Surname-based Connections, Informal Institutions, and Local Innovation *

Po-Hsuan Hsu † Jian Xie‡ Bernard Yeung§ Qinhong Yu¶

July 23, 2026

Abstract

Surname-based ties facilitate local social networks and collaboration, particularly where formal institutions are weak. We hypothesize that these ties foster local innovation because innovation depends on teamwork and the coordination of complementary resources that surname-based networks help mobilize. Using a county-year panel from China, we find that counties with stronger surname-based connections among business owners exhibit significantly higher patenting activity. An instrumental-variable strategy based on historical peasant rebellions in imperial China–which increased the importance of surname-based collaboration in local culture–supports a causal interpretation. More importantly, patenting rises disproportionately among firms controlled by business owners sharing more common surnames, allowing us to attribute local increases in innovation directly to surname-based ties. Consistent with these ties substituting for formal institutions, the innovation-enhancing effect weakens with better access to finance, higher-quality public services, and stronger supply-chain cooperation, and diminishes over time as China's formal institutions strengthen. We further show that stronger surname-based connections are associated with greater local entrepreneurship and faster economic development, highlighting the broader role of informal social networks in shaping economic growth.

Keywords: Social Network, Informal Institutions, Surname Homogeneity, Innovation, Entrepreneurs, Economic Development, China

JEL codes: O3, O43, O17, E26, L26, Z13

*We thank Ying Bai, Bishnupriya Gupta, Wonjoon Kim, Renliang Liu, Dennis Novy, Yancheng Qiu, Ao Wang, Mingang Zhang, Maggie Zhou, and participants at the 2024 SUFE-SDU-Warwick Economics Workshop, the 2024 IESR-Warwick Workshop, the seminar at the Li Anmin Institute of Economic Research, the Greater Bay Area Workshop on Applied Microeconomics and Public Policy, the CUHK Brownbag Seminar, the SUSTech Brownbag Seminar, the 3rd Conference on Global Innovation and Intellectual Property, the AIEA seminar, and the 12th Hong Kong Economic Association Biennial Conference, for their helpful comments. All errors are our own. The authors contribute equally as first authors. We therefore list authorship by alphabetical order.

†College of Technology Management, National Tsing Hua University. Email: pohsuanhsu@mx.nthu.edu.tw.

‡College of Business, Southern University of Science and Technology. Email: xiej@sustech.edu.cn.

§NUS Business School, National University of Singapore and College of Business, Southern University of Science and Technology. Corresponding author. Email: byeung@nus.edu.sg.

¶College of Business, Southern University of Science and Technology. Email: 12431486@mail.sustech.edu.cn.

1 Introduction

Local economic development depends on many factors, among which technological innovation is a central driver of long-run growth (Solow, 1957; Prescott, 1986; Romer, 1986, 1990; Griliches, 1987). Historically, innovation is inherently collaborative, requiring individuals to combine complementary knowledge and resources (Jones, 2005; Mokyr, 2005). Consequently, social networks that facilitate trust and cooperation can play an important role in shaping regional innovation (Feldman, 1999; Carlino and Kerr, 2015).

However, unlike routine tasks, innovation takes a long time to deliver. It involves high uncertainty and information asymmetry, making outcomes difficult to verify and thus hard to contract for within or between organizations (Aghion and Tirole, 1994a,b). This contractibility problem is even more damaging to innovation collaborations in emerging countries, where formal institutions are weak, e.g., underdeveloped financial markets and law enforcement (Ayyagari et al., 2011). In such environments, informal institutional factors such as social networks may substitute for formal institutions and mitigate hindrances to innovation, such as information asymmetry, agency issues, and insufficient protection of property rights.[1]

In this paper, we focus on surname-based connections, an important phenomenon in Chinese sociology (Du et al., 1992; Du, 2019; Bai and Kung, 2022). In Chinese culture, sharing the same surname creates a natural social or emotional connection between individuals (Du et al., 1992; Du, 2019). In addition, a surname functions as social capital in a network because it reflects social identity and carries the reputations of all people who bear it (Tadelis, 1999; Clark and Cummins, 2014), thereby further shaping social pressure. Thus, we aim to capture social connections through observable surnames for our analysis. It is worth noting that, given China's long history, two people sharing the same surname do not necessarily mean they are distant relatives, even though they may share the same ancestry many thousands or hundreds of years ago. In other words, the role and influence of such social connections are beyond those of kinship or clanship.

The formation and significance of common-surname networks are well documented in the literature (e.g., Freedman, 1958; Peng, 2004; Angelucci et al., 2010; Tan et al., 2021; Huang et al., 2023). McPherson et al. (2001) document that surname-based connections can foster trust, social cohesion, and mutual understanding while also facilitating more effective communication. Moreover, prior studies have shown how social connections and trust lead to better innovation outcomes at country and firm levels (Faleye et al., 2014; Zacchia, 2020; Chang and Wu, 2021; Xie et al., 2022). These discussions motivate us to hypothesize that surname-based connections facilitate and lubricate local collaborations, thereby increasing local innovation output in China, as informal institutions have

[1] In environments where formal institutions are lacking, sustaining innovation and economic growth requires informal institutions. As defined by North (1990), informal institutions are a set of rules, procedures, and norms that constrain individual behavior to maximize collective benefit. Some pioneering works have emphasized the role of informal institutions (Guiso et al., 2006; Greif and Tabellini, 2010) in conferring legitimacy, providing guidance, and incentivizing firms and individuals when formal institutions are absent or underdeveloped, thereby shaping economic outcomes.

played an important role in local communities when formal institutions are underdeveloped (Xu and Yao, 2015; Allen et al., 2005).[2]

We focus on China's context for two key reasons. First, China's low level of foreign immigration, compared with large countries such as the United States, Germany, France, and the U.K., has resulted in a more ethnically homogeneous population. Thus, regional surname homogeneity could be a clean proxy for social connections, with less risk of confounding from ethnic diversity. Second, China has experienced marked regional and temporal variation in institutional development, which offers us an opportunity to pinpoint the substitution between informal and formal institutions, especially from the perspective of local innovation. In particular, the central planning experiment in China since 1949 had dismantled all the market and economic institutions we know in a market economy, until China began rebuilding a market-based economic system starting in 1978, particularly after it entered the WTO in 2001.[3]

Using registration data for the universe of Chinese private enterprises, we measure surname-based connections by calculating annual county-level surname homogeneity among business owners over 2000–2019, using an entropy-based approach. This results in time-varying surname homogeneity variables, allowing us to implement panel regression analysis with county- and year-fixed effects and controlling for geographic, transportation, demographic, and socioeconomic factors. We find that higher surname homogeneity is positively associated with Chinese patent applications filed by local private firms, which we use as a proxy for local innovation. Specifically, a one standard deviation increase in surname homogeneity is associated with a 21.1% increase in patent filings relative to the sample mean, an economically significant magnitude. This positive relationship is more pronounced in regions characterized by stronger kinship culture.

There are two key concerns regarding our main findings. The first is the integrity of our baseline results: the robustness of our dependent variables and the baseline regression results. We began by checking the robustness of our patenting variable. Patent applications in China may not accurately capture meaningful innovation activity. To address this concern, we replace the outcome variable with the number of US patents filed by Chinese firms in each county to the United States Patent and Trademark Office (USPTO). These patents plausibly reflect more significant innovations, as obtaining USPTO approval entails a more stringent examination process. We also use the number of eventually granted patents (in China) and their forward citations as alternative outcome variables to capture patent quality. These measures confirm a positive, statistically significant relationship with surname homogeneity. Taken together, these alternative measures yield results consistent with our baseline findings, indicating that surname homogeneity among business owners robustly predicts

[2]We believe this hypothesis is intellectually intriguing because part of the literature (e.g., Posch et al. (2026)) suggests that surname homogeneity proxies for insularity and a lack of diversity, which may hinder local innovation. Different from their use of surname homogeneity to measure ethnic diversity, our use mainly reflects social connections given the ethnically homogeneous population in China.

[3]Over the year, China tried to rebuild market-oriented institutions, particularly after entering the WTO in 2001 (Halverson, 2004). Its ranking in institutional development on the WIPO Global Innovation Index improved from 79th in 2016 to 44th in 2024.

local innovation in the Chinese economy.

We then conduct a battery of additional robustness checks to validate our baseline results. First, we add additional control variables to account for potential confounders, including economic development, industry structure and homogeneity, clan culture, agricultural diversity, and the presence of treaty ports. Second, we use the Herfindahl–Hirschman Index as an alternative measure of surname-based connections. Third, beyond employing alternative patenting measures, we address concern that a government patent subsidy policy may have encouraged patent trading and inflated patent application counts, by excluding the sub-period affected by these policies. We also lag the independent variables by one to five years to account for the possibility that changes in social connections influence local innovation over a longer horizon. Fourth, we adopt three approaches to mitigate concerns about serial dependence in surname homogeneity. Finally, we sequentially exclude three types of special samples: counties where ethnic minority populations exceed 10%, centrally administered municipalities, and special economic zones. Our results remain robust across all specifications.

The second key concern is endogeneity arising from omitted variable bias in surname-based connections. We address this concern in several ways. First, as discussed earlier, we exploit within county variation by including county fixed effects in our baseline regressions, which absorbs time invariant local unobservables, e.g., natural resources. Second, we incorporate a rich set of controls to account for time-varying confounders, e.g., transportation, demographic, and socioeconomic factors. Third, we implement an instrumental-variable (IV) strategy that exploits cross-county variation in proximity to historical peasant rebellions from the Ming Dynasty (1368–1644) to the Qing Dynasty (1644–1911). Prior studies show that exposure to violence can increase the importance of local social cohesion (Gilligan et al., 2014; Bellows and Miguel, 2009; Bauer et al., 2016). In the Chinese context, historical peasant rebellions may have weakened formal institutions and increased local insecurity, thereby generating different trajectories in the importance of surname-based trust for collaboration. We therefore use geographic proximity to these rebellions as a plausible source of exogenous variation,[4] and present a robust surname connection-innovation relation in which proximity to peasant rebellions explains differential time trends in local surname-based connections.[5]

To provide more direct evidence for a positive effect of surname-based connection on local innovation, we further perform "surname-specific" analyses. In particular, we decompose county-level innovation by the surname of business owners. We hypothesize that the relationship between patenting and the number of firms whose owners share the same surname becomes stronger as surname

[4]In times of rebellion and social disorder, people tend to rely more heavily on trust and cohesion within family, kinship, and clan networks. Thus, if a peasant rebellion altered local social norms, its effects would likely propagate through these networks, which are bounded by family, kinship, and clan ties. Although we can never perfectly establish the exclusion restriction, we argue that it is likely to hold for the following reasons: first, rebellions occurred centuries ago and are therefore unlikely to directly affect formal institutional environments today (Bauer et al., 2016). Second, as clanship has been the most long-lasting core in local community in Chinese history, local people's perceptions of surname-based connections today are shaped by their ancestors' experiences with clanship. Thus, if peasant rebellions were to alter any informal institutions, then such an effect should affect surname-based networks primarily. Third, we have included an extensive list of demographic and economic control variables at the county level to mitigate the influences of all other informal institutions.

[5]Our results are also robust to an alternative IV based on the distance to the nearest Zhu Xi academy. The corresponding results are available upon request.

homogeneity increases. Our estimation is implemented in a surname-county-year panel, in which we regress patent applications by business owners with a specific surname on the number of such firms, interacted with surname homogeneity. This setting also allows us to control for surname-by-county fixed effects, which absorb all time-invariant heterogeneity specific to each surname-county pair, year-by-surname fixed effects that account for surname-specific time-varying shocks (e.g., population growth and immigration), county-by-year fixed effects that account for time-varying county-level unobservables. We find that the relationship between surname connections and patenting becomes stronger as surname homogeneity increases, consistent with a causal interpretation linking surname-based connections to innovation.

We next investigate the mechanism underlying our baseline results – surname-based connections foster local innovation by serving as informal institutions. Such connections may promote corporate innovation through three channels. First, they ease financial constraints by providing network-based access to funding, bypassing barriers to bank and other formal lenders' finance. Second, in areas with weak public services, such networks offer quasi-public goods, such as shared development of infrastructure, security, enforcement of rights, and social support, thereby reducing operational uncertainty. Finally, they lower coordination costs (Davis, 2006), enhance resource mobilization, and enable large-scale collective action (Xu and Yao, 2015), effectively substituting for underdeveloped market mechanisms and improving supply chain efficiency.

We test this mechanism by interacting surname homogeneity with a proxy for formal institutional development. We construct this proxy using principal component analysis based on variables that capture the three channels, including bank density, the ratio of total loans and deposits to GDP, public service expenditure, intellectual property rights protection, factor market development, and product market development.[6] Our results indicate that the positive relationship between surname-based connections and innovation diminishes as overall formal institutions strengthen, and as each individual dimension of formal institutional development improves. These results suggest that the effect of social connection is stronger in areas with weaker institutional environments, where innovation collaborations rely more on social networks. We also find that the effect of informal institutions is strongest in the early 2000s, when China's formal institutions were less developed, and declines over time, underscoring the compensatory role of informal institutions.

Finally, we examine the potential consequences of social networks for economic development and entrepreneurial activity. Using nighttime light intensity as a proxy for economic development and the number of newly registered firms as a measure of entrepreneurial activity, we find that surname homogeneity positively explains nighttime light intensity and the number of newly registered firms. Our results suggest that such social connections not only foster local innovation but also contribute to broader local economic activity and entrepreneurial dynamism.

[6]Bank density is defined as the number of banks normalized by population. Details on the other variables and data sources are provided in Section 3.

This paper makes four contributions. First, we contribute to the literature on regional innovation and development by identifying informal institutions as an important driver of local economic outcomes. While prior work emphasizes formal institutions, we show that social connections, proxied by surname homogeneity among business owners, systematically shape innovation and entrepreneurship.[7] Our measure strongly predicts patenting, firm entry, and broader economic development across counties. These results suggest that when formal institutions develop slowly, informal networks can serve as an alternative mechanism for coordination and enforcement, supporting innovation.

Second, we contribute to the literature on diversity, homogeneity, and innovation by showing that their relations depend critically on institutional context.[8] In contrast to Posch et al. (2026), who find that surname diversity fostered innovation in the historical United States, we document a positive relationship between surname homogeneity and innovation in China. Rather than being contradictory, the two findings suggest that the effects of surname structure depend on the underlying social and institutional environment. During 1850–1940, rapid immigration made surname diversity in the United States a proxy for heterogeneous skills, knowledge, and cultures, which facilitated idea recombination and innovation. By contrast, China's much larger and less geographically mobile population means that even within surname networks, members are likely to possess diverse skills and knowledge, while shared surnames foster trust and cooperation, lowering coordination costs and facilitating innovation. Consistent with this interpretation, we find that the positive effect of surname homogeneity weakens as formal institutions strengthen, suggesting that surname networks substitute for formal institutions in supporting innovation. Our findings, therefore, highlight that the relationship between surname homogeneity and innovation depends not only on the diversity of knowledge embedded in social networks but also on the institutional environment in which those networks operate.

Third, in connection with the above point, we provide evidence on the mechanisms through which social connections operate. We show that surname-based networks alleviate frictions associated with weak formal institutions by improving access to finance and public services, and by facilitating transactions in underdeveloped factor markets. These findings extend the literature on informal institutions by documenting concrete channels through which social structure substitutes for formal governance. Our identification leverages a comprehensive panel covering the universe of China's private sector, allowing for more robust inference than prior studies based on limited cross-sections.

[7] Those studies focus on intellectual property rights protection (Ang et al., 2014; Fang et al., 2017), international trade (Hu and Jefferson, 2009; Liu and Qiu, 2016; Chen et al., 2017; Liu and Ma, 2020; Liu et al., 2021), subsidies and policies (Ding and Li, 2015; Fang et al., 2018; Shao and Xiao, 2019; Wei et al., 2021; Zheng and Li, 2020), productivity growth (Fang et al., 2020; Köbig et al., 2022), and infrastructure development (Wang et al., 2018; Gao and Zheng, 2020; Hanley et al., 2022; Liang et al., 2025). For a more comprehensive analysis, please refer to Wei et al. (2017).

[8] Existing evidence is mixed: while some studies emphasize the innovation-enhancing role of diversity (Kerr and Lincoln, 2010; Alesina et al., 2016; Fiszbein, 2022; Posch et al., 2026), others suggest that homogeneity may foster trust and lower transaction costs, and thereby promote development and innovation (Easterly and Levine, 1997; Alesina et al., 2003; Alesina and La Ferrara, 2005; Habyarimana et al., 2007; Ashraf and Galor, 2011). Moreover, Dinesen et al. (2020) conduct a meta-analysis of 87 studies and document an average negative relationship between ethnic diversity and social trust.

Fourth, we document the aggregate economic relevance of social connections for technological development. While prior work demonstrates productivity gains from inventor networks at the individual and firm levels, evidence at the regional level remains scarce.[9] Using granular county-level data, we show that surname-based networks increase patenting and entrepreneurial activity, with meaningful implications for local growth.[10] Our surname-county-year panel further enables surname-specific tests that sharpen identification and provide a scalable framework for studying network effects. More broadly, we introduce surname sharing as a tractable proxy for localized trust, enabling measurement at fine geographic levels and facilitating more precise analysis of its economic consequences.

The rest of the paper is organized as follows. Section 2 introduces the concept of Chinese surnames and explains the relationships between surname-based connections and local innovation. Section 3 details our data, variables, and measurement methods. Section 4 presents our empirical approach and results, including the baseline estimation, robustness checks, and the IV analysis. Section 5 presents the surname-specific analyses which strengthen our causal interpretation. Section 6 presents the mechanisms. Section 7 presents the analysis across different periods. Section 8 reports the results on the overall economic impact. Finally, Section 9 concludes the paper.

2 Conceptual Framework

2.1 Introduction to Chinese Surnames

Most Chinese surnames date back over two thousand years and have diverse origins, including geographic locations, occupations, tribal affiliations, official titles, natural features, names conferred by emperors, and phonetic transliterations of non-Han ethnic names. Ruofu and Yida (1996) compile a total of 11,969 Chinese surnames across all historical periods. Among them, the 504 surnames included in "*The Hundred Surnames*" (*Baijiaxing*, compiled during the Song Dynasty, 960-1279 AD) are significantly more common than others. These 504 surnames comprise 444 single-character and 60 double-character surnames. Over the past two thousand years, Chinese surnames have typically been passed down through the patriarchal line, signifying family lineage and serving as a connection to ancestors and extended family or clan members. This tradition reinforces the importance of filial piety, family ties, and social identity. It is rare for Chinese individuals to change their surnames, except during periods of persecution or other life-threatening events.

A complex interplay of factors, including geographic features, historical migrations, socio-political events, cultural characteristics, and fertility rates, shapes the spatial distribution of Chinese sur-

[9] Faleye et al. (2014) and Chang and Wu (2021) find that the connectedness of CEOs and directors positively influences corporate innovation.

[10] Prior studies on this direction include Knack and Keefer (1997), Zak and Knack (2001), Barro and McCleary (2003), Guiso et al. (2006), Guiso et al. (2008), Xie et al. (2022), Eesley and Lee (2023).

names. In some regions of China, specific surnames dominate entire villages or areas, shaped by geographic isolation or strong patrilineal clan traditions that maintain genealogical records and emphasize shared ancestry. In contrast, regions with higher migration and more inclusive cultural norms tend to display greater surname diversity. Figure 1 illustrates the surname distribution of entrepreneurs in two example counties (Hanjiang and Yingjiang) in 2000. As illustrated in Figure 1a, the surnames "Chen" and "Huang" dominate, accounting for a substantial proportion of the population in Hanjiang. By comparison, Figure 1b demonstrates a comparatively more diverse distribution of surnames in Yingjiang. The difference in surname homogeneity between the two counties may be attributed to Hanjiang's relatively rugged terrain, stronger clan culture, and lower historical migration levels compared to Yingjiang. The distribution of surnames varies not only across counties but also over time within the same county.



(a) Hanjiang in Fujian Province

(b) Yingjiang in Anhui Province

Figure 1: The Distribution of Entrepreneurs' Surnames in Two Counties, 2000

Individuals who share the same surname often demonstrate high levels of mutual support, trust, and understanding for the following reasons. First, historically, individuals with the same surname in a small village often belonged to the same extended family or clan, making surname-based networks effectively kinship networks that shaped economic interactions in China (Greif and Tabellini, 2010). Second, even as explicit kinship ties have attenuated, shared surnames continue to foster perceived common ancestry (Du et al., 1992; Du, 2019; Tan et al., 2021) and may confer reputation-based social capital (Tadelis, 1999; Clark and Cummins, 2014).

Many studies have documented the impact of surname-based networks in promoting mutual support among individuals in China. For example, Xu and Yao (2015) find that village leaders from the two largest surname-based families are more effective at organizing community activities. Similarly, Huang et al. (2023) show that greater surname homogeneity among a firm's board members reduces agency costs and mitigates information asymmetry. Tan et al. (2021) also use the sharing of directors' surnames as a proxy for measuring social ties in organizations. Importantly, this phenomenon is not unique to China. For instance, Cruz et al. (2017) show that people are more likely to support candidates who share the same surname as them during elections in the Philippines. Buonanno and Vanin (2017) find that regions in Italy with more concentrated surnames are associated with lower crime rates, highlighting the role of surname-based networks in fostering social cohesion. Angelucci et al. (2010) demonstrate that surname-based networks influence secondary school

enrollment in Mexico.

2.2 Surname-based Connections and Local Innovation

We propose that the above surname-network effects shape innovation. We proxy these connections at the county level using surname homogeneity among business owners. The relationship between such surname homogeneity and regional corporate innovation, however, is theoretically ambiguous, reflecting both positive and negative forces.

On the positive side, surname homogeneity among business owners can reflect the strength of social networks as informal institutions. Particularly, in regions with underdeveloped formal institutions, such informal institutions can act as substitutes, fostering mutual trust and collaborations (Freedman, 1958; Tsai, 2004). We argue that they can promote corporate innovation through several mechanisms.

First, social networks can substitute for formal institutions in financing innovation (Kinnan and Townsend, 2012). Business owners can raise capital through their social networks when formal financial institutions are unwilling or unable to provide financing. These networks are characterized by long-term relational ties or high levels of mutual trust, which significantly reduce information asymmetry and transaction costs in borrowing and lending. Moreover, these networks tend to offer more flexible terms and personalized support, enabling entrepreneurs to invest in R&D without the stringent requirements typically imposed by formal lenders (Nordman, 2016).

Second, informal institutions can partially substitute for formal institutions in the provision of public goods, such as public safety, infrastructure maintenance, property rights protection, and social security. For example, Buonanno and Vanin (2017) find that surname-based networks reduce crime rates in Italy. These effects arise because such networks facilitate resource pooling, mutual aid, dispute resolution, and risk-sharing among members, thereby fostering a safer and more supportive environment. When governments are inefficient in delivering public services, the role of these social networks becomes more salient. Given the importance of infrastructure and protection of rights for innovation, we expect the public services provision associated with these networks to benefit local innovation activity.

Finally, social networks can complement formal market mechanisms, as they are often associated with higher levels of trust, lower transaction costs, and smoother collaboration. In regions with underdeveloped or inefficient factor and product markets, opportunistic behavior can undermine arm's-length transactional relationships. Individuals often turn to these informal institutions to organize production, facilitate trade, and coordinate economic activities. By doing so, they enhance the efficiency of supply chains and promote the flow of resources,[11] which are conducive to local

[11] As argued by Lim et al. (2024), Chinese communities in Java promote local economic development through their commercial networks

innovation.

On the negative side, high surname homogeneity might hinder innovation for several reasons. First, homogeneous societies often share traditions, beliefs, and norms. In such settings, individuals tend to amplify or reinforce preexisting beliefs, making them reluctant to generate new ideas or explore alternative perspectives. As Alesina and La Ferrara (2005) explain, diversity is more productive at higher levels of development because advanced economies can exploit skill complementarities and manage distributional conflicts through robust institutions. Consistent with this view, Ashraf and Galor (2011) show that cultural homogeneity fostered development in the agricultural era with weak formal institutions, but became a drag on income per capita during industrialization.

Second, homogeneous communities tend to be more conservative and risk-averse, making them less conducive to innovation, which thrives on uncertainty, experimentation, and disruption (Morck and Yeung, 2003). Very importantly, high surname homogeneity often reflects social insularity, with limited exposure to outside influences and fewer opportunities for collaboration, learning, and the adoption of external knowledge or complementary skills (Posch et al., 2026).

Finally, high surname concentration may lead to entrenched power structures, in which a few dominant family clans exercise significant control over resources and decision-making (Morck and Yeung, 2004). Such hierarchies can undermine merit-based competition, discouraging the emergence of innovative ideas that may challenge the status quo.

The above discussion suggests two contradictory relationships between surname homogeneity among business owners and local innovation. Surname homogeneity can be beneficial where formal institutions are weak, as it eases financing constraints, provides quasi-public goods, and supports factor markets. In contrast, where formal institutions are well developed, surname homogeneity may no longer enhance innovation, and may even hinder it. Thus, empirical evidence is called for.

3 Data and Variables

3.1 Surname Homogeneity among Business Owners

Our data on business owners' surnames come from the registration records covering the universe of Chinese enterprises maintained by the State Administration for Industry and Commerce (SAIC). Following Bai et al. (2024), we identify a business owner's surname based on the legal representative's surname. We restrict the analysis to private firms, as legal representatives of state-owned or foreign-controlled firms are unlikely to be founders. From 2000 to 2019, approximately 49 million private firms were registered, representing about 86% of all firms.

We focus on private business owners rather than the general population for two reasons. First, surname homogeneity among business owners more directly captures the connectedness of key

decision-makers, who are the primary drivers or supporters of innovation. Second, comprehensive surname data for the general population are unavailable: the 2005 mini-census includes surnames for only 2.5 million individuals, far fewer than in the firm registry. Reassuringly, Figure A.2 in the Online Appendix shows that surname homogeneity among private business owners is highly correlated with population-level surname homogeneity within the same regions, based on the 2005 mini-census.

We compute surname homogeneity among business owners at the county level, including counties, county-level districts, and county-level cities, rather than at higher administrative levels. It is worth noting that the county level is the smallest administrative unit for which all key variables required in this study are available. Our use of the county as the unit of analysis enables us to exploit rich granular variation across China, which likely provides greater statistical power.

We measure the surname homogeneity among business owners using a modified entropy-based approach. Originally proposed by Shannon (1948) to quantify diversity in information theory, this approach has since been widely applied across various scientific disciplines. In economics, Buonanno and Vanin (2017) and Posch et al. (2026) use this approach to calculate surname diversity in their analyses. Given that our study focuses on surname homogeneity, the conceptual opposite of diversity, we adapt their formula by reversing its sign. Specifically, the surname homogeneity among business owners at the county level is calculated as follows:

$$SurnameHomogeneity_{c,t} = \sum_{k} p_{c,t,k} log_2(p_{c,t,k}) \tag{3.1}$$

where $p_{c,t,k}$ is the fraction of business owners with surname k in county c in year t. We consider business owners of all existing firms in year t. The functional form of $\log_2(p_{c,t,k})$ reflects the idea that rare surname groups contribute more "surprise" compared to more common surname groups. In general, fewer surnames and more concentrated fractions ($p_{c,t,k}$) result in higher homogeneity (i.e., lower diversity). In an extreme case in which all business owners in a county share the same surname in a given year, this measure equals 0, representing the highest possible value, since the measure is negative when $p_{c,t,k} < 1$. In contrast, a larger number of surnames and more evenly distributed fractions ($p_{c,t,k}$) lead to lower homogeneity (or higher diversity).

To assess whether this measure captures the strength of surname-based social connections, we implement validation tests to show that our surname homogeneity is positively related to the ratio of extended families, the share of family managers, and the share of family board members in Section A.2 of the Online Appendix. Another commonly used measure of homogeneity is based on the Herfindahl-Hirschman Index (HHI). However, it has a much more skewed distribution compared to the entropy-based approach. Therefore, we use entropy as our baseline measure and use HHI as a robustness check.

3.2 County-level Corporate Innovation

We measure local innovation using the number of patent applications filed by private firms in each county and year, normalized by the local population in 2000. Patent applications are widely used as a measure of innovation in the literature (e.g., Liu et al., 2021; Liu and Qiu, 2016; Aghion et al., 2009; Hashmi, 2013; Hu and Jefferson, 2009).[12] We consider patent applications filed between 2001 and 2020, as China's accession to the WTO in 2001 has significantly affected the country's innovation landscape. Limitations notwithstanding, patent application data generally reflects innovative activities, as it captures the effectiveness of innovation efforts, including both observable and unobservable inputs (Liu and Qiu, 2016).

In addition, we focus on the applications of invention patents, excluding utility models and design patents, as the latter are less novel than the former. Moreover, we focus exclusively on patent applications by private firms rather than all patent applications within a county, as our measure of surname homogeneity is based solely on the business owners of these firms and thus should correspond to their patent applications. Moreover, this choice prevents us from including innovation outcomes from universities and government labs, which are much less related to social networks.

We acknowledge other measures of innovation activity. Our results remain robust when granted patents are used, as to be discussed in Section 4.2. Moreover, given concerns about the quality of Chinese patents, we also use the number of US patents filed by Chinese firms in each county that are ultimately granted by the United States Patent and Trademark Office (USPTO). These data are collected from PatentsView.[14]

3.3 Control Variables

We include several sets of control variables in our analysis.

Geographic controls. Our geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Average ruggedness and distance to the coast are computed for each county using GIS software. The provincial capital dummy identifies counties located in provincial capitals, while the special economic zone dummy indicates whether a county falls within a designated special economic zone. Provincial capitals typically have greater access to resources and attract more non-local businesses, while special economic zones are designed to promote economic development, including encouraging innovations. These geographic characteristics may influence corporate innovation and surname homogeneity. As they are time-invariant, we account for them by interacting each with year fixed effects. Specifically, for each geographic characteristic, we create a set of interaction terms between that characteristic and the

[12] Data on Chinese patents are sourced from the Global Patent Database (IncoPat), which provides comprehensive information on applicants and the dates of application.[13]

[14] https://patentsview.org

year indicators. This allows the relation between geographic characteristics and the outcome to vary flexibly across years, thereby controlling for differential time trends associated with geography.

Transportation controls and the number of private firms. Transportation controls consist of a high-speed railway station dummy and an airport dummy. The high-speed railway station dummy identifies counties with a high-speed railway station in a given year, while the airport dummy indicates whether the county is located in a city with an airport. We use city-level data for airports, as airports typically serve a broader region than a county. The influence of infrastructure on innovation is well-documented in the literature (Wang et al., 2018; Ma et al., 2021; Hanley et al., 2022). Data on high-speed railway stations are sourced from China Railway, while airport data are hand-collected. We also control for the number of private firms, calculated from the registration records from the universe of Chinese enterprises. This variable is relevant because it correlates with both surname homogeneity among business owners and local innovation.

Demographic controls. Our demographic controls include population, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Population reflects a county's scale, while average years of schooling and urbanization rate capture education and economic development levels, respectively. The fraction of minorities and migration rate may be correlated with surname homogeneity. These variables are sourced from the 2000 census. Similar to geographic controls, these variables are also interacted with year fixed effects since they are time-invariant.

Socioeconomic controls. Finally, we control for several socioeconomic variables that might influence corporate innovation, including a marketization index, bank density, the number of universities, and the number of elite universities (Project 211 universities). The marketization index, which varies by city, is sourced from the China Market Index Database.[15] It reflects the relationship between the government and the market, as well as the development of the market. It has been used in prior studies such as Li et al. (2011), Fan et al. (2013), and Xie and Yuan (2025). Bank density, defined as the number of banks normalized by population, serves as a measure of financial development. The data are sourced from the China Banking and Insurance Regulatory Commission.[16] The number of universities and the number of elite universities are used to measure educational development. The number of elite universities is basically time-invariant and is interacted with year fixed effects, similar to geographic controls.

3.4 Variables for Heterogeneous Institutional Development

To deepen our understanding of the relationship between innovation and informal institutions, proxied by surname homogeneity among business owners, we consider three sets of measures of institutional development. We then combine these variables into an index of formal institutional development using principal component analysis, and use this index as a moderating variable in the

[15] https://cmi.ssap.com.cn/
[16] https://www.cbirc.gov.cn/en/view/pages/index/index.html

analysis.

Financial Development. We use two proxies to measure financial development. The first proxy is bank density, as mentioned in Section 3.3, which varies across counties. The second proxy is the ratio of total loans and deposits from formal financial institutions to GDP, which captures the overall scale of financial intermediation (King and Levine, 1993; Beck et al., 2000; World Bank, 2015). The data is sourced from China's Provincial Statistical Yearbooks and varies at the provincial level.

Quality of Public Administration. We also consider two proxies for the quality of public administration. The first is public service expenditure, normalized by GDP, measured at the provincial level in 2010. Data on public service expenditure are obtained from the China Statistical Yearbooks and are available only from 2007 onward. The second proxy is the intellectual property rights protection index, which is a subcomponent of the broader marketization index from the China Market Index Database.[17] This index is also available only at the provincial level.

Development of Factor Markets. This measure is based on the indices of the factor and product market development. Li et al. (2009) provide details on how these indices are constructed.[18] These indices are only available at the provincial level and are subcomponents of the broader marketization index. In contrast to the broader marketization index, which captures the relationship between the market and the government, these indices specifically focus on the factor and product markets.

3.5 Other Variables

In Section 4.2, we conduct robustness checks by controlling for additional variables. City-level GDP per capita is sourced from the prefecture-level city statistical yearbooks. To allow for the possibility that surname homogeneity may be correlated with clan culture, we include a control for clan culture, measured by the number of genealogies per 1,000 people. Genealogy data are sourced from the Shanghai Library Chinese Genealogy Knowledge Service Platform.[19] Agricultural diversity is known to affect innovation (Fiszbein, 2022). To accommodate this possibility, we include it as a control. We measure agricultural diversity based on the caloric content of five crops. Data on the caloric content of each crop are from the Food and Agriculture Organization (FAO) of the United Nations.[20]

In Section 4.4, we conduct an IV analysis based on distance to the nearest historical peasant rebellion during the Ming and Qing dynasties, using data from Dincecco and Wang (2018). Specifically, for each county, we compute its distance to all historical peasant rebellions during the Ming and

[17] https://cmi.ssap.com.cn/

[18] The factor market development index measures how efficiently capital, labor, foreign investment, and technology are allocated through market mechanisms and whether these factors can flow toward more productive firms and regions. The product market development index captures the extent to which goods can be freely priced and traded across regions. Specifically, it measures whether the prices of goods are determined by market demand and supply, as well as the extent of regional trade barriers.

[19] https://jiapu.library.sh.cn/#/

[20] https://data.apps.fao.org/

Qing dynasties and use the shortest distance as a measure of exposure to historical rebellions. Our results are robust to using alternative functional forms that account for exposure to all historical rebellions.

In Section 8, we further consider economic development and entrepreneurship. We measure economic development using average nighttime light intensity. The data originate from DMSP/OLS (Defense Meteorological Satellite Program/Operational Linescan System) and NPP-VIIRS (National Polar-orbiting Partnership/Visible Infrared Imaging Radiometer Suite). We measure entrepreneurial activity in each county using the number of new business registrations of Chinese firms from SAIC.

Finally, the summary statistics for all the variables discussed in this section are presented in Table A.1 in the Online Appendix.

4 Empirical Strategy and Results

4.1 Baseline Estimation

To investigate the relationship between surname homogeneity among business owners and corporate innovation at the county level, we employ the following econometric specification:

$$Y_{c,t+1} = \alpha + \beta SurnameHomogeneity_{c,t} + X'_{c,t}\gamma + \mu_c + \delta_t + \varepsilon_{c,t}, \tag{4.1}$$

where $Y_{c,t+1}$ is the number of patent filings by private firms in county c in year $t+1$, normalized by 1,000 local population. It covers the period from 2001 to 2020. We winsorize the outcome variable at the 5th and 95th percentiles to mitigate the influence of outliers.[21] We have considered alternative measures of local innovation including patent filings normalized by the number of private firms in the county, unscaled number of patent filings, the number of granted patents, and the number of US patents in our robustness checks.

$SurnameHomogeneity_{c,t}$, as in Equation 3.1, is the surname homogeneity among business owners in county c in year t, using data from all existing private firms in year t. The time-series variation in our main explanatory variable ($SurnameHomogeneity_{c,t}$) allows us to control for county fixed effects in our specification, which allows us to subsume time-invariant county-level omitted variables and focus exclusively on within-county changes in innovation output driven by within-county changes in surname homogeneity. By doing so, we effectively eliminate all time-invariant county-level unobservables. This approach aligns with that of Xu and Yao (2015), leveraging the strengths of a panel data structure.[22]

[21] Our results remain robust when the outcome variable is winsorized at the 1st and 99th percentiles, as shown in Table A.7 in the Online Appendix.

[22] We acknowledge that our explanatory variable ($SurnameHomogeneity_{c,t}$) is fairly stable in time and may exhibit serial dependence. We shall address the concern in Section 4.2.

$X'_{c,t}$ indicates control variables. There are four sets of controls, as discussed in Section 3.3: geographic controls; transportation controls and the number of private firms; demographic controls; and socioeconomic controls. Time-invariant controls, including geographic variables, demographic variables, and the number of elite universities, are interacted with year dummies to account for the time-varying effects of these fixed characteristics.

μ_c represents the county fixed effects; δ_t represents the year fixed effects; and $\varepsilon_{c,t}$ is the error term. Tibet and Qinghai are excluded from the estimation due to the presence of numerous omitted variables. We use standard errors clustered two-way by county and year to account for potential correlation in the error terms within the same county across years and within the same year across different counties.

Table 1 presents the baseline estimates. Column (1) begins with a specification that includes only county and year fixed effects, which prevents the "bad controls" issue. In column (2), we add geographic controls, and in column (3), we incorporate transportation controls and the number of private firms. Column (4) further includes demographic variables, while column (5) incorporates socioeconomic variables, comprising the full set of controls, as explained in Section 3.3. Column (5) represents our preferred specification, corresponding to Equation 4.1. The estimates indicate that surname homogeneity is positively correlated with patent filings, with results statistically significant at the 1% level across all five columns. According to our preferred specification in column (5), a one-unit increase in surname homogeneity results in a 0.028 increase in patent filings per 1,000 people. In other words, a one-standard-deviation increase in surname homogeneity is associated with a 21.1% rise in patent filings,[23] relative to the sample period's mean value. This represents an economically meaningful magnitude.

For completeness, we report in column (1) of Online Appendix Table A.8 the coefficients on the time-varying controls from our preferred specification (column (5) of Table 1).[24] Time-invariant controls are included through their interactions with year fixed effects.[25] We also show in column (2) of Table A.8 the relationship between surname homogeneity among business owners and time-varying controls.

4.2 Robustness Checks

To ensure the robustness of our results, we consider an extensive list of approaches to address potential concerns. We briefly discuss them in the main text and leave all detailed discussions to Online Appendix Section A.3. Our robustness checks include (a) more control variables for potential omitted variable issues; (b) an alternative definition of surname homogeneity; (c) alternative

[23] $(0.028 \times 0.587)/0.078 = 21.1\%$

[24] Column (1) of Online Appendix Table A.8 and column (5) of Table 1 are based on the same regression. The only difference is that we report the coefficients on time-varying controls in column (1) of Online Appendix Table A.8.

[25] The corresponding coefficients are not reported because the full set of interaction coefficients would substantially increase the length of the tables and make them difficult to read.

Table 1: Baseline Results

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.106***	0.057***	0.041***	0.029***	0.028***
	(0.012)	(0.006)	(0.005)	(0.006)	(0.006)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.627	0.657	0.685	0.703	0.719
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

measures for innovation outcomes; (d) the serial dependence in surname homogeneity; and (e) the special regions.

More controls for potential omitted variable issues. Although our baseline estimation includes many controls and fixed effects, we acknowledge that the omitted-variable problem may persist. To address this issue and ensure robustness, we further include the following additional variables that could confound our estimation: (i) economic development; (ii) industry homogeneity; (iii) industry structure; (iv) clan culture; (v) agricultural diversity; and (vi) exposure to foreign culture. We do not include these variables in our baseline regressions due to potential issues with bad-control or collinearity concerns. When we include those additional control variables, the positive relation between surname homogeneity and local innovation remains strong and significant.

Another measure of surname homogeneity: HHI. We also employ an alternative measure of concentration, the Herfindahl-Hirschman Index (HHI) as below:

$$SurnameHomogeneity_{c,t} = \sum_{k} p_{c,t,k}^2, \tag{4.2}$$

where $p_{c,t,k}$ is the fraction of business owners with surname k in county c in year t. As this measure is highly skewed with many outliers, we winsorize the data at the 5% and 95% levels. We find that this measure of surname homogeneity continues to predict corporate innovation positively.

Alternative measures for innovation outcomes. Patent applications in China may not accurately capture meaningful innovation. To address this concern, we conduct several additional analyses by

replacing the dependent variable in our baseline specification in Equation 4.1 with (i) the number of patents filed by Chinese firms in each county that are ultimately granted by the United States Patent and Trademark Office (USPTO); (ii) the number of patents filed by those firms which are eventually granted in China; (iii) the total number of forward citations received by those patents.

Additionally, firms may file low-quality patent applications for patent subsidy or signaling. Wei et al. (2021) document that China's patent subsidy policy–the InnoCom program introduced in 2008–encouraged patent trading among firms and may have artificially inflated patent application counts, thereby weakening their validity as a proxy for corporate innovation. To address this concern, we restrict the sample to the pre-policy period from 2001 to 2008.

Furthermore, to account for the time lag between innovation and patent applications, we also consider the cumulative number of patent filings over longer horizons. Specifically, we replace the outcome variable with the cumulative number of patent filings from year $t+1$ through years $t+2$, $t+3$, $t+4$, and $t+5$.

Across these alternative measures and empirical approaches, the results provide robust and consistent evidence that surname homogeneity predicts local innovation outcomes.

Serial dependence in surname homogeneity. Another concern is serial dependence in surname homogeneity, as within-county variation evolves slowly over time. We address this concern in three ways. First, we employ Newey-West standard errors to account for serial correlation. Second, we adopt a long-difference specification that effectively mitigates serial dependence by examining the relationship between changes in surname homogeneity and changes in patent filings over a 20-year period. Third, we implement sparse sampling by restricting the sample to every third or fourth year between 2000 and 2018, which reduces persistence in the independent variable. Across these specifications, we find consistent results.

Special regions. We exclude three types of special samples that could potentially affect our results. The first type consists of counties with a significant minority population. The second type includes municipalities directly administered by the central government (i.e., Beijing, Shanghai, Tianjin, and Chongqing). The third type comprises special economic zones (i.e., Shenzhen, Zhuhai, Shantou, Xiamen, and Hainan), whose distinct institutional arrangements may foster greater innovation through alternative approaches. The results from excluding these three types of special samples remain consistent.

4.3 Heterogeneity by Kinship Culture

The effect of surname homogeneity is likely to be stronger in regions with a deeper kinship culture, where shared surnames are more deeply embedded in clan institutions and kinship norms. In such settings, surname homogeneity is more likely to reflect meaningful social ties, mutual obliga-

tions, and informal enforcement, thereby strengthening its explanatory power for local innovation. Following the literature that uses genealogy records to measure local clan and kinship culture (Chen et al., 2022; Xie and Yuan, 2025), we classify counties as having stronger kinship culture if they have at least one genealogy book, and as having weaker kinship culture otherwise. This classification divides the sample into two groups of roughly similar size.

Table 2 reports the heterogeneity results. The coefficient of surname homogeneity is positive and statistically significant in both strong and weak subsamples, but it is consistently larger in regions with stronger kinship culture. In the preferred specification in column (5), the coefficient is 0.046 in regions with genealogy records, compared with 0.016 in regions without genealogy records. The difference between the two coefficients is statistically significant: the t-statistic is 2.91, with a two-sided p-value of approximately 0.004. These results support the interpretation that although surname homogeneity promotes innovation partly through kinship-based social ties, its overall effect extends beyond kinship.

Table 2: Heterogeneity Analysis by Kinship Culture

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Panel A: Regions with Stronger Kinship Culture					
Surname homogeneity	0.118***	0.074***	0.055***	0.057***	0.046***
	(0.016)	(0.011)	(0.009)	(0.010)	(0.009)
Observations	25,920	25,820	25,820	25,820	25,080
R^2	0.601	0.635	0.666	0.684	0.709
Mean dependent vars.	0.074	0.074	0.074	0.074	0.074
Panel B: Regions with Weaker Kinship Culture					
Surname homogeneity	0.106***	0.056***	0.040***	0.014**	0.016***
	(0.012)	(0.007)	(0.006)	(0.006)	(0.005)
Observations	27,815	27,795	27,795	27,795	27,415
R^2	0.650	0.683	0.707	0.728	0.741
Mean dependent vars.	0.081	0.081	0.081	0.081	0.082
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: County-level OLS regressions. We classify counties as having stronger or weaker kinship culture based on the presence of genealogy books. The outcome variable is the number of patent filings by private firms in a county, normalized by 1000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

4.4 Instrumental Variable Estimation

To address remaining endogeneity concerns, we employ an instrumental-variable (IV) strategy that exploits cross-county variation in proximity to historical peasant rebellions from the Ming to the Qing Dynasties.[26] Prior studies in other countries show that exposure to violence can increase local social cohesion (Gilligan et al., 2014), promote collective action (Bellows and Miguel, 2009), and strengthen local cooperation (Bauer et al., 2016). In the context of China, historical peasant rebellions could weaken formal institutions and increase local insecurity, thereby raising the importance of surname-based connections that can provide protection, informal enforcement, and coordination. Historically, sharing the same surname has served as a visible and persistent basis for local social connectedness in China. We therefore propose that historical peasant rebellions, by increasing the need for trusted relationships, may have strengthened same-surname networks.[27] We use proximity to these rebellions as a source of plausibly exogenous variation in the strength of surname-based connections.

Peasant rebellions likely satisfy the exclusion restriction as they occurred centuries ago and are therefore unlikely to directly affect formal institutional environments today (Bauer et al., 2016). We acknowledge that we can never rule out the possibility that peasant rebellions also influence other informal institutions unrelated to surname-based connections, which may then violate the exclusion restriction. Nevertheless, as clanship has long been an enduring feature of local communities in Chinese culture (Zhang, 2020; Xie and Yuan, 2025), modern surname-based connections embedded in local habits and social norms may be shaped by ancestral experiences with clan organization. Thus, if peasant rebellions were to change any informal institutions, then they most likely impacted surname-based networks first. Also, our extensive list of county-level demographic and economic control variables helps mitigate the influence of other informal institutions.

Our measure of surname-based connections is constructed for the period 2000–2019 and is allowed to vary over time. Historical peasant rebellions may have shaped the long-run formation of local informal social networks. These historically rooted social structures may, in turn, generate different trajectories of surname-based connections among business owners during the period of market development. Therefore, we interact it with year to examine whether historical proximity to peasant rebellions predicts differential time trends in surname homogeneity.[28] Accordingly, we estimate the following first-stage regression:

$$SurnameHomogeneity_{c,t} = \alpha + \beta Distance_c \times Year_t + X'_{c,t}\gamma + \mu_c + \delta_t + \varepsilon_{c,t} \tag{4.3}$$

[26] We focus on these two relatively recent dynasties because rebellions occurring in much earlier periods are likely to have weaker long-run effects.

[27] In times of rebellion and social disorder, people tend to rely more heavily on trust and cohesion within family, kinship, and clan networks. Thus, if a peasant rebellion altered local social norms, its effects would likely propagate through these networks, which are bounded by family, kinship, and clan ties.

[28] For example, if the distance is 100 and the sample year is 2000, the value of the instrument is 100×2000.

where $Distance_c$ denotes the minimum distance from county c to the nearest peasant rebellion during the Ming Dynasty to the Qing Dynasty, using data from Dincecco and Wang (2018). This approach has been used in other studies when the underlying variable used to construct the instrument is time-invariant, while the explanatory variable is time-varying (Manacorda and Tesei, 2020; Guriev et al., 2021; Fenske et al., 2023). In such settings, researchers often interact time-invariant variables with time trends to generate time-series variation for identification. The control variables and fixed effects are identical to those included in our baseline specification in Equation 4.1.

Table 3: IV: Historical Peasant Rebellions

Second-stage dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.342***	0.491***	0.451**	0.593***	0.549***
	(0.070)	(0.158)	(0.163)	(0.182)	(0.170)
First-stage dependent variable	Surname homogeneity				
	(1)	(2)	(3)	(4)	(5)
Distance to rebellion × Year	-0.498***	-0.246***	-0.213***	-0.246***	-0.246***
	(0.041)	(0.042)	(0.061)	(0.061)	(0.060)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.832	0.867	0.869	0.898	0.899
Kleibergen-Paap F	47.662	14.912	12.300	16.354	16.884
tF confidence interval	[0.19, 0.49]	[0.04, 0.95]	[-0.06, 0.96]	[0.09, 1.10]	[0.08, 1.02]
p-value	0.000	0.034	0.081	0.022	0.021

Notes: County-level 2SLS regressions. The outcome variable in the second stage is the number of patent filings by private firms in a county, normalized by 1,000 local population. Time spans from 2001 to 2020 for the outcome variable. The IV is the interaction between the minimum distance (in 10,000 km) to historical peasant rebellions from the Ming Dynasty to the Qing Dynasty and year. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table 3 reports the IV estimates. The first-stage results show that the coefficients on the instrument are statistically significant at the 1% level across all specifications. The negative coefficients indicate that counties closer to historical rebellions experienced a relatively stronger evolution of surname homogeneity over time. In column (5), our preferred specification with the full set of controls, the Kleibergen–Paap F-statistic is 16.884. Although it is above the conventional threshold of 10, conventional t-ratio inference for IV can still over-reject, and the associated confidence intervals for the second-stage coefficients can under-cover when instruments are not sufficiently strong (Lee et al., 2022). We therefore follow Lee et al. (2022) and construct 95% tF confidence intervals for the

second-stage IV estimates using the first-stage F-statistic. The results are reported at the bottom of Table 3. In column (5), the 95% tF confidence interval excludes zero, and the corresponding tF-adjusted p-value is below 0.05, suggesting that our inference is robust to concerns about weak instruments. In the second stage, surname homogeneity has a positive, statistically significant effect on innovation across all columns. The consistency of these estimates across specifications reinforces both the positive direction and the robustness of our baseline results.

The IV estimates are larger in magnitude than the corresponding OLS estimates, which may be due to the following reasons. First, the gap may reflect a local average treatment effect (LATE). The IV strategy identifies the causal effect for complier counties, those in which surname-based connections evolved in response to proximity to historical peasant rebellions. In these counties, informal institutions are more economically salient, so increases in surname homogeneity more plausibly capture effective surname networks that reduce transaction costs and facilitate trust, cooperation, and mutual support. Consequently, innovation responds more strongly, leading the IV estimates to exceed the OLS estimates.

Second, measurement error likely attenuates the OLS estimates. Surname homogeneity may imperfectly capture the strength of surname-based connections. Such classical measurement error would bias the OLS estimates toward zero. The inclusion of county fixed effects may further exacerbate this attenuation, as identification relies solely on limited within-county variation in a slow-moving variable.

Third, reverse causality may bias the OLS estimates downward. Innovation-intensive counties may attract outside entrepreneurs or skilled workers, mechanically lowering observed surname concentration among business owners even as innovation increases. In contrast, the IV strategy isolates exogenous variation in surname-based connections driven by exposure to historical peasant rebellions, thereby purging this reverse-causality channel and yielding larger estimated effects.

5 Surname-County Level Analysis

To provide more direct evidence for a positive relation between surname homogeneity and local innovation, we resort to surname-specific analyses. We further decompose the data to construct the following surname-county level variables for each year: (i) the number of firms controlled by a given surname within a county, and (ii) the number of patent filings generated by firms whose owners share this specific surname. Naturally, for a given surname in a given county, controlling a larger number of firms is mechanically associated with a greater volume of patent filings. Given the facilitating role of social networks, we hypothesize that greater surname homogeneity disproportionately raises patenting by business owners from larger surname groups.

To examine this hypothesis, we estimate the following specification using a surname-county-year

panel:

$$Y_{c,s,t+1} = \alpha + \beta_1 SurnameHomogeneity_{c,t} \times Firms\ controlled\ by\ surname\ s_{c,s,t} + \beta_2 Firms\ controlled\ by\ surname\ s_{c,s,t} + \mu_{s,c} + \delta_{s,t} + \gamma_{c,t} + \varepsilon_{c,s,t}, \tag{5.1}$$

where $Y_{c,s,t+1}$ is the number of patent filings by private firms controlled by surname s in county c in year $t+1$. The outcome variable is winsorized at the 1st and 99th percentiles to mitigate the influence of outliers. Our results are also robust to alternative functional forms that address outliers without winsorization, such as the log transformation and the IHS transformation; these results are reported in Online Appendix Tables A.31 and A.32. Because the outcome variable contains many zeros, we additionally estimate Poisson regressions, with the results reported in Online Appendix Table A.33.

Firms controlled by surname $s_{c,s,t}$ is the number of private firms controlled by surname s in county c in year t, measured in thousands; $\mu_{s,c}$ denotes surname-by-county fixed effects, which absorb all time-invariant heterogeneity specific to each surname-county pair; $\delta_{s,t}$ denotes surname-by-year fixed effects that account for surname-specific time-varying shocks (e.g., population growth and immigration); and $\gamma_{c,t}$ denotes county-by-year fixed effects, which absorb all control variables included in the baseline county-level specification as well as all time-varying county-level unobservables. This rich list of fixed effects helps us effectively mitigate the concerns about potential confounders. Our standard errors are two-way clustered by county and year.

Our main interest is in the coefficient on $SurnameHomogeneity_{c,t} \times$ *Firms controlled by surname* $s_{c,s,t}$, the interaction term. A significantly positive coefficient would indicate that the connection-innovation relationship is stronger among owners who share a more dominant surname. This pattern suggests that the baseline effect is surname-specific and supports a causal interpretation by helping rule out alternative explanations.

The results are reported in Table 4. In column (1), we include county, year, and surname fixed effects only. In column (2), we include county-by-surname and year-by-surname fixed effects. Column (3) presents our preferred specification, which controls for all high-dimensional fixed effects. County-by-year fixed effects absorb surname homogeneity, which varies at the county-year level, in this column. Across columns, the coefficients on the interaction terms are consistently positive, suggesting that stronger social networks enhance the innovation gains for business owners who share a majority surname. Specifically, in column (3), the coefficients on the interaction term and on *Firms controlled by surname* $s_{c,s,t}$ are jointly significant, with an F-statistic of 19.72. The coefficient on the interaction term is 2.113, indicating that for each one-unit increase in county-level surname homogeneity, the effect of the number of firms controlled by owners with the same surname on patent filings increases by 2.113 patents. This represents a sizable increment relative to the coefficient on the number of firms (14.746), suggesting that the positive correlation between the number of firms

controlled by a given surname and those firms' patent filings is strengthened when social networks are stronger.

Table 4: Surname-County Level Analysis

Dependent variable	Patent filings by firms controlled by surname s		
	(1)	(2)	(3)
Surname homogeneity $\times$ Firms controlled by surname s	1.917***	1.608***	2.113***
	(0.231)	(0.272)	(0.251)
Firms controlled by surname s	14.222***	11.785***	14.746***
	(1.506)	(1.794)	(1.674)
Surname homogeneity	0.115***	0.422***	
	(0.016)	(0.045)	
County FE	Yes	No	No
Year FE	Yes	No	No
Surname FE	Yes	No	No
Surname-County FE	No	Yes	Yes
Year-Surname FE	No	Yes	Yes
County-Year FE	No	No	Yes
Observations	8,791,391	8,613,486	8,613,445
R^2	0.218	0.436	0.483
Mean dependent vars.	0.159	0.162	0.162

Notes: Surname-county level OLS regressions. The outcome variable is the number of patent filings by firms controlled by a given surname in each county-year. Firms controlled by surname denotes the number of firms controlled by surname s in county c in year t, measured in thousands of firms. The time span is from 2001 to 2020 for the outcome variable. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

6 Mechanisms

6.1 Formal Institutions: County-Level Analysis

To provide corroborative evidence that surname homogeneity among business owners promotes local innovation, and consistent with the view that social networks can substitute for weak formal institutions, we examine these theoretically motivated mechanisms. Social networks may improve members' access to finance, provide public goods, reduce the risk of property-rights violations, lower coordination costs, and facilitate the organization of production. However, the role of such networks in supporting innovation is likely to weaken as formal institutions in these dimensions become stronger.

We therefore construct an index of formal institutional development using principal component analysis. The index is based on variables measuring bank density, the ratio of total loans and deposits to GDP, the quality of public administration, intellectual property rights protection, factor market development, and product market development, as described in Section 3. Specifically, we estimate the following equation:

$$Y_{c,t+1} = \alpha + \beta_1 SurnameHomogeneity_{c,t} \times FormalInstitution_{c,t} + \beta_2 SurnameHomogeneity_{c,t} + \beta_3 Formal_{c,t} + X'_{c,t}\gamma + \mu_c + \delta_t + \varepsilon_{c,t} \tag{6.1}$$

where $FormalInstitution_{c,t}$ represents the index of formal institutional development in county c in year t. Other notations are consistent with those in our baseline estimation in Equation 4.1.

Table 5: Mechanism: Formal Institution

Dependent variable	Patent filings per 1000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity × Formal institution	-0.031**	-0.032**	-0.024**	-0.027***	-0.025***
	(0.012)	(0.012)	(0.009)	(0.008)	(0.008)
Surname homogeneity	0.134***	0.108***	0.080***	0.081***	0.076***
	(0.024)	(0.022)	(0.019)	(0.017)	(0.016)
Formal institution	-0.139*	-0.155**	-0.110**	-0.131***	-0.125***
	(0.069)	(0.066)	(0.052)	(0.044)	(0.041)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	50,953	50,839	50,839	50,839	49,870
R^2	0.640	0.665	0.691	0.707	0.720
Mean dependent vars.	0.072	0.072	0.072	0.072	0.072

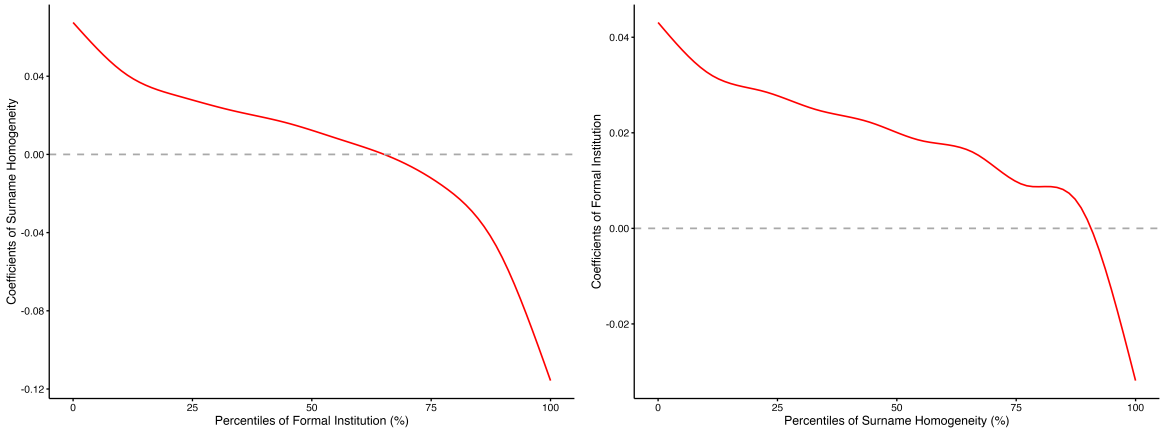
Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1000 local population. The time span is from 2001 to 2019 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table 5 presents the results. The interaction between surname homogeneity and formal institutions is negative and statistically significant across all columns. This suggests that the positive relationship between surname homogeneity and local innovation becomes weaker when formal institutions are stronger.

In column (5), our preferred specification, the F-statistic for the joint significance of surname homogeneity, formal institutions, and their interaction is 12.86, with a p-value of 0.000, indicating that these variables are jointly significant. We note that the marginal effect of formal institutions on local innovation depends on both the coefficient on formal institutions and the coefficient on the interaction term listed in the first row. At the mean value of surname homogeneity, which is -5.699, this marginal effect is -0.125 + (-5.699)*(-0.025) = 0.0175, which is positive. Therefore, although the coefficient on formal institutions alone is negative, the overall marginal effect of formal institutions

is positive at the sample mean of surname homogeneity.

Using the estimates from column (5), Figure 2a plots the implied effect of surname homogeneity across different levels of formal institutional development. For approximately 65% of the observations, surname homogeneity is positively associated with local innovation. However, this association becomes negative once formal institutions are sufficiently strong. This pattern suggests that surname homogeneity plays a positive role in local innovation primarily when formal institutions are underdeveloped. Once formal institutions become sufficiently developed, the role of surname homogeneity may diminish. Similarly, Figure 2b presents the implied effect of formal institutions across different levels of surname homogeneity, again based on the estimates from column (5). In more than 90% of the observations, formal institutions positively predict local innovation.



(a) Coefficients of Surname Homogeneity

(b) Coefficients of Formal Institution

Figure 2: Coefficients of Surname Homogeneity with Respect to Formal Institutions and Coefficients of Formal Institutions with Respect to Surname Homogeneity

Note: These two figures illustrate how the estimated effect of surname homogeneity varies across percentiles of formal institutional strength, and how the estimated effect of formal institutions varies across percentiles of surname homogeneity.

To provide more detailed evidence on how surname networks substitute for different dimensions of formal institutions, we estimate Equation 6.1 separately by replacing the index of formal institutional development with measures of financial development, the quality of public administration, and market development, respectively. The details are provided in Section A.5 of the Online Appendix. Overall, we find that the role of surname networks declines when financial development is stronger, public administration is of higher quality, and markets are more developed. These findings are consistent with the view that surname networks substitute for formal institutions when such institutions are weak.

6.2 Formal Institutions: Surname-County Level Analysis

Finally, to complete our analysis of the substitution between formal and informal institutions, as captured by surname-based networks, we revisit our finding that patenting increases with the number of firms owned by individuals sharing the same surname as surname homogeneity rises (Section 5). We hypothesize that this relationship attenuates with stronger formal institutions, as local innovation relies less on surname-based ties and the marginal benefit of a surname controlling more firms correspondingly declines. To test this possibility, we estimate a surname-county level regression analogous to Equation 5.1, adding an interaction between the index of formal institutional development and *Firms controlled by surname* $s_{c,s,t}$.

Table 6 reports the results. The coefficient on *Firms controlled by surname* $s_{c,s,t}$ is 15.535, based on the estimates in column (3). This implies that, when both the formal institution index and surname homogeneity are equal to zero, an increase of 1,000 private firms controlled by surname s in a county-year is associated with 15.535 additional patent filings by those firms in the following year. This magnitude is modest, reflecting the fact that the vast majority of firms do not file patents. The estimated interaction coefficients on *Surname homogeneity* and *Firms controlled by surname* $s_{c,s,t}$ are significantly positive, which suggests that patent filings of firms controlled by majority surnames increase more as surname homogeneity becomes stronger and is consistent with Table 4.

In addition, we find that formal institutions promote innovation as suggested in columns (1) and (2). The estimated interaction coefficients on *Formal institution* and *Firms controlled by surname* $s_{c,s,t}$ are consistently negative, indicating that the positive association between the number of firms controlled by majority surnames and their patent filings diminishes as formal institutions become stronger. When the formal institution index increases by one unit, this positive association decreases by 0.055. Put differently, the incremental innovation gains associated with majority surnames are smaller in counties with more developed formal institutional environments. Given the mean value of *Firms controlled by surname* $s_{c,s,t}$, which is 0.024, a one-unit increase in the formal institution index is associated with an increase in the outcome variable of approximately $0.100 + 0.024 \times (-0.059) = 0.099$ in column (2).

By contrast, when surname homogeneity increases by one unit, the positive association increases by 2.164, consistent with our previous findings. These results not only further confirm our baseline estimates, but also provide evidence for the mechanism through which surname homogeneity enhances innovation activity via informal institutions.

Our empirical evidence, together with well-established evidence (e.g., Posch et al. (2026)), indicates that the relationship between surname homogeneity and innovation could be affected by the demographic environment in which those networks evolve. In societies with high migration and population mobility, surname homogeneity is likely to reflect insular social networks and reduced exposure to diverse ideas, dampening innovation. In contrast, in large populations with limited migration, shared surnames can strengthen trust, facilitate cooperation, and lower the costs of information exchange among individuals with common ancestry. In such settings, surname homogeneity can enhance knowledge diffusion and collective innovation, ultimately contributing to technological progress and economic growth.

Table 6: Formal Institutions: Surname-County Level Analysis

Dependent variable	Patent filings by firms controlled by surname s		
	(1)	(2)	(3)
Firms controlled by surname s	13.852***	12.416***	15.535***
	(1.512)	(1.769)	(1.722)
Surname homogeneity	0.084***	0.347***	
	(0.018)	(0.046)	
Surname homogeneity $\times$ Firms controlled by surname s	1.749***	1.629***	2.164***
	(0.243)	(0.265)	(0.258)
Formal institution	0.064***	0.100***	
	(0.011)	(0.021)	
Formal institution $\times$ Firms controlled by surname s	-0.098	-0.059	-0.055
	(0.061)	(0.053)	(0.046)
County FE	Yes	No	No
Year FE	Yes	No	No
Surname FE	Yes	No	No
Surname-County FE	No	Yes	Yes
Year-Surname FE	No	Yes	Yes
County-Year FE	No	No	Yes
Observations	7,983,034	7,821,118	7,821,086
R^2	0.216	0.433	0.477
Mean dependent vars.	0.149	0.152	0.152

Notes: Surname-county level OLS regressions. The outcome variable is the number of patent filings by firms controlled by a given surname in each county-year. Firms controlled by surname denotes the number of firms controlled by surname s in county c in year t, measured in thousands of firms. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

7 Time-series Variation of the Relation

As China's overall institutional environment improves over time, the importance from informal institutions may decrease in more recent decades. Between 2001 and 2020, China experienced significant advancements in formal institutions. From Figure A.3a to Figure A.3f in the Online Appendix, we show the changes in the six time-variant proxies for formal institutions in our heterogeneity analyses, reflecting improvements in formal institutions over the past 20 years. Additionally, the Global Innovation Index published by WIPO reflects China's institutional development over time: China ranked 79th in institutional development in 2016, improved to 62nd in 2020, and reached 44th in 2024. These trends highlight the development of China's formal institutions alongside its economic growth.[29]

Given that the positive correlation between surname homogeneity and local innovation is tied to the role of social networks in fostering innovation, we expect this correlation to diminish over time as formal institutions strengthen. To test this hypothesis, we examine how the correlation between surname homogeneity and local innovation evolves over time using the following econometric specification:

$$Y_{c,t+1} = \alpha + \sum_{2000}^{2019} \beta_t SurnameHomogeneity_{c,t} \times Year_t + X'_{c,t}\gamma + \mu_c + \delta_t + \varepsilon_{c,t}, \tag{7.1}$$

where $Year_t$ denotes year dummies, and β_t represents the corresponding year-specific coefficients. Other notations are as previously defined.

This specification allows the correlation between surname homogeneity and local innovation to vary flexibly by year. We plot the year-specific coefficients in Figure 3, which correspond to those reported in column (5) of Online Appendix Table A.34. As shown in the figure, the correlation is positive in the early 2000s, gradually declines over time, and turns negative after 2010. This pattern suggests that the role of social networks in promoting local innovation diminishes over time as formal institutions improve, to the extent that it switches to play a marginally significant negative role since 2011.

[29] Studies show that since China's accession to the World Trade Organization in 2001, the country has strived to advance its institutional development, although the effects could vary sectorally and over time. WTO obligations induced legal and regulatory reforms, strengthened rule-based governance, and promoted sectoral liberalization. These developments enhance firm-level incentives for innovation (Gao et al., 2023).

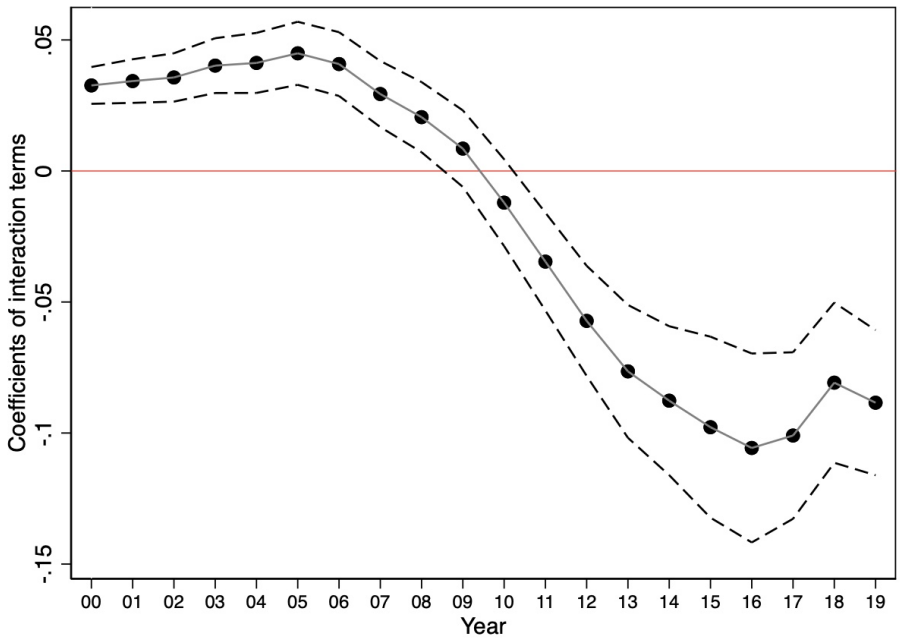


Figure 3: The Correlation over Time

Notes: This figure shows the correlation between surname homogeneity and local innovation over time. We plot the coefficients and the 95% confidence intervals, as reported in column (5) of Online Appendix Table A.34.

8 Overall Economic Impact

Our basic premise is that surname homogeneity among local business owners reflects strong social networks as informal institutions, which foster local innovation by private firms. These surname-based connections enhance economic performance by fostering trust, reducing coordination and transaction costs, and promoting entrepreneurial market entry through improved information sharing, risk mitigation, and stronger social capital.

To test this proposition, we employ Equation 4.1 from our baseline specification, using surname homogeneity as an explanatory variable, but replace the outcome variable with economic development and entrepreneurial activity measures, respectively. Specifically, we measure economic development using nighttime light intensity, and entrepreneurial market entry using the number of newly registered private firms (normalized per 1,000 of the local population). These variables are measured at the county level and vary across years.[30] Other settings remain the same as in our baseline estimation.

Panel A of Table 7 presents the results for the economic development outcome, while Panel B presents the results for entrepreneurial market entry. Across both panels, we find that economic development and entrepreneurial market entry are positively associated with surname homogeneity among business owners. These findings confirm that social networks can be conducive to a variety of economic outcomes, complementing our analysis of their impact on local innovation.

[30] They are also winsorized at the 5% and 95% percentiles to maintain consistency with the baseline estimation.

Table 7: Overall Economic Impact

Panel A: Economic Development					
Dependent variable	Nighttime light				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	129.681***	59.474***	51.208***	35.436***	33.765***
	(18.604)	(15.010)	(14.057)	(11.579)	(10.632)
Observations	53,355	52,995	52,995	52,995	52,355
R^2	0.983	0.984	0.984	0.985	0.986
Mean dependent vars.	932.758	930.153	930.153	930.153	937.111
Panel B: Entrepreneurship					
Dependent variable	# of new firms per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.665***	0.275***	0.209***	0.069	0.050
	(0.076)	(0.054)	(0.051)	(0.059)	(0.050)
Observations	53,695	53,575	53,575	53,575	52,455
R^2	0.820	0.834	0.838	0.857	0.864
Mean dependent vars.	1.646	1.644	1.644	1.644	1.648
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: County-level OLS regressions. The outcome variable in Panel A is nighttime light. The outcome variable in Panel B is the number of newly established private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

9 Conclusion

Leveraging a comprehensive dataset encompassing the universe of Chinese private firms, we compute surname homogeneity among business owners at the county level and use it as a proxy for local social networks. Our analysis reveals that the proxy positively predicts patent filings by private firms at the county level. This relation holds after controlling for county fixed effects and a wide range of control variables, and it has a causal interpretation under an instrumental-variable approach based on proximity to historical peasant rebellions. Our surname-county level analysis shows that surname homogeneity disproportionately benefits business owners from large surname groups, and this effect remains robust to the inclusion of county-by-surname fixed effects, year-by-surname fixed effects, and county-by-year fixed effects. This surname-county level analysis confirms that our baseline effect is surname-specific.

We investigate the mechanisms through which surname-based connections facilitate local innovation. In contexts where formal institutions are weak, such networks can help improve access to finance, enable public service delivery, and support supply chain coordination. However, as formal institutions strengthen, the marginal role of surname-based connections tends to decline. It may even become a hindrance to local innovation if they resist change or entrench existing interests. Our empirical findings support this conditional effect.

Finally, we find that surname homogeneity among local business owners contributes to better economic development and greater entrepreneurial activity, confirming the overall importance of this social capital for economic growth. Overall, our large-scale panel study highlights how informal institutions shape the landscape of innovation activity and entrepreneurship, and eventually influence local economies.

Our findings, though based on the Chinese context, have significant implications for other developing countries. We demonstrate how surname-based social connections promote innovation across three key dimensions of formal institutions, thereby enhancing our understanding of how firms innovate in the absence of robust formal institutions. Additionally, we reveal a dynamic interplay between informal and formal institutions: the role of informal institutions can be crucial during the earlier stages of economic development, even though their role may diminishe later as formal institutions strengthen. This underscores the potential of informal networks in promoting innovation and economic progress when formal structures are still evolving.

References

Aghion, Philippe and Jean Tirole, "The management of innovation," *The Quarterly Journal of Economics*, 1994, *109* (4), 1185–1209.

_ and _, "Opening the black box of innovation," *European Economic Review*, 1994, *38* (3-4), 701–710.

_, Richard Blundell, Rachel Griffith, Peter Howitt, and Susanne Prantl, "The effects of entry on incumbent innovation and productivity," *The Review of Economics and Statistics*, 2009, *91* (1), 20–32.

Alesina, Alberto and Eliana La Ferrara, "Ethnic diversity and economic performance," *Journal of Economic Literature*, 2005, *43* (3), 762–800.

_, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg, "Fractionalization," *Journal of Economic Growth*, 2003, *8*, 155–194.

_, Johann Harnoss, and Hillel Rapoport, "Birthplace diversity and economic prosperity," *Journal of Economic Growth*, 2016, *21*, 101–138.

Allen, Franklin, Jun Qian, and Meijun Qian, "Law, finance, and economic growth in China," *Journal of Financial Economics*, 2005, *77* (1), 57–116.

Ang, James S, Yingmei Cheng, and Chaopeng Wu, "Does enforcement of intellectual property rights matter in China? Evidence from financing and investment choices in the high-tech industry," *Review of Economics and Statistics*, 2014, *96* (2), 332–348.

Angelucci, Manuela, Giacomo De Giorgi, Marcos A Rangel, and Imran Rasul, "Family networks and school enrolment: Evidence from a randomized social experiment," *Journal of Public Economics*, 2010, *94* (3-4), 197–221.

Ashraf, Quamrul and Oded Galor, "Cultural diversity, geographical isolation, and the origin of the wealth of nations," 2011.

Ayyagari, Meghana, Asli Demirgüç-Kunt, and Vojislav Maksimovic, "Firm innovation in emerging markets: The role of finance, governance, and competition," *Journal of Financial and Quantitative Analysis*, 2011, *46* (6), 1545–1580.

Bai, Chong-En, Ruixue Jia, Hongbin Li, and Xin Wang, "Entrepreneurial reluctance: Talent and firm creation in China," *The Economic Journal*, 2024, p. ueae094.

Bai, Ying and James Kai sing Kung, "Surname distance and technology diffusion: the case of the adoption of maize in late imperial China," *Journal of Economic Growth*, 2022, *27* (4), 569–607.

Barro, Robert J and Rachel M McCleary, "Religion and economic growth across countries," *American Sociological Review*, 2003, *68* (5), 760–781.

Bauer, Michal, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts, "Can war foster cooperation?," *Journal of Economic Perspectives*, 2016, *30* (3), 249–274.

Beck, Thorsten, Ross Levine, and Norman Loayza, "Finance and the sources of growth," *Journal of Financial Economics*, 2000, *58* (1–2), 261–300.

Bellows, John and Edward Miguel, "War and local collective action in Sierra Leone," *Journal of Public Economics*, 2009, *93* (11-12), 1144–1157.

Buonanno, Paolo and Paolo Vanin, "Social closure, surnames and crime," *Journal of Economic Behavior & Organization*, 2017, *137*, 160–175.

Carlino, Gerald and William R Kerr, "Agglomeration and innovation," *Handbook of Regional and Urban Economics*, 2015, *5*, 349–404.

Chang, Ching-Hung and Qingqing Wu, "Board networks and corporate innovation," *Management Science*, 2021, *67* (6), 3618–3654.

Chen, Zhiwu, Chicheng Ma, and Andrew J Sinclair, "Banking on the Confucian clan: why China developed financial markets so late," *The Economic Journal*, 2022, *132* (644), 1378–1413.

Chen, Zhiyuan, Jie Zhang, and Wenping Zheng, "Import and innovation: Evidence from Chinese firms," *European Economic Review*, 2017, *94*, 205–220.

Clark, Gregory and Neil Cummins, "Surnames and social mobility in England, 1170–2012," *Human Nature*, 2014, *25* (4), 517–537.

Cruz, Cesi, Julien Labonne, and Pablo Querubin, "Politician family networks and electoral outcomes: Evidence from the Philippines," *American Economic Review*, 2017, *107* (10), 3006–3037.

Davis, Lewis S, "Growing apart: The division of labor and the breakdown of informal institutions," *Journal of Comparative Economics*, 2006, *34* (1), 75–91.

Dincecco, Mark and Yuhua Wang, "Violent conflict and political development over the long run: China versus Europe," *Annual Review of Political Science*, 2018, *21* (1), 341–358.

Dinesen, Peter Thisted, Merlin Schaeffer, and Kim Mannemar Sønderskov, "Ethnic diversity and social trust: A narrative and meta-analytical review," *Annual Review of Political Science*, 2020, *23* (1), 441–465.

Ding, Xuedong and Jun Li, *Incentives for innovation in China: Building an innovative economy*, Routledge, 2015.

Du, Ruofu, Yida Yuan, Juliana Hwang, Joanna Mountain, and L Luca Cavalli-Sforza, "Chinese surnames and the genetic differences between north and south China," *Journal of Chinese Linguistics Monograph Series*, 1992, (5), 1–93.

Du, Xingqiang, "What's in a surname? The effect of auditor-CEO surname sharing on financial misstatement," *Journal of Business Ethics*, 2019, *158* (3), 849–874.

Easterly, William and Ross Levine, "Africa's growth tragedy: policies and ethnic divisions," *The Quarterly Journal of Economics*, 1997, pp. 1203–1250.

Eesley, Charles and Yong Suk Lee, "In institutions we trust? Trust in government and the allocation of entrepreneurial intentions," *Organization Science*, 2023, *34* (2), 532–556.

Faleye, Olubunmi, Tunde Kovacs, and Anand Venkateswaran, "Do better-connected CEOs innovate more?," *Journal of Financial and Quantitative Analysis*, 2014, *49* (5-6), 1201–1225.

Fan, Joseph PH, TJ Wong, and Tianyu Zhang, "Institutions and organizational structure: The case of state-owned corporate pyramids," *The Journal of Law, Economics, & Organization*, 2013, *29* (6), 1217–1252.

Fang, Jing, Hui He, and Nan Li, "China's rising IQ (Innovation Quotient) and growth: Firm-level evidence," *Journal of Development Economics*, 2020, *147*, 102561.

Fang, Lily H, Josh Lerner, and Chaopeng Wu, "Intellectual property rights protection, ownership, and innovation: Evidence from China," *The Review of Financial Studies*, 2017, *30* (7), 2446–2477.

Fang, Lily, Josh Lerner, Chaopeng Wu, and Qi Zhang, "Corruption, government subsidies, and innovation: Evidence from China," 2018.

Feldman, Maryann P, "The new economics of innovation, spillovers and agglomeration: A review of empirical studies," *Economics of Innovation and New Technology*, 1999, *8* (1-2), 5–25.

Fenske, James, Namrata Kala, and Jinlin Wei, "Railways and cities in India," *Journal of Development Economics*, 2023, *161*, 103038.

Fiszbein, Martin, "Agricultural diversity, structural change, and long-run development: Evidence from the United States," *American Economic Journal: Macroeconomics*, 2022, *14* (2), 1–43.

Freedman, Maurice, "Lineage Organisation in Southeast China: Fukien and Kwangtung," *London: University of*, 1958.

Gao, Henry, Damian Raess, and Ka Zeng, *China and the WTO: A twenty-year assessment*, Cambridge University Press, 2023.

Gao, Yanyan and Jianghuai Zheng, "The impact of high-speed rail on innovation: An empirical test of the companion innovation hypothesis of transportation improvement with Chinaâs manufacturing firms," *World Development*, 2020, *127*, 104838.

Gilligan, Michael J, Benjamin J Pasquale, and Cyrus Samii, "Civil war and social cohesion: Lab-in-the-field evidence from Nepal," *American Journal of Political Science*, 2014, *58* (3), 604–619.

Greif, Avner and Guido Tabellini, "Cultural and institutional bifurcation: China and Europe compared," *American Economic Review*, 2010, *100* (2), 135–140.

Griliches, Zvi, "R&D and productivity: Measurement issues and econometric results," *Science*, 1987, *237* (4810), 31–35.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, "Does culture affect economic outcomes?," *Journal of Economic perspectives*, 2006, *20* (2), 23–48.

_ , _ , and _ , "Trusting the stock market," *the Journal of Finance*, 2008, *63* (6), 2557–2600.

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya, "3g internet and confidence in government," *The Quarterly Journal of Economics*, 2021, *136* (4), 2533–2613.

Habyarimana, James, Macartan Humphreys, Daniel N Posner, and Jeremy M Weinstein, "Why does ethnic diversity undermine public goods provision?," *American Political Science Review*, 2007, *101* (4), 709–725.

Halverson, Karen, "China's WTO accession: economic, legal, and political implications," *BC Int'l & Comp. L. Rev.*, 2004, *27*, 319.

Hanley, Douglas, Jiancheng Li, and Mingqin Wu, "High-speed railways and collaborative innovation," *Regional Science and Urban Economics*, 2022, *93*, 103717.

Hashmi, Aamir Rafique, "Competition and innovation: The inverted-U relationship revisited," *Review of Economics and Statistics*, 2013, *95* (5), 1653–1668.

Hu, Albert Guangzhou and Gary H Jefferson, "A great wall of patents: What is behind China's recent patent explosion?," *Journal of Development Economics*, 2009, *90* (1), 57–68.

Huang, Baibing, Endong Yang, and Yang Zhang, "Board surname sharing and investment efficiency: Evidence from Chinese state-owned enterprises," *Corporate Governance: An International Review*, 2023, *31* (4), 597–624.

Jones, Charles I, "Growth and ideas," in "Handbook of Economic Growth," Vol. 1, Elsevier, 2005, pp. 1063–1111.

Kerr, William R and William F Lincoln, "The supply side of innovation: H-1B visa reforms and US ethnic invention," *Journal of Labor Economics*, 2010, *28* (3), 473–508.

King, Robert G. and Ross Levine, "Finance, entrepreneurship and growth: Theory and evidence," *Journal of Monetary Economics*, 1993, *32* (3), 513–542.

Kinnan, Cynthia and Robert Townsend, "Kinship and financial networks, formal financial access, and risk reduction," *American Economic Review*, 2012, *102* (3), 289–293.

Knack, Stephen and Philip Keefer, "Does social capital have an economic payoff? A cross-country investigation," *The Quarterly Journal of Economics*, 1997, *112* (4), 1251–1288.

Köbig, Michael, Kjetil Storesletten, Zheng Song, and Fabrizio Zilibotti, "From imitation to innovation: Where is all that Chinese R&D going?," *Econometrica*, 2022, *90* (4), 1615–1654.

Lee, David S, Justin McCrary, Marcelo J Moreira, and Jack Porter, "Valid t-ratio Inference for IV," *American Economic Review*, 2022, *112* (10), 3260–3290.

Li, Kai, Heng Yue, and Longkai Zhao, "Ownership, institutions, and capital structure: Evidence from China," *Journal of Comparative Economics*, 2009, *37* (3), 471–490.

_ , Tan Wang, Yan-Leung Cheung, and Ping Jiang, "Privatization and risk sharing: Evidence from the split share structure reform in China," *The Review of Financial Studies*, 2011, *24* (7), 2499–2525.

Liang, Yousha, Kang Shi, Junjie Tang, Hanyi Tao, and Juanyi Xu, "Knowledge Diffusion and Local Innovation: Evidence from China," 2025.

Lim, Gedeon J, Quoc-Anh Do, and Sebastian Ellingsen, "Communities of Commerce: The Legacy of Chinese Immigration on Java, Indonesia," *Indonesia (November 15, 2024)*, 2024.

Liu, Qing and Hong Ma, "Trade policy uncertainty and innovation: Firm level evidence from China's WTO accession," *Journal of International Economics*, 2020, *127*, 103387.

_ and Larry D Qiu, "Intermediate input imports and innovations: Evidence from Chinese firms' patent filings," *Journal of International Economics*, 2016, *103*, 166–183.

_ , Ruosi Lu, Yi Lu, and Tuan Anh Luong, "Import competition and firm innovation: Evidence from China," *Journal of Development Economics*, 2021, *151*, 102650.

Ma, Liya, Dongxiao Niu, and Weizeng Sun, "Transportation infrastructure and entrepreneurship: Evidence from high-speed railway in China," *China Economic Review*, 2021, *65*, 101577.

Manacorda, Marco and Andrea Tesei, "Liberation technology: Mobile phones and political mobilization in Africa," *Econometrica*, 2020, *88* (2), 533–567.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 2001, *27* (1), 415–444.

Mokyr, Joel, "Long-term economic growth and the history of technology," in "Handbook of Economic Growth," Vol. 1, Elsevier, 2005, pp. 1113–1180.

Morck, Randall and Bernard Yeung, "Agency problems in large family business groups," *Entrepreneurship Theory and Practice*, 2003, *27* (4), 367–382.

_ and _ , "Family control and the rent–seeking society," *Entrepreneurship Theory and Practice*, 2004, *28* (4), 391–409.

Nordman, Christophe Jalil, "do family and kinship networks support entrepreneurs?," *IZA World of Labor*, 2016.

North, Douglass C, "Institutions, institutional change and economic performance," *Cambridge University*, 1990.

Peng, Yusheng, "Kinship networks and entrepreneurs in China's transitional economy," *American Journal of Sociology*, 2004, *109* (5), 1045–1074.

Posch, Max, Jonathan Schulz, and Joseph Henrich, "How cultural diversity drives innovation: Surnames and patents in us history," *Journal of Political Economy*, 2026, *134* (4), 000–000.

Prescott, Edward C, "Theory ahead of business-cycle measurement," in "Carnegie-Rochester Conference Series on Public Policy," Vol. 25 Elsevier 1986, pp. 11–44.

Romer, Paul M, "Increasing returns and long-run growth," *Journal of Political Economy*, 1986, *94* (5), 1002–1037.

_ , "Endogenous technological change," *Journal of Political Economy*, 1990, *98* (5, Part 2), S71–S102.

Ruofu, Du and Yuan Yida, "Encyclopedia of Chinese Surnames," 1996.

Shannon, Claude Elwood, "A mathematical theory of communication," *The Bell System Technical Journal*, 1948, *27* (3), 379–423.

Shao, Yuchen and Chengrui Xiao, "Corporate tax policy and heterogeneous firm innovation: Evidence from a developing country," *Journal of Comparative Economics*, 2019, *47* (2), 470–486.

Solow, Robert M, "Technical change and the aggregate production function," *The Review of Economics and Statistics*, 1957, *39* (3), 312–320.

Tadelis, Steven, "What's in a Name? Reputation as a Tradeable Asset," *American Economic Review*, 1999, *89* (3), 548–563.

Tan, Youchao, Jason Xiao, Cheng Colin Zeng, and Hong Zou, "What's in a name? The valuation effect of directors' sharing of surnames," *Journal of Banking & Finance*, 2021, *122*, 105991.

Tsai, Kellee S, "Imperfect substitutes: the local political economy of informal finance and microfinance in rural China and India," *World Development*, 2004, *32* (9), 1487–1507.

Wang, Xu, Zhuan Xie, Xiaobo Zhang, and Yiping Huang, "Roads to innovation: Firm-level evidence from People's Republic of China (PRC)," *China Economic Review*, 2018, *49*, 154–170.

Wei, Shang-Jin, Jianhuan Xu, Ge Yin, and Xiaobo Zhang, "Bureaucratic bean counting and patent subsidies: A welfare evaluation of China's pro-innovation program," 2021.

_ **, Zhuan Xie, and Xiaobo Zhang**, "From "made in China" to "innovated in China": Necessity, prospect, and challenges," *Journal of Economic Perspectives*, 2017, *31* (1), 49–70.

World Bank, *Global Financial Development Report 2015/2016: Long-Term Finance*, World Bank, 2015.

Xie, Fei, Bohui Zhang, and Wenrui Zhang, "Trust, incomplete contracting, and corporate innovation," *Management Science*, 2022, *68* (5), 3419–3443.

Xie, Jian and Song Yuan, "The cultural origins of family firms," *Journal of Comparative Economics*, 2025.

Xu, Yiqing and Yang Yao, "Informal institutions, collective action, and public investment in rural China," *American Political Science Review*, 2015, *109* (2), 371–391.

Zacchia, Paolo, "Knowledge spillovers through networks of scientists," *The Review of Economic Studies*, 2020, *87* (4), 1989–2018.

Zak, Paul J and Stephen Knack, "Trust and growth," *The Economic Journal*, 2001, *111* (470), 295–321.

Zhang, Chuanchuan, "Clans, entrepreneurship, and development of the private sector in China," *Journal of Comparative Economics*, 2020, *48* (1), 100–123.

Zheng, Shilin and Zhaochen Li, "Pilot governance and the rise of China's innovation," *China Economic Review*, 2020, *63*, 101521.

A Online Appendices

A.1 Summary Statistics

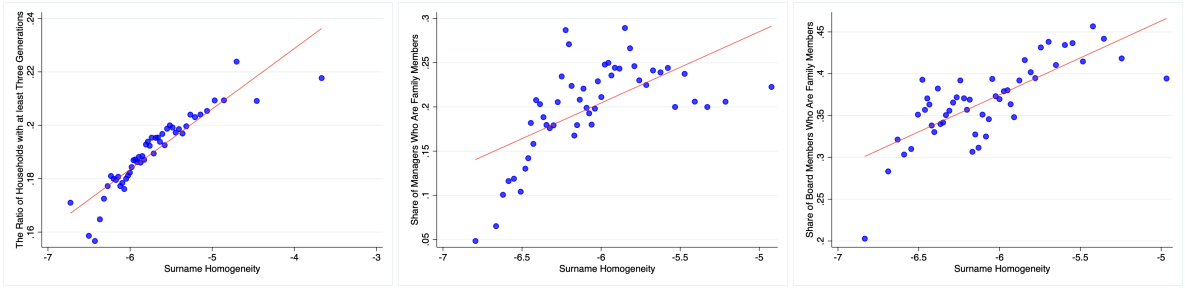
Table A.1: Summary Statistics

	Count	Mean	SD	Min	Max
Patent filings per 1,000 people	54,100	0.077	0.166	0.000	0.663
Surname homogeneity	56,970	-5.699	0.587	-7.867	0.000
Ruggedness	56,679	2.484	2.478	0.017	17.180
Distance to coast	56,805	569,223.880	578,025.046	154.065	3648153.000
Province capital	56,805	0.105	0.306	0.000	1.000
Special economic zone	56,805	0.016	0.125	0.000	1.000
High-speed railway station	57,372	0.086	0.281	0.000	1.000
Airport	57,372	0.481	0.500	0.000	1.000
Number of private firms	57,372	4,517.843	13,626.536	0.000	519,106.000
Population	56,805	462,731.441	355,062.483	8,815.000	6445777.000
Average schooling years	56,805	7.555	1.262	1.849	11.848
Urbanization rate	56,805	27.383	24.540	2.260	98.910
Fraction of minorities	56,805	13.463	25.550	0.000	99.278
Immigration rate	56,805	-0.000	0.127	-0.567	0.904
Marketization index	56,007	9.142	3.492	1.217	19.874
Banks density	56,700	0.143	0.114	0.000	2.182
Number of universities	57,372	1.081	2.714	0.000	36.000
Number of 211 universities	57,372	0.047	0.442	0.000	18.000
Distance to rebellion	57,372	110.979	203.590	0.013	1,299.982
Patent filings per firm	54,256	0.007	0.010	0.000	0.038
USPTO patent grants per 100,000 people	56,805	0.011	0.044	0.000	0.194
Granted patent per 1,000 people	54,100	0.020	0.041	0.000	0.160
Patent citations per 1,000 people	56,805	0.110	0.229	0.000	0.897
Formal Institution	51,300	2.863	1.509	0.000	10.472
Financial development	57,372	2.665	0.773	1.413	7.578
Public service expenditure	38,248	0.034	0.019	0.008	0.125
Intellectual property rights protection	51,908	5.442	8.086	0.040	53.510
Development of the factor market	57,372	6.718	3.073	0.801	16.713
Development of the product market	57,372	7.905	1.715	-0.247	10.609
Nighttime light	56,406	920.838	1,198.420	22.038	4,424.641
New established firms per 1,000 people	54,060	1.636	1.970	0.101	7.518
HHI	56,970	0.038	0.010	0.024	0.063
Industry homogeneity	56,805	-5.052	0.379	-5.550	-4.146
Genealogy density	57,372	0.026	0.119	0.000	2.422
Agricultural diversity	56,805	1.521	0.491	0.000	2.232
Treaty ports	57,372	0.162	0.369	0.000	1.000

A.2 Validation Test for Surname Homogeneity Measures

We posit that surname homogeneity among business owners within a county could indicate the strength of social connections, as discussed in Section 2. To validate this measure, we use three other county-level variables that potentially reflect this strength. The first variable is the ratio of extended families, measured as the number of households with at least three generations divided by the total number of households. The prevalence of extended families, where multiple generations co-reside, typically indicates strong kinship ties, in contrast to nuclear families, which emphasize smaller, more independent household units. The second variable is the share of managers who are family members, and the third variable is the share of board members who are family members. These two variables measure family members' participation in the firm and are strongly correlated with local kinship culture (Xie and Yuan, 2025).

Figure A.1 illustrates a strong positive relationship between surname homogeneity among business owners and all three variables. This suggests that regions with higher surname homogeneity tend to have stronger kinship-based informal institutions.



(a) Extended Families (b) Family Managers (c) Family Board Members

Figure A.1: Validation Test for the Measure of Surname Homogeneity

Notes: Data on the number of households with at least three generations are from the 2000 Population Census. Data on the share of managers who are family members and the share of board members who are family members come from the China Private Enterprise Survey 2010.

A.3 Robustness Checks - Details

To ensure the robustness of our results, we conduct an extensive set of additional analyses that address concerns related to: (a) omitted variables and the inclusion of additional controls; (b) the measure of surname homogeneity; (c) the validity of our measure of innovation outcomes; (d) serial dependence in surname homogeneity; and (e) the influence of special regions.

a. More controls for potential omitted variable issues

Our baseline estimation includes many controls: county and year fixed effects, geographic and transportation controls, the number of private firms, and demographic and socioeconomic variables. Nevertheless, the omitted-variable problem may persist. To address this issue and ensure robustness, we incorporate the following additional variables that could confound our estimation: (i) economic

development; (ii) industry homogeneity; (iii) industry structure; (iv) clan culture; (v) agricultural diversity; and (vi) exposure to foreign culture.

We elaborate our reasons to include each control as follows. First, higher levels of innovation may be driven by overall economic development. Based on our baseline specification, we further control for city-level GDP per capita, a measure of economic development, in Table A.9.[31] Second, we account for industry homogeneity, which could affect innovation outcomes. The method for measuring industry homogeneity is analogous to that used for surname homogeneity, but instead of the fraction of specific surnames, we use the fraction of firms sorted by specific industries. The results are shown in Table A.10. Third, we include industry structure as a control, measured by the fraction of firms of different sizes.[32] High fractions of smaller firms could correlate with both surname homogeneity and innovation. The results are reported in Table A.11. Fourth, we include the dynamic effects of clan culture, proxied by the number of genealogies per 1,000 people, interacted with year dummies.[33] This control is included to address the possibility that our results are driven by clan culture rather than surname homogeneity. The corresponding results are displayed in Table A.12. Fifth, we control for agricultural diversity interacted with year dummies, as it may influence innovation (Fiszbein, 2022).[34] The results are displayed in Table A.13. Finally, we include a treaty port dummy interacted with year dummies, as treaty ports may have influenced cultural characteristics and innovation through exposure to foreign cultures. The results are presented in Table A.14. As evident across all these tables, our findings remain robust, indicating that the estimation is unlikely to be significantly affected by omitted-variable bias.

b. Another measure of surname homogeneity: HHI

In our baseline estimation, we quantify surname homogeneity using an entropy-based measure, with a well-established theoretical foundation in information theory. To ensure the robustness of our results, we also employ an alternative and widely used measure of concentration, the Herfindahl-Hirschman Index (HHI) as below:

$$SurnameHomogeneity_{c,t} = \sum_{k} p_{c,t,k}^2 \tag{A.1}$$

where $p_{c,t,k}$ is the fraction of business owners with surname k in county c in year t. A higher value of this measure indicates greater surname concentration. However, this measure is highly skewed, resulting in several outliers. To mitigate the influence of these outliers, we winsorize the data at the 5% and 95% levels. The results, presented in Table A.15, demonstrate that this measure of surname

[31] It may be a bad control, as innovation could also influence economic development.

[32] (1) firms with registered capital smaller than 100,000 CNY, (2) firms with capital between 100,000 and 1,000,000 CNY, (3) firms with capital between 1,000,000 and 5,000,000 CNY, and (4) firms with capital between 5,000,000 and 10,000,000 CNY.

[33] The number of genealogies in each county, normalized by 1,000 local residents, serves as the measure of clan culture. Since it is time-invariant, this variable is interacted with year dummies.

[34] We measure agricultural diversity by considering the caloric contribution of five key crops: rice, wheat, maize, sweet potato, and white potato. This diversity is quantified using an entropy-based functional form, analogous to our measure of surname homogeneity, as outlined in Equation 3.1.

homogeneity continues to predict corporate innovation positively.

c. Alternative measures for innovation outcomes

A potential concern is whether our outcome variable accurately captures local innovation. To address this concern, we conduct several additional analyses using our baseline specification in Equation 4.1.

First, we replace the outcome variable with the number of patents filed by Chinese firms in each county that are ultimately granted by the United States Patent and Trademark Office (USPTO). These patents plausibly reflect more significant innovations, as obtaining USPTO approval entails a more stringent examination process. Appendix Table A.16 shows that greater surname homogeneity among business owners in a Chinese county is strongly positively associated with its number of USPTO-granted patents.

Second, instead of using the number of patent applications, we employ two alternative measures of innovation outcomes: the number of patents eventually granted in China and the total number of patent citations, both normalized per 1,000 residents.[35] Consistent with the baseline results in Table 1, both measures exhibit a positive and statistically significant association with surname homogeneity. The corresponding estimates are reported in Appendix Tables A.17 and A.18.

Third, firms may file patent applications indiscriminately. Wei et al. (2021) document that China's patent subsidy policy–the InnoCom program introduced in 2008–encouraged patent trading among firms and may have artificially inflated patent application counts, thereby weakening their validity as a proxy for corporate innovation. To address this concern, we restrict the sample to the pre-policy period from 2001 to 2008. As shown in Appendix Table A.19, our baseline results remain robust.

Finally, although innovation precedes patent applications, China's first-to-file system incentivizes rapid filing. Accordingly, our baseline specification uses surname homogeneity among business owners in year t to explain the number of patent filings in year $t+1$. Patent filings, however, may respond with a lag; that is, they may be affected by surname homogeneity in earlier years. To account for such potential dynamic effects, we consider the cumulative number of patent filings over longer horizons: from year $t+1$ to $t+2$, from $t+1$ to $t+3$, from $t+1$ to $t+4$, and from $t+1$ to $t+5$. For this analysis, we replicate the baseline model specified in Equation 4.1, replacing the outcome variable with the cumulative number of patent filings over these extended periods. The results, presented in Tables A.20, A.21, A.22, and A.23, correspond to the four respective timeframes. All these results consistently confirm our baseline findings.

Taken together, these findings provide robust and consistent evidence that surname homogeneity among business owners is a strong predictor of innovation outcomes at the county level in China.

d. Serial dependence in surname homogeneity

[35] Our baseline regressions use the number of patent applications filed as a proxy for corporate innovation, as patent grants involve uncertain and potentially long time lags. Total patent citations are measured as of the end of 2020.

Another major concern is serial dependence in surname homogeneity, as within-county variation evolves slowly over time. Because the measure is constructed from the stock of existing private firms, substantial firm overlap across adjacent years may induce persistence in the variable. We address this concern in three ways.

First, we employ Newey-West standard errors to account for serial correlation. The results are reported in Table A.24. Second, we adopt a long-difference specification that effectively mitigates serial dependence by examining the relationship between changes in surname homogeneity and changes in patent filings over a 20-year period. Specifically, we replace $Y_{c,t+1}$ in Equation 4.1 with ΔY_c, which is the change in patent filings in county c from 2001 to 2020, and similarly replace $SurnameHomogeneity_{c,t}$ with $\Delta SurnameHomogeneity_c$, which is the change in surname homogeneity among business owners in county c from 2000 to 2019. We include the same set of controls as in Equation 4.1, using their values in 2000 for all time-varying controls, ensuring they are predetermined. The results are presented in Table A.25.

Third, we implement sparse sampling by restricting the sample to every third year between 2000 and 2018 (2000, 2003, 2006, 2009, 2012, 2015, and 2018), which reduces persistence in the independent variable. The results are reported in Table A.26. We also consider an alternative sparse-sampling scheme using four-year intervals (2000, 2004, 2008, 2012, and 2016), with results reported in Table A.27. Across these specifications, the estimates remain consistent with our baseline findings.

e. Special regions

We exclude three types of special samples that could potentially affect our results. The first type consists of counties with a significant minority population. According to the 2005 population census, over 90% of China's population is ethnic Han. However, some counties have a high proportion of minorities, resulting in a distinct cultural and socioeconomic environment. Therefore, we exclude counties with minority populations exceeding 10%. The second type includes municipalities directly administered by the central government–Beijing, Shanghai, Tianjin, and Chongqing. These municipalities function as provinces, with an administrative level above that of other cities, potentially influencing policy implementation and economic activity. The third type comprises special economic zones (Shenzhen, Zhuhai, Shantou, Xiamen, and Hainan), whose distinct institutional arrangements may foster greater innovation through alternative approaches.

The estimates excluding these three types of special samples are presented in Tables A.28, A.29 and A.30, respectively. As shown, our results remain robust across all these tables. Notably, the magnitudes in Table A.28 are larger compared to our baseline estimation, suggesting that minority populations may have influenced our baseline results.

A.4 Figures

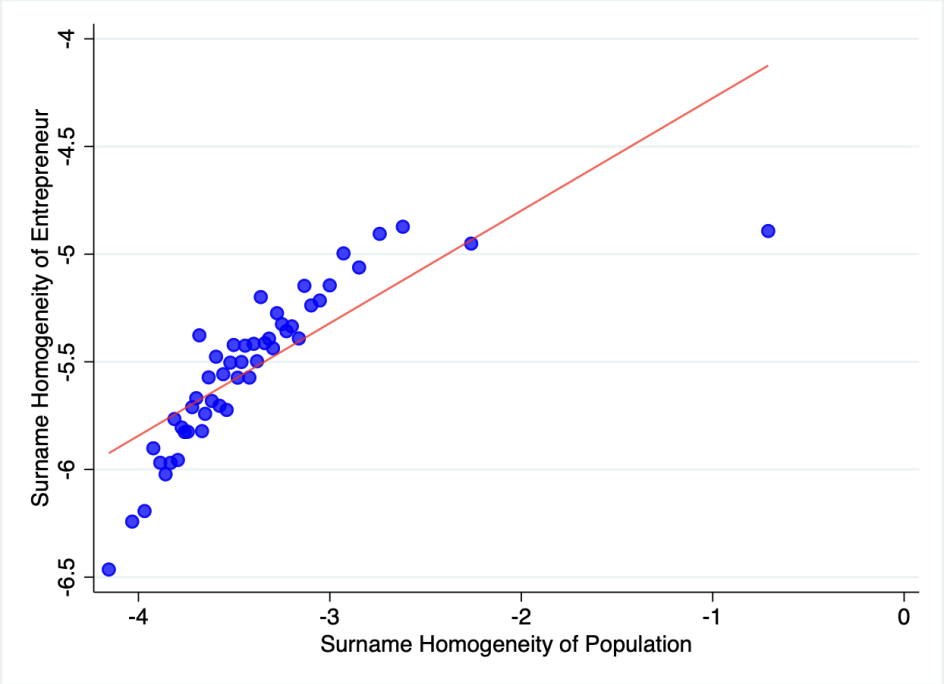
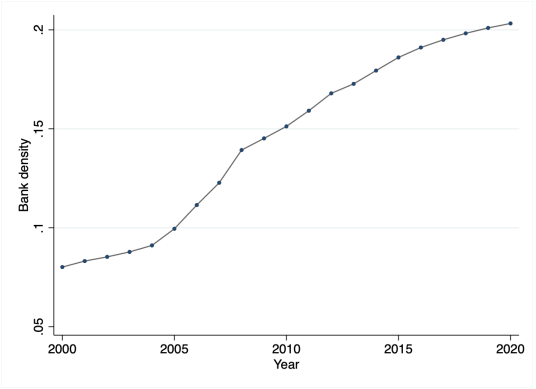
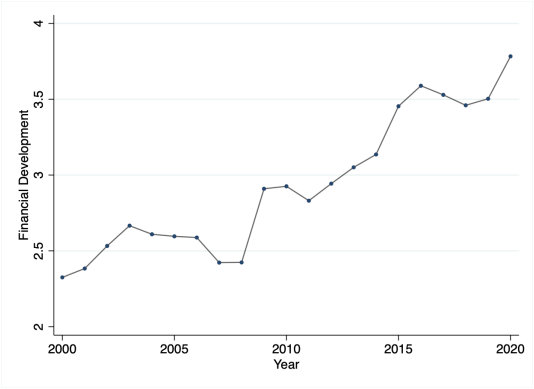


Figure A.2: Surname Homogeneity of the Population and Surname Homogeneity among Business Owners in 2005

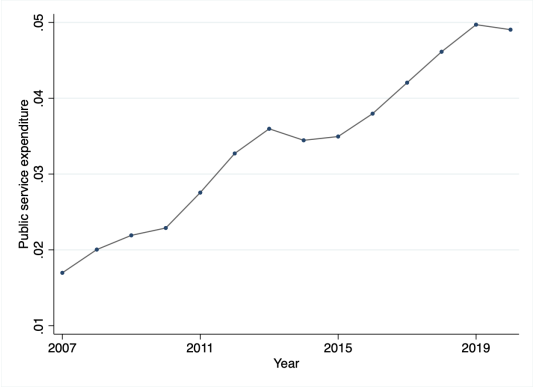
Notes: This figure shows the relationship between surname homogeneity of the population and surname homogeneity among business owners in 2005. Data on surname homogeneity of the population is from the 2005 mini-census.



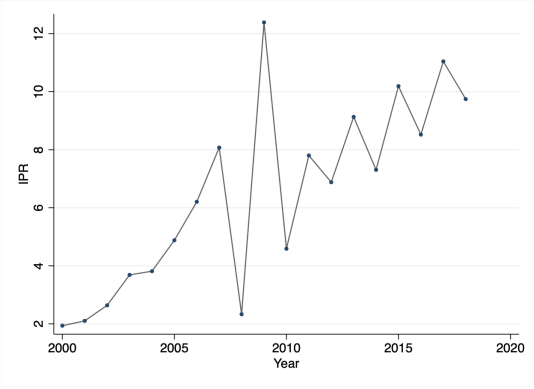
(a) Bank Density



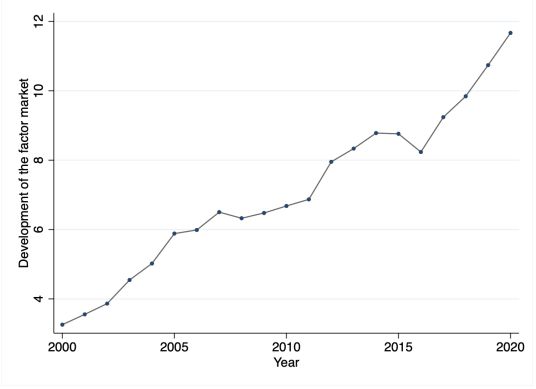
(b) Financial Development



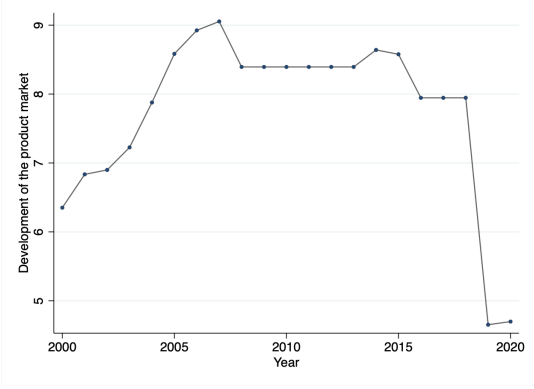
(c) Public Service Expenditure



(d) Intellectual Property Rights Protection



(e) Factor Market



(f) Product Market

Note: These six figures show the changes in six measures of the strength of formal institutions over time in China, including bank density, financial development, public service expenditure, the intellectual property rights protection index, the index of factor market development, and the index of product market development. We compute their average values for each year.

A.5 Mechanisms - Details

A.5.1 Financial Development

Social networks can enhance members' financial accessibility, thereby fueling innovation. As discussed in Chen et al. (2022), kinship networks and financial markets act as competing substitutes in China. The R&D process requires significant capital, and entrepreneurs can raise funds from their social networks when access to formal financial institutions is limited. However, the role of such networks in financing innovation is likely to diminish as financial development progresses. If this is indeed the case, then the financing mechanism of surname-based connections is confirmed.

We are thus motivated to test this mechanism by estimating Equation 6.1, replacing $Formal_{c,t}$ with two proxies for local financial development. The first proxy is bank density, defined as the number of banks per county population. The second proxy is the ratio of total loans and deposits from formal financial institutions to the province's GDP. These two variables capture the availability of formal financial services. The first variable varies at the county level, while the second variable varies at the province level.

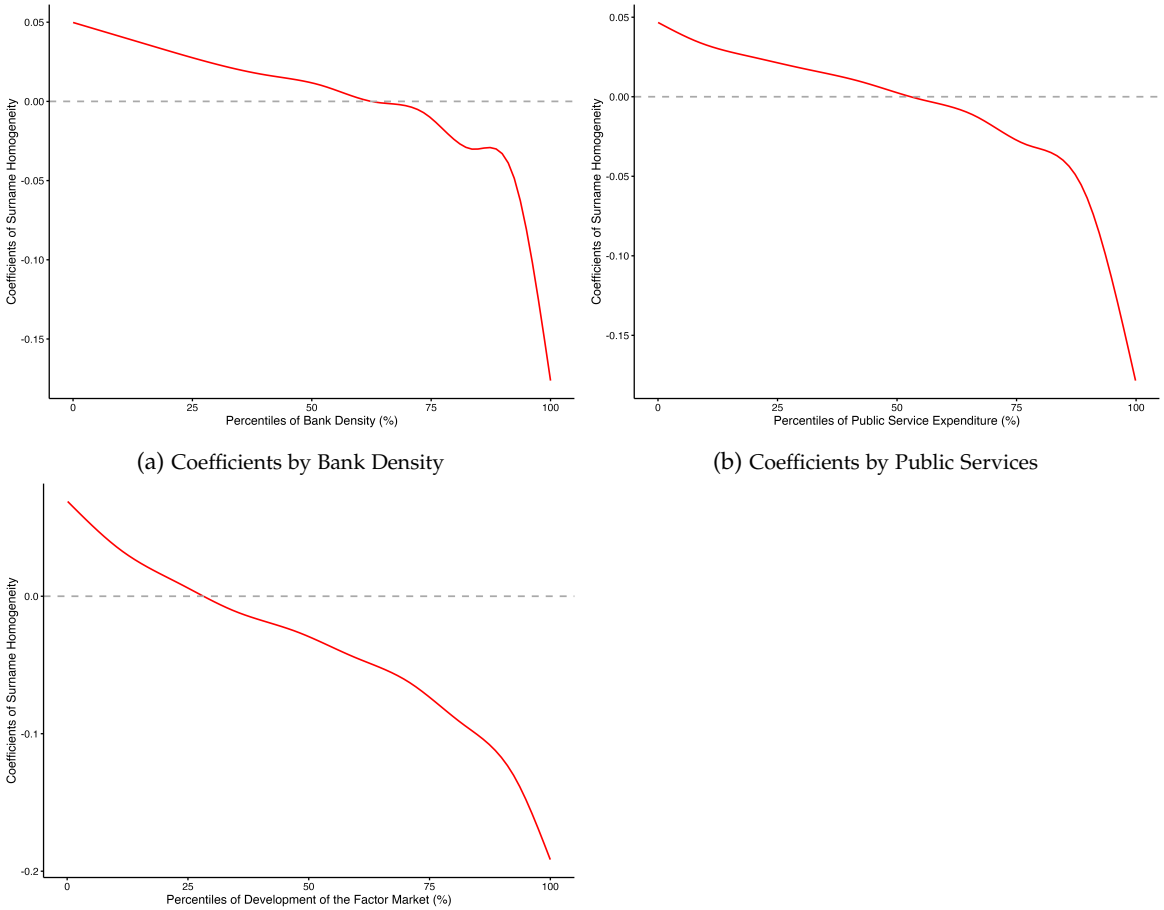
Table A.2 presents the results based on the two proxies of financial development, with Panel A and Panel B reporting the results for each proxy separately. In both panels, the three key independent variables are jointly significant. For example, for column (5), panel A of Table A.2, the F-statistic for the joint significance of surname homogeneity, bank density, and their interaction is 50.03. Also, the interaction between surname homogeneity and financial development is negative and significant in all columns, indicating that the positive relationship between surname homogeneity and local innovation weakens with financial development and may eventually become negative. We plot, in Figure A.4a, the overall coefficients of surname homogeneity with respect to bank density based on the estimates from column (5), Panel A of Table A.2. The turning point occurs at a bank density of 0.149 (0.054 / 0.363). For 64.2% of the observations (where bank density is below 0.149), surname homogeneity positively correlates with local innovation; for the remaining 35.8%, the correlation is negative.[36]

[36] Bank density itself predicts innovation positively: at the mean level of surname homogeneity (-5.699), its coefficient is 0.433 (-1.636 + (-0.363) $\times(-5.699)$).

Table A.2: Mechanism: Financial Development

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Panel A: Heterogeneity Analysis Based on Bank Density					
Surname homogeneity × Bank density	-0.529***	-0.517***	-0.424***	-0.371***	-0.363***
	(0.073)	(0.068)	(0.060)	(0.058)	(0.058)
Surname homogeneity	0.127***	0.083***	0.064***	0.054***	0.054***
	(0.015)	(0.009)	(0.008)	(0.008)	(0.008)
Bank density	-2.434***	-2.421***	-1.947***	-1.675***	-1.636***
	(0.439)	(0.413)	(0.356)	(0.340)	(0.334)
Observations	53,635	53,515	53,515	53,515	52,495
R^2	0.673	0.696	0.712	0.720	0.723
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078
Panel B: Heterogeneity Analysis Based on Loans & Deposits					
Surname homogeneity × Loans & deposits	-0.084***	-0.099***	-0.073***	-0.062***	-0.055***
	(0.009)	(0.010)	(0.009)	(0.008)	(0.007)
Surname homogeneity	0.294***	0.270***	0.200***	0.168***	0.151***
	(0.034)	(0.027)	(0.024)	(0.021)	(0.019)
Financial	-0.499***	-0.596***	-0.441***	-0.388***	-0.351***
	(0.062)	(0.060)	(0.055)	(0.052)	(0.045)
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.642	0.677	0.695	0.710	0.725
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: County-level OLS regressions. "Bank density" is the number of banks per capita. "Loans & deposits" is measured by the ratio of total loans and deposits from formal financial institutions to GDP. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.



(a) Coefficients by Bank Density

(b) Coefficients by Public Services

(c) Coefficients by Factor Markets

Figure A.4: Coefficients of Surname Homogeneity with Respect to the Strength of Formal Institutions

Note: These figures show the coefficients of surname homogeneity in percentiles by bank density, public service expenditure, and factor market development, respectively.

A.5.2 Quality of Public Administration

Social networks can provide public services and reduce the risk of losing property rights (Easterly and Levine, 1997; Zhang, 2020). If this is the case, it could be another potential channel through which surname homogeneity facilitates local innovation, as better public services, such as protecting rights, maintaining safe environments and infrastructure, benefit innovation. Moreover, the effectiveness of this channel likely depends on the quality of public administration: as it improves, the role of surname homogeneity in promoting local innovation should diminish. Such a weakened relation supports the notion that social networks offer public services.

We measure the quality of public administration using two proxies: annual public service expenditure normalized by GDP, measured at the provincial level in 2010, and the intellectual property rights protection index.[37] We also employ a specification similar to Equation 6.1, but replace the measure of financial access with that of public administration quality.

Table A.3 reports the regression results, with Panels A and B presenting the results for each proxy separately. The coefficients for surname homogeneity, public service expenditure, and their interaction term are jointly significant, as indicated by an F-statistic of 22.86 based on the estimation in column (5). The coefficients for the interaction terms are negative, suggesting that the positive influence of surname homogeneity on innovation diminishes as public service provision improves. In such cases, local firms increasingly rely on formal public infrastructure for the provision of public services rather than support from social networks. We plot the overall coefficients of surname homogeneity with respect to public service expenditure based on the estimates from column (5) of Table A.3. As shown in Figure A.4b, the turning point occurs when public service expenditure equals 0.03 (calculated as 0.072/2.368). In 52.8% of observations where public service expenditure is below the threshold, surname homogeneity correlates positively with local innovation; in the remaining 47.2% of observations, where public service expenditure exceeds the threshold, the correlation turns negative. We also find that the role of social networks declines as intellectual property rights protection strengthens, suggesting that social networks serve as a substitute when public administration quality is weaker.

[37] The data on public service expenditure is available only from 2007.

Table A.3: Mechanism: Quality of Public Administration

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Panel A: Heterogeneity Analysis Based on Public Service Expenditure					
Surname homogeneity × Public	-2.716***	-3.676***	-3.154***	-2.618***	-2.368***
	(0.487)	(0.506)	(0.423)	(0.392)	(0.355)
Surname homogeneity	0.208***	0.144***	0.119***	0.072***	0.072***
	(0.023)	(0.019)	(0.016)	(0.021)	(0.020)
Public	-17.014***	-21.253***	-18.151***	-14.378***	-12.767***
	(2.946)	(2.986)	(2.469)	(2.296)	(2.036)
Observations	34,931	34,853	34,853	34,853	34,125
R^2	0.763	0.782	0.786	0.792	0.798
Mean dependent vars.	0.115	0.115	0.115	0.115	0.116
Panel B: Heterogeneity Analysis Based on Intellectual Property Rights Protection					
Surname homogeneity × IPR	-0.010***	-0.009***	-0.007***	-0.006***	-0.006***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Surname homogeneity	0.087***	0.054***	0.041***	0.032***	0.031***
	(0.013)	(0.007)	(0.006)	(0.006)	(0.006)
IPR	-0.055***	-0.048***	-0.036***	-0.029***	-0.028***
	(0.009)	(0.007)	(0.005)	(0.004)	(0.004)
Observations	51,048	50,934	50,934	50,934	49,870
R^2	0.670	0.694	0.709	0.726	0.738
Mean dependent vars.	0.072	0.072	0.072	0.072	0.072
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: County-level OLS regressions. "Public" indicates the annual public service expenditure normalized by GDP in 2010. "IPR" indicates the intellectual property rights protection index. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

A.5.3 Development of Factor Markets

Social networks can reduce coordination costs and facilitate the organization of production. When factor markets are underdeveloped, informal institutions such as social networks may help secure input factors and supply chains as well as facilitate product sales, thereby improving operational efficiency. However, this function will likely reduce as factor markets become more established and operate effectively.

We explore this possibility by examining the interaction between surname homogeneity and the strength of market forces. Specifically, we consider two indices for *Formal* in Equation 6.1: the factor market development index and the product market development index, both at the province level and have been explained in Section 3.4. Our estimation follows a specification similar to that in Equation 6.1. The results for the two indices are reported separately in Panels A and B of Table A.4. In all columns of both tables, the coefficients for surname homogeneity, the index of market development, and their interaction terms are jointly significant. For example, based on the regression in column (5), Panel A of Table A.4, the F-statistic is 24.55. This column shows that the interaction terms' coefficients are significant and negative, suggesting that the positive effect of surname homogeneity diminishes as market mechanisms improve. Based on the estimates in this column, Figure A.4c illustrates this negative relationship. The turning point occurs when the index of factor market development reaches 4.79 (computed as $0.091/0.019$). In 27.6% of observations where the index is below the threshold, surname homogeneity is positively associated with innovation; in the remaining 72.4%, the association becomes negative.

Table A.4: Mechanism: Development of the Factor Market

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Panel A: Heterogeneity Analysis Based on Factor Market					
Surname homogeneity × Factor market	-0.028***	-0.028***	-0.023***	-0.021***	-0.019***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Surname homogeneity	0.172***	0.134***	0.106***	0.098***	0.091***
	(0.028)	(0.020)	(0.018)	(0.015)	(0.015)
Factor market	-0.156***	-0.160***	-0.127***	-0.121***	-0.109***
	(0.019)	(0.019)	(0.017)	(0.016)	(0.015)
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.663	0.690	0.704	0.717	0.731
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078
Panel B: Heterogeneity Analysis Based on Product Market					
Surname homogeneity × Product market	-0.004	-0.001	-0.004	-0.001	-0.001
	(0.005)	(0.005)	(0.004)	(0.003)	(0.002)
Surname homogeneity	0.131***	0.066*	0.065**	0.038	0.033*
	(0.028)	(0.033)	(0.025)	(0.023)	(0.018)
Product market	-0.019	-0.004	-0.017	-0.000	0.003
	(0.030)	(0.031)	(0.022)	(0.019)	(0.015)
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.628	0.658	0.685	0.704	0.720
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Notes: County-level OLS regressions. "Factor market" indicates the index of development of the factor market. "Product market" indicates the index of development of the product market. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.5: Baseline Results: Normalized by Number of Private Firms

Dependent variable	Patent filings per firm				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.004***	0.003***	0.002***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	54,256	53,615	53,615	53,615	52,495
R^2	0.573	0.582	0.586	0.591	0.597
Mean dependent vars.	0.007	0.007	0.007	0.007	0.007

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by number of private firms. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.6: Baseline Results: Number of Patent Filings (Unnormalized)

Dependent variable	Patent filings				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	55.702***	30.976***	22.830***	12.868***	11.645***
	(5.828)	(3.431)	(2.862)	(2.968)	(2.795)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	54,256	53,615	53,615	53,615	52,495
R^2	0.637	0.671	0.698	0.715	0.731
Mean dependent vars.	36.754	36.763	36.763	36.763	37.011

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.7: Baseline Results: Winsorized at 1% and 99%

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.204***	0.093***	0.041***	0.045***	0.043***
	(0.028)	(0.013)	(0.010)	(0.013)	(0.012)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.537	0.570	0.638	0.659	0.679
Mean dependent vars.	0.123	0.123	0.123	0.123	0.125

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

A.6 Coefficients of Controls

Table A.8: Coefficients of Controls

Dependent variable	Patent filings per 1,000 people	Surname Homogeneity
	(1)	(2)
Surname homogeneity	0.028***	
	(0.006)	
Firm number	0.000***	-0.000**
	(0.000)	(0.000)
High speed train station	0.043***	0.030***
	(0.009)	(0.010)
Airport	-0.029***	-0.061***
	(0.006)	(0.018)
University	0.028***	0.022**
	(0.005)	(0.009)
Bank density	0.527***	0.033
	(0.049)	(0.111)
Market index	-0.006***	-0.002
	(0.002)	(0.004)
Other controls	Yes	Yes
County FE	Yes	Yes
Year FE	Yes	Yes
Observations	52,495	52,495
R^2	0.719	0.897
Mean dependent vars.	0.078	-5.731

Notes: County-level OLS regressions. The outcome variable in column (1) is the number of patent filings by private firms in a county, normalized by 1,000 local population. The outcome variable in column (2) is surname homogeneity among business owners. Time-variant controls include the number of private firms, a high-speed railway station dummy, an airport dummy, the number of universities, bank density, and the marketization index. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

A.7 Tables of Robustness Checks

Table A.9: Control for Economic Development

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.060***	0.034***	0.025***	0.021***	0.021***
	(0.009)	(0.005)	(0.005)	(0.005)	(0.005)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
GDP per capita	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,315	53,195	53,195	53,195	52,495
R^2	0.694	0.706	0.717	0.727	0.739
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.10: Control for Industry Homogeneity

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.048***	0.013***	0.010**	0.010*	0.013**
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
Industry homogeneity	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,555	53,435	53,435	53,435	52,315
R^2	0.655	0.676	0.695	0.708	0.722
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.11: Control for Industry Structure

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.104***	0.057***	0.041***	0.028***	0.027***
	(0.011)	(0.006)	(0.005)	(0.006)	(0.006)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
Industry structure	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,533	53,413	53,413	53,413	52,293
R^2	0.631	0.658	0.686	0.703	0.719
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. We control the fraction of firm size: (1) smaller than 100,000; (2) between 100,000 to 1,000,000; (3) between 1,000,000 to 5,000,000; (4) between 5,000,000 to 10,000,000. The size is measured by the registered capital in CNY. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.12: Control for Clan Culture

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.105***	0.056***	0.040***	0.029***	0.028***
	(0.012)	(0.006)	(0.005)	(0.006)	(0.005)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
Clan culture	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.631	0.661	0.688	0.707	0.722
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.13: Control for Agricultural Diversity

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.099***	0.052***	0.037***	0.022***	0.019***
	(0.010)	(0.006)	(0.005)	(0.006)	(0.006)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
Agricultural diversity	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.631	0.660	0.687	0.706	0.724
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.14: Control for Treaty Port Dummy

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.099***	0.053***	0.039***	0.029***	0.028***
	(0.011)	(0.006)	(0.005)	(0.006)	(0.006)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
Treaty port dummy	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.641	0.666	0.690	0.706	0.721
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.15: Another Measure of Surname Homogeneity: HHI

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
HHI	4.293***	1.928***	1.325***	0.747***	0.783***
	(0.448)	(0.207)	(0.260)	(0.260)	(0.218)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.615	0.654	0.684	0.702	0.718
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners, measured based on the Herfindahl-Hirschman Index (HHI). Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels in all columns except column (2). In column (2), we report standard errors clustered only at the county level because two-way clustering fails. * p<0.1; ** p<0.05; *** p<0.01.

Table A.16: Baseline Results: Patents Granted in the U.S.

Dependent variable	Patent filings per 100,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.007***	0.006***	0.005***	0.003**	0.004**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.457	0.459	0.462	0.465	0.469
Mean dependent vars.	0.011	0.011	0.011	0.011	0.012

Notes: County-level OLS regressions. The outcome variable is the number of patents granted by the USPTO to Chinese firms located in Chinese counties, normalized per 100,000 local residents. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.17: Baseline Results: Granted Patents as the Outcome Variable

Dependent variable	Granted Patents per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.022***	0.012***	0.010***	0.007***	0.006***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.663	0.689	0.701	0.726	0.739
Mean dependent vars.	0.020	0.020	0.020	0.020	0.020

Notes: County-level OLS regressions. The outcome variable is the number of granted patents by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.18: Baseline Results: Patent Citations as the Outcome Variable

Dependent variable	Patent citations per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.094***	0.055***	0.048***	0.037***	0.035***
	(0.024)	(0.005)	(0.011)	(0.009)	(0.008)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.662	0.688	0.691	0.718	0.727
Mean dependent vars.	0.116	0.115	0.115	0.115	0.116

Notes: County-level OLS regressions. The outcome variable is the number of patent citations by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels except column (2). * p<0.1; ** p<0.05; *** p<0.01.

Table A.19: The Correlation Across Different Periods: 2001-2008

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.017***	0.010***	0.002	0.003*	0.003*
	(0.004)	(0.003)	(0.002)	(0.002)	(0.001)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	21,491	21,443	21,443	21,443	20,995
R^2	0.667	0.677	0.735	0.744	0.748
Mean dependent vars.	0.009	0.009	0.009	0.009	0.010

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. Time spans from 2001 to 2008 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.20: Baseline Results: T+1 to T+2

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.211***	0.117***	0.085***	0.063***	0.059***
	(0.024)	(0.013)	(0.011)	(0.012)	(0.011)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	51,048	50,934	50,934	50,934	49,870
R^2	0.653	0.682	0.710	0.727	0.743
Mean dependent vars.	0.153	0.153	0.153	0.153	0.154

Notes: County-level OLS regressions. The outcome variable is the cumulative number of patent filings by private firms in a county in the following two years, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels in all columns except column (1). In column (1), we report standard errors clustered only at the county level because two-way clustering fails. * p<0.1; ** p<0.05; *** p<0.01.

Table A.21: Baseline Results: T+1 to T+3

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.316***	0.179***	0.132***	0.103***	0.096***
	(0.014)	(0.020)	(0.017)	(0.017)	(0.016)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	48,361	48,253	48,253	48,253	47,245
R^2	0.673	0.701	0.729	0.745	0.760
Mean dependent vars.	0.228	0.227	0.227	0.227	0.229

Notes: County-level OLS regressions. The outcome variable is the cumulative number of patent filings by private firms in a county in the following theree years, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels in all columns except column (1). In column (1), we report standard errors clustered only at the county level because two-way clustering fails. * p<0.1; ** p<0.05; *** p<0.01.

Table A.22: Baseline Results: T+1 to T+4

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.419***	0.243***	0.181***	0.148***	0.137***
	(0.049)	(0.019)	(0.024)	(0.022)	(0.022)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	45,674	45,572	45,572	45,572	44,620
R^2	0.692	0.719	0.747	0.762	0.776
Mean dependent vars.	0.299	0.299	0.299	0.299	0.301

Notes: County-level OLS regressions. The outcome variable is the cumulative number of patent filings by private firms in a county in the following four years, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels in all columns except column (2). In column (2), we report standard errors clustered only at the county level because two-way clustering fails. * p<0.1; ** p<0.05; *** p<0.01.

Table A.23: Baseline Results: T+1 to T+5

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.517***	0.307***	0.227***	0.194***	0.177***
	(0.062)	(0.036)	(0.032)	(0.028)	(0.027)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	42,987	42,891	42,891	42,891	41,995
R^2	0.711	0.737	0.766	0.780	0.793
Mean dependent vars.	0.368	0.367	0.367	0.367	0.370

Notes: County-level OLS regressions. The outcome variable is the cumulative number of patent filings by private firms in a county in the following five years, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.24: Baseline Results: Newey-West Standard Errors

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.106***	0.057***	0.041***	0.029***	0.028***
	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.627	0.657	0.685	0.703	0.719
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. Newey-West standard errors are in parentheses. * p<0.1; ** p<0.05; *** p<0.01.

Table A.25: Long Difference Approach

Dependent variable	Change in Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Change in Surname homogeneity	0.115***	0.069***	0.043***	0.032***	0.033***
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
Observations	2,678	2,678	2,678	2,678	2,622
R^2	0.146	0.229	0.348	0.397	0.424
Mean dependent vars.	0.186	0.186	0.186	0.186	0.188

Notes: County-level OLS regressions. The outcome variable is the change in patent filings per 1,000 people over the period from 2001 to 2020. The explanatory variable is the change in surname homogeneity among business owners over the period from 2000 to 2019. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. Robust standard errors are in parentheses. * p<0.1; ** p<0.05; *** p<0.01.

Table A.26: Restricting the Sample to the Years 2000, 2003, 2006, 2009, 2012, 2015, and 2018

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.097***	0.054***	0.038***	0.028**	0.026**
	(0.018)	(0.009)	(0.007)	(0.008)	(0.008)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	18,806	18,764	18,764	18,764	18,372
R^2	0.617	0.650	0.683	0.704	0.721
Mean dependent vars.	0.072	0.072	0.072	0.072	0.072

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. We restrict the Sample to the Years 2000, 2003, 2006, 2009, 2012, 2015, and 2018. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.27: Restricting the Sample to the Years 2000, 2004, 2008, 2012, and 2016

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.092**	0.055**	0.038**	0.033**	0.030**
	(0.024)	(0.013)	(0.010)	(0.010)	(0.010)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	13,432	13,402	13,402	13,402	13,122
R^2	0.606	0.637	0.673	0.690	0.708
Mean dependent vars.	0.067	0.067	0.067	0.067	0.067

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. We restrict the sample to the years 2000, 2004, 2008, 2012, and 2016. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.28: Excluding Cities with Minority Rates Above 10%

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.154***	0.110***	0.083***	0.061***	0.046***
	(0.008)	(0.014)	(0.011)	(0.009)	(0.007)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	40,700	40,580	40,580	40,580	39,860
R^2	0.643	0.679	0.700	0.723	0.740
Mean dependent vars.	0.092	0.091	0.091	0.091	0.092

Notes: County-level OLS regressions. The outcome variable is the cumulative number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels in all columns except column (1). In column (1), we report standard errors clustered only at the county level because two-way clustering fails. * p<0.1; ** p<0.05; *** p<0.01.

Table A.29: Excluding Municipalities Directly under the Central Government

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.099***	0.049***	0.036***	0.027***	0.027***
	(0.012)	(0.006)	(0.005)	(0.006)	(0.005)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	52,015	51,915	51,915	51,915	51,275
R^2	0.610	0.645	0.672	0.689	0.704
Mean dependent vars.	0.072	0.072	0.072	0.072	0.073

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.30: Excluding Special Economic Zones

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity	0.105***	0.056***	0.039***	0.030***	0.028***
	(0.012)	(0.006)	(0.005)	(0.006)	(0.006)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	52,875	52,755	52,755	52,755	51,935
R^2	0.622	0.652	0.684	0.701	0.716
Mean dependent vars.	0.076	0.076	0.076	0.076	0.076

Notes: County-level OLS regressions. The outcome variable is the number of patent filings by private firms in a county, normalized by 1,000 local population. The time span is from 2001 to 2020 for the outcome variable. The explanatory variable is the surname homogeneity among business owners. Geographic controls include average ruggedness, distance to the coast, a provincial capital dummy, and a special economic zone dummy. Transportation controls and firm number consist of a high-speed railway station dummy, an airport dummy, and the number of private firms. Demographic controls include population size, average years of schooling, urbanization rate, fraction of minorities, and migration rate. Socioeconomic controls include the marketization index, bank density, number of universities, and number of elite universities. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

A.8 Surname-County Level Analysis

Table A.31: Surname-County Level Analysis: Logarithm

Dependent variable	Log(1+Patent filings by firms controlled by surname s)		
	(1)	(2)	(3)
Surname homogeneity $\times$ Firms controlled by surname s	0.601***	0.438***	0.620***
	(0.107)	(0.119)	(0.107)
Firms controlled by surname s	4.730***	3.503***	4.566***
	(0.698)	(0.784)	(0.712)
Surname homogeneity	0.044***	0.166***	
	(0.006)	(0.018)	
County FE	Yes	No	No
Year FE	Yes	No	No
Surname FE	Yes	No	No
Surname-County FE	No	Yes	Yes
Year-Surname FE	No	Yes	Yes
County-Year FE	No	No	Yes
Observations	8,791,391	8,613,486	8,613,445
R^2	0.260	0.482	0.529
Mean dependent vars.	0.067	0.068	0.068

Notes: Surname-County level OLS regressions. The outcome variable is the number of patent filings by firms controlled by a given surname in each county-year, transformed by taking the natural logarithm of one plus the count. Firms controlled by surname s denotes the number of firms controlled by surname s, measured in thousands of firms. The time span is from 2001 to 2020 for the outcome variable. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.32: Surname-County Level Analysis: IHS

Dependent variable	IHS(Patent filings by firms controlled by surname s)		
	(1)	(2)	(3)
Surname homogeneity × Firms controlled by surname s	0.801***	0.596***	0.822***
	(0.127)	(0.142)	(0.128)
Firms controlled by surname s	6.182***	4.641***	5.958***
	(0.832)	(0.937)	(0.856)
Surname homogeneity	0.055***	0.208***	
	(0.008)	(0.022)	
County FE	Yes	No	No
Year FE	Yes	No	No
Surname FE	Yes	No	No
Surname-County FE	No	Yes	Yes
Year-Surname FE	No	Yes	Yes
County-Year FE	No	No	Yes
Observations	8,791,391	8,613,486	8,613,445
R^2	0.255	0.478	0.525
Mean dependent vars.	0.084	0.085	0.085

Notes: Surname-County level OLS regressions. The outcome variable is the number of patent filings by firms controlled by a given surname in each county-year, transformed using the inverse hyperbolic sine (IHS) transformation. Firms controlled by surname s denotes the number of firms controlled by surname s, measured in thousands of firms. The time span is from 2001 to 2020 for the outcome variable. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.33: Surname-County Level Analysis: Poisson Regressions

Dependent variable	Patent filings by firms controlled by surname s		
	(1)	(2)	(3)
Surname homogeneity × Firms controlled by surname s	0.200*	0.176*	0.194
	(0.112)	(0.101)	(0.156)
Firms controlled by surname s	1.339*	1.119*	1.218
	(0.747)	(0.659)	(1.014)
Surname homogeneity	-1.266***	-1.094***	
	(0.400)	(0.296)	
County FE	Yes	No	No
Year FE	Yes	No	No
Surname FE	Yes	No	No
Surname-County FE	No	Yes	Yes
Year-Surname FE	No	Yes	Yes
County-Year FE	No	No	Yes
Observations	8,353,552	1,812,176	1,516,249
Pseudo R^2	0.605	0.757	0.788
Mean dependent vars.	0.430	1.979	2.365

Notes: Surname-County level Poisson regressions. The outcome variable is the number of patent filings by firms controlled by a given surname in each county-year. Firms controlled by surname s denotes the number of firms controlled by surname s, measured in thousands of firms. The time span is from 2001 to 2020 for the outcome variable. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.

Table A.34: The Correlation over Time

Dependent variable	Patent filings per 1,000 people				
	(1)	(2)	(3)	(4)	(5)
Surname homogeneity × Year 2000	0.093***	0.051***	0.040***	0.035***	0.033***
	(0.011)	(0.008)	(0.007)	(0.006)	(0.006)
Surname homogeneity × Year 2001	0.101***	0.053***	0.042***	0.036***	0.034***
	(0.012)	(0.009)	(0.008)	(0.007)	(0.006)
Surname homogeneity × Year 2002	0.107***	0.055***	0.043***	0.037***	0.036***
	(0.013)	(0.009)	(0.008)	(0.007)	(0.007)
Surname homogeneity × Year 2003	0.119***	0.060***	0.047***	0.041***	0.040***
	(0.015)	(0.010)	(0.009)	(0.008)	(0.008)
Surname homogeneity × Year 2004	0.126***	0.061***	0.048***	0.041***	0.041***
	(0.016)	(0.011)	(0.010)	(0.009)	(0.008)
Surname homogeneity × Year 2005	0.130***	0.061***	0.048***	0.042***	0.045***
	(0.017)	(0.012)	(0.010)	(0.009)	(0.009)
Surname homogeneity × Year 2006	0.126***	0.054***	0.042***	0.037***	0.041***
	(0.018)	(0.013)	(0.011)	(0.009)	(0.009)
Surname homogeneity × Year 2007	0.114***	0.041***	0.029**	0.027***	0.029***
	(0.019)	(0.013)	(0.011)	(0.009)	(0.009)
Surname homogeneity × Year 2008	0.096***	0.024*	0.015	0.017*	0.021**
	(0.019)	(0.013)	(0.011)	(0.009)	(0.009)
Surname homogeneity × Year 2009	0.078***	0.005	-0.002	0.004	0.009
	(0.020)	(0.014)	(0.011)	(0.010)	(0.010)
Surname homogeneity × Year 2010	0.045**	-0.026*	-0.030**	-0.019*	-0.012
	(0.021)	(0.015)	(0.012)	(0.011)	(0.011)
Surname homogeneity × Year 2011	0.011	-0.058***	-0.059***	-0.044***	-0.035***
	(0.021)	(0.016)	(0.013)	(0.012)	(0.012)
Surname homogeneity × Year 2012	-0.018	-0.084***	-0.083***	-0.067***	-0.057***
	(0.022)	(0.017)	(0.014)	(0.013)	(0.013)
Surname homogeneity × Year 2013	-0.040*	-0.105***	-0.101***	-0.088***	-0.076***
	(0.022)	(0.018)	(0.015)	(0.015)	(0.014)
Surname homogeneity × Year 2014	-0.059**	-0.128***	-0.117***	-0.100***	-0.088***
	(0.023)	(0.020)	(0.016)	(0.016)	(0.016)
Surname homogeneity × Year 2015	-0.068***	-0.139***	-0.123***	-0.111***	-0.098***
	(0.023)	(0.022)	(0.018)	(0.018)	(0.018)
Surname homogeneity × Year 2016	-0.073***	-0.151***	-0.127***	-0.120***	-0.106***
	(0.023)	(0.023)	(0.018)	(0.019)	(0.018)
Surname homogeneity × Year 2017	-0.069***	-0.152***	-0.121***	-0.116***	-0.101***
	(0.023)	(0.022)	(0.017)	(0.017)	(0.017)
Surname homogeneity × Year 2018	-0.067***	-0.144***	-0.108***	-0.096***	-0.081***
	(0.022)	(0.023)	(0.018)	(0.017)	(0.016)
Surname homogeneity × Year 2019	-0.067***	-0.150***	-0.107***	-0.105***	-0.088***
	(0.022)	(0.020)	(0.015)	(0.015)	(0.015)
Geographic controls	No	Yes	Yes	Yes	Yes
Transportation & firm num	No	No	Yes	Yes	Yes
Demographic controls	No	No	No	Yes	Yes
Socioeconomic controls	No	No	No	No	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	53,735	53,615	53,615	53,615	52,495
R^2	0.667	0.697	0.710	0.720	0.732
Mean dependent vars.	0.078	0.078	0.078	0.078	0.078

Notes: County-level OLS regressions. The explanatory variables are the interactions between surname homogeneity among business owners and year dummies. We use two-way clustering of standard errors at the county and year levels. * p<0.1; ** p<0.05; *** p<0.01.