newly developed residential segregation of surnames (e.g., Greens tend to live next to other Greens) within counties and with the strength of family ties (Alesina and Giuliano, 2011; Raz, 2023; Logan and Parman, 2017). The former demonstrates that our proxy captures the actual physical proximity of people with different surnames (who we show are informationally homogeneous) while the latter captures people's tendency to build strongly homophilic networks with relatives. As we'll highlight below, a large literature already connects such intensive or tight kin networks to lower trust in strangers and greater moral parochialism (Enke, 2019; Schulz et al., 2019; Henrich, 2020; Alesina and Giuliano, 2014). Our approach confirms empirically that

informational heterogeneity tends to accompany the formation of broader, more diverse social connections.

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 based on their textual similarity to both previous and subsequent patents. Breakthrough patents have low similarity to previous patents but high similarity to subsequent ones.

To estimate the effect of social structure, measured by surname entropy, on innovation, we employ an instrumental variable (IV) strategy, building on and adapting the approach developed by Burchardi et al. (2019). This strategy leverages historical immigration patterns as a significant determinant of surname entropy in U.S. counties.

After births, deaths and marriages, migration represents the next most important driver of counties' surname composition. Crucially, immigration does not monotonically increase surname entropy. Its impact depends on the preexisting surname distribution in a county. In other words, the arrival of the same set of immigrants (e.g., lots of 'Notaros') can increase surname entropy in some counties while decreasing it in others (e.g., depending on whether 'Notaros' are initially relatively rare or common). We hypothesize that this relationship between immigration and surname heterogeneity affects both the informational and social channels (detailed in section 2). When individuals carrying *locally* rare surnames arrive, they enhance surname heterogeneity, and in turn, may create opportunities for diverse social interactions, knowledge acquisition, and the cultivation of trust towards individuals with differing cultural or family backgrounds. On the other hand, an inflow of individuals bearing locally common surnames decreases our measure of surname heterogeneity. 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 orientation 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 plausibly exogenous variation in surname distributions across counties arising from historical shocks to migrations that date 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 entropy raises patents and breakthrough patents (per 1,000 people) by 132-144% and 90-129% relative to their sample means, respectively.

These results hold across a broad battery of robustness checks. Here, we list the most important of our checks:

1. To scrutinize the potential for reverse causality, i.e., an increase in a county's innovation leading to increased surname entropy, we perform a falsification exercise by regressing past patents on future surname entropy. 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.

2. 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.

3. To confront the concern that a direct effect of immigration, which is not channeled through social structure, confounds our estimates, we control for exogenous variation in the number of recent immigrants, which we construct following the approach pioneered by Burchardi et al. (2021). We find that controlling for immigration has minimal impact on the estimates for surname entropy, reinforcing the interpretation that it is social structure rather than immigration per se that is driving our results.

4. To consider the possibility that our results are region-specific, especially in light of regional variation in factors such as racial discrimination (Cook, 2014; Coluccia et al., 2023), we re-estimate our analyses in each of the United States' major census regions: the Northeast, Midwest, South and West. We find consistently positive coefficients.

5. 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 the social structure 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 stable and highly significant across all specifications.

6. To consider the impact that formal schooling might have on our results, we conducted a set of OLS regressions in the 1940 cross-section, adding the data on years of formal

schooling across counties (formal schooling measures are not available before this). The results reveal that adding formal schooling to the regression has little impact on the size or significance of the coefficient for our primary explanatory variable, surname entropy. Of course, more formal schooling is indeed associated with more innovation, on both our measures. Interestingly, education interacts powerfully with surname entropy, suggesting that the impact of education depends on the social structure, with a more diverse social structure potentially amplifying the impact of formal schooling.

Finally, we conclude our analysis by substantiating our proposed mechanism—that social structure influences heterogeneous social interactions in ways that catalyze the recombination of ideas, technologies and techniques.

1. Using our IV setup, we show that greater surname entropy leads to less residential segregation of surnames, weaker family ties and greater occupational diversity. This shows how social structure fosters both more heterogeneous social interactions (based on segregation and family ties) and greater informational diversity (occupations).

2. Similarly, we confirm that greater surname entropy at the county-level results in greater surname entropy at the patent-level and a greater number of technology codes per patent. The former demonstrates how social structure fosters more heterogeneous interactions at the level of inventions, while the latter suggests that social structures foster more complex recombinations. We supplement these IV analyses with OLS regressions that link greater *patent-level* surname entropy to both more technology codes per patent and an increase in breakthrough patents. These analyses support the view that social interactions among diverse inventors encourage inventions that are both more complex and impactful.

3. Using our IV setup, we confirm that the effects of surname entropy on innovation are geographically localized, largely impacting the focal county but not spilling over into nearby counties. This observation is consistent innovation being driven by people meeting day-to-day and interacting face-to-face.

4. To assess the relative importance of the informational vs. the social components ensnared in our main explanatory variable, surname entropy, we run a battery of OLS regressions controlling for the number of distinct surnames that go into our entropy calculations. We find that both components are associated with social structure's impact on innovation.

Taken together, our results indicate that social structure, captured using surnames within counties, contributes to explaining the patterns of innovation captured by the U.S. patents from the mid-19th to the mid-20th century. These patterns support the hypothesis that social interactions among heterogeneous individuals foster innovation through recombination.

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). In this section, focusing on those lines of research most closely linked to our current efforts, we discuss research that explores how innovation is influenced by urbanization, geography, kin networks, social proximity, diversity and immigration.

Several long-running lines of interrelated research have linked innovation to cities, population density, agglomeration and geographic proximity (Carlino and Kerr, 2015; Akcigit et al., 2017; Glaeser, 2011; Agrawal et al., 2008; Jacobs, 1985; Packalen and Bhattacharya, 2015; Feldman and Audretsch, 1999; Carlino et al., 2007). Research in this area emphasizes the impact 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 more diverse social interactions (Kerr, 2010, 2008; AlShebli et al., 2018). Our efforts extend these observations and insights more broadly—across the entire U.S. and back to the mid-19th century—while offering a viable approach to measuring social structure and heterogeneous social interactions across many contexts.

Our focus on social structure based on surnames naturally connects our enterprise with efforts to understand the impact of nuclear family ties, cousin marriage and kin-based institutions on economic outcomes. Both measures of family ties and kinship intensity or 'tightness' have been linked to key aspects of psychology (e.g., impersonal trust and moral universalism) that may influence interactions among strangers and offer the kind of psychological differences (e.g., in cognitive styles) that are conducive to creating novel recombinations (Henrich, 2020; Schulz et al., 2019; Alesina and Giuliano, 2014, 2015). Evidence also links family ties, kinship intensity and cousin marriage practices to income, economic prosperity and innovation (Bahrami-Rad et al., 2022; Alesina and Giuliano, 2010; Enke, 2019; Ghosh et al., 2023).

A related, but nascent, literature examines how particular social institutions and orga-

nizations, by facilitating heterogeneous social interactions, propel more rapid innovation. For example, the closure of saloons during Prohibition reduced patenting rates (Andrews, 2023), suggesting that environments and organizations conducive to social interactions among strangers or acquaintances may inadvertently spur innovation. Similar mechanisms operate today, as illustrated by evidence suggesting that the spread of coffee shops has caffeinated 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). Finally, 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 efforts also intersect 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 entropy 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 entropy (Barrai et al., 1996). Diverging from Ashraf and Galor (2013)'s global perspective on world history, we focus on a specific historical episode within the context of a single country: the United States from 1850 to 1940. This allows us to investigate regionally fine-grained changes in surname entropy over time in a panel setting. Using surname-fixed effects, we empirically establish that our results are not confounded by specific surnames, making genetic influences unlikely to play a substantial role. That is, when comparing people with the same surnames, those located in counties with more diverse social structures 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.[1]

[1]Beyond innovation, previous studies have also highlighted the positive effects of birthplace or country-

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 propelled by social interaction that arguably underlies much innovation and highlight supporting lines of evidence. Next, we conceptualize—and provide evidence—how social structure, captured by the entropy of surnames, is a major determinant of the type of social interactions that fuel innovation. Finally, we discuss how we compute surname entropy across counties from 1850 to 1940

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 entropy complements Buonanno and Vanin (2017) who, using it as a measure of social closure, focus on crime.

and how we use U.S. patents to measure innovation.

2.1 From Social Structure to Innovation

2.1.1 How Social Interactions Drive Innovation

In 1933, marking an important step on the road to modern radar, three men in Washington D.C.—Taylor, 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). We propose that recombinative innovation driven by social interactions involves two crucial dimensions: informational and social-psychological. Informationally, it requires diversity in a population's skills, knowledge, and perspectives to foster novel idea combinations. Social-psychologically, individuals must engage and share ideas; without such exchanges, even a diverse group cannot innovate. Therefore, conceptually, the ability of local populations to innovate should hinge on the diversity of social interactions—the extent to which diverse individuals freely exchange ideas—and social structures that support this diversity.

The plausibility of this hypothesis is reinforced by three distinct strands of research. First, a significant body of work posits a central role of recombination in innovation. Second, 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. Acemoglu et al. (2016) analyze the connections among 1.8 million U.S. patents, showing how the production of new patents depends on progress in other associated technological areas. 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.[2]

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.[3] 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 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.

These considerations highlight the importance of social interactional diversity for recombinative innovation. We now detail how the social structure created by surname groups determines key aspects of the social interactional diversity of local populations.

[2]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.

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

2.1.2 Surname Structure Shapes Social Interactional Diversity

Individuals sharing the same surname are more likely to be related, either through their genealogy or broader cultural ties. Thus, the distribution of surnames may be well-suited to capture key aspects of social structure relevant to both the informational and social-psychological dimensions of social interactional diversity.

Informationally, we propose—and provide evidence in Section 2.5.1—that surnames, typically patrilineally inherited in the U.S., serve as markers of different kinds of knowledge, skills and perspectives. These attributes originate from learning within familial, regional, or professional networks and persist across generations because of intergenerational cultural transmission (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Bisin and Verdier, 1998). This hypothesis is consistent with recent findings in the social mobility literature, which demonstrate that surnames capture unique skills, socialization, and know-how (Clark, 2014; Güell et al., 2015; Bell et al., 2019; Barone and Mocetti, 2021). Although very frequent surnames may capture family- or profession-specific traditions to a lesser extent, they nevertheless still encapsulate unique knowledge. For example, "Smith" and "Johnson", the most frequent surnames in the 1880 census, show a distinct pattern. The Smiths outnumber the Johnsons by a factor of about 1.65 times. Yet, among individuals who reported blacksmith as their occupation, there are 2.46 times more Smiths than Johnsons. This correlation suggests the surname "Smith" has a long-standing association with metalworking and blacksmithing, a trend that persists in 1880, showing it to be a relatively more common surname among metalworkers.

Socially and psychologically, we hypothesize—and test empirically in Section 6—that a higher concentration of surnames indicates the presence of stronger family or culturally homogeneous networks, and a greater reluctance to interact and connect with outsiders. The family is perhaps the most fundamental of human institutions and a key source of social structure, shaping the socialization process, the transmission of information, and even individuals' psychologies (Henrich, 2020). Previous work on the impact of kinship on sociality has shown that weaker kinship ties are associated with increased openness towards strangers, evidenced by higher market integration, division of labor, impersonal trust, cooperation, and lower nepotism (Enke, 2019; Alesina and Giuliano, 2014; Schulz et al., 2019; Bahrami-Rad et al., 2022).

2.2 Operationalizing Social Structure with Surnames

We capture the social structure represented by surnames using an entropy measure. To motivate this measure and conceptually link it to recombinative innovations, consider a

subpopulation consisting of N individuals partitioned into K surname groups, each of size N_k such that $\sum_{k=1}^{k=K} N_k = N$. Each group carries unique information (e.g., skills, know-how, metaphors), labeled as $s_k \in \{a, b, c, d, ...\}$ and $s_k \neq s_h$ for all $k \neq h$. When individuals from different surname groups meet, the likelihood of recombinative innovation increases. Information theory (Shannon, 1948) tells us that the average informational content (or the innovation potential) of such a population in which people randomly meet is

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

where $p_k = \frac{n_k}{N}$ is the probability that a person with group affiliation k is drawn and $\log_2 p_k$ is the informational content embedded in this individual (expressed in bits). 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 entirely predictable, resulting in an entropy of 0, and thus, no recombinations can arise through social interaction. In contrast, entropy for a population with a fixed number of groups is maximized if all groups are of equal size. In such a population, individuals engaging in a random social interactions are 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.[4]

From an informational perspective, the link between surname entropy and recombinative innovation is evident. Surname entropy reveals the innovative potential of local populations, provided that people interact randomly and surnames are indicative of distinct knowledge. In Section 2.5.1, we provide evidence that surnames in U.S. history indeed capture distinct information. However, this perspective overlooks the role of social-psychological factors: people do not interact randomly, and the social structure may

[4]In economics, a Herfindahl measure, which population geneticists call Isonymy, 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 ($\rho = 0.83$, Table B1).

affect individuals' readiness to engage with others. In the mechanisms Section 6, we report findings that highlight the role of these social-psychological factors. Our findings suggest that an increase in surname entropy leads to weaker family ties and a greater openness to interact with those outside one's surname group.

2.3 U.S. Surname Entropy 1850-1940

To calculate surname entropy, we draw on 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.

To address potential biases from misspellings and Anglicization of surnames, we apply the *metaphone* phonetic algorithm, as outlined by Philips (1990), to standardize name strings. This method reduces the risk that our analysis of the impact of surname entropy on innovation is biased by such variations. For instance, the algorithm treats phonetically similar names—like "Heinrich" and its Anglicized form "Henrich"—as identical ("HNRX"), thus countering concerns that discrimination-induced name changes in less innovative regions could spuriously suggest a link between low surname entropy and lack of innovation.

We implement the Philips (1990) phonetic algorithm, *metaphone*, to deal with misspellings in the name string. This transformation also mitigates concerns that anglicization of surnames might affect our results. For example, one might be concerned that discrimination against certain groups in certain parts of the country sometimes led people to anglicize their last names. If that happened in places that were less open (and hence probably less innovative), then this might result in a positive association between surname entropy and innovation. The *metaphone* algorithm mitigates this concern for similarly sounding name changes. For example, one of the authors' ancestors were named "Heinrich" and changed their names at some point to "Henrich". The *metaphone* algorithm groups both names to "HNRX".

Following Burchardi et al. (2021), we also obtain the variables `age` and `yrimmig` (the year of immigration) to estimate surname entropy for the mid-decades 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the mid-decade.

Figure B1 maps surname entropy across U.S. counties in the year 1940. It shows both the raw data and the residual variation that is orthogonal to log population size. [5] Clear geographical patterns emerge. While counties in California and most of the Northeast

[5] Figure B2 displays the variation in surname entropy conditional on log county population from 1850 to 1930.

score high on surname entropy, Utah and the Southern states score substantially lower, i.e., they are more homogeneous with regard to surnames.

2.4 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.

2.5 Surnames Capture Both Distinct Information and Social Behavior

We now empirically establish that our measure of surname entropy captures both the informational dimension—i.e., surnames are indicative of occupation and ancestral origins—and the social-psychological aspect—i.e., surname entropy correlates with residential

segregation of surnames and the strength of family ties. Our endeavor here is not to establish causal linkages—we will do this in the mechanism Section 6—but to reveal the kinds of empirical relationships one would expect if our measure of social structure indeed captures both captures a population's diversity of social interactions, i.e., the degree that exchange of ideas among diverse people occurs.

2.5.1 Distinct Information

To gain a more systematic sense of the degree to which surnames reflect unique knowledge in the U.S., we calculate 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)[6], and patent technology categories (columns 8 and 9).[7] For example, in 1880, two people with the same surname 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

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

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

Table 1: Surnames cluster in occupations, birthplaces, and patent technology categories

	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. We use the 31 subnational regional origins for German immigrants recorded by the Census (the variable `bp1d` with codes 45301 to 45361). Column 8 (9) includes all inventors of patents issued from 1880 to 1889 (1940 to 1949). We use the main technology class on the patent.

inner-German region (out of 31 regions). Finally, columns 8 and 9 show that surnames cluster on the fine-grained technology categories on patents.

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).[8]

[8] 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 entropy 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 entropy indicator that further differentiates along country of birth. Table B1 shows that the main surname entropy 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 entropy indicator and indicators that are based on (i) phonetically uncorrected surnames, (ii) surnames of men only, (iii) surnames of household heads only, and (iv) surnames of whites only.

2.5.2 Social Behavior

Moving now to focus on the social-psychological component of social structure, we analyze the correlation between surname entropy and (i) the residential segregation of surnames within countries as well as (ii) the strength of family ties in 1940.

The first measure captures the residential segregation of same-surname groups, importantly reflecting segregation beyond what would be expected from random choices of residence. We construct a measure of next-door neighbor segregation closely following the methodology of Logan and Parman (2017), except that we focus on surname-based segregation rather than racial segregation. This segregation measure's construction involves two steps: first, we create a measure for each surname in each county during each time period to quantify its degree of segregation. Second, we calculate the average degree of surname segregation at the county level by averaging across all surname-specific measures (see the data Appendix A for details).

Crucially, this indicator is constructed to account for the surname distribution at the county level. Positive values signify segregation beyond what would occur by chance, while negative values imply that neighbors are more diverse than what chance alone would predict, indicating a *preference for diversity*. Assessing the degree of segregation in relation to a county's surname pattern against that expected by chance allows us to capture people's settlement preferences. Thus, an association between surname entropy and segregation is not merely driven by more diverse counties being also less segregated. Rather, it reflects differences in fine-grained settlement preferences while holding background diversity constant.

Appendix Figure B3 illustrates the geographic distribution of surname-based residential segregation in 1940. As depicted in Figure 1, this variation is highly negatively correlated with surname entropy ($\rho < -0.68$), indicating that individuals with the same surname in counties with low surname entropy tend to live in closer geographic proximity.

The second measure, the strength of family ties (Raz, 2023), captures the size, homogeneity, and stability of households as social units. It represents the first principal component of four underlying county-level time-period 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. Higher values for the strength of family ties indicate a preference for surrounding oneself with relatives (see Appendix A for all details on how we compute this metric).

Appendix Figure B4 illustrates the geographic distribution of the strength of family ties in 1940, and Figure 1 demonstrates that this variation is highly negatively correlated with surname entropy, particularly when controlling for log population size ($\rho \approx -0.70$).

This suggests that individuals in counties with low surname entropy tend to have strongly homophilic networks with relatives.

3 Empirical Strategy

To estimate the causal effect of surname entropy on innovation, we examine the following equation:

$$Y_i^t = \beta \text{ Surname entropy}_i^t + \alpha_{s(i),t} + \alpha_i + \varepsilon_i^t \tag{2}$$

where Surname entropy$_i^t$ denotes the surname entropy in county i in period t (years: 1900, 1905, 1910, 1915, 1920, 1925, 1930, 1940). Y_i^t represents the outcome of interest, typically the number of patents or breakthrough patents filed in county i in the five-year period starting in year t, normalized by the county's population in 1900. The coefficient β is our main interest. $\alpha_{s(i),t}$ and α_i are state-period and county fixed effects, respectively. These fixed effects enable the estimation of β from changes within the same county over time, while controlling for both persistent and time-varying differences across states. The error term is denoted by ε_i^t. In our most restrictive model specifications, we include county-specific linear time trends, represented by $\alpha_i \times t$. These trends control for any inherent county-specific trend in Y_i^t, allowing us to utilize only the variation in the growth rate of patenting over time within each county.

The main concern with the OLS estimate of β is that reverse causality or unobserved factors that co-determine surname entropy and innovation might induce a spurious correlation. For instance, a highly innovative county may attract a more diverse set of migrants, which increases surname entropy. Similarly, highly-skilled individuals, more prone to innovation, might prefer more diverse counties. In both instances, a correlation between surname entropy and innovation could be observed even if no causal relationship exists.

We observe that migration, along with births, deaths and marriages, plays a significant role in changing counties' surname entropy. This observation allows us to isolate variations in surname entropy that are independent of any determinants of innovation, by leveraging methodological advances in immigration studies. We adapt the instrumental variable (IV) approach of Burchardi et al. (2019) to isolate quasi-random variation in surname stocks across counties, and use these stocks to compute an instrument for surname entropy. This approach enables us to estimate the local average treatment effect (LATE) of changes in surname entropy on innovation, specifically those induced by immigration to the U.S., as

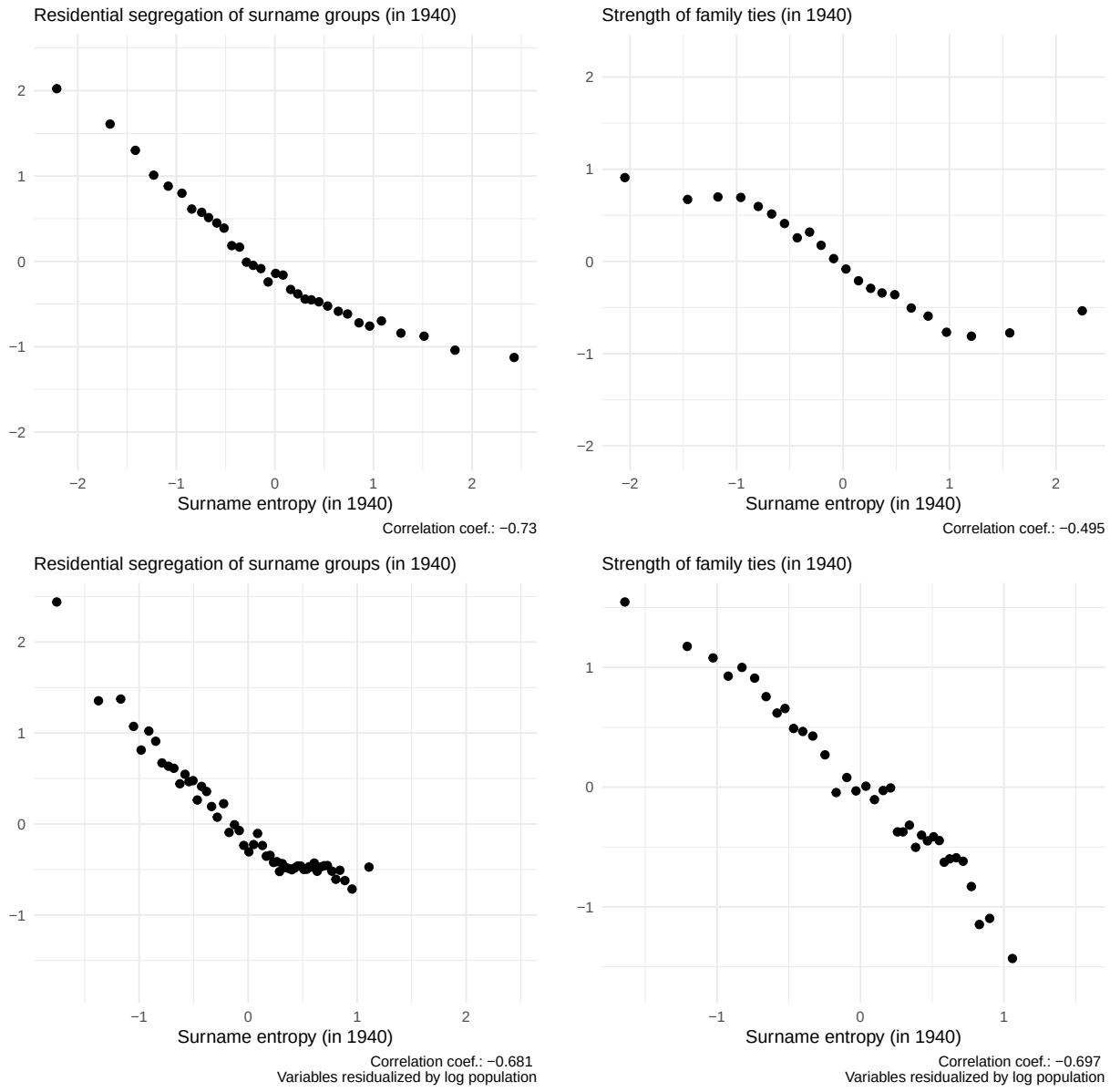


Figure 1: Relationships between surname entropy and social behaviors

Notes: The figures show binscatter plots of the relationships between surname entropy and surname residential segregation (left plots) and strength of family ties (right plots). Top: Raw data; bottom: residualized by log county population. An observation is a county in 1940. The segregation measure is constructed adapting the Logan and Parman (2017) procedure to surnames. The segregation measure is constructed following Raz (2023). The sources and construction of all variables are explained in Appendix Section A. Binscatter plot created using the R package written by Cattaneo et al. (2019).

opposed to changes stemming from births, deaths, marriage, or domestic migration.

It is important to note that our IV strategy, which relies on immigration-induced changes in surname composition, does not imply a simple, monotonically increasing relationship between immigration and surname entropy. Immigration can both decrease and increase surname entropy, depending on the existing local surname distribution. For example, if the 'Smiths' immigrate to a county with few or no 'Smiths', surname entropy increases. Conversely, if they move to a county where the 'Smiths' are already common, entropy decreases. Consistent with this distinction, our estimates remain almost unchanged when controlling for immigration (see Section 5.3). Essentially, our strategy capitalizes on how immigration affects surname composition, using quasi-random variation in immigration as an instrument for surname entropy (see Section 3.3 for identification details).

We hypothesize that the relationship between immigration and surname entropy influences both the informational and social-psychological dimensions of diverse interactions. The arrival of individuals with locally rare surnames increases surname entropy, potentially fostering opportunities for diverse social interactions, knowledge acquisition, and trust-building with people from different cultural and familial backgrounds. This may occur because a rise in surname entropy not only limits individuals' ability to fulfill their needs within their own (shrinking) group but also opens up avenues for valuable interactions outside their immediate circle. Consistent with contact theory (e.g., Allport, 1954; Bursztyn et al., 2024), we expect that increased mutually beneficial engagement with outsiders will foster greater openness to associate with strangers.

Conversely, an influx of individuals with locally common surnames, who are culturally and genealogically related to the dominant groups within counties, reduces surname entropy. This reduction potentially reinforces knowledge within culturally homogeneous groups, strengthens intra-group ties, and fosters a low-trust mentality towards outsiders. This hypothesis aligns with previous work showing that high fragmentation fosters trust, whereas polarization creates intergroup antagonism (Bazzi et al., 2019). We will examine this hypothesis in more detail in the mechanisms section 6.

3.1 Construction of the Instrument

The construction of the instrument requires two steps. First, we isolate quasi-random variation in the stock $N_{k,i}^{t}$ of each surname k residing in county i in period t. This isolation is based on specific historical migration patterns, which influence the distribution of surnames across counties. Second, we compute the instrument for surname entropy by

calculating entropy based on these (predicted) quasi-random stocks of surnames, denoted as $\hat{N}^t_{k,i}$. 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. This approach posits that a combination of push factors (such as economic or political conditions in the immigrants' origin countries) and pull factors (like economic opportunities in the destination counties) jointly determines the allocation of immigrants with specific surnames to counties. The historical interactions of these two factors arguably create quasi-random variation in surname stocks that persists over time.

Empirically, the push factor is represented by the total number of immigrants with a particular surname entering the U.S. during a specific period. The pull factor, meanwhile, is indicated by the attractiveness of a county in the same period, operationalized as the proportion of immigrants choosing to settle in that county out of all immigrants entering the U.S. These factors, and their interaction, have varied over time, and we can trace their impact back to 1880, leading to quasi-random variation in a county's surname distribution.

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

$$N^t_{k,i} = \delta_i + \delta_{k,r(i)} + \sum_{\tau=1880}^{t-1} b^\tau \underbrace{I^\tau_{k,-r(i)}}_{\text{Push}} \underbrace{\frac{I^\tau_{i,-k}}{I^\tau_{-k}}}_{\text{Pull}} + \sum_{\tau=1880}^{t-1} d^\tau \frac{I^\tau_{i,-k}}{I^\tau_{-k}} + u_{i,k}, \tag{3}$$

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 $I^\tau_{k,-r(i)}$ 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 $\frac{I^\tau_{i,-k}}{I^\tau_{-k}}$, where $I^\tau_{i,-k}$ is the total number of migrants who settle in county i during this period and who do not have surname k, and $I^\tau_{-k} = \sum_i I^\tau_{i,-k}$ is the total number of migrants who settled in the U.S. during the same period and who do not have surname k.[9]

[9]We follow Burchardi et al. (2019) and estimate equation (3) 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

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 (3) 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 (3) also includes the term $\sum_{\tau=1880}^{t-1} d^\tau \frac{I^\tau_{i,-k}}{I^\tau_{-k}}$, 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 (3) we estimate the coefficients $\hat{b}^\tau$ for each period τ and then 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 (3), 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 (3). 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 Entropy In step 2, we compute the instrument for surname entropy by applying the entropy formula on the predicted stock

$r(i)$. 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.

of each surname $\hat{N}^t_{k,i}$:

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

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.

3.2 IV Estimating Equations

We implement our IV procedure using 2SLS. The second-stage equation is given by equation (2). The first-stage equation is given by:

$$\text{Surname entropy}^t_i = \gamma \, \widehat{\text{Surname entropy}}^t_i + \mu_{s(i),t} + \mu_i + v^t_i \tag{4}$$

where i indexes counties, s states, and t periods. $\text{Surname entropy}^t_i$ is county i's surname entropy in t; and $\widehat{\text{Surname entropy}}^t_i$ is county i's predicted surname entropy in t, as described above. State-period fixed effects are denoted by $\mu_{s(i),t}$, and μ_i are county fixed effects.

In addition, our most demanding specifications include county-specific linear time trends, such that β in equation (2) captures the relationship between deviations in the changes in surname entropy and innovation within counties over time relative to their overall trend. Comparing the estimates of these specifications to the baseline estimates 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.

3.3 Identification

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

$$\left(I^\tau_{k,-r(i)} \frac{I^\tau_{i,-k}}{I^\tau_{-k}} \right)^\perp \perp \varepsilon^t_i.$$

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 surname entropy (Step 2). Here we detail threats to identification and how we address them.

Reverse causality. 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. 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.

Burchardi et al. (2019) argue that substantial variation in push-pull factors over time and space makes this unlikely. Focusing on surname groups rather than country of origin, this concern is further reduced, as idiosyncratic factors are likely even more important at this finer level of aggregation. 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 (3)). Consequently, the IV estimates do not reflect unobserved persistent factors that had already manifested themselves in immigrants' past settlement decisions. Additionally, in robustness checks, we control for the number of recent immigrants, ensuring that our estimates do not capture counties' current attractiveness as a destination for immigrants (see Section 5.3). Second, in our preferred specification, we control for county-specific linear time trends. To the degree that these linear time-trends capture county-specific 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 entropy. We do not find any evidence for reverse causality, i.e., a shock to surname entropy is statistically unrelated to previous-period innovation (see Section 5.1). Therefore, it is unlikely that our estimates are driven by persistent unobserved confounders.

Settlement preferences. 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 surname entropy and innovation would be due to the settlement preferences of individuals with high innovative capacity and not due to surname entropy 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 immigrants to counties does not

rest on individual preferences.[10] 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 5.5, we further address this concern by devising a specification with surname-county-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.

Immigration. A major source of variation in surname entropy stems from immigration. This raises the concern that immigration confounds our estimates through channels other than surname entropy. We address this potential issue by orthogonalizing our instrument to counties' immigration history. As a sensitivity check, in Section 5.3, we directly control for the number of immigrants (applying an IV strategy to address the endogeneity in immigration). The results exhibit remarkable stability. Controlling for immigration is feasible because, even though we confirm that immigration influences the surname composition of counties, there is no straightforward, monotonically increasing relationship between immigration and surname entropy conceptually. The degree to which immigrants influence surname entropy is contingent upon the specific surname composition of immigrants compared to the local (county) population.

3.4 Zero-stage Estimates

We report the zero-stage estimates of equation (3) in Table B2. 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 instrument for surname entropy. In total, we estimate equation (3) 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. [11]

[10] 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 immigrants 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 $r(i)$.

[11] Qualitatively, our results parallel those of Burchardi et al. (2019), who estimate the push-pull factor at

Using the estimated models shown in Table B2, 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 entropy by applying the entropy formula to the predicted surname stocks.

The predicted values of surname entropy for each period from 1900 to 1940, net of county and state-period fixed effects, are depicted in Figure 2. The maps demonstrate that our instrument picks up substantial variation both over time and across counties.

4 Results

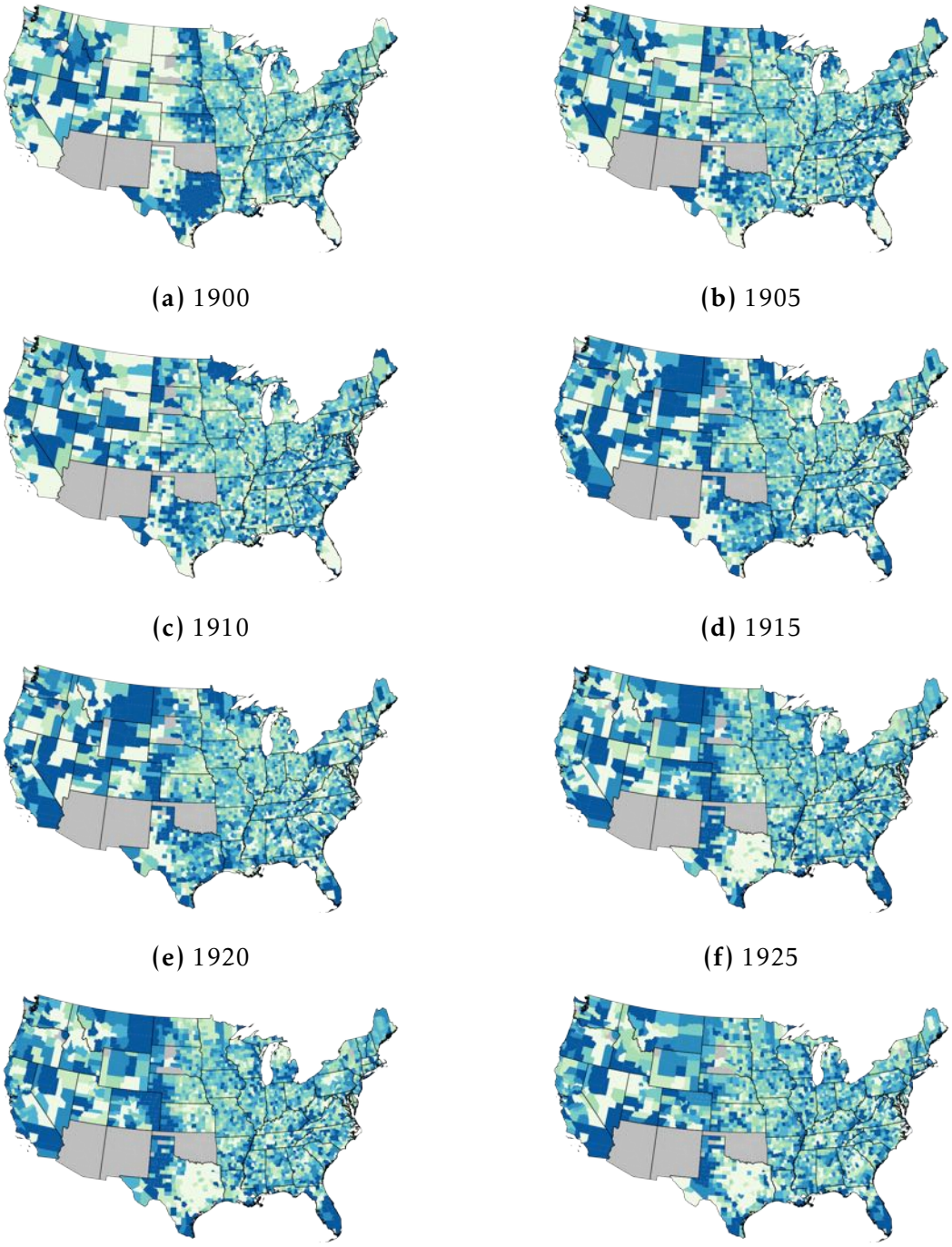
Table 2 reports OLS, reduced-form, second-stage and first-stage estimates. Starting with the first-stage estimates reported in Panel D, we find that the instrument is strongly correlated with actual surname entropy, with a Kleibergen-Paap F-statistic of around 104 in our baseline specification in columns 2 and 5. The F-statistic shrinks to roughly 92 when we add county-specific linear time trends (columns 3 and 6). Across specifications, the point estimates imply that a one standard deviation increase in the instrument is associated with roughly 0.65 to 0.67 standard deviation greater surname entropy. Taken together, the first-stage relationship of the instrument with surname entropy is highly significant, and the F-statistics of the excluded instrument in all specifications exceed conventional thresholds commonly used in the literature.

Figure B6 shows binscatter plots that show the first-stage relationship between the instrument and actual surname entropy, 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 entropy to innovation. Table 2 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 either period fixed effects (column 1 and 4) or state-period fixed effects (columns 2 to 3 and 5 to 6). In addition, specifications reported in columns 3 and 6 control for county-specific linear time trends. The reported standard errors are clustered at the state level.

The least-squares estimates reveal a significantly positive relationship between surname entropy and both patents and breakthrough patents. In columns 2 and 5, a one standard

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.



(a) 1900 **(b)** 1905

(c) 1910 **(d)** 1915

(e) 1920 **(f)** 1925

(g) 1930 **(h)** 1940

Figure 2: Predicted surname entropy (residuals)

Notes: This figure maps residualized instrumented surname entropy for each of the eight periods. We regress the instrument for surname entropy on county and state-year fixed effects, and calculate the residuals. This visualization depicts the instrument used in the regressions reported in Table 2. 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 2: Panel estimates of the effect of surname entropy on innovation

	Patents per 1,000 people (mean = 1.05, sd = 1.73)			Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname entropy	1.370***	1.303***	1.461***	0.157***	0.140***	0.117***
	(0.166)	(0.232)	(0.170)	(0.019)	(0.017)	(0.020)
Panel B: Reduced-form estimates						
Surname entropy (push-pull IV)	0.955***	0.921***	0.993***	0.113***	0.104***	0.077***
	(0.180)	(0.223)	(0.193)	(0.014)	(0.017)	(0.015)
Panel C: Instrumental-variable estimates						
Surname entropy	1.423***	1.388***	1.519***	0.168***	0.157***	0.118***
	(0.174)	(0.234)	(0.183)	(0.019)	(0.019)	(0.019)
Kleibergen-Paap *F*-statistic	119.613	104.522	91.652	119.613	104.522	91.652
Panel D: First-stage estimates			Surname entropy			
Surname entropy (push-pull IV)	0.671***	0.663***	0.654***	0.671***	0.663***	0.654***
	(0.061)	(0.065)	(0.068)	(0.061)	(0.065)	(0.068)
Within R^2	0.706	0.688	0.682	0.706	0.688	0.682
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	22,073	22,073	22,073	22,073	22,073	22,073

Notes: The table reports least-squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (2) and first-stage estimates for equation (4). An observation is a county in a period from 1900 to 1940. The endogenous variable is county-level surname entropy 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.

deviation increase in a county's entropy is associated with approximately 1.3 more patents (per 1,000 people), and about 0.14 more breakthrough patents.

To explore the stability of this relationship over time, we extend our analysis to earlier periods. In Appendix Figure B7, we present OLS estimates back to the earliest census wave for which surname data is available. In our main analysis, we cannot extend beyond the year 1900 due to the limitations of our push-pull IV approach, which relies on immigration data not available before 1880. Appendix Figure B7 displays our findings. The top row's plots present coefficients of regressions of (breakthrough) patents per capita in 1900 on surname entropy interacted with period dummies, controlling for county and state-period effects. We normalize by 1900 county population to be consistent with our main analysis. We find consistently positive, significant relationships between surname entropy and both innovation outcomes in the latter 19th century, with a decreasing trend in the relationship's magnitude over time, except for 1850, when few patents were issued. The bottom row repeats the analysis with (breakthrough) patents normalized by 1850 population as dependent variables. We again find positive and and almost always statistically significant, with a less pronounced decrease in the effect's magnitude over time.

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 entropy based on the zero stage). To visualize these relationships, Figure B8 shows partial correlation plots. Finally, Panel C presents the IV estimates. The coefficients for surname entropy 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 entropy increases the number of patents (per 1'000 inhabitants) by about 1.39 and breakthrough patents by about 0.16. The coefficients are similar when we additionally control for linear time trends (columns 3 and 6). This stability bolsters our confidence that the instrument for surname entropy is orthogonal to persistent or gradually growing county-level confounding factors.

Overall, the estimates indicate that surname entropy significantly boosts both innovation quantity and quality. The notable similarity between the least-squares estimates in Panel A and the IV estimates in Panel C could imply that the former are potentially unbiased, suggesting that surname entropy might be relatively exogenous within our empirical framework. This exogeneity could stem from the unique, granular nature of surname changes, which are often influenced by unique historical events at the individual surname level. Specifically, the complex and nonlinear effects of immigration on surname entropy, coupled with the pre-existing surname distribution, underscore this point. Given

that other factors affecting surname entropy, such as birth and death rates, change too gradually to significantly impact our results, it reinforces the idea that our variation source is highly detailed and specific. To examine the validity of these findings, we will now proceed with a series of sensitivity analyses and robustness checks.

5 Robustness and Sensitivity Checks

We test the robustness of our estimates using eight different approaches: (1) a placebo test that regresses past innovation on surname entropy; (2) models controlling for population size to address potential scale effects; (3) regressions with immigration controls to assess the role of immigrants in our estimates; (4) analysis of heterogeneity across the four major census regions; (5) the use of surname fixed effects to determine the impact of surname-specific characteristics; (6) controlling for years of schooling to explore whether the results are confounded by education; and (7) employing log-transformed dependent variables.

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 potentially may increase surname entropy (in case those migrants are sufficiently different from the existing population). 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 3.1 for details). Moreover, we have examined specifications that include county-specific linear time trends, which absorb the effects of trending unobserved factors associated with innovation and migration.

To further test the validity of our instrument, we conduct a placebo exercise to determine whether contemporaneous surname entropy affects past innovation activity. This analysis aims to identify any evidence of reverse causality, such as innovative counties attracting immigrants and thus increasing surname entropy. In Table 3, columns 1-2 and 6-7, we regress innovation measures from two periods prior ($t-2$) and one period prior ($t-1$) against current period (t) surname entropy. Columns 3 and 8 revisit our original specification, regressing innovation with same-period surname entropy (as reported in Table 2, column 5).

Our results, spanning least-square (Panel A), reduced-form (Panel B), and IV estimates (Panel C), consistently show no significant positive relationship between earlier innovation and subsequent surname entropy, both for patents (columns 1 and 2) and breakthrough

Table 3: Robustness I: Placebo test and persistence

	Patents per 1,000 people					Breakthrough patents per 1,000 people				
	$t-2$ (1)	$t-1$ (2)	t (3)	$t+1$ (4)	$t+2$ (5)	$t-2$ (6)	$t-1$ (7)	t (8)	$t+1$ (9)	$t+2$ (10)
Panel A: Least-squares estimates										
Surname entropy	-0.025	0.161	1.461***	0.806***	0.109	-0.011	-0.006	0.117***	0.137***	-0.006
	(0.380)	(0.129)	(0.170)	(0.119)	(0.214)	(0.048)	(0.023)	(0.020)	(0.024)	(0.031)
Panel B: Reduced-form estimates										
Surname entropy (push-pull IV)	-0.010	0.113	0.993***	0.517***	0.059	-0.012	-0.014	0.077***	0.103***	0.007
	(0.230)	(0.102)	(0.193)	(0.107)	(0.113)	(0.035)	(0.020)	(0.015)	(0.019)	(0.014)
Panel C: Instrumental-variable estimates										
Surname entropy	-0.016	0.197	1.519***	0.815***	0.092	-0.019	-0.025	0.118***	0.162***	0.010
	(0.381)	(0.191)	(0.183)	(0.134)	(0.183)	(0.056)	(0.033)	(0.019)	(0.026)	(0.022)
Observations	16,489	16,489	22,073	16,489	16,489	16,489	16,489	22,073	16,489	16,489
Kleibergen-Paap F-statistic	71.454	61.534	91.652	73.921	83.840	71.454	61.534	91.652	73.921	83.840
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear time trends	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Observations	16,489	16,489	22,073	16,489	16,489	16,489	16,489	22,073	16,489	16,489

Notes: The table reports least-squares, reduced-form, and instrumental-variable (IV) estimates of the leads and lags of innovation outcomes on surname entropy for the specifications described in equation (2). Columns 1–2 and 6–7 use the two-period and one-period lag of the dependent variables, respectively. Columns 3 and 8 repeat the baseline specification (contemporaneous values of the dependent variables). Columns 4–5 and 9–10 use the one-period and two-period lead of the dependent variables, respectively. An observation is a county in a period. 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.

patents (columns 6 and 7). However, when patenting in period t is regressed on surname entropy in the same period, the coefficients grow in size and become significantly positive (columns 3 and 8). This lack of evidence for reverse causality strengthens our confidence in the identification strategy.

Additionally, we explore the lasting effects of surname entropy on innovation by analyzing one-period (innovation in $t+1$, columns 4 and 9) and two-period leads ($t+2$, columns 5 and 10) in relation to surname entropy in period t. The results indicate a sustained impact of entropy on patenting in the subsequent period ($t+1$), but no significant relationship is found for the following period ($t+2$).

5.2 Sensitivity to Scale Effects

In our baseline analysis, we base our per capita dependent variables on the population size in 1900, following the approach in the literature (e.g., Burchardi et al., 2021). This choice accounts for the likelihood that population growth, potentially endogenous due to innovative regions attracting more people, could impact the results. However, this approach may not fully capture the scale effects associated with an increasing population

(Romer, 1990). To address this, we explore the robustness of our estimates by including population size in our specification.

The results of this analysis are reported in Appendix Table B3. To address endogeneity concerns, our reduced-form and IV specifications use predicted population, constructed similarly to our instrument for surname entropy. By leveraging the historical push-pull interactions from the zero stage analysis (as discussed in Section 3), we estimate the predicted surname stocks in each county for a given period. Aggregating these stocks provides us with quasi-random estimates of county populations at specific points in time.

Comparing these findings with our baseline results, as presented in Table 2, we find that our results are robust to controlling for predicted population. The estimates are very similar, reinforcing that our baseline findings are not driven by scale effects.

5.3 Sensitivity to Immigration

Another concern is the potential confounding effect of immigration on our estimates, beyond its impact on surname entropy. Previous research indicates that immigrants can significantly fuel innovation due to factors like higher skills, entrepreneurial spirit, or unique patentable knowledge (Moser et al., 2014; Abramitzky and Boustan, 2017; Sequeira et al., 2020; Burchardi et al., 2021). However, our analysis addresses this concern by using an instrument that is orthogonal to counties' historical immigration patterns, ensuring that our estimates are not biased by past immigration trends (details on the instrument's construction are in Section 3.1). Additionally, our IV estimates are robust to controlling for changes in population due to migration (see Section 5.2).

In our further analysis, we control directly for the number of recent immigrants. While our instrument is based on changes in surname entropy induced by recent immigration, we can separate the effects of recent immigration from those of surname entropy, as they vary independently. The change in entropy due to immigration also depends on the existing surname distribution in a region (see Section 3).

To address the endogeneity of immigration, we adopt a shift-share approach, which utilizes patterns where immigrants tend to settle near individuals from their country of origin (Altonji and Card, 1991; Card, 2001). Adapting this to our study, we predict the migration inflow for each surname into each county between t and $t-1$, based on the proportion of people with that surname already residing there in $t-1$. This method gives us an estimate of each surname's migration inflow into a county by the end of the t. A concern with this approach is that previous-period surname stocks are endogenous. Following Burchardi et al. (2021), we therefore use *predicted*, not actual, previous-period

surname stocks to mitigate endogeneity concerns. These predicted stocks are derived from the historical push-pull approach. We sum these predicted inflows to estimate total migration into a county, excluding data from the period ending in 1940 due to the absence of immigration year information, which is essential for calculating predicted immigration.

Table 4 presents the results. In line with previous work, we find a positive effect of immigration on innovation in all specification. We also find that controlling for immigration has minimal impact on the estimates for surname entropy. This holds true when actual immigration numbers are included in the least-square regressions (Panel A) and when predicted immigration is used in the reduced-form and IV regressions (Panel B and C, respectively). For example, the IV estimate for patents per 1,000 people in column 1 of Panel C is 1.38 without the immigration control, compared to 1.28 with the immigration control in column 2. The estimates for breakthrough patents are similarly consistent, being 0.14 and 0.12 in columns 5 and 6 of Panel C, without and with the immigration control, respectively. These findings suggest that the influence of immigration on our measures of innovation is not significantly confounded by factors other than surname entropy.

5.4 Estimates for Major U.S. Regions

To further understand how the relationship between surname entropy and innovation varies across U.S. regions, we analyze data by region—specifically, the Midwest, Northeast, South, and West. This analysis, presented in Table B4, is also insightful for examining the influence of immigration. Notably, the South experienced relatively low immigration during this period. Therefore, finding similar effects in the South would provide additional evidence that our results reflect broader dynamics beyond just immigration patterns.

Across all OLS, reduced-form and IV specifications, we observe that the region-specific estimates are uniformly positive and statistically significant. The range of IV coefficients for patents (Panel C, column 3) spans from 1.09 to 3.8, and for breakthrough patents (column 6), from 0.13 to 0.38. The larger coefficients for the West are particularly interesting, possibly reflecting the region's unique historical and immigration context. However, the precision of these estimates is not sufficient to conclusively assert regional differences.

5.5 Surname-level analysis

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

Table 4: Robustness II: Controlling for immigration

	Patents per 1,000 people (mean = 1.08, sd = 1.68)				Breakthrough patents per 1,000 people (mean = 0.12, sd = 0.27)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname entropy	1.350***	1.348***	1.291***	1.289***	0.136***	0.136***	0.119***	0.135***
	(0.125)	(0.125)	(0.172)	(0.191)	(0.021)	(0.021)	(0.017)	(0.019)
Immigration		0.012*	0.013**	0.004**		0.001	0.001	0.000
		(0.006)	(0.006)	(0.002)		(0.001)	(0.001)	(0.001)
Panel B: Reduced-form estimates								
Surname entropy (push-pull IV)	0.931***	0.857***	0.853***	0.818***	0.097***	0.081***	0.075***	0.079***
	(0.145)	(0.153)	(0.186)	(0.187)	(0.012)	(0.012)	(0.013)	(0.017)
Immigrants (shift-share IV)		0.331***	0.290***	0.140***		0.075***	0.058***	0.050***
		(0.090)	(0.080)	(0.051)		(0.014)	(0.010)	(0.011)
Panel C: Instrumental-variable estimates								
Surname entropy	1.377***	1.275***	1.272***	1.302***	0.144***	0.120***	0.112***	0.125***
	(0.143)	(0.151)	(0.191)	(0.199)	(0.020)	(0.018)	(0.016)	(0.020)
Immigrants (shift-share IV)		0.309***	0.286***	0.119***		0.073***	0.058***	0.048***
		(0.081)	(0.071)	(0.041)		(0.014)	(0.010)	(0.010)
Kleibergen-Paap *F*-statistic	130.836	120.095	112.618	78.022	130.836	120.095	112.618	78.022
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Period fixed effects	✓	✓			✓	✓		
State-Period fixed effects			✓	✓			✓	✓
County-specific linear time trends				✓				✓
Observations	19,324	19,324	19,324	19,324	19,324	19,324	19,324	19,324

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (2). Columns 2 – 4 and 6 – 8 control for the actual (Panel A) and predicted (Panel B and C) number of recent immigrants (between $t-1$ and t). Predicted number of immigrants is based on the shift-share approach described in section 5.3. An observation is a county in a period from 1900 to 1930. The endogenous variable is county-level surname entropy in t. In columns 1 to 3, the dependent variable is the number of patents filed in the county in the five-year period starting in t divided by county population size in 1900. In columns 4 to 6, the dependent variable is the number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900. Standard errors are clustered at the state level. 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.

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.

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

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

where i indexes counties, s states, t census years (including the midyears), and k surnames. As before, $\text{Surname entropy}_i^t$ is county i's surname entropy in t, and $\widehat{\text{Surname entropy}}_i^t$ is county i's predicted surname entropy in t. $Y_{i,k}^t$ is now the number of (breakthrough) patents filed by people with surname k residing in county i in the five-year period starting in t (in terms of 1,000 residents with the same surname in the year 1900). For example, 18,351 individuals with the surname 'Johnson' resided in Cook County (IL) in 1900 and filed about 69 patents and 1 breakthrough patent between 1900 and 1904. Therefore, while the surname entropy remains defined at the county-period level, the innovation outcomes vary at the surname-county-period level.[12] Crucially, this shift allows us to include county-surname fixed effects and surname-period fixed effects (denoted by $\alpha_{i,k}$ and $\alpha_{k,t}$, respectively), thus non-parametrically controlling for any traits specific to individuals with a particular surname in a given county or time period (i.e., traits specific to all 'Johnson' in Cook County or in 1940). The remaining parameters and variables are as in equations (5) and (6). As before, the coefficient of interest is β. Standard errors are clustered in two ways, on states and surnames.

The results reported in Table 5 show that the estimates are large, highly significant and comparable to those reported in Table 2 across all specifications, despite changing the unit of observation. Crucially, the surname fixed effects ensure that the estimates are independent of any unobserved surname-specific characteristics.

Using a similar approach, 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. As reported in Table B9 in Appendix Section B.1, we find that all results hold with this patent class fixed effects specification.

[12] Consequently, the number of observations increases because they are now determined by the total number of unique surnames in a given county. Consistent with the county-level 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: Robustness III: Surname fixed effects

	Patents per 1,000 people (mean = 1.17, sd = 31.37)			Breakthrough patents per 1,000 people (mean = 0.16, sd = 11.02)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname entropy	1.421***	1.443***	1.539***	0.222***	0.224***	0.095**
	(0.250)	(0.253)	(0.297)	(0.062)	(0.062)	(0.044)
Panel B: Reduced-form estimates						
Surname entropy (push-pull IV)	0.904***	0.913***	0.919***	0.152***	0.153***	0.084**
	(0.198)	(0.201)	(0.190)	(0.042)	(0.042)	(0.037)
Panel C: Instrumental-variable estimates						
Surname entropy	1.719***	1.738***	1.815***	0.290***	0.291***	0.166**
	(0.333)	(0.339)	(0.335)	(0.078)	(0.079)	(0.075)
Kleibergen-Paap *F*-statistic	172.726	173.776	129.807	172.726	173.776	129.807
County-Surname fixed effects	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓
Surname-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	28,236,803	28,236,803	28,236,803	28,236,803	28,236,803	28,236,803

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (6). An observation is a surname in a given county in a period from 1900 to 1940. Observations are weighted based on their population shares within counties. In columns 1 to 3, the dependent variable is number of patents filed by individuals with surname n residing in county in i in the five-year period starting in t divded by surname population size in 1900. In columns 4 to 6, the dependent variable is number of breakthrough patents filed by individuals with surname n residing in county in i in the five-year period starting in t divded by surname population size in 1900. Standard errors are clustered two-way clustered at the state and surname 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.

5.6 Education

We also explore whether formal education may explain our results. While the diversity of knowledge, customs, and traditions captured by surname entropy provides a rich source of ideas, transforming these ideas into patentable innovations may require educated individuals. For example, Alesina et al. (2016) and Burchardi et al. (2021) demonstrate that immigrants' skill levels are an important driver of innovations.

To gain insights into formal education, we estimate simple OLS specifications based on the cross-section of U.S counties in the year 1940, the only census wave in our sample that reports years of schooling. Appendix Table B5 presents the estimates. As a benchmark, we first regress our two dependent variables on surname entropy (columns 1 and 4). Then, in columns 2 and 5, we control for the average years of schooling in counties, and in columns 3 and 6, we examine the interaction between surname entropy and years of schooling.

Across all specifications, we find a significant association between surname entropy and patenting. Years of schooling are also related to patenting, however its inclusion

hardly affects the coefficients on surname entropy. Notably, the interaction between the two independent variables is highly significant. These findings support the idea that a more diverse social structure may fuel patenting, especially among more highly educated individuals.

5.7 Log-transformed Dependent Variables

Our findings are robust to using log-transformed innovation outcomes. Since right skewness is present in per-capita patents and breakthrough patents, we also take the log of these measures as our dependent variables. We add one to the (breakthrough) patents to avoid dropping counties no patents. The estimates using transformed outcomes are reported in Appendix Table B6.

Additionally, we use log patents and log breakthrough patents as dependent variables (not normalized by population). These estimates are reported in Appendix Table B7.

6 Mechanisms

Conceptually, our hypothesis posits that greater surname entropy leads to the production of both more patents and breakthrough patents. This is because innovation often emerges from the recombination of ideas, insights, and techniques that arise from social interactions among people with diverse skills, expertise, cultural backgrounds, and perspectives. If true, in addition to the causal link between surname entropy and innovation established above, surname entropy should also be associated with:

1. **More diverse social interactions at the county level**: Using our instrumental variable setup for the period 1900 to 1940, we show that greater surname entropy results in (1) less residential segregation of surnames (geographically more diverse mixing), (2) weaker family ties, and (3) greater occupational diversity. This establishes a relationship between our primary independent variable and three outcomes that should all contribute to more diverse social interaction.

2. **More diverse social interactions at the patent level, greater patent complexity, and more breakthrough patents**: We demonstrate that (1) county-level surname entropy increases surname entropy at the individual patent level (IV analysis), (2) greater patent-level surname entropy is associated with more breakthrough patents (OLS analysis), (3) greater county-level surname entropy leads to more technology

classes per patent (IV analysis), suggesting greater complexity, and (4) greater patent-level surname entropy is associated with more technology classes per patent (OLS analysis). The first result suggests that county-level heterogeneity actually results in more diverse interactions at the patent level. The latter three results support the idea that greater surname heterogeneity leads to greater innovativeness.

3. **Geographically localized effects on patenting, without much spillover to neighboring counties**. Using our IV setup, we demonstrate that greater surname entropy in a county shows little spillover to neighboring counties, suggesting the localized impacts of social interactions.

4. **Both the informational and social components of social structure play important roles**. Here, using OLS, we regress our innovation measures on both the number of distinct surnames in each county-period (logged), proxying for informational content, and on our surname entropy measure, normalized by dividing it by our new predictor (number of surnames, logged). The former aims to capture the informational content and the latter with social component. We find that both components play important roles in explaining innovation.

6.1 Local Diverse Social Interactions at County Level

In our main analysis, we establish a causal link from surname entropy to a pair of patent-based measures of innovation. We argue that the social structure, as captured by surname entropy, affects innovation by fostering social interactions among diverse individuals within counties. To more clearly establish this, we use our instrumental variable (IV) approach to demonstrate that surname entropy affects (1) residential segregation, (2) the strength of family ties, and (3) occupational diversity. The first measure captures the residential segregation of same-surname groups, importantly reflecting segregation beyond what would be expected from random choices of residence, thus allowing us to investigate how surname entropy impacts surname segregation beyond random assortment. The second measure, the strength of family ties, captures the size, homogeneity, and stability of households as social units. It represents the first principal component of four underlying county-level time-period 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. Higher values for the strength of family ties indicate a preference for surrounding oneself with relatives. Finally, our measure of occupational diversity is the entropy of occupational

codes recorded in the census data. Our analyses of these three outcomes suggest that places with greater surname entropy are those where diverse people, with different occupations, are more likely to live in physical proximity. Let's consider each of these results in turn.

6.1.1 Greater Surname Entropy, Less Residential Segregation

To explore the link between surname entropy and residential proximity, we construct a measure of next-door neighbor segregation closely following the methodology of Logan and Parman (2017), except that we focus on surname-based segregation rather than racial segregation. This segregation measure's construction involves two steps: first, we create a measure for each surname in each county during each time period to quantify its degree of segregation. Second, we calculate the average degree of surname segregation at the county level by averaging across all surname-specific measures (see the data Appendix A for details).

Crucially, this indicator is constructed to account for the surname distribution at the county level. Positive values signify segregation beyond what would occur by chance, while negative values imply that neighbors are more diverse than what chance alone would predict, indicating a *preference for diversity*. Assessing the degree of segregation in relation to a county's surname pattern against that expected by chance allows us to capture people's settlement preferences. Thus, an association between surname entropy and segregation is not merely driven by more diverse counties being also less segregated. Rather, it reflects differences in fine-grained settlement preferences while holding background diversity constant

Appendix Figure B3 illustrates the geographic distribution of surname-based residential segregation in 1940. As depicted in Appendix Figure 1, this variation is highly negatively correlated with surname entropy ($\rho < -0.68$), indicating that individuals with the same surname in counties with low surname entropy tend to live in closer geographic proximity.

In Section 2.5.2, we showed that surname entropy is strongly negatively associated with residential segregation and the strength of family ties. Here, we provide causal evidence highlighting the role of surname entropy in capturing social-psychological variation.

In our causal framework, we first estimate our baseline instrumental variable (IV) specification presented in equation (2), but with the dependent variable replaced by our surname residential segregation variable. The results are reported in columns 1 and 2 of Table 6. We present least-squares estimates in Panel A, reduced-form estimates in Panel B, and IV estimates in Panel C. Across all three panels, the coefficients on surname entropy indicate that greater surname entropy is associated with less residential segregation than

Table 6: Mechanism I: Effects on proxy measures of social interactional diversity

	Residential segregation of surname groups				Strength of family ties				Occupational entropy	
	County		Surname-county		County		Surname-county		County	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Least-squares estimates										
Surname entropy	-0.010***	-0.007***			-0.026	-0.213*			0.342***	0.338***
	(0.001)	(0.001)			(0.090)	(0.109)			(0.045)	(0.087)
Surname share of county population			0.036***	0.036***			0.250***	0.254***		
			(0.002)	(0.002)			(0.009)	(0.010)		
Panel B: Reduced-form estimates										
Surname entropy (push-pull IV)	-0.006***	-0.004***			-0.083*	-0.215***			0.217***	0.225***
	(0.001)	(0.001)			(0.046)	(0.057)			(0.034)	(0.054)
Surname share of county population (push-pull IV)			0.012***	0.005***			0.039***	0.034***		
			(0.003)	(0.002)			(0.005)	(0.006)		
Panel C: Instrumental-variable estimates										
Surname entropy	-0.011***	-0.007***			-0.144	-0.377***			0.372***	0.391***
	(0.001)	(0.002)			(0.089)	(0.112)			(0.053)	(0.086)
Surname share of county population			0.067***	0.037***			0.351***	0.361***		
			(0.015)	(0.011)			(0.047)	(0.039)		
Kleibergen-Paap *F*-statistic	115.660	92.023	68.272	124.486	110.294	97.720	60.847	78.628	104.602	91.689
County fixed effects	✓	✓			✓	✓			✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
County-Surname fixed effects			✓	✓			✓	✓		
Surname-Period fixed effects			✓	✓			✓	✓		
County-specific linear time trends		✓		✓		✓		✓		✓
Observations	13,812	13,812	10,578,595	10,578,595	22,056	22,056	10,450,814	10,450,814	22,070	22,070

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates. Columns 1-2, 5-6 and 9-10 report estimates for the specifications described in equation (2) with residential segregation of surname groups, the strength of family ties and occupational entropy as the outcome variables. An observation is a county in a period from 1900 to 1940. The residential segregation of surname groups variable is constructed following Logan and Parman (2017). The strength of family ties variable is constructed following Raz (2023). Standard errors are clustered at the state level. Columns 3-4 and 7-8 report estimates for the specifications described in equation (8). Observation are surnames within counties in from 1900 to 1940 and are weighted by their population shares within counties. Residential segregation of surname groups and strength of family ties are constructed at the surname-county level. Standard errors are two-way clustered on states and surnames and reported in parentheses. All variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

would be expected by chance alone. These findings not only hold when comparing counties over time within states (including both county and state-period fixed effects) but also remain robust to accounting for county-specific linear time trends (column 2). A one-standard-deviation increase in surname entropy is associated with an approximate 0.01 standard deviation decrease in residential segregation.

The structure of the data enables 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 residential segregation. To achieve this, we compute the population shares of surname groups within counties and their push-pull predicted counterparts. We refine our surname segregation measure to now specifically capture the degree of segregation among individuals in a county who share the same surname. We then analyze this within-surname residential segregation by regressing it on the instrumented shares of individual surnames within a county. Formally, we estimate the following equations, where equation (7) represents the first stage and equation (8) constitutes the second stage:

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

$$\text{Residential segregation}_{i,k}^{t} = \beta\, \text{Surname population share}_{i,k}^{t} + \alpha_{s(i),t} + \alpha_{i,k} + \varepsilon_{i,k}^{t} \quad (8)$$

where i indexes counties, k names, s states, and t census years (excluding the midyears, because we cannot compute residential segregation for those years; see Appendix A for details). Surname population share$_{i,k}^{t}$ is surname group k's population share in county i in t; and $\widehat{\text{Surname population share}}_{i,k}^{t}$ surname group k's push-pull predicted population share in i and t. Residential segregation$_{i,k}^{t}$ is surname group k's residential segregation in county i in t. Equations (7) and (8) also include state-period fixed effects, $\mu_{s(i),t}$ and $\alpha_{s(i),t}$, and surname-county fixed effects, $\mu_{i,k}$ and $\alpha_{i,k}$. These fixed effects allow us to focus on changes in counties' surname shares over time while controlling for persistent surname-county-specific and time-varying state-specific factors.

In Table 6, columns 3 and 4 report the results. The least-squares estimates in Panel A reveal a positive and highly significant relationship between the surname share of the county population and residential segregation. As the surname share increases, residential segregation of individuals with that surname also rises, beyond what would be expected based on the mere presence of more people with that surname under random assortment. In essence, greater dominance relative to the overall county population results in a disproportionate increase in residential segregation. Panels B and C, featuring reduced-form and IV estimates, corroborate this association and suggest a causal connection. The IV coefficients indicate that a shift in a surname share by one standard deviation increases residential segregation by approximately 0.04 standard deviations. These analyses suggest that as counties become more diverse, individuals become less inclined to live next to those who share their surnames—possibly reflecting a change in people's preferences

6.1.2 Greater Surname Entropy, Weaker Family Ties

Following the same approach used above for residential segregation, we also analyze the relationship between surname entropy and the strength of family ties (see Appendix A for all details on how we compute this metric). Appendix Figure B4 illustrates the geographic distribution of the strength of family ties in 1940, and Appendix Figure 1 demonstrates that this variation is highly negatively correlated with surname entropy, particularly when controlling for log population size ($\rho \approx -0.70$). This suggests that individuals in counties with low surname entropy tend to have strongly homophilic networks with relatives.

Again, alongside our county-level analysis, the data structure allows us to include specifications that directly focus on the impact of the relative size of surname (or family) groups within a county on the strength of family ties. To accomplish this, 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.

In Table 6, columns 5 and 6 deliver the results of the county-level analysis. Broadly, the coefficients reveal the expected negative relationship between surname entropy and the strength of family ties, though holding constant county-specific linear time trends increases the size and precision of these coefficients (resulting in conventional significance in two cases). For our strictest IV specification, a one-standard-deviation increase in surname entropy leads to a decrease in the strength of family ties of 0.38 of a standard deviation.

As in our analysis of residential segregation, columns 7 and 8 report the estimates from least-squares (Panel A), reduced-form (Panel B), and IV regressions (Panel C) of the strength of family ties on surname population shares within counties. The positive relationship between surname share and the strength of family ties at the surname-county level is robust across all six specifications, revealing the largest coefficients under our strictest IV specification, indicating that a one-standard-deviation increase in surname share results in a 0.36 standard deviation increase in the strength of family ties.

In sum, these results indicate that as surname entropy increases and surname shares within a county decrease, kinship ties weaken. This aligns with our hypothesis that smaller group sizes limit individuals' ability to meet their needs within their family, creating more opportunities for exchanges with unrelated individuals outside of family networks. As a result, greater diversity fosters engagement with non-kin individuals, potentially cultivating a culture of impersonal trust and promoting recombinant innovation.

6.1.3 Greater Surname Entropy, Greater Occupational Diversity

In Section 2.5.1, we established that surname entropy captures at least some important aspects of the informational heterogeneity of a county's population. Here, we provide additional evidence underscoring the role of surname entropy in capturing informational diversity. Specifically, our findings suggest that surname entropy influences the occupational composition within a county. We hypothesize that the heightened informational diversity in counties with high surname entropy will manifest in increased occupational entropy. As a population incorporates a broader range of knowledge, we anticipate a

corresponding increase in the diversity of occupations.

Table 6 reports the results of regressions examining the relationship between surname and occupational entropy. Panels A, B, and C all present positive, significant, and substantial coefficients. In our strictest IV specification, presented in column 10, the analysis reveals that a one-standard-deviation increase in surname entropy causes a 0.39 standard deviation increase in occupational entropy. Of course, as we have shown in Section 2.5.1, occupational diversity is only one aspect of information heterogeneity captured by surnames.

6.2 Diverse Social Interactions at the Patent Level

To further establish the mechanisms underlying our main result—the relationship between county-level surname entropy and patents per capita—we conduct four analyses of patent-level outcomes. First, using our instrumental variable (IV) setup, we demonstrate that greater surname entropy at the county level is associated with greater surname entropy at the patent level. This establishes that higher surname entropy in a county generates more diverse interactions at the patent level. Second, using a least-squares approach, we show that greater patent-level surname entropy is associated with more breakthrough patents. We have already demonstrated that higher surname entropy at the county level leads to more breakthrough patents per capita. Third, returning to our IV approach, we find that greater surname entropy at the county level predicts a higher number of technology classes per patent, suggesting greater complexity (Akcigit et al., 2013). Fourth, we document a correlation between patent-level surname entropy and the number of technology classes per patent. These latter three results support the view that diverse social interactions foster greater innovativeness. We now consider each set of analyses in turn.

In Table 7, columns 1 and 2 report the estimates from the second stage of IV regressions that use our instrument for county-level surname entropy and, as the dependent variable, surname entropy at the individual patent level. Our strictest specification, which includes fixed effects for each county, state-period, and each patent technology class-period, suggests that a one-standard deviation increase in county-level entropy results in an increase in surname entropy at the patent level of 0.06 standard deviations.

Next, columns 3 and 4 regress our measure of breakthrough patents on patent-level surname entropy using a least-squares approach. The coefficients indicate a positive relationship between patent-level surname entropy and the likelihood of a patent being classified as a breakthrough. In column 4, which includes controls for patent technology class fixed effects interacted with time fixed effects, a one-standard-deviation increase in

Table 7: Mechanism II: Patent-level results

	Surname entropy of patent		Breakthrough patent indicator (×100)		Tech classes per patent			
	IV		LS		IV		LS	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Surname entropy	0.104**	0.063*			0.204**	0.124*		
	(0.047)	(0.036)			(0.080)	(0.070)		
Surname entropy of patents			0.908***	0.165***			0.042***	0.021***
			(0.133)	(0.048)			(0.004)	(0.003)
Kleibergen-Paap *F*-statistic	11.453	11.717			11.453	11.717		
County fixed effects	✓	✓			✓	✓		
State-Period fixed effects	✓	✓			✓	✓		
Patent technology class-Period fixed effects		✓		✓		✓		✓
County-Period fixed effects			✓	✓			✓	✓
Observations	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459	1,451,459

Notes: An observation corresponds to a patent from 1900 to 1944. Columns 1-2 and 5-6 report IV estimates for a specification similar to equation (6), but with two patent-level dependent variables: surname entropy of the patent and the number of patent technology classes. Columns 3-4 and 7-8 report least-squares estimates with surname entropy of patents as the independent variable and two dependent variables: the breakthrough patent indicator multiplied by 100, and the number of technology classes per patent. Columns 2, 4, 6, and 8 additionally include patent technology class fixed effects interacted with period fixed effects. Observations are weighted by the inverse of the number of authors multiplied by the weight used to assign patents to counties within their 1900 borders. Standard errors are clustered two-ways by states and patent technology class and are reported in parentheses. All entropy variables are standardized to have a mean of zero and a unit variance. The means of the breakthrough patent indicator and the number of tech classes per patent are 17.12 and 2.49, respectively. The sources and construction of all variables are detailed in Appendix A. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

patent entropy is associated with a 0.17 standard deviation increase in the likelihood of a patent being a breakthrough.

To measure the complexity of each patent, we follow other authors (Strumsky et al., 2011; Akcigit et al., 2013; Fiszbein, 2022) by counting the number of technology classes assigned to each patent. There are 408 patent classes in our sample from 1900 to 1944. Examples are "Geometrical Instruments", "Stoves and Furnace", and "Chemistry: Electrical and Wave Energy". Then, using our IV setup in columns 5 and 6, we demonstrate that greater surname entropy at the county level leads to a higher number of technology classes per patent. The IV estimates suggest that a one-standard-deviation increase in surname entropy results in a 0.12-0.20 increase in the number of technologies per patent, signifying a 5-8% enhancement relative to the sample mean.

Finally, in columns 7 and 8, we conclude the patent-level analysis by regressing the number of technology classes per patent on patent-level surname entropy. The coefficients reveal a positive relationship. In column 8, a one-standard-deviation increase in patent entropy is associated with approximately a 0.02 standard deviation increase in the number of technologies per patent.

6.3 Localized social interactions

People acquire inspiration, knowledge and ideas from others they frequently observe and interact with in their daily lives. This suggests that physical proximity and local diversity will play a big role (Jaffe et al., 1993; Carlino and Kerr, 2015). To explore this, we study the impact of the county-level surname entropy in neighboring counties. If local social interactions are the most important, the impact of the social interactional diversity in neighboring counties should have small or negligible impacts. Thus, we compute surname entropy among individuals residing in surrounding regions at successively further distances from our focal county. Specifically, for each county i at time t, we pool the individuals in surrounding counties within 100 miles, compute their surname entropy and construct a separate instrument for these individuals, excluding i itself. We repeat this exercise for individuals between 100 and 200 miles and between 200 and 300 miles.[13]

Table 8 presents the results. For patents per capita, as shown in columns 1 to 4, the coefficients for surname entropy outside the county are mostly small and positive, but all are poorly estimated. In contrast, the coefficients for surname entropy within the focal county remain large and do not decrease in magnitude when considering the surname entropy of surrounding counties. For breakthrough patents, as detailed in columns 5 to 8, the small and non-significant coefficients display a mix of both positive and negative values, indicating that the entropy of surrounding counties has no detectable impact on breakthroughs. Overall, consistent with our theoretical framework, these findings suggest that the causal link between surname entropy and innovation is geographically localized.

6.4 The Components of Social Structure

Our main explanatory variable, surname entropy, combines the informational component—the diversity of information embedded in surnames within a county—with the social-psychological component—the concentration of these surnames. To assess the relative importance of the two components, we run a battery of OLS regressions controlling for the number of distinct surnames that go into our entropy calculations. Appendix Table B8 reports the estimates. We find that both components are associated with social structure's impact on innovation.

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

Table 8: Mechanism III: Limited spatial spillovers

	Patents per 1,000 people (mean = 0.99, sd = 1.60)				Breakthrough patents per 1,000 people (mean = 0.12, sd = 0.26)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname entropy	1.151***	1.153***	1.155***	1.239***	0.125***	0.125***	0.126***	0.101***
	(0.236)	(0.237)	(0.240)	(0.185)	(0.016)	(0.016)	(0.016)	(0.016)
Surname entropy (< 100 miles)	0.183	0.139	0.147	0.234	-0.002	-0.016	-0.015	0.014
	(0.165)	(0.167)	(0.163)	(0.176)	(0.029)	(0.028)	(0.029)	(0.044)
Surname entropy (100 < 200 miles)		0.234	0.228	-0.291		0.073	0.073	0.014
		(0.324)	(0.322)	(0.175)		(0.044)	(0.045)	(0.086)
Surname entropy (200 < 300 miles)			0.113	-0.151			0.013	0.000
			(0.191)	(0.144)			(0.034)	(0.039)
Panel B: Reduced-form estimates								
Surname entropy (push-pull IV)	0.787***	0.786***	0.786***	0.820***	0.091***	0.091***	0.091***	0.069***
	(0.230)	(0.231)	(0.231)	(0.201)	(0.018)	(0.018)	(0.018)	(0.014)
Surname entropy (push-pull IV, < 100 miles)	0.169	0.170	0.170	0.189*	0.009	0.009	0.007	-0.005
	(0.110)	(0.112)	(0.108)	(0.102)	(0.026)	(0.027)	(0.026)	(0.023)
Surname entropy (push-pull IV, 100 < 200 miles)		0.031	0.029	0.015		0.004	-0.002	0.016
		(0.140)	(0.131)	(0.101)		(0.020)	(0.020)	(0.024)
Surname entropy (push-pull IV, 200 < 300 miles)			-0.011	0.009			-0.027	-0.022
			(0.111)	(0.085)			(0.024)	(0.022)
Panel C: Instrumental-variable estimates								
Surname entropy	1.193***	1.191***	1.182***	1.272***	0.141***	0.141***	0.135***	0.108***
	(0.278)	(0.271)	(0.289)	(0.207)	(0.021)	(0.021)	(0.023)	(0.018)
Surname entropy (< 100 miles)	0.160	0.204	0.159	0.264	-0.031	-0.031	-0.064	-0.068
	(0.609)	(0.511)	(0.437)	(0.378)	(0.125)	(0.115)	(0.098)	(0.082)
Surname entropy (100 < 200 miles)		-0.180	-0.001	-0.315		-0.001	0.130	0.087
		(0.957)	(1.528)	(0.486)		(0.146)	(0.250)	(0.112)
Surname entropy (200 < 300 miles)			-0.462	0.033			-0.335	-0.121
			(1.775)	(0.401)			(0.365)	(0.112)
F-statistic: Surname entropy	56.259	58.415	45.582	39.172	56.259	58.415	45.582	39.172
F-statistic: Surname entropy (< 100 miles)	21.231	16.160	13.308	21.314	21.231	16.160	13.308	21.314
F-statistic: Surname entropy (100 < 200 miles)		7.293	9.698	19.331		7.293	9.698	19.331
F-statistic: Surname entropy (200 < 300 miles)			3.217	3.604			3.217	3.604
County fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
County-specific linear time trends				✓				✓
Observations	21,430	21,430	21,430	21,430	21,430	21,430	21,430	21,430

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

7 Conclusion

Focusing on the United States, during the period when it rose to dominate global innovation (1850-1940), we study the impact of social structure 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 social structure, we use 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 employ an instrumental variable approach, using immigrant flows to extract an exogenous component of surname entropy to examine the effect of surname entropy on our innovation outcomes. This analysis suggests that greater surname entropy causes faster innovation. Second, 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 county with greater surname entropy. Our analysis closes by showing that surname entropy lowers residential segregation, weakens family ties and fosters heterogeneous interactions among inventors at the level of individual patents.

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.

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

Akcigit, Ufuk, 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) "Immigration and the rise of American ingenuity", *American Economic Review*, 107 (5), pp. 327–331.

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

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

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

Alesina, Alberto & Paola Giuliano (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.

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

Allport, Gordon W. (1954) *The nature of prejudice*, The nature of prejudice. Oxford, England, Addison-Wesley, Pages: xviii, 537.

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.

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

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

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.

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*.

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

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

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

Carlino, Gerald & William R. Kerr (2015) *Agglomeration and Innovation*, 5, p. 349–404, Elsevier.

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.

Coluccia, Davide M, Gaia Dossi & Sebastian Ottinger (2023) "Racial discrimination and lost innovation".

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.

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.

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

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

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

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

Galor, 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.

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.

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.

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

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

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

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. (2008) "Ethnic Scientific Communities and International Technology Diffusion", *The Review of Economics and Statistics*, 90 (3), pp. 518–537.

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.

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

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.

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*.

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.

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

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.

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).

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

For Online Publication

Appendix to "How Social Structure Drives Innovation: Surname Diversity and Patents in U.S. History"

Max Posch, Jonathan Schulz, and Joseph Henrich

A Data Sources and Construction

Surname entropy

To construct county-level surname entropy 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 1900 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 1900 as the reference year because this is the first year of our panel data set in our main analysis.

Counties harmonized with very few people in a given census year may exhibit very low surname entropy (notably counties in Texas in 1900), potentially due to small sample bias. To address this, we winsorize all surname entropy variables at 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 entropy 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 entropy 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 entropy. 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 (3), 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 and occupational diversity, as well as 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 occupational diversity, we draw on the variable occ1950, dropping all observations with value greater or equal than 979. We transform the data from each period to 1900 U.S. counties using the M4 weights from the Ferrara et al. (2021) cross-walks. We use the variables age and yrimmig to estimate these variables for the midyears 1895, 1905, 1915, and 1925 by removing all individuals who were born or immigrated after the midyear.

Following the methodology in Raz (2023), we construct the strength of family ties measure from the full-count census data for all census waves from 1860 to 1940. The strength of family ties is determined by the first principal component of four underlying variables: (i) the divorce-to-marriage ratio, (ii) the share of elderly people living without a relative, (iii) the share of people living with at least one non-relative, and (iv) the mean size of families. The variables age and yrimmig are also used to estimate the strength of family ties for the midyears by removing all individuals who arrived after the midyear.

We 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.

Additionally, we calculate the strength of family ties within surname groups across counties, requiring a positive number of married and elderly individuals within the surname groups of each county. However, some surname groups do not meet this criterion, preventing the construction of the principal component for these groups and reducing the number of observations.

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

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

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

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

Construction of the innovation measures

We use the *Comprehensive Universe of 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 harmonize the data to 1900 U.S. counties.

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

In our least-squares analysis, depicted in Appendix Figure B7, we use patent issuing years rather than filing years and normalize patents by the population size in 1850 because filing years are inconsistently recorded in the CUSP dataset before 1870.

For county-period observations, we winsorize innovation outcome variables at the 99% level from the upper tail to lessen the impact of outlier counties with an exceptionally large number of patents. We do not winsorize the surname-county-period-level outcomes, as the number of breakthrough patents filed by inventors with a specific surname in a given county during a specified period is typically small. We also present results using non-winsorized, log-transformed patent counts (see Appendix Table B6).

The surname-county-period-level outcomes count the number of (breakthrough) patents filed by inventors with surname k residing in county i during period t, normalized by the 1900 surname population within county i. Constructing these variables requires inventor surnames. The CUSP data includes inventor names as a string variable containing the surname, first name, and sometimes middle names or initials. Identifying surnames from this variable can be challenging due to inconsistent name order. We use punctuation marks such as semicolons, colons, or commas to identify surnames. When the string variable starts with initials followed by a token of two or more characters, or when it ends with a whitespace followed by "DE", "DU", "DE LA", "DI", "DEL", "DELLA", "VAN", "VON",

"LE", "LA", or "ST", we identify the surnames accordingly. For the remaining entries, we tokenize the string based on whitespace, keeping the first and last tokens, typically representing the first name and surname. To determine the surname, we compare the frequencies of all name combinations from pooled census years 1900, 1910, 1920, 1930, and 1940, identifying the surname based on the most common constellation. For example, for the tokens "JOHN" and "PETER", we identify the surname based on whether there were more individuals named "JOHN PETER" or "PETER JOHN". Finally, we clean the surname variable following the described steps.

B Additional Tables and Figures

Table B1: Correlations between baseline surname entropy and alternative surname entropy measures

HHI surname	Surname, uncorrected	Surname, men	Surname, household heads	Surname, whites	Surname-race	Surname-country of birth
0.83	0.99	1.00	0.99	0.99	0.98	0.98

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

Table B2: 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)
$I^{1880}_{n,-r(i)} \times \frac{I^{1880}_{-n,i}}{I^{1880}_{-n}}$	3.177***	3.161***	3.674***	3.696***	4.088***	4.200***	4.610***	4.717***
	(0.095)	(0.104)	(0.109)	(0.089)	(0.097)	(0.117)	(0.092)	(0.089)
$I^{1895}_{n,-r(i)} \times \frac{I^{1895}_{-n,i}}{I^{1895}_{-n}}$	0.927***	1.323***	1.719***	2.146***	2.425***	3.394***	3.753***	3.873***
	(0.242)	(0.328)	(0.281)	(0.162)	(0.172)	(0.207)	(0.253)	(0.316)
$I^{1900}_{n,-r(i)} \times \frac{I^{1900}_{-n,i}}{I^{1900}_{-n}}$		-5.152*	-3.206	-5.028*	-4.033	-9.957***	-9.810***	-7.854***
		(3.051)	(3.495)	(2.594)	(2.783)	(2.983)	(3.225)	(2.899)
$I^{1905}_{n,-r(i)} \times \frac{I^{1905}_{-n,i}}{I^{1905}_{-n}}$			14.533***	18.658***	21.489***	26.922***	30.084***	32.477***
			(0.993)	(1.150)	(1.345)	(1.392)	(1.240)	(1.052)
$I^{1910}_{n,-r(i)} \times \frac{I^{1910}_{-n,i}}{I^{1910}_{-n}}$				17.646***	20.004***	27.313***	29.980***	31.991***
				(2.864)	(3.145)	(3.264)	(3.197)	(2.802)
$I^{1915}_{n,-r(i)} \times \frac{I^{1915}_{-n,i}}{I^{1915}_{-n}}$					8.868***	15.501***	17.807***	20.289***
					(1.391)	(1.981)	(1.460)	(1.599)
$I^{1920}_{n,-r(i)} \times \frac{I^{1920}_{-n,i}}{I^{1920}_{-n}}$						0.977	2.001*	6.371***
						(2.009)	(1.090)	(1.460)
$I^{1925}_{n,-r(i)} \times \frac{I^{1925}_{-n,i}}{I^{1925}_{-n}}$							26.904***	32.465***
							(1.237)	(1.500)
$I^{1930}_{n,-r(i)} \times \frac{I^{1930}_{-n,i}}{I^{1930}_{-n}}$								-34.132***
								(3.240)
Observations	5,346,601	6,604,833	7,012,154	7,284,327	7,364,850	8,067,190	8,143,606	8,980,212
R^2	0.719	0.695	0.706	0.696	0.700	0.660	0.702	0.692
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 (3), 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.

Table B3: Robustness: Controlling for population size

	Patents per 1,000 people (mean = 1.05, sd = 1.73)			Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname entropy	1.318***	1.261***	1.439***	0.147***	0.133***	0.115***
	(0.174)	(0.237)	(0.171)	(0.018)	(0.017)	(0.020)
Population	0.328***	0.307***	0.396**	0.061***	0.052***	0.033
	(0.085)	(0.081)	(0.166)	(0.015)	(0.013)	(0.031)
Panel B: Reduced-form estimates						
Surname entropy (push-pull IV)	0.909***	0.883***	0.979***	0.103***	0.097***	0.076***
	(0.187)	(0.228)	(0.203)	(0.015)	(0.018)	(0.019)
Population (push-pull IV)	0.226**	0.179*	0.086	0.046**	0.032**	0.008
	(0.106)	(0.099)	(0.181)	(0.019)	(0.015)	(0.057)
Panel C: Instrumental-variable estimates						
Surname entropy	1.358***	1.326***	1.466***	0.154***	0.146***	0.114***
	(0.185)	(0.243)	(0.196)	(0.019)	(0.020)	(0.026)
Population (push-pull IV)	0.215**	0.194**	0.215	0.045**	0.034**	0.018
	(0.094)	(0.089)	(0.169)	(0.018)	(0.014)	(0.057)
Kleibergen-Paap *F*-statistic	110.932	100.800	87.534	110.932	100.800	87.534
Panel D: First-stage estimates			Surname entropy			
Surname entropy (push-pull IV)	0.669***	0.666***	0.668***	0.669***	0.666***	0.668***
	(0.064)	(0.066)	(0.071)	(0.064)	(0.066)	(0.071)
Population (push-pull IV)	0.008	-0.012	-0.088	0.008	-0.012	-0.088
	(0.017)	(0.017)	(0.069)	(0.017)	(0.017)	(0.069)
Within R^2	0.706	0.688	0.684	0.706	0.688	0.684
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	22,073	22,073	22,073	22,073	22,073	22,073

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (2) and first-stage estimates for equation (4) while additionally controlling for number of distinct surnames in a county. An observation is a county in a period from 1900 to 1940. The endogenous variable is county-level surname entropy 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.

Table B4: Robustness: Regional heterogeneity in the effect of surname entropy on innovation

	Patents per 1,000 people			Breakthrough patents per 1,000 people		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname entropy × Region = Midwest	1.820***	1.596***	1.749***	0.254***	0.219***	0.151***
	(0.314)	(0.273)	(0.291)	(0.073)	(0.067)	(0.056)
Surname entropy × Region = Northeast	2.947***	2.834*	0.430	0.862***	0.732	-0.063
	(0.867)	(1.538)	(0.503)	(0.297)	(0.509)	(0.318)
Surname entropy × Region = South	3.349***	3.469**	1.513***	0.693**	0.773**	0.023
	(1.103)	(1.427)	(0.484)	(0.260)	(0.329)	(0.021)
Surname entropy × Region = West	2.592***	2.436***	3.368***	0.399***	0.313***	0.305**
	(0.589)	(0.716)	(1.144)	(0.111)	(0.104)	(0.121)
Panel B: Reduced-form estimates						
Surname entropy (push-pull IV) × Region = Midwest	1.217***	1.233***	1.246***	0.157***	0.171***	0.094*
	(0.220)	(0.206)	(0.289)	(0.043)	(0.049)	(0.050)
Surname entropy (push-pull IV) × Region = Northeast	2.007***	1.782**	0.456***	0.597**	0.545*	0.143**
	(0.747)	(0.728)	(0.124)	(0.246)	(0.277)	(0.055)
Surname entropy (push-pull IV) × Region = South	1.870**	1.900*	1.013*	0.411*	0.459	0.122*
	(0.899)	(1.099)	(0.551)	(0.226)	(0.278)	(0.069)
Surname entropy (push-pull IV) × Region = West	1.481***	1.390***	2.098***	0.237***	0.175***	0.208***
	(0.277)	(0.324)	(0.512)	(0.059)	(0.040)	(0.054)
Panel C: Instrumental-variable estimates						
Surname entropy × Region = Midwest	1.792***	1.750***	1.748***	0.238***	0.243***	0.132*
	(0.305)	(0.303)	(0.367)	(0.060)	(0.074)	(0.070)
Surname entropy × Region = Northeast	3.129***	3.789**	1.087***	0.919***	1.158*	0.341*
	(0.958)	(1.684)	(0.307)	(0.337)	(0.652)	(0.172)
Surname entropy × Region = South	3.221***	3.304**	1.847***	0.707**	0.798**	0.222**
	(1.066)	(1.428)	(0.671)	(0.282)	(0.366)	(0.086)
Surname entropy × Region = West	2.667***	2.667***	3.838***	0.433***	0.336***	0.380**
	(0.574)	(0.785)	(1.233)	(0.124)	(0.107)	(0.148)
F-statistic: 1st coefficient	156.129	71.996	69.147	156.129	71.996	69.147
F-statistic: 2nd coefficient	29.185	4.253	3.297	29.185	4.253	3.297
F-statistic: 3rd coefficient	41.699	9.336	6.508	41.699	9.336	6.508
F-statistic: 4th coefficient	40.978	6.952	7.205	40.978	6.952	7.205
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	22,073	22,073	22,073	22,073	22,073	22,073

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

Table B5: Robustness: Education

	Patents per 1,000 people (mean = 0.75, sd = 1.28)			Breakthrough patents per 1,000 people (mean = 0.15, sd = 0.27)		
	(1)	(2)	(3)	(4)	(5)	(6)
Surname entropy	0.510***	0.453***	0.392***	0.099***	0.087***	0.075***
	(0.076)	(0.072)	(0.054)	(0.016)	(0.015)	(0.011)
Average years of schooling		0.194***	0.309***		0.039***	0.062***
		(0.045)	(0.079)		(0.008)	(0.015)
Surname entropy × Average years of schooling			0.308***			0.060***
			(0.066)			(0.013)
R^2	0.400	0.408	0.447	0.351	0.359	0.392
State fixed effects	✓	✓	✓	✓	✓	✓
Observations	2,820	2,820	2,820	2,820	2,820	2,820

Notes: The table reports OLS estimates of regressions of number of patents and breakthrough patents filed between 1940 and 1944 per 1940 population on surname entropy and individuals' average years of schooling in 1940. The unit of observation is a county in 1940 harmonized to 1900 borders. Standard errors are clustered on states and reported in parentheses. 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 B6: Robustness: Log-transformed innovation outcomes (I/II)

	Log Patents per 1,000 people			Log Breakthrough patents per 1,000 people		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname entropy	0.467***	0.445***	0.430***	0.141***	0.130***	0.083***
	(0.055)	(0.079)	(0.056)	(0.018)	(0.024)	(0.013)
Panel B: Reduced-form estimates						
Surname entropy (push-pull IV)	0.327***	0.312***	0.290***	0.100***	0.095***	0.060***
	(0.060)	(0.075)	(0.063)	(0.019)	(0.023)	(0.011)
Panel C: Instrumental-variable estimates						
Surname entropy	0.487***	0.470***	0.443***	0.150***	0.143***	0.092***
	(0.052)	(0.076)	(0.059)	(0.020)	(0.026)	(0.013)
Kleibergen-Paap *F*-statistic	119.613	104.522	91.652	119.613	104.522	91.652
Panel D: First-stage estimates			Surname entropy			
Surname entropy (push-pull IV)	0.671***	0.663***	0.654***	0.671***	0.663***	0.654***
	(0.061)	(0.065)	(0.068)	(0.061)	(0.065)	(0.068)
Within R^2	0.706	0.688	0.682	0.706	0.688	0.682
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	22,073	22,073	22,073	22,073	22,073	22,073

Notes: The table reports least squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation (2) and first-stage estimates for equation (4). An observation is a county in a period from 1900 to 1940. The endogenous variable is county-level surname entropy in t. In columns 1 to 3, the dependent variable is log number of patents filed in the county in the five-year period starting in t divided by county population size in 1900 (plus one, to avoid dropping observations with zero patents). In columns 4 to 6, the dependent variable is log number of breakthrough patents filed in the county in the five-year period starting in t divided by county population size in 1900 (plus one). Standard errors are clustered at the state level. 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 B7: Robustness: Log-transformed innovation outcomes (II/II)

	Log Patents			Log Breakthrough patents		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Least-squares estimates						
Surname entropy	0.758***	0.694***	0.455***	0.341***	0.281***	0.181***
	(0.054)	(0.035)	(0.041)	(0.057)	(0.041)	(0.033)
Panel B: Reduced-form estimates						
Surname entropy (push-pull IV)	0.566***	0.529***	0.337***	0.265***	0.215***	0.123***
	(0.044)	(0.057)	(0.041)	(0.030)	(0.027)	(0.023)
Panel C: Instrumental-variable estimates						
Surname entropy	0.844***	0.798***	0.516***	0.396***	0.324***	0.188***
	(0.055)	(0.041)	(0.041)	(0.060)	(0.043)	(0.036)
Kleibergen-Paap *F*-statistic	12.301	15.845	16.881	12.301	15.845	16.881
Panel D: First-stage estimates			Surname entropy			
Surname entropy (push-pull IV)	0.671***	0.663***	0.654***	0.671***	0.663***	0.654***
	(0.061)	(0.065)	(0.068)	(0.061)	(0.065)	(0.068)
Within R^2	0.706	0.688	0.682	0.706	0.688	0.682
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	22,073	22,073	22,073	22,073	22,073	22,073

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

Table B8: Contributions of the informational and social-psychological components

	Patents per 1,000 people (mean = 1.05, sd = 1.73)			Breakthrough patents per 1,000 people (mean = 0.13, sd = 0.28)		
	(1)	(2)	(3)	(4)	(5)	(6)
Surname entropy	0.996***	0.896***	1.275***	0.083***	0.063**	0.097**
	(0.200)	(0.267)	(0.176)	(0.031)	(0.024)	(0.042)
Log Number of distinct surnames	0.393***	0.433**	0.207*	0.077***	0.082***	0.022
	(0.144)	(0.164)	(0.114)	(0.028)	(0.027)	(0.033)
County fixed effects	✓	✓	✓	✓	✓	✓
Period fixed effects	✓			✓		
State-Period fixed effects		✓	✓		✓	✓
County-specific linear time trends			✓			✓
Observations	22,071	22,071	22,071	22,071	22,071	22,071

Notes: The table reports least squares estimates. An observation is a county in a period from 1900 to 1940. To be consistent with the entropy formula, where we use log to the base of 2, we transform the number of distinct surname using log to the base of 2. Standard errors are clustered at the state level. All independent variables are standardized to mean zero and unit variance. The sources and construction of all variables are explained in Appendix A. ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

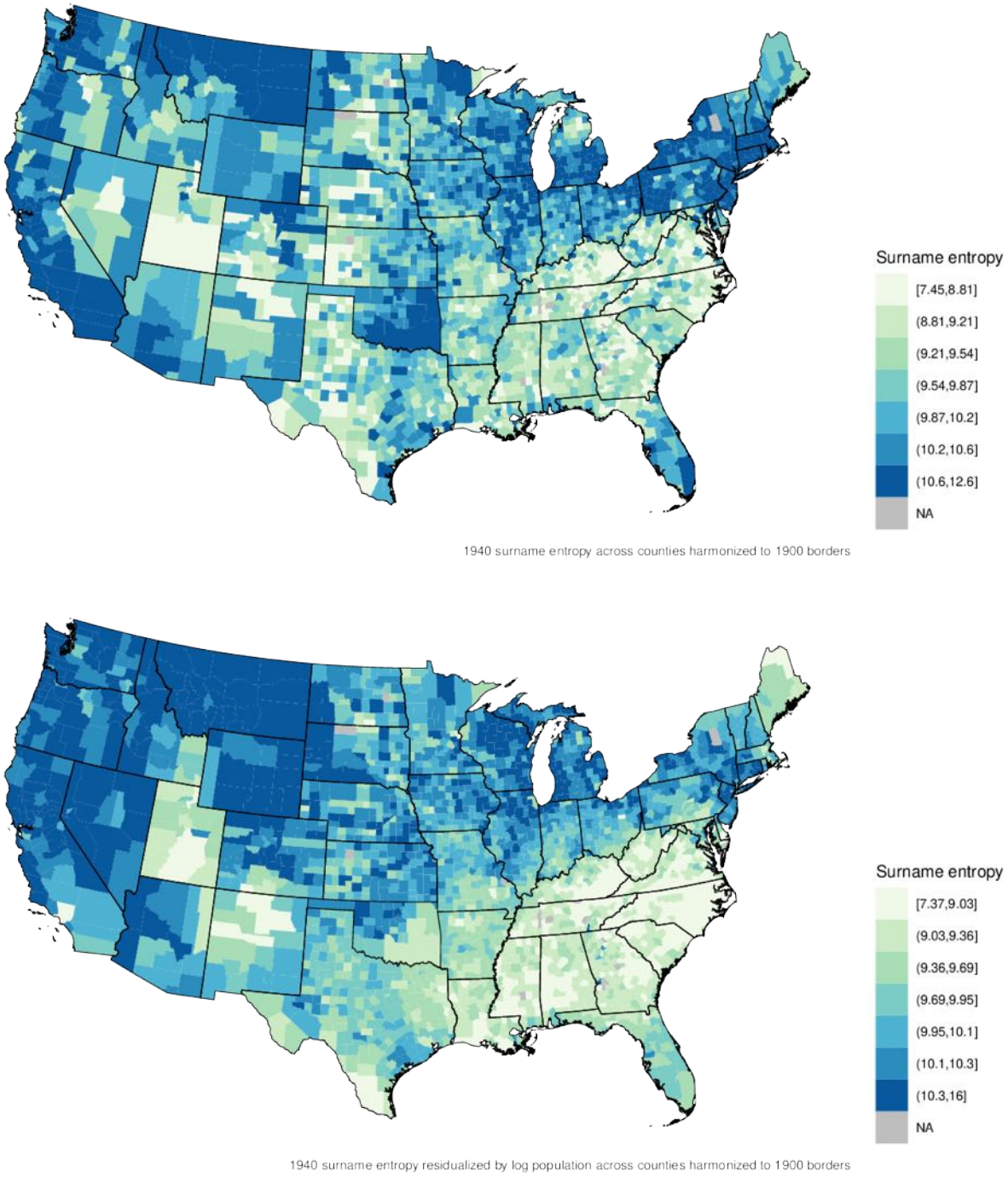
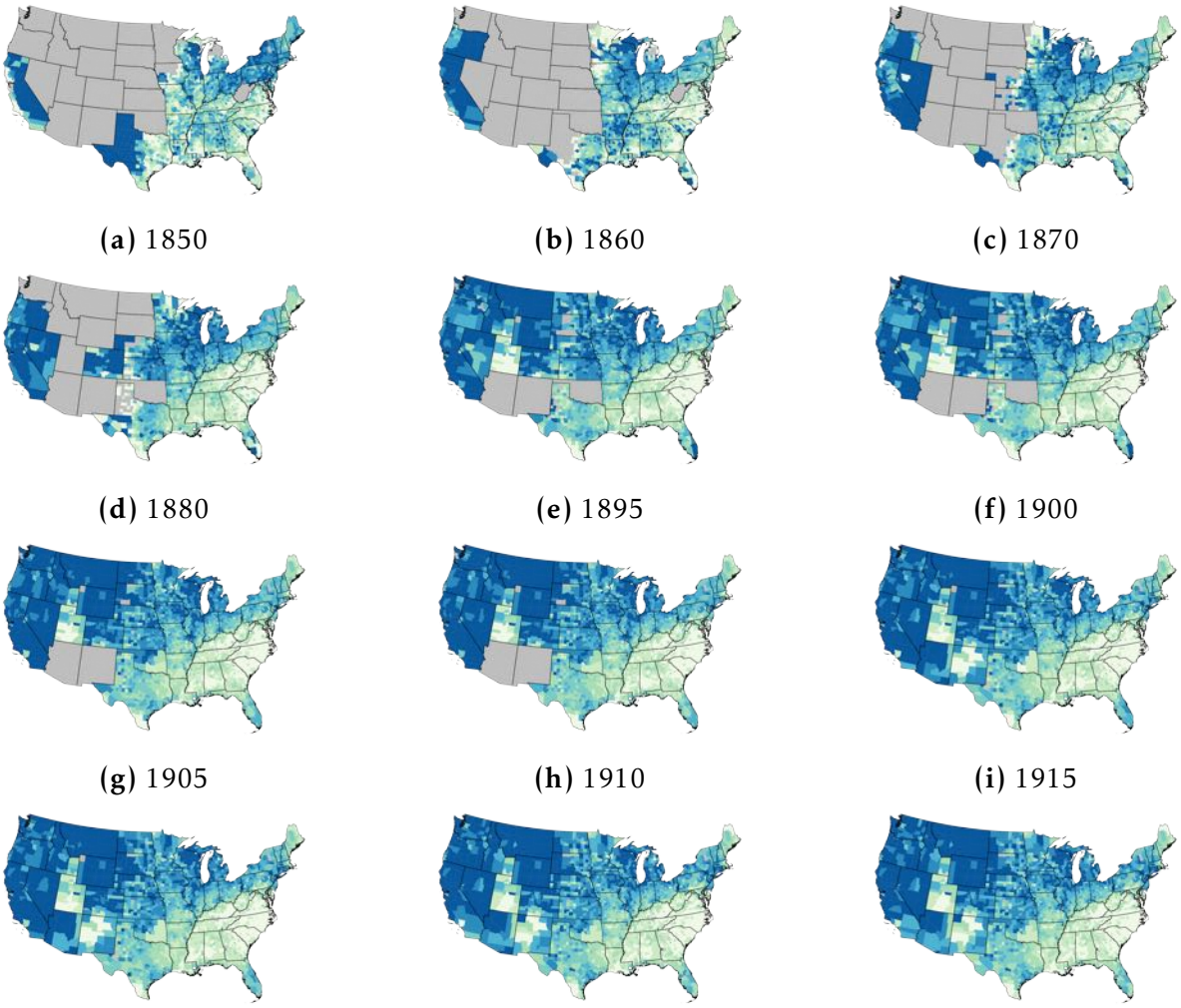


Figure B1: Surname entropy in 1940

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



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

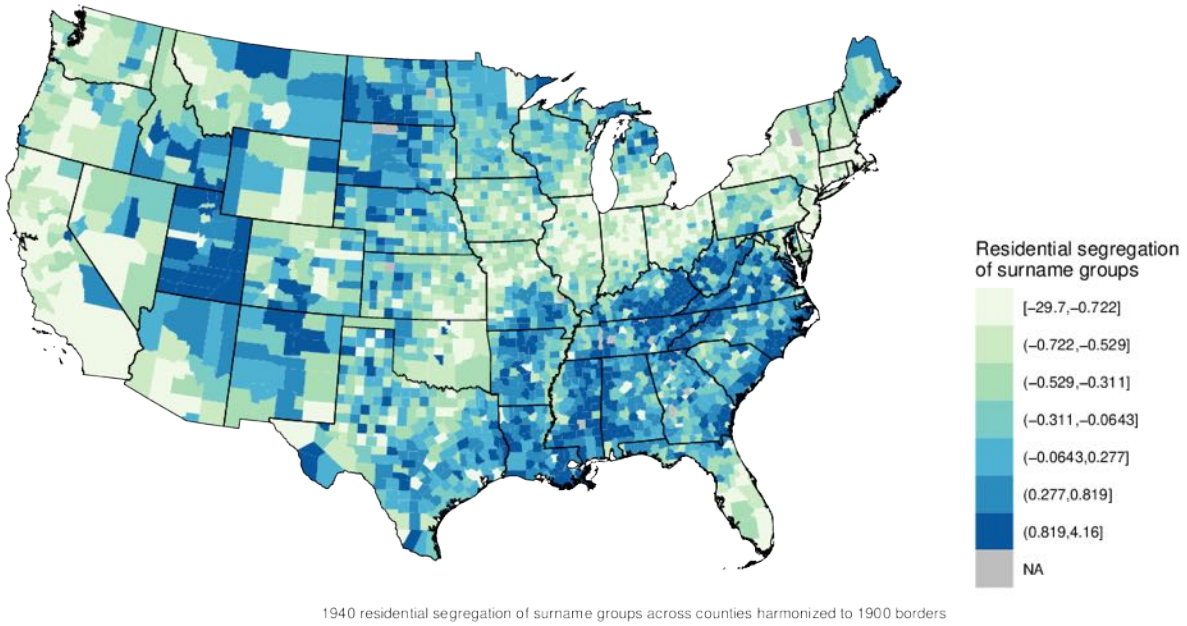
(d) 1880 **(e)** 1895 **(f)** 1900

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

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

Figure B2: Surname entropy from 1850 to 1930

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



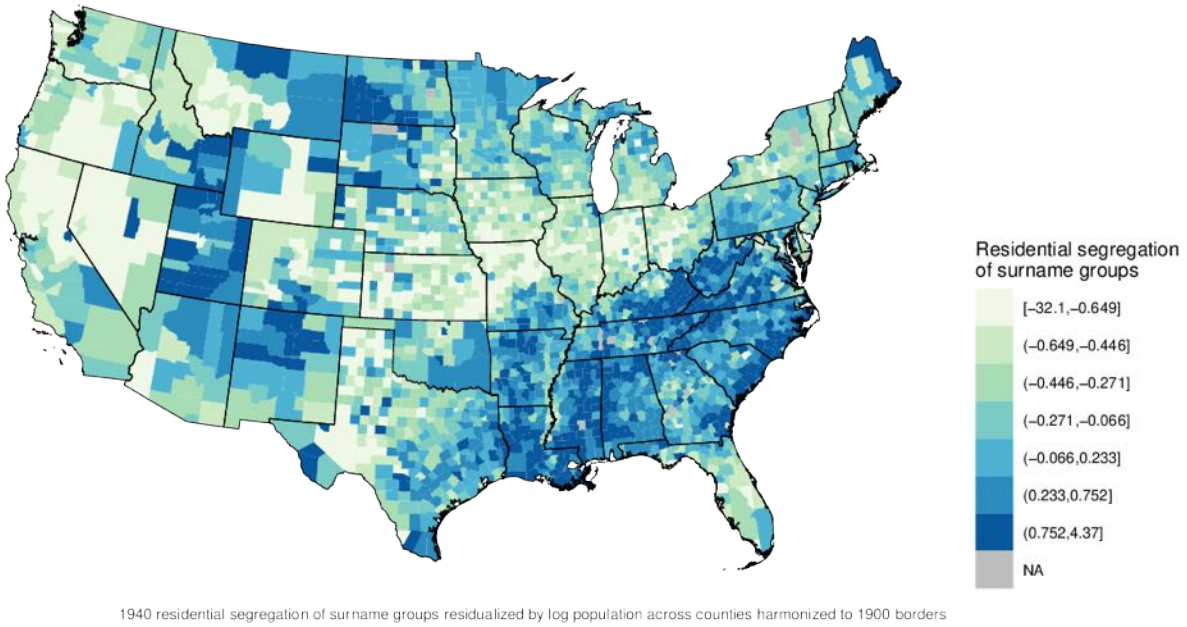
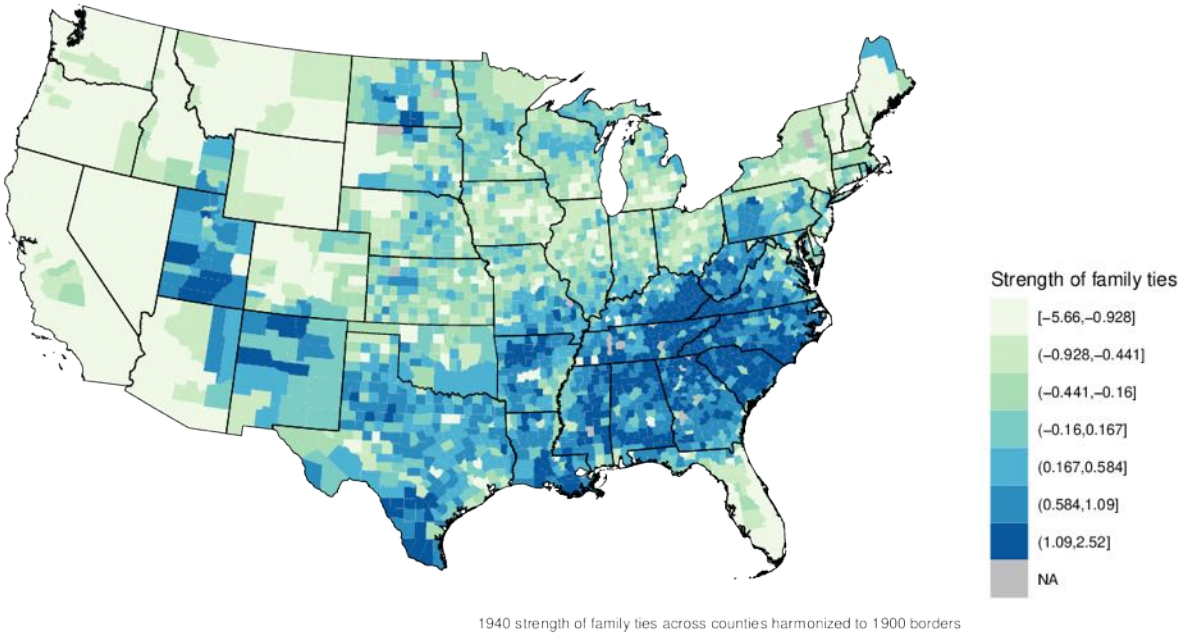


Figure B3: Residential segregation of surname groups in 1940

Notes: The figures show the geographic variation in residential segregation of surname groups in 1940 across counties harmonized to 1900 borders. Top: Raw data; bottom: residualized by log county population. The segregation measure is constructed adapting the Logan and Parman (2017) procedure to surnames. The sources and construction of all variables are explained in Appendix Section A.



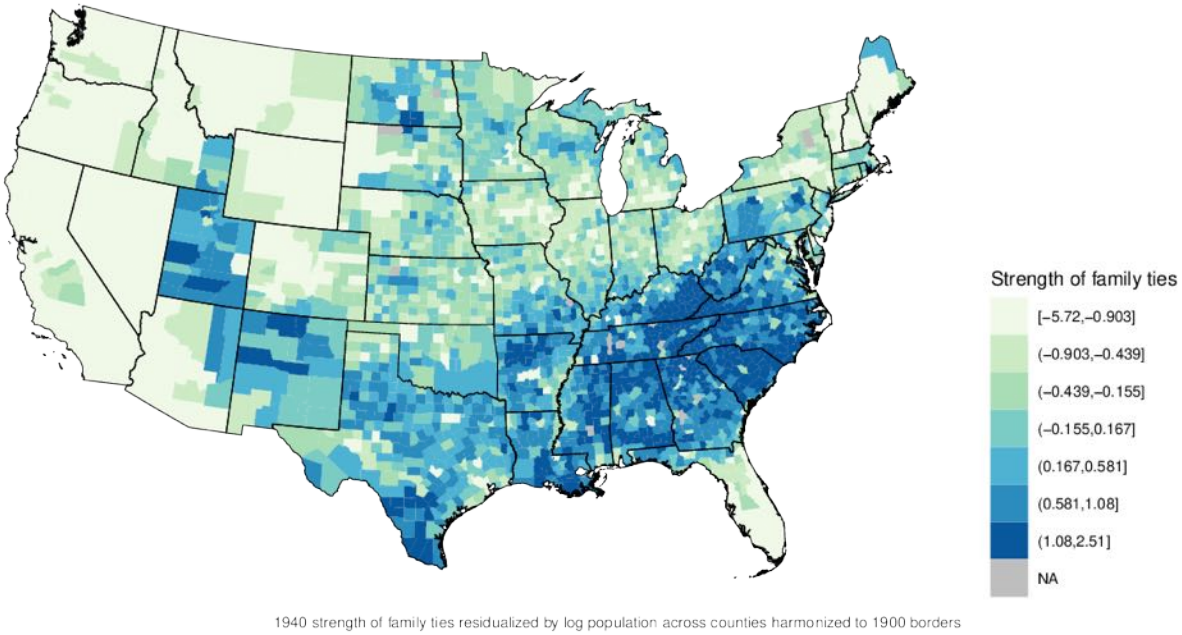
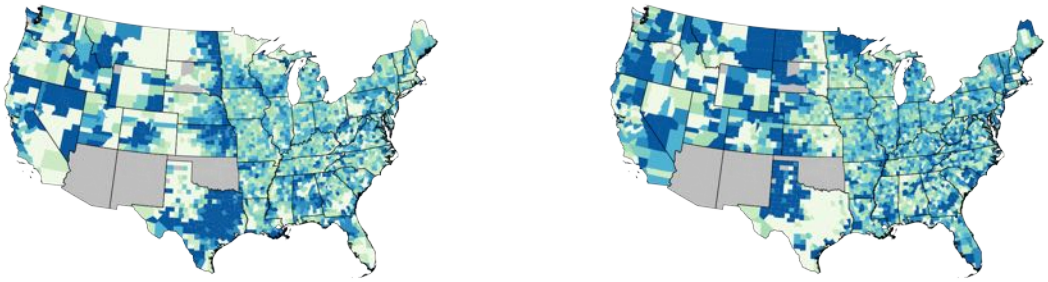


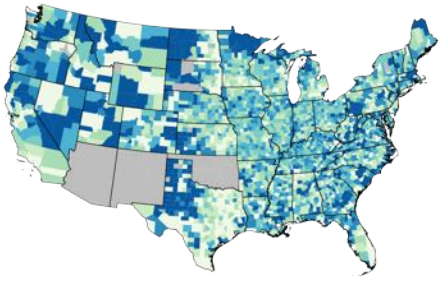
Figure B4: Strength of family ties in 1940

Notes: The figures show the geographic variation in strength of family ties in 1940 across counties harmonized to 1900 borders. Top: Raw data; bottom: residualized by log county population. The segregation measure is constructed following Raz (2023). The sources and construction of all variables are explained in Appendix Section A.

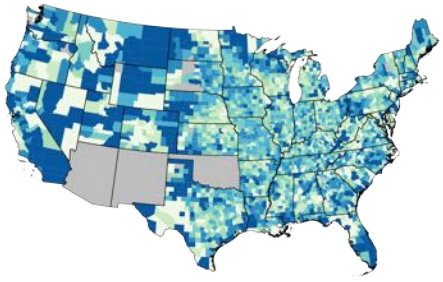


(a) 1900

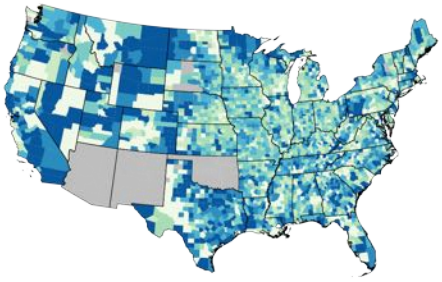
(b) 1905



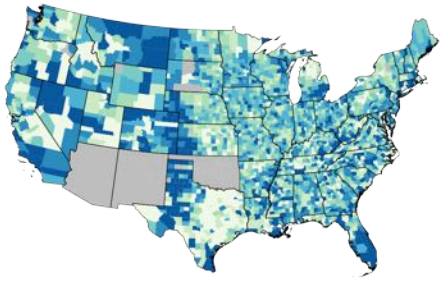
(c) 1910



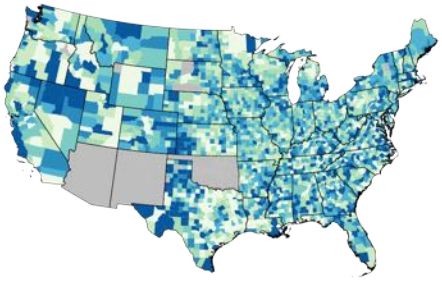
(d) 1915



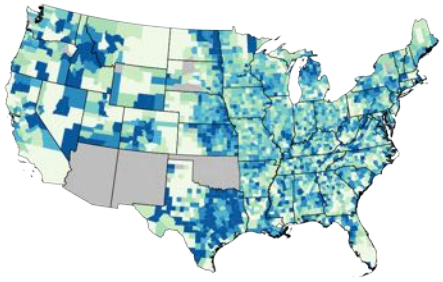
(e) 1920



(f) 1925



(g) 1930



(h) 1940

Figure B5: Predicted surname entropy (residuals)

Notes: This figure maps residualized instrumented surname entropy for each of the eight periods. We regress the instrument for surname entropy on county and state-year fixed effects, and county specific linear time trends, and calculate the residuals. This visualization depicts the instrument used in the regression in Table 2. 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.

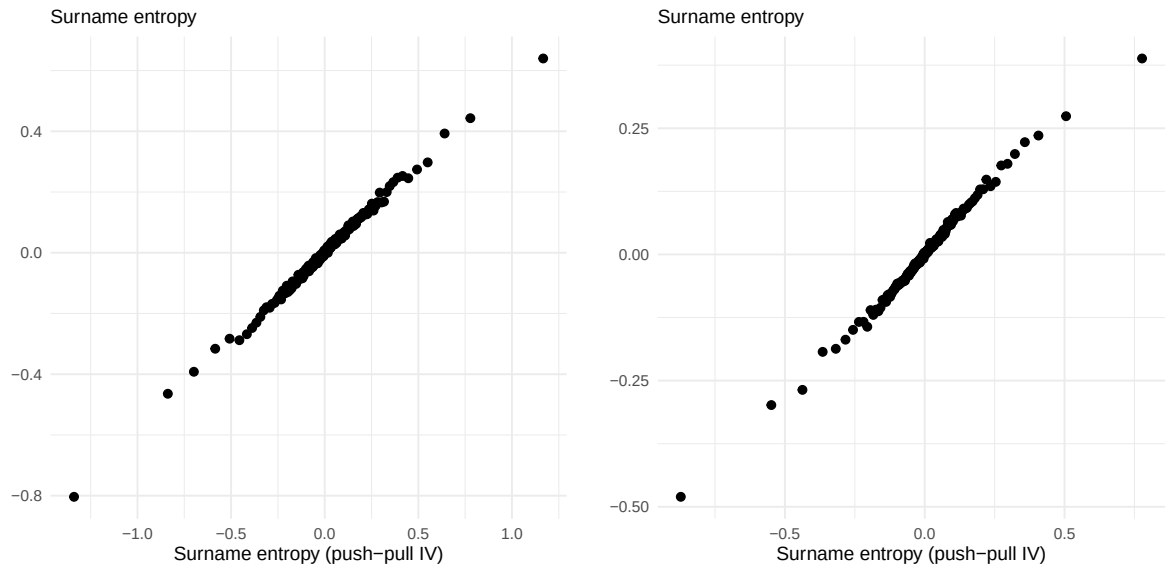


Figure B6: First Stage: Binned scatter plots of surname entropy (pull-push IV) and actual surname entropy from 1900 to 1940

Notes: County-level data from 1900 to 1940 (including midyears). Observations are residualized by county fixed effects and state-period fixed effects (left plot) and additionally county-specific time trends (right plot). Binscatter plot created using the R package written by Cattaneo et al. (2019).

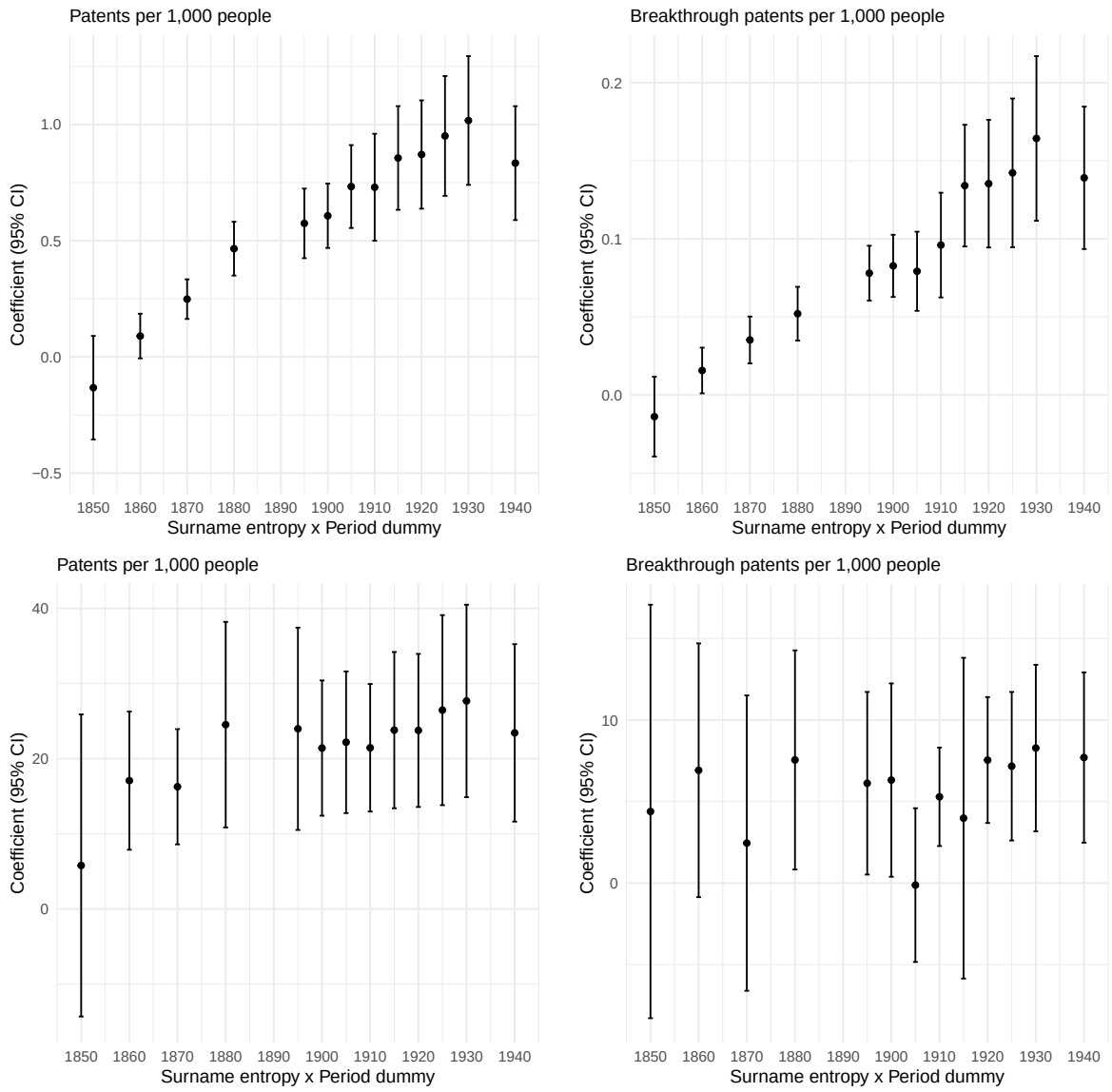


Figure B7: Least-squares estimates in years 1850-1940

Notes: The top figures display coefficients from a regression analysis of the number of (breakthrough) patents normalized by county population in 1900 (the same dependent variable as in the main analysis) on surname entropy, interacted with period dummies, conditional on county fixed effects and state-period fixed effects. The bottom figures show the coefficients for (breakthrough) patents normalized by county population in 1850. Standard errors are clustered at the state level. Relatively few patents were issued in 1850 compared to the latter years. Surname entropy is standardized to have a mean of 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.

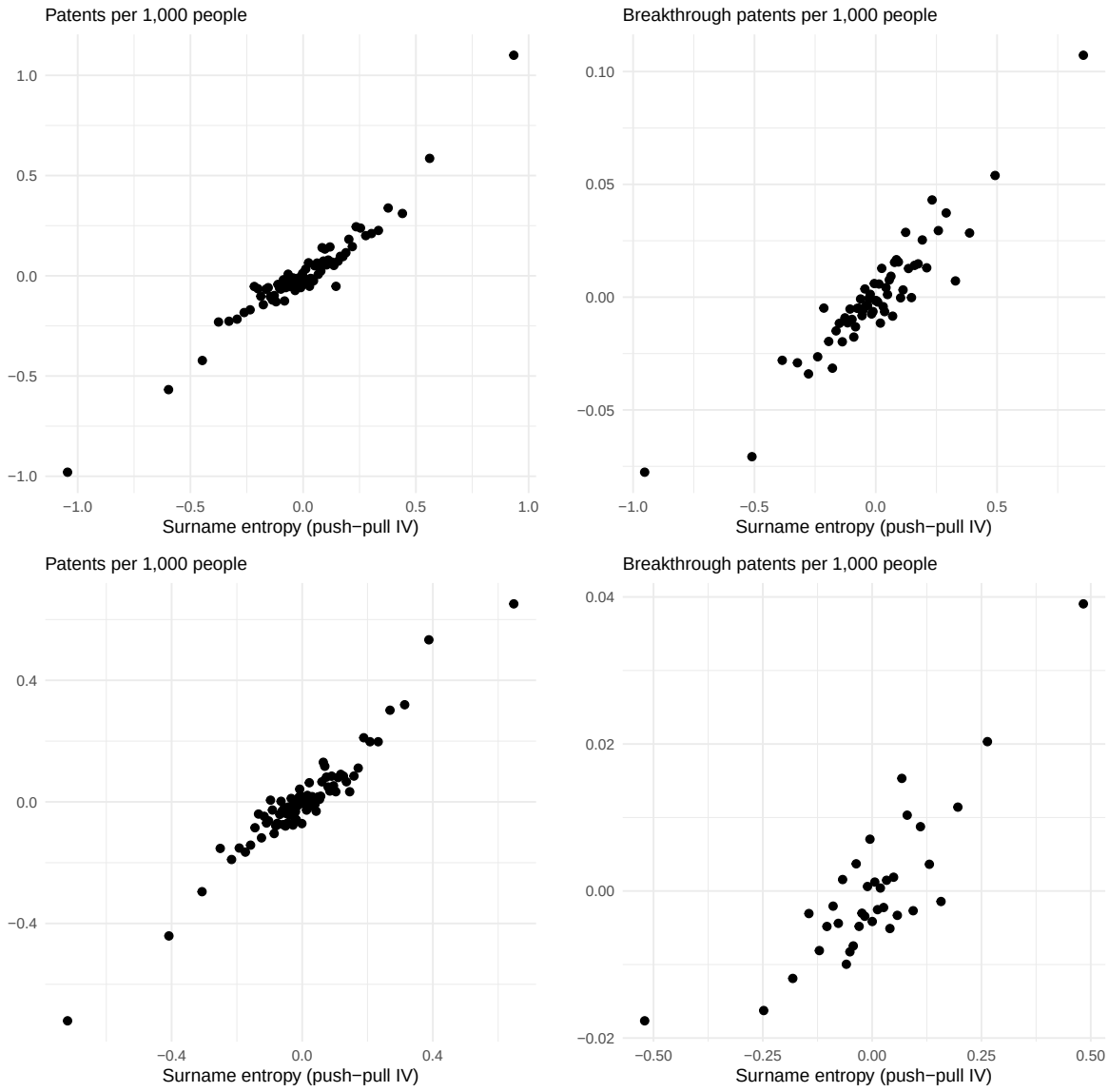


Figure B8: Reduced-from relationships: Binned scatter plots of surname entropy (pull-push IV) and innovation outcomes from 1900 to 1940

Notes: County-level data from 1900 to 1940 (including midyears). Observations are residualized by county fixed effects and state-period fixed effects (top plots), and additionally county-specific time trends (bottom plots). Binscatter plot created using the R package written by Cattaneo et al. (2019).

B.1 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.

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

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

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, $\text{Surname entropy}_i^t$ is county i's surname entropy in t, and $\widehat{\text{Surname entropy}}_i^t$ is county i's predicted surname entropy in t. $Y_{i,c}^t$ now is the log number of (breakthrough) patents 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 entropy 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. Standard errors are clustered in two ways, on states and patent class.

The results are reported in Table B9 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 statistically significant in all specifications. Thus, we conclude that differences across technological categories do not affect our results.

Table B9: Robustness: Patent technology class fixed effects

	Patents per 1,000 people				Breakthrough patents per 1,000 people			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Least-squares estimates								
Surname diversity	0.003***	0.003***	0.003***	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 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	0.004***	0.004***	0.004***	0.003***	0.000***	0.000***	0.000***	0.000**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Kleibergen-Paap F-statistic	114.358	114.358	114.358	113.439	114.358	114.358	114.358	113.439
County fixed effects	✓			✓	✓			✓
State-Period fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
County-Patent class fixed effects		✓	✓	✓		✓	✓	✓
Patent class-Period fixed effects			✓	✓			✓	✓
County-specific linear time trends				✓				✓
Observations	8,405,616	8,405,616	8,405,616	8,405,616	8,405,616	8,405,616	8,405,616	8,405,616

Notes: The table reports least-squares, reduced-form, and instrumental-variable (IV) estimates for the specifications described in equation 10. An observation is a patent class in a given county in a period from 1900 to 1940. 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 county population size 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.