The Rise of American Ingenuity: Innovation and Inventors of the Golden Age*

Ufuk Akcigit John Grigsby Tom Nicholas

January 11, 2017

We examine the golden age of U.S. innovation by undertaking a major data collection exercise linking historical U.S. patents to state and county-level aggregates and matching inventors to Federal Censuses between 1880 and 1940. We identify a causal relationship between patented inventions and long-run economic growth and outline a basic framework for analyzing key macro and micro-level determinants. We find a positive relationship between innovation and drivers of regional performance including population density, financial development and geographic connectedness. We also explore the impact of social structure measured by slavery and religion. We then profile the characteristics of inventors and their life cycle finding that inventors were highly educated, positively selected through exit early in their careers, made time allocation decisions such as delayed marriage, and tended to migrate to places that were conducive to innovation. Father's income was positively correlated with becoming an inventor, though not when controlling for the child's education. We show there were strong financial returns to technological development. Finally, we document an inverted-U shaped relationship between inequality and innovation but also show that innovative places tended to be more socially mobile. Our new data help to address important questions related to innovation and long-run growth dynamics.

Keywords: Patents, innovation, inventors, growth, demographics, census, earnings, migration.

JEL classification: O31, O40, N11, N12.

*Authors' Addresses: Akcigit: University of Chicago, NBER, and CEPR (uakcigit@uchicago.edu). Grigsby: University of Chicago (jgrigsby@uchicago.edu). Nicholas: Harvard Business School (tnicholas@hbs.edu). We would like to thank Philippe Aghion, Harun Alp, Price Fishback, David Galenson, Rick Hornbeck, Ali Hortacsu, Xavier Jaravel, Taylor Jaworski, Ben Jones, Chad Jones, Fatih Karahan, Pete Klenow, Alan Marco, Casey Mulligan, Rob Shimer, Andrei Shleifer, Stefanie Stantcheva, Nancy Stokey, John Van Reenen, and Dietrich Vollrath, and the seminar and conference participants at the NBER Summer Institute Economic Growth (EFJK) Group, Development of the American Economy Group, Firms in Macroeconomics Conference, College de France, the USPTO, the University of Houston, the Fuji Conference in Tokyo, Queen's University, Koc University, and Northwestern for very helpful feedback and discussions. We thank the Minnesota Population Center for access to the Census data and William Kerr, Richard Hornbeck, and Price Fishback for sharing various datasets with us. Akcigit gratefully acknowledges the National Science Foundation, the Alfred P. Sloan Foundation, and the Ewing Marion Kauffman Foundation for financial support. Nicholas received funding from the Division of Research and Faculty Development at Harvard Business School.

1 Introduction

At the heart of the 25-year old endogenous growth literature is the basic premise that innovation and technological progress are engines of long-run economic growth (e.g., Romer (1990), Aghion and Howitt (1992)). Due to data limitations, however, evidence from longer time horizons is missing from this literature even though innovations from history such as light bulbs, air conditioners and storage batteries have tremendous influence on the way we currently live, consume, and produce (Gordon (2016)). Learning from the creators of these types of inventions has the potential to shed light on several current debates in the innovation and growth literatures. An abundance of modern data has helped to inform theoretical perspectives on key aspects of the growth process including the impact of firm entry and exit, reallocation, the role of cities, the distribution of talent, and the relationship between inequality, social mobility, innovation and growth (e.g., Klette and Kortum (2004), Akcigit and Kerr (2016), Acemoglu et al. (2015), Hsieh and Moretti (2015), Hsieh et al. (2013), Jones and Kim (2014), Aghion et al. (2015a)).

This paper develops a number of facts about the environment in which many of the essential technologies that we use today were created, the life cycle of inventors who developed these inventions and how innovation affected economic inequality and social mobility. We construct a large-scale micro-level historical dataset on inventors that can be mapped to both state and county-level economic aggregates and to socio-economic variables on individuals.

We use comprehensive data covering United States patents geocoded to the level of individual inventors and data matching inventors in patent records to Federal Censuses from 1880 to 1940. Typically such data has only been available historically for specific sub-samples of inventors (e.g., Lamoreaux and Sokoloff (1999)), superstar inventors (e.g., Khan and Sokoloff (2004)) or for broader populations in modern time periods (Aghion et al. (2015b), Bell et al. (2015)). The new data allow us to examine who became an inventor and the types of environments that were most conducive to innovation during a critical period of US economic growth.

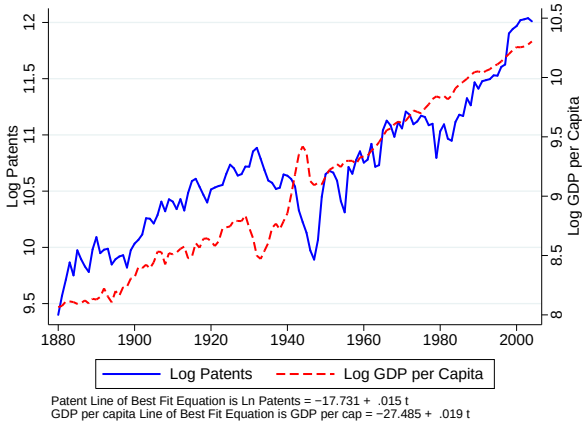
While the present paper presents numerous new facts about long run growth and innovation in the United States since the late nineteenth century, it is not unique in its goal. Kaldor (1961) presented six stylized facts around which the theory of economic growth developed. Jones and Romer (2010) updated Kaldor's facts to reflect the subsequent fifty years of data, providing the empirical foundations for modern growth theory. Both papers have facilitated informed discussion and permitted important breakthroughs in our understanding of the process of economic growth. Our paper has a similar goal in mind: to establish the fundamental facts regarding the process of innovation and its relationship to long run growth.

As an overview of the underlying innovation data, Figure 1 (Panel A) plots the time-series of log patents filed at the USPTO. It shows that innovative activity (proxy measured by patenting) has been growing over time. In keeping with the predictions of the large theoretical literature highlighting the central role of technological progress in endogenous growth we find a positive association between our innovation measures and output growth over the long run.

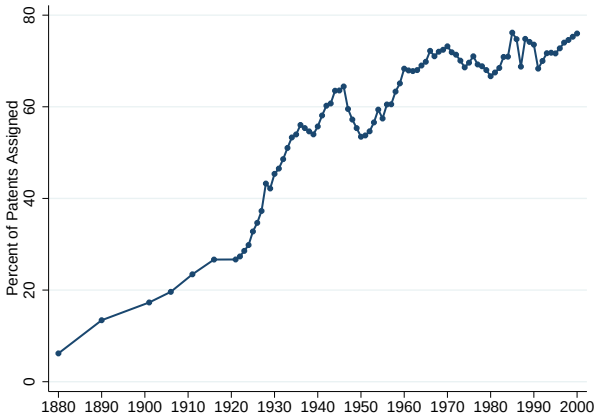
Panel B shows that at the beginning of our time period most inventors in the US developed technologies outside the boundaries of firms with the rising share of patents assigned over time

Figure 1: Long-Run History of Total Patents Filed in the USPTO

Panel A: Total Patents



Panel B: Share of Corporate Patents



Notes: Panel A: Excluding the exceptional years around World War II (1930-1950), a regression of annualized 5-year GDP per capita growth rates on log patents yields a coefficient of 0.639, which is statistically different from 0 at the 10% level. The correlation between these five year growth rates and log patents is 0.3915. Correlation between log patents and log GDP per capita (levels) is 0.8897, while the estimated linear relationship between the two is 0.7511 (with heteroskedasticity robust standard error 0.0325). Panel B: The share of US patents assigned at their grant date proxies for the amount of corporate patenting. Sources: USPTO, Maddison, Bureau of Economic Analysis, Klein (2013).

reflecting the development of R&D labs inside the modern corporation (Mowery and Rosenberg (1999)). The time period we cover is central to recent debates on innovation and growth. Indeed, we analyze the years that Gordon (2016) associates with the second industrial revolution, which produced major innovations like electricity and the motor vehicle. We also span the 1930s, which Field (2003) identifies as the most innovative decade of the twentieth century.

Roadmap We analyze the innovation process through the experiences of inventors over their life cycles. Some famous case studies can be useful to motivate our approach. Born to a poor family in rural Ohio, Thomas Edison (1847-1931) faced tight financing constraints in his early career. He ultimately relocated to New Jersey, building the Menlo Park Lab in 1876, a pioneering research laboratory where creative inventors could collaborate to develop new technical ideas. To develop his technologies further Edison accessed capital from a group of financiers, including J.P. Morgan. The investment bank Drexel, Morgan & Co. (which later became J.P. Morgan & Co.) provided loans, acted as a financial intermediary for Edison's firm, and provided wealth management (Brunner and Carr (2007)). Edison was granted 1,093 US patents. His experience suggests the importance of access to external finance, population density and human interactions in the innovation process. Edison's career also exemplifies social mobility and the potential for strong financial returns to innovation.

As another example, Nikola Tesla (1856-1943), a Serbian immigrant, demonstrates the contribution of international migration to US technological progress. Tesla was granted 112 US patents, the timing of which also illustrates how productivity could change over the life cycle of an inventor's career. 64 of Tesla's patents (57% of the total) were granted within a decade of his arrival in America in 1884 at age 28, where he worked at the Edison Machine Works in New

York (Carlson (2013)). But his patenting rate attenuated sharply later on. He acquired just 16 patents (14% of the total) after he was age 45. Tesla held uncommon personal preferences. He believed that private relationships detracted from productive research time, and never built a family. His decision not to marry shows that inventors faced trade-offs when it comes to time allocation.

As a final example, Melvin De Groote (1895-1963), one of the most prolific inventors in US history, received two degrees in Chemical Engineering. His highly-educated background was crucial for turning his creativity into valuable innovations. De Groote was granted 925 US patents, mostly developing novel methods to separate crude oil emulsion into its oil and water components.[1] De Groote moved from his state of birth in West Virginia to various places in the US where innovative firms were located.[2]

The above case studies hint that myriad factors, such as immigration, social mobility, access to finance, human interactions, education, time allocation, might spur innovation. Our new dataset can document these patterns in a systematic fashion, and thereby inform theories of endogenous growth. To organize our exposition, we set forth a very simple model of innovation. Consider a firm or agent deciding its investment in costly innovation activity. The agent can choose the probability p of discovering a new invention in order to maximize its expected profits. That is, the agent solves

$$\max_{p} \quad \left\{ p \cdot \Pi(\text{Macro factors}, \text{Micro Factors}) - C(p|\text{Micro Factors}, \text{Macro Factors}) \right\}$$

where $\Pi(\cdot)$ is a function returning the expected profit the firm can receive if it successfully innovates. This payoff function will depend on both surrounding micro and macro factors, such as personal income or local market size. Likewise $C(p|\cdot)$ represents the cost choosing an innovation probability p given the micro and macro factors that could impact the cost of innovation, such as education or the cost of borrowing. Naturally, there may be many factors that affect the costs and benefits of innovation. To address these systematically, we provide a framework to consider the innovation process, outlined in Figure 2.

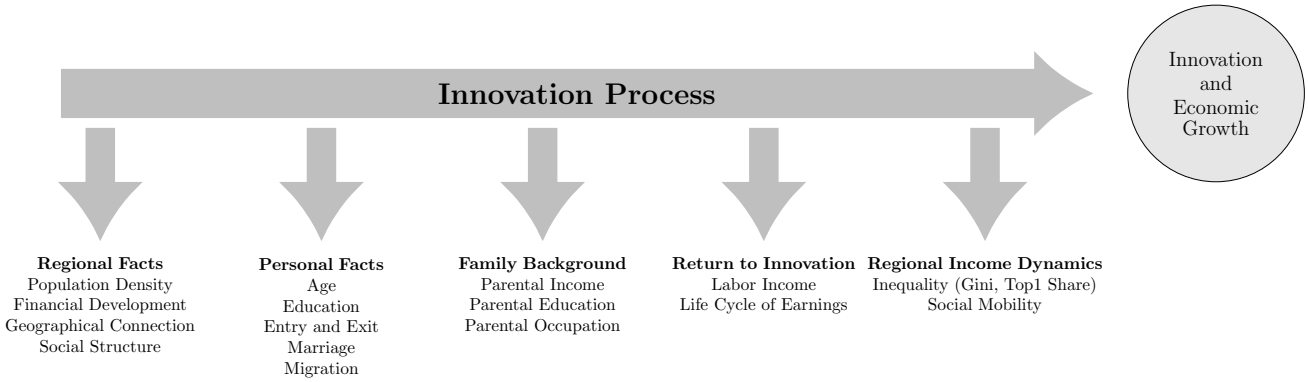
As a broad summary of our approach, we first establish a link between innovation and growth at the state level, and then consider the environment in which an inventor operates. We examine various characteristics of the inventors' state of residence including population density, financial development, geographical connectedness and social structure (measured using data on slave-owning families and religiosity). For our macro-level analysis we first present state-level results; we then perform checks for robustness of the findings at the county level.

Next we examine the basic demographic facts of inventors: their life cycle, education, and migration decisions. For example, we examine the family background of inventors, which may affect both the costs and benefits of innovation through factors such as relaxed credit constraints,

[1] De Groote is also noted as an innovator in food processing. For example, his inventions for extracting flavor from vanilla beans without using alcohol as a solvent were important to the production of ice cream.

[2] De Groote was a consultant or employee to numerous corporations including Hachmeister Lind Chemical of Pittsburgh, Procter and Gamble in Cincinnati, and Petrolite Corporation in Webster Groves, Missouri where he was Vice President and Director of Research.

Figure 2: Roadmap of the Analysis



or improved social networks, among others. We then provide evidence on the private rewards to innovation by studying the wages of successful inventors from labor income data in the 1940 Census. Finally, we investigate the societal consequences of innovation by establishing the correlation between patenting activity and income inequality or social mobility at the state level.

Our analysis uncovers the following stylized facts about the process of innovation and technological development during the golden age of US economic growth and beyond:

Regional Facts

Fact 1. More inventive states and sectors **grew faster** on average.

Fact 2. **Densely-populated** states were more inventive.

Fact 3. **Financially-developed** states were more inventive.

Fact 4. **Geographically-connected** states were more inventive.

Fact 5. States associated with **slavery** were less inventive but **religiosity** is not robustly correlated with inventiveness.

Personal Facts

Fact 6. Inventors were **more educated** on average and were most productive between the age of **36 and 55**.

Fact 7. Inventors were **positively selected** through exit early in their careers, but were less productive and more likely to exit late in their careers.

Fact 8. The patents of **new inventors received more citations** on average, and were more likely to be in the top decile of the citations distribution.

Fact 9. Inventors **delayed marriage** and had **fewer children**.

Fact 10. Inventors were more likely to **have migrated** from their state of birth. They moved to states that were more conducive to innovation.

Family Background

Fact 11. **Father's income** was correlated with becoming an inventor. This effect disappears once **child's education** is controlled for.

The Return to Innovation

Fact 12. Successful patentees had substantially higher **labor income**, even controlling for demographics, occupation, and education.

Fact 13. Inventor's income was highly correlated with the **quality of invention** measured by patent citations.

Inequality and Social Mobility

Fact 14. Broad measures of **income inequality**, such as 90/10 ratio, Gini coefficient, were negatively correlated with innovation.

Fact 15. However, the **top-1% income share** has a U-shaped relationship with innovation.

Fact 16. Innovation was strongly positively correlated with **social mobility**.

More specifically, our analysis of regional facts begins by identifying the relationship between innovation and long-run economic growth, as postulated by the active endogenous growth literature. We study the relationship between innovation and long-run growth across states over 100 years between 1900 and 2000. Our results show that the link has been strongly positive and economically sizable. Estimates suggest that if two states had the same initial GDP per capita in the beginning of the period and one state innovated four times more than the other (Massachusetts vs Wyoming, for instance), this could lead to 30% higher GDP per capita in the innovative state after 100 years. We verify a positive relationship in the data at the sectoral level. We also use an instrumental variables approach exploiting an historical shift in innovation activity during World War II to show that these estimates could be causal.

Our analysis of the characteristics of innovative states sheds important light on the macroeconomic environments where inventors operated. We find that innovative regions were more densely populated, were more financially developed, and were better connected to other parts of the country by transport links. We also test the extent to which the social structure of regions was correlated with their inventiveness given the notion that societal attitudes towards innovation affects "openness" to technological disruption (e.g., Florida (2002), Acemoglu et al. (2014)). We find that states associated with high levels of slave ownership before the Civil War were less inventive. Although we find some evidence of a negative relationship between religiosity and inventiveness, it is not as robust as that found for modern time periods (Benabou et al. (2013) and Benabou et al. (2015)). We relate these findings to large theoretical and empirical literatures in macroeconomics. In particular we highlight the modern drivers of innovation and long run economic growth which retain their importance in a historical context, and show which factors exhibit different patterns in recent and historical data.

We then transition from the macro environment to the micro-level. We find that inventors were not uneducated amateurs; rather they were typically highly-educated individuals and were most productive between ages 35-55. We report a range of summary statistics, indicating, for example, that the vast majority of inventors were white males.

To investigate personal characteristics further we benchmark the profile of inventors using different groups and sub-groups of the US population. First, we find that, just like individuals in other high-skill occupations, inventors tended to postpone marriage relative to the average

low-skill person in the population. Second, we find that inventors were significantly more likely to have migrated from their state of birth than *both* high-skill and low-skill persons. Third, we document that conditional on moving, inventors tended to relocate to more densely-populated and better financially-developed places.

Due to the richness of our patent data, where we can track inventors longitudinally, we also present novel life cycle facts. We find that inventors produced their highest quality inventions (measured by patent citations) early in their careers and that the early exit of low productivity inventors led to positive selection. Moreover, we find that the probability of exit was high as inventors faced obsolescence later in their careers. We relate these central findings to the empirical literature on productivity and firm dynamics (e.g., Haltiwanger (2012)) and theoretical models in which firm entry and exit play important roles in the growth development process (e.g., Acemoglu et al. (2015), Jovanovic (1982) and Hopenhayn (1992)).

We add to our analysis of career patterns by investigating the family background of inventors which reveals important differences between the extensive and intensive margins. While we find that some inventors gained privileged access into a career as an inventor, this effect operated only in the upper tail of the father income distribution. Entry was quite democratic, in keeping with the argument that rapid innovation and growth in the US economy at this time was facilitated by open intellectual property rights institutions (Khan (2009)). On the intensive margin we find no effect of father's income on patent productivity or quality.

Innovation is a process where inventors invest in costly effort *ex ante* in the expectation of gaining *ex post* returns, financial or otherwise (e.g., Stern (2004)). Yet, the financial returns to inventors in particular are rarely measurable even for modern time periods (e.g., Astebro (2003)). Subject to some limitations associated with the income data from the 1940 Census, (which we discuss in Section 2) we show that inventors had high incomes, even after controlling for their observable characteristics. We find that inventors had three times higher labor income on average. Indeed, inventors had a steeper earnings profile over their life cycle. Fully 73% of inventors were in the top decile of the overall income distribution. We identify strong returns to the quality of innovation: inventors with higher citation-adjusted patents received higher wage income.

Finally, in a key component of our analysis, we study regional income dynamics as an outcome measure in relation to prior-period patenting activity. We focus on various measures of inequality: the 90/10 ratio, the Gini coefficient and the top-1% income share. We also construct a measure of social mobility using information in the 1940 Census that focuses on the fraction of those with a low-skill father who themselves have a high-skill occupation. We find that innovative regions in the US had lower income inequality measured as the 90/10 ratio or the Gini coefficient, yet the top income share features a U-shaped relationship with state innovation. In general we find that the most innovative states had higher levels of social mobility.

Overall, our analysis uses novel historical microdata linked to regional aggregates to provide key macro and micro-level facts to inform critical questions in the study of technological progress and long-run economic growth. The remainder of the paper is organized as follows. Section 2 outlines our data, Section 3 presents the empirical results. Section 4 codifies our findings in

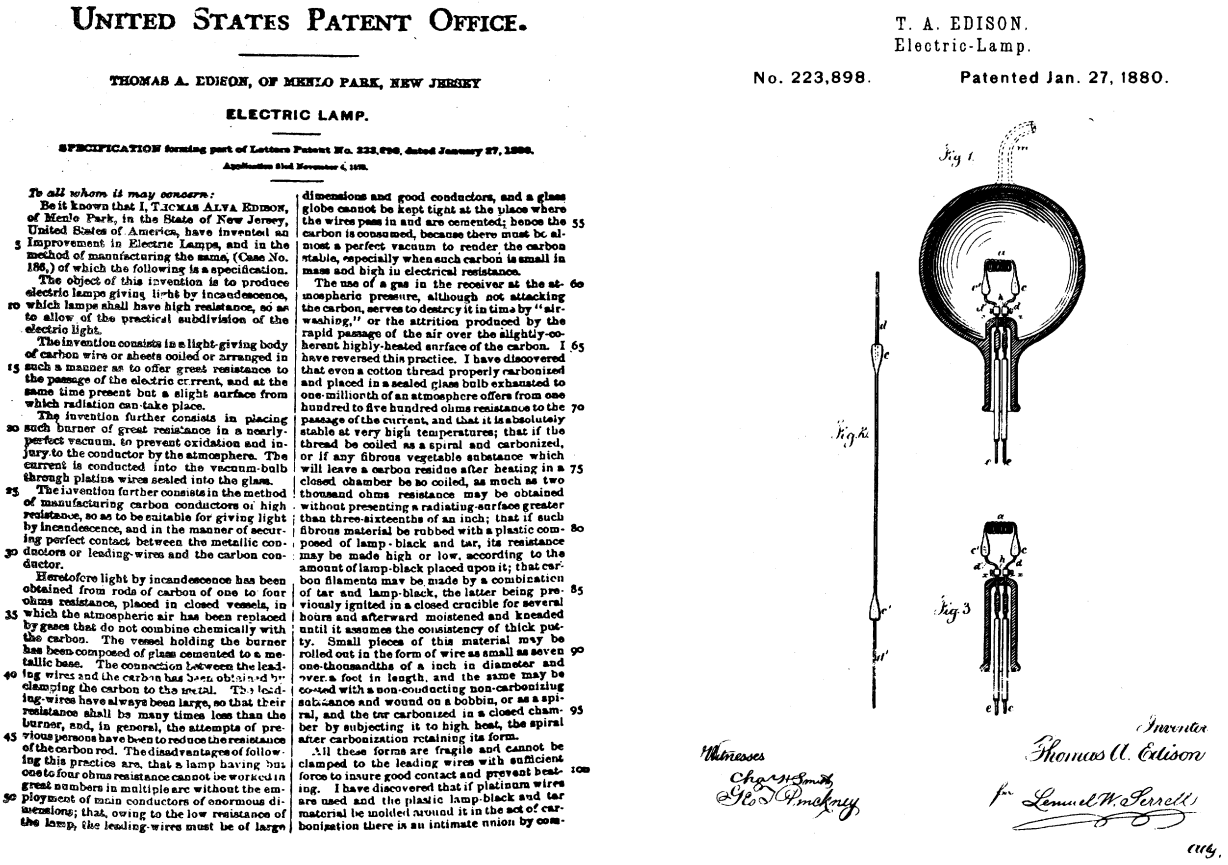
relation to existing studies, and Section 5 concludes. Appendices A to D provide a detailed description of the data we use, our matching methodology and additional robustness checks.

2 Data Construction

Patents, Technology Areas and Citations

Patents are a commonly used measure of innovation in the empirical literature on technological change. A patent entry shows the name and address of an inventor: it shows the surname, first name, middle initial(s) where relevant, state, city, county, and country of the applicant when the patent was granted. For example, Figure 3 shows the famous USPTO patent 223,898 granted to Thomas Edison in Menlo Park on January 27, 1880. Patents also represent transferable property rights. If a patent was assigned to another individual or firm, the assignee is also recorded on the document. Edison's patent was unassigned at the grant date. As shown in Figure 1 (Panel B) the share of patent assignments in the US grew over time.

Figure 3: Pages of USPTO Patent Number 223,898



UNITED STATES PATENT OFFICE.

THOMAS A. EDISON, OF MENLO PARK, NEW JERSEY

ELECTRIC LAMP.

SPECIFICATION forming part of Letters Patent No. 223,898, dated January 27, 1880.

Application filed November 4, 1879.

To all whom it may concern:

Be it known that I, THOMAS ALVA EDISON, of Menlo Park, in the State of New Jersey, United States of America, have invented an Improvement in Electric Lamps, and in the method of manufacturing the same, (Case No. 186,) of which the following is a specification.

The object of this invention is to produce electric lamps giving light by incandescence, which lamps shall have high resistance, so as to allow of the practical subdivision of the electric light.

The invention consists in a light-giving body of carbon wire or sheets coiled or arranged in such a manner as to offer great resistance to the passage of the electric current, and at the same time present but a slight surface from which radiation can take place.

The invention further consists in placing such burner of great resistance in a nearly-perfect vacuum, to prevent oxidation and injury to the conductor by the atmosphere. The current is conducted into the vacuum-bulb through platina wires sealed into the glass.

The invention further consists in the method of manufacturing carbon conductors of high resistance, so as to be suitable for giving light by incandescence, and in the manner of securing perfect contact between the metallic conductors or leading-wires and the carbon conductor.

Heretofore light by incandescence has been obtained from rods of carbon of one to four ohms resistance, placed in closed vessels, in which the atmospheric air has been replaced by gases that do not combine chemically with the carbon. The vessel holding the burner has been composed of glass cemented to a metallic base. The connection between the leading wires and the carbon has been obtained by clamping the carbon to the metal. The leading-wires have always been large, so that their resistance shall be many times less than the burner, and, in general, the attempts of previous persons have been to reduce the resistance of the carbon rod. The disadvantages of following this practice are, that a lamp having but one to four ohms resistance cannot be worked in great numbers in multiple arc without the employment of main conductors of enormous dimensions; that, owing to the low resistance of the lamp, the leading-wires must be of large dimensions and good conductors, and a glass globe cannot be kept tight at the place where the wires pass in and are cemented; hence the carbon is consumed, because there must be almost a perfect vacuum to render the carbon stable, especially when such carbon is small in mass and high in electrical resistance.

The use of a gas in the receiver at the atmospheric pressure, although not attacking the carbon, serves to destroy it in time by "air-washing," or the attrition produced by the rapid passage of the air over the slightly-coherent highly-heated surface of the carbon. I have reversed this practice. I have discovered that even a cotton thread properly carbonized and placed in a sealed glass bulb exhausted to one-millionth of an atmosphere offers from one hundred to five hundred ohms resistance to the passage of the current, and that it is absolutely stable at very high temperatures; that if the thread be coiled as a spiral and carbonized, or if any fibrous vegetable substance which will leave a carbon residue after heating in a closed chamber be so coiled, as much as two thousand ohms resistance may be obtained without presenting a radiating-surface greater than three-sixteenths of an inch; that if such fibrous material be rubbed with a plastic composed of lamp-black and tar, its resistance may be made high or low, according to the amount of lamp-black placed upon it; that carbon filaments may be made by a combination of tar and lamp-black, the latter being previously ignited in a closed crucible for several hours and afterward moistened and kneaded until it assumes the consistency of thick putty. Small pieces of this material may be rolled out in the form of wire as small as seven one-thousandths of a inch in diameter and over a foot in length, and the same may be coated with a non-conducting non-carbonizing substance and wound on a bobbin, or as a spiral, and the tar carbonized in a closed chamber by subjecting it to high heat, the spiral after carbonization retaining its form.

All these forms are fragile and cannot be clamped to the leading wires with sufficient force to insure good contact and prevent heating. I have discovered that if platinum wires are used and the plastic lamp-black and tar material be molded around it in the act of carbonization there is an intimate union by com-

T. A. EDISON.
Electric-Lamp.

No. 223,898. Patented Jan. 27, 1880.

Fig. 1.

Fig. 2.

Fig. 3.

Witnesses
Chas. G. Smith
Geo. H. Pomeroy

Inventor
Thomas A. Edison
per Lemuel W. Serrell
atty.

We use a comprehensive collection of over 6 million US patents, which allows us to gain unique insights into the characteristics of US inventive activity over longer time horizons, by linking patents to both regional economic and social aggregates and by matching patents at the inventor-level with information in the decennial Federal Censuses from 1880 to 1940. The construction of the patent data are described in Appendix B.

In addition to the raw patent data we use two datasets to augment the information available from the original patent documents. First, we use the USPTOs classification of patents to isolate

the technology area of inventions. This classification is consistent over time because whenever a new classification is introduced, existing patents are retroactively re-classified. Patents list several technological components, and we observe main classes and subclasses for each invention. Second, we use historical patent citations to identify the most influential technological development. Our data include 3.8 million citations to patents granted between 1880 and 1940 from the population of patents granted between February 1947 (when front page citations began to be systematically recorded) and September 2008. Following Hall et al. (2001), we adjust citations to account for bias due to truncation or aggregate fluctuations in citation propensity.[3]

Several aspects of patenting are worth highlighting in the context of our linkage between inventors and the Census records. First, access to patenting was widespread. The cost of obtaining a US patent was very low by international standards. Lerner (2002) estimates that in 1875 to hold a patent to a full term of 17 years in the United States cost just 5% of the amount in the United Kingdom, 11% of the UK amount in 1900 and 10% of the UK amount in 1925. Moreover, inventors were actively encouraged to innovate and file for patents by the way the application process was configured. They could mail documents to the US patent office in Washington, DC through the extensive network of post offices connecting the country (Khan (2009), Acemoglu et al. (2016a)). Inventors could also use a large network of intermediaries (patent agents and lawyers) to administer the patenting process (Lamoreaux and Sokoloff (1999)).

Second, although patents could be sold because a market for technology had flourished since the middle of the nineteenth century (e.g., Lamoreaux and Sokoloff (1999), Akcigit et al. (2016a)) the location of the original inventor is still recorded on the patent document. US patent law stipulated that the "first and true inventor" be listed in the patent application even if the patent was assigned to another individual or firm at its grant date. Even as the law of the of employee-inventor evolved to make pre-invention employee assignment agreements legally enforceable, the original inventor can be identified on the patent (Fisk (2009)).

Third, at least for the early years of our study the timing of a patent being filed by an applicant and the timing of a patent grant—when we observe their location—were quite close. In 1880 an average of 170 days elapsed between the filing and grant date. In 1900 343 days elapsed. For this time period there is a reasonable alignment between the patent grant year and when an individual was observed in the Census year. By 1910, however, the average patent pendency period was almost a year-and-a-half (536 days). In 1930 it had extended to over 1,000 days and it was still over 800 days in 1940.[4] We would therefore expect to see more measurement error in our matching *ceteris paribus* for later years. In terms of the further link between the timing of the idea behind an invention and its fruition Schmookler (1966) reports that it took about one-and-a-half years for an invention by an independent inventor to be produced.

[3]Details of the citation adjustment are provided in Appendix B.1.

[4]By comparison, the average difference between a patent application and grant date for US patents granted between 2008 and 2015 was 1,278 days.

Census Data

The release of the complete-count Census data by the Minnesota Population Center (IPUMS) provides an opportunity to examine a number of questions related to the historical development of innovation in the United States. We use the decennial Censuses in 1880, 1900, 1910, 1920, 1930, and 1940.[5] We start our patent-Census matching exercise in 1880 because that is the first year a reasonable number of patent observations become available. Around 11,400 patents were granted to inventors residing in the US in that particular year. We end in 1940 because that is the latest Census available given the Census Bureau's 72 year lag release rule.

We view these data analogously to modern studies using administrative records such as Bell et al. (2015) who uncover major new facts about the nature of US innovation. Not only is it possible to link the historical Census data with patent records (as we show below) but data on the entire population permits analysis of inventor life cycles relative to other sub-groups of individuals—for instance with different occupational skill levels. This type of systematic information across large groups of individuals for the entire United States has never been available before for long historical horizons. Nevertheless, although the Censuses present an especially useful source of data, it is also worthwhile understanding some of the potential limitations.

First, the quality of the Census records varies over time. While the Census included quality control procedures in an effort to ensure consistent enumeration, a lot depended on the way the Censuses were generally administered. For example, the 1920 Census was conducted in the winter (January 1st 1920) whereas the 1910 Census had been conducted in the spring (April 15th). Winter enumeration had a large effect on seasonal occupations like agricultural labor and movement back and forth to cities. Although we show in Appendix C that our match rate is lower for 1920 than for other years, the level of underenumeration is not sufficient to bias our results. Dorn (1937) estimates underenumeration in the native white population (which would be most relevant to the inventors in our dataset) of between 1% and 1.1%.

Second, beyond standard variables like the name and location of individuals, the information contained in the Censuses varies widely over our period of interest. Although a number of variables are commonly recorded across Census years such as age, race, gender and marital status, in other instances variables are recorded in one year only to be dropped in another. As an example, occupation is listed in 1880 but not in 1900 or 1910. Generally, a wider array of variables are available in later years. Beginning in 1920, for instance, enumerators asked specifically about education (school attendance) and home ownership.

Third, some variables are subject to measurement error. An advantage of the 1940 Census is that it questioned individuals about income. Prior to the availability of these data, researchers routinely imputed incomes by assigning individuals the median income in their reported occupational category (e.g., Abramitzky et al. (2014)). But, the income data needs to be carefully interpreted. Enumerators were instructed to report incomes of greater than $5,000 a year at

[5]For a full description of the Census datasets, the variables they contain, and our attempts to clean them, see Appendix A. The 1890 Census was largely destroyed in a fire in 1921 and others records from this Census were destroyed under intransigent Federal record management polices in place at the time. Only a limited set of 1890 Census schedules survived.

$5,000+. Hence, the data are "top-coded." For example, Melvin De Groote, a superstar inventor we profiled earlier in the paper, reports this level of income in the 1940 Census (average income was $1,368). Furthermore, Petro (2016) finds that "if a farmer worked for himself and sold his crops, he did not report that money" in the 1940 Census. By the same token we assume that inventors selling their inventions would not have reported this as income.

Approach to Matching Patent and Census Records

The main challenge associated with matching inventors, as listed on patent documents, to individuals listed in the Census, is the absence of a unique identifier across datasets. The first Social Security number was issued in 1936. Although a supplementary question was asked in the 1940 Census about whether a person had a Social Security number, the number itself was not recorded. Social Security numbers were not included in patent documents at the time.

In both the patent and Census datasets we observe variables denoting surname, first name, initial, state, city and county. This vector of information provides a basis for our matching. Of course, the challenge of matching observations without unique identifiers is self-evident. For example, in their study of intergenerational occupational mobility, Long and Ferrie (2013) tracked fathers and sons across the 1851 Census and the 1881 Census using the proximity of the name and birth year. In a recent paper, Feigenbaum (2015) estimates intergenerational income mobility using a machine-learning approach to match individuals from the 1915 Iowa State Census and the 1940 US Federal Census. He matches on first name, surname, middle initial, state of birth, and year of birth finding approximately a 59 percent match rate.

Unlike these prior studies we do not observe year of birth in the patent records, but we can still limit the likelihood of matching "false positives" by restricting our analysis to only those observations where we match precisely across a range of our matching variables. We proceed in two steps. First, we adopt a "basic" matching approach where the criterion for matching is that the inventor listed on the patent has the same first name and surname as the individual in the Census, and lives in the same state. Naturally, this leads to repeated individuals in some cases.

Therefore, we next adopt a "refined" matching approach. In addition to the criterion in our basic match we require additionally that individuals listed on the patent document and individuals in the Census reside in the same county. Then, if there are still many observations for a given inventor, we first check if there is an inventor which has the same middle initial in both the patent and Census datasets, and we keep that inventor if there is a match. We then keep only Census inventors who live in the same city or township as is listed on the patent document, if one exists. Next, we ask if there is any matched inventor between 16 and 85 years old. If so, we keep that inventor only. Finally, we repeat the age refinement, keeping only matched inventors between 18 and 65 year old, if one exists. In other words, to be in our final dataset requires that individuals match systematically on surname, first name, (where relevant initial), state, and county. Although there are still data matching issues we cannot overcome—for example, sometimes the Census uses registration areas (e.g., Precincts or Districts) rather than cities, making it impossible for us to identify the right individual by location—overall our

matching rates are encouraging. We match an average of 46 percent of patentees in the Census with a high of 62 percent in 1880 and a low of 34 percent in 1920. A detailed description of the matching process is provided in Appendix C.

Additional Data Sources

Finally, our analysis uses additional important datasets.

1. **Demographics**. Intercensal population estimates are provided by the Census Bureau. We also use the Census data to construct measures of slave ownership (from the 1860 Census), and we construct measures of religiosity from the Census of Religious Bodies.
2. **State-level Output**. State output data is taken from the Bureau of Economic Analysis (BEA) for 1929 through the present day. In addition, Klein (2013) provides estimates of gross state products in 1880, 1890, 1900, and 1910 and Martin (1939) provides estimates of state incomes from 1919 to 1938.[6]
3. **Sector-level Output**. Sector value added and full-time equivalent employment data come from the BEA.[7]
4. **Financial Development Measures**. To establish proxies for financial development, we use Federal Deposit Insurance Corporation (FDIC) data. The FDIC dataset, from the University of Michigan's ICPSR repository (number 0007) provides the number of deposits, banks, and bank suspensions at a county level from 1920-1936.
5. **Transportation Cost**. Obtained from Donaldson and Hornbeck (2016).
6. **Scientific Research and Development Contracts**. In our instrumented growth regressions (detailed below) we use data obtained from the Library of Congress on Office of Scientific Research and Development (OSRD) contracts for technological development efforts during World War II.

2.1 Summary Statistics

As a precursor to the main analysis we present descriptive statistics on our data, which we discuss in more detail in subsequent sections. In keeping with our approach of examining US innovation from macroeconomic and microeconomic perspectives, we structure these data to characterize inventiveness at both the state and individual inventor levels.

In Table 1, we compare the most and least innovative 10 states on an average patents per capita basis between 1880 and 1940 whereas in Table 2 we report the summary statistics for all states. In Table 3, we present a statistical profile of inventors based on information in the Censuses and we compare inventor characteristics with those of the population as a whole.

As Table 1 shows, the top most and least inventive states differed along numerous observable dimensions. The top 10 inventive states were more populous and much more densely populated.

[6] These data were provided to us courtesy of Price Fishback.

[7] Downloaded from http://www.bea.gov/industry/gdpbyind_data.htm. We use data between 1948 and 1986, as the SIC codes change then. The match from USPTO classes to SIC codes is done using files provided by Bill Kerr and assigns patent classes to the SIC code which manufactures the highest share of patents in that class.

Table 1: The Characteristics of Inventive States

	Top 10 Inventive States	Bottom 10 Inventive States
Av. Population (000s)	2716	1437
Population Density (Pop per km^2)	38.85	11.35
GDP per Capita	623.5	238.6
Av. Patents per 10,000 People	5.8	0.7
Av. Patents Granted	1571	96.94
Av. Inventors per 10,000 People	3.33	0.57
% Interstate Migrant	39.80%	19.96%
% International Migrant	20.64%	1.68%
Migrant Inventors per 10,000 Interstate Migrants	29.26	6.01
Migrant Inventors per 10,000 International Migrants	22.10	7.78
Percent White	96.2%	66.8%
Percent Black	2.8%	30.5%
% Over 35 Years Old	38.5%	27.7%
% Under 35 Years Old	61.5%	72.3%
No Schooling	2.88%	3.95%
Less Than High School	65.16%	74.69%
High School	20.77%	12.32%
Some College	6.25%	5.79%
4+ Years College	4.94%	3.24%
High Skill Occupation	9.3%	4.7%
Medium Skill Occ.	27.4%	19.3%
Low Skill Occ.	63.3%	75.9%
Employment Rate	54.69%	54.33%
Mean Wage Income	1140.7	681.4
Median Wage Income	947.6	459.4
Male Labor Force Participation	89.05%	88.21%
Female Labor Force Participation	33.31%	28.40%
% of Minorities with High Skill Job	2.85%	0.86%

Notes: Innovation measured as the average number of patents per capita between 1880 and 1940. The top 10 states are: California, Connecticut, Delaware, Illinois, Massachusetts, Nevada, New Jersey, New York, Ohio, and Rhode Island. The bottom 10 states are: Alabama, Arkansas, Georgia, Mississippi, New Mexico, North Carolina, North Dakota, Oklahoma, South Carolina, and Tennessee. Source: 1880 through 1940 Historical Census Data, USPTO patent records.

Table 2: State-Level Summary Statistics

	Mean	Std. Dev.	Min	Max
Patents per 10,000: Average between 1880 and 1940	2.48	1.73	0.40	7.37
Patents per 10,000: Average between 1920 and 1930	2.79	2.04	0.40	11.21
Patents per 10,000: Average 1940-2004	1.93	1.31	0.33	7.39
% Living in Urban Area (1940)	47.60	15.35	22.58	82.55
% Living on Farm (1940)	23.63	13.30	2.00	57.41
Female LFP Rate (1940)	27.85	5.51	18.90	40.45
Percent of Families Owning Slaves (1860)	0.11	0.16	0.00	0.49
Deposits per Capita (1920)	277.08	133.52	0.00	785.67
Bank Failure Rate (1920)	1.83	2.27	0.00	9.46
Mean outgoing transport cost (1880)	10.80	2.72	7.84	18.13
90-10 Wage Income Ratio (1940)	12.30	2.98	6.93	18.00
Gini coefficient of wage income (1940)	0.44	0.04	0.38	0.53
Share of Wage Income Held by Top 1% (1940)	6.31	1.57	4.11	11.30
Percent with low-skill father who have high-skill occupation (1940)	2.14	0.85	0.47	4.32

Sources: Historical Census Data, USPTO patent records, FDIC, Donaldson and Hornbeck (2016).

Indeed, the top 10 inventive states contained 11 (55%) of the largest 20 cities in the United States in 1880 and 9 (45%) of the the largest 20 cities in 1940. By contrast, the bottom 10 inventive states contained none. The inventive states were distinctly more urban.

Additionally, the most inventive states had higher GDP per capita, and average and median wage income suggesting a positive relationship between inventive activity and the level of economic development. The population in the inventive states was also more educated, implying a link between human capital and inventiveness. The most inventive states also had larger migrant populations, especially international migrants, who accounted for about one-fifth of the population in the top 10 inventive states but less than one-fiftieth of the population in the bottom 10 inventive states. The population of the most inventive states was largely white, whereas almost a third of the population in the least inventive states were black. Notably, seven of the bottom 10 inventive states were associated with slavery in the antebellum era.

In terms of similarities, the states shared a largely common age structure. The employment rate was almost exactly the same across these state groupings and both the male and female labor force participation rates were roughly equivalent. However, the distribution of occupations was different in that the inventive states contained a larger share of individuals with a high-skill job. These states also appear to have been more "open" to external influences insofar as this can be inferred from minorities being more likely to be employed in a high-skill job.

Table 3: The Characteristics of Inventors

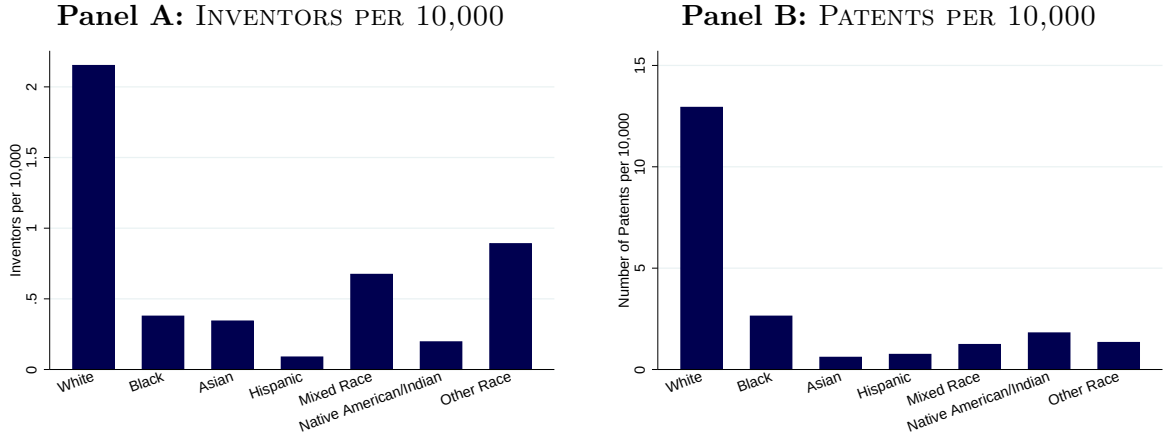
	Inventors	Full U.S.
Percent White	97.9%	89.4%
Percent Black	1.8%	9.1%
Percent Male	97.9%	51.0%
Single	16.1%	27.7%
Married	80.2%	65.4%
Percent 19-25	8.4%	22.6%
Percent 26-35	23.8%	27.5%
Percent 36-45	31.0%	22.5%
Percent 46-55	24.1%	16.6%
Percent 56-65	12.7%	10.8%
Prob. Child: $\leq$ 35 yrs old	72.9%	80.0%
Prob. Child: $>$ 35 yrs old	80.9%	89.7%
Av. # Children: $\leq$ 35 yrs old	1.9	2.3
Av. # Children: $>$ 35 yrs old	3.2	4.7
Percent Interstate Migrant	58.8%	42.8%
Percent International Migrant	21.1%	17.4%
Percent Born in Great Britain	5.19%	3.46%
Percent Born in Germany	4.0%	2.67%
Percent Born in Other Europe	8.72%	8.27%
Percent Born in Canada	2.56%	1.73%
Percent Born in Other Countries	0.65%	1.24%
Percent Of Population	0.02%	99.98%

Notes: We use all matched census records to construct this table. Age, race, marital status, and migrant status are reported for all years. Fertility is reported only in 1910 and 1940. Source: 1880 through 1940 Historical Census Data, USPTO patent records.

Meanwhile, Table 3 shows inventors were more likely to be married, middle-aged, white males, who had fewer children early and were migrants relative to the population as a whole. Figure

4 highlights again that our data consists overwhelmingly of white male inventors. Our time period roughly coincides with phases Goldin (2006) identifies (i.e., from the 1880s to the 1950s) where women's involvement in the labor market was generally restricted to positions like office and clerical work. There were some exceptional female inventors, but not very many. For example, Khan and Sokoloff (2004) found only one female inventor in their list of 400 superstar US inventors listed in the *Dictionary of American Biography* who were born before 1886.

Figure 4: Patenting Behavior by Race



Notes: Figure plots the number of inventors (panel A) and patents (panel B) per 10,000 census individuals by race, aggregating all census years together. Source: 1880 through 1940 Historical Census Data, USPTO patent records.

Finally, the base of Table 3 shows that inventors represented a very small share of the population.[8] Yet, when knowledge diffuses rapidly, inventors developing breakthrough inventions can have a large influence on economic growth (e.g., Comin and Hobijn (2010)). During the US golden age and major epochs of economic development more generally, the technological ingenuity and innovative capabilities of the minority tended to matter the most (e.g., Mokyr and Voth (2012), Squicciarini and Voigtlaender (2015)).

3 Empirical Analysis

3.1 Regional Facts

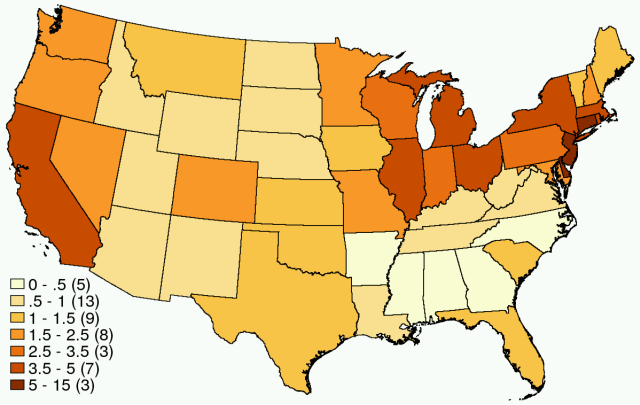
Following the road map outlined in Figure 2 we start by discussing the regional facts that we observe in our data. In particular, we first exhibit the strong positive relationship between innovation and economic growth at the state level. We then consider the relationship between innovation and population density, financial development, geographic connectivity at the state level, and social structure measured by association with slavery and religiosity. As checks on our state-level results, where possible we carry out the analysis at the county level for robustness. The county-level results are reported in subsections labeled "Robustness" within each fact.

[8] By comparison, in all years for which occupation data is reported in the Census, 0.46% of the working age population was a doctor or a lawyer.

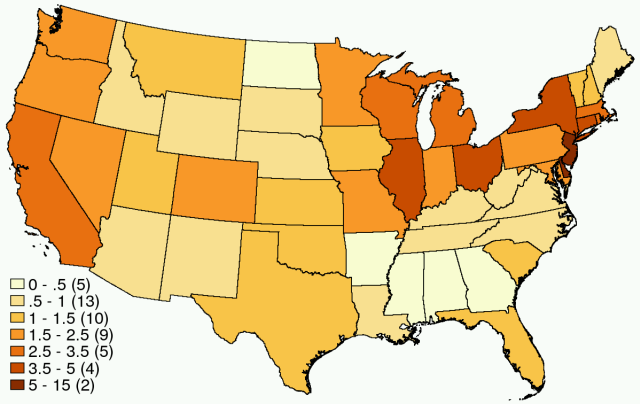
Figures 5A and 5B illustrate the geography of inventiveness defined as patents and inventors per 10,000 people in 1940. Both figures reveal concentrations of activity in rust-belt manufacturing areas, which mirrors the distribution of industrial activity at the time (Glaeser (2011)). California also stands out as a center of innovation and this holds for most of the years we observe. This is not caused by sparse population counts mechanically inflating the patent and inventor counts. While Los Angeles ranked as the 36th largest city in the US in 1900, it was ranked number 10 in 1920 and number 5 in 1940. Wyoming was innovative in several of our snapshot years, perhaps because of developments related to the evolution of the Union Pacific Railroad. However, as noted above, Wyoming under-performed relative to other states in GDP per capita growth over the very long run. Figures A-6 and A-7 in Appendix D show that the geographic patterns displayed in Figure 5 are remarkably stable across our six census years.

Figure 5: The Geography of Inventiveness

Panel A: Patents per 10,000 People



Panel B: Inventors per 10,000 People



Notes: Figure maps the number of patents (panel A) or inventors (panel B) per 10,000 residents in each state of the mainland US in 1940. Darker colors represent more inventive activity per resident. Patent data come from the USPTO's historical patent files, while population counts are calculated using the US Census. Appendix D reports similar maps in different decennial census years.

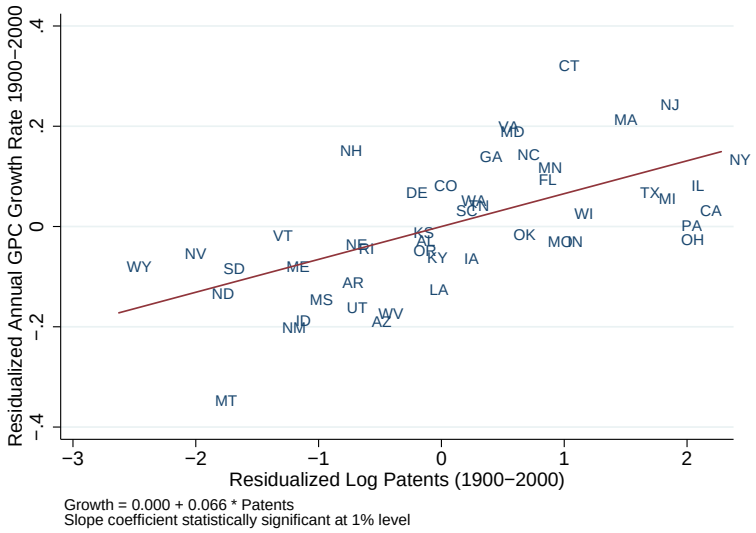
Fact 1. ***More inventive states and sectors grew faster on average.***

The long-standing endogenous growth literature builds on the premise that long-run growth is driven by innovation and technological progress. Although this idea is intuitive, providing empirical support for this premise has been challenging due to data limitations on historical innovations. In fact, even though a large literature has studied the empirical determinants of macro-level economic growth (e.g., Barro (1991)), to our knowledge no study has documented a causal empirical relationship between innovation and growth for the US over the long run.

Figure 6 shows the basic correlation between a proxy measure of innovation (patents) and economic growth. To account for scale effects we plot variables residualized against 1900 log GDP per capita. The relationship is strongly positive.

Table 4 reports coefficients from growth regressions controlling for the long-run effects of initial conditions and population density. The dependent variable in these regressions is the annualized growth rate in state-level GDP per capita between 1900 and 2000. We find that the log of patents

Figure 6: Innovation and Long-run Growth: US States between 1900-2000



Notes: Figure plots the total number of patents granted to inventors in each state between 1900 and 2000 on the horizontal axis, and the annualized growth rate in state GDP per capita between 1900 and 2000 on the vertical axis. Both horizontal and vertical axes plot the variables of interest residualized against 1900 log GDP per capita, to control for conditional convergence. Source: BEA Historical Regional Economic Accounts, and Klein (2013)

granted between 1900 and 2000 had a consistently positive and statistically significant effect in columns 1 and 2. These results are robust in columns 3 and 4 to measuring the growth rate using the approach established by Davis et al. (1996) in the employment literature that corrects for any potential bias associated with transitory shocks to growth and mean reversion.[9]

Table 4: Innovation and Long Run Growth: US States between 1900-2000

	Annualized Growth Rate		DHS Growth Rate	
	(1)	(2)	(3)	(4)
Log Patents	0.066***	0.054***	0.031***	0.026***
	(0.013)	(0.012)	(0.008)	(0.007)
Initial GDP per Capita	-0.877***	-0.891***	-0.324***	-0.330***
	(0.036)	(0.036)	(0.025)	(0.026)
Population Density		1.145*		0.517*
		(0.588)		(0.304)
Observations	48	48	48	48
Mean Growth	2.154	2.154	1.552	1.552
Std. Dev. of Growth	0.417	0.417	0.159	0.159

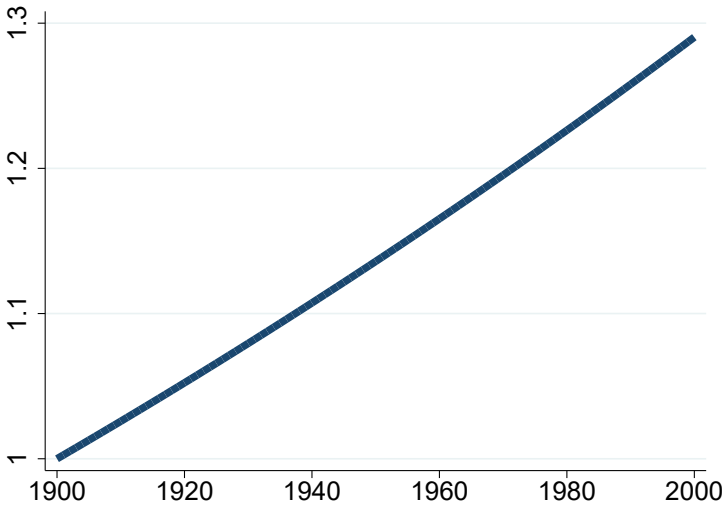
Notes: Table reports estimated coefficients from a regression in which the dependent variable is the state-level annualized growth rate in real GDP per capita from 1900-2000. White heteroskedasticity robust standard errors reported in parentheses. DHS growth rate refers to the growth rate measure as proposed by Davis, Haltiwanger, and Schuh. Output data provided by Klein (2013) and the Bureau of Economic Analysis. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level.

The economic magnitude of these estimates is especially informative. As an example, consider Massachusetts (MA) versus Wyoming (WY). As shown in Figure 6, Massachusetts had 4 times as many patents per capita than Wyoming during the twentieth century. Assume MA and WY

[9]Figure A-5 in appendix D shows that this strong positive relationship between long run growth and innovation holds for historical output calculated using the methodology of Martin (1939).

had the same initial GDP per capita in 1900 and identical population densities. Our estimated coefficients imply that the gap between MA and WY would have increased dramatically as illustrated in Figure 7. By the end of the century, MA would be 30% richer than WY just because of the differences in their innovativeness.

Figure 7: GDP Per Capita Ratio: Massachusetts/Wyoming

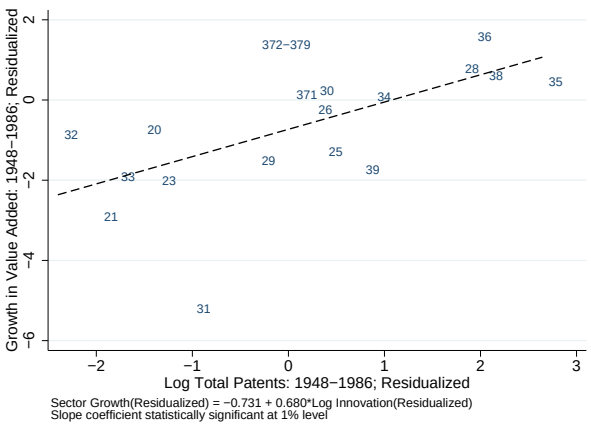


Sector-level Analysis

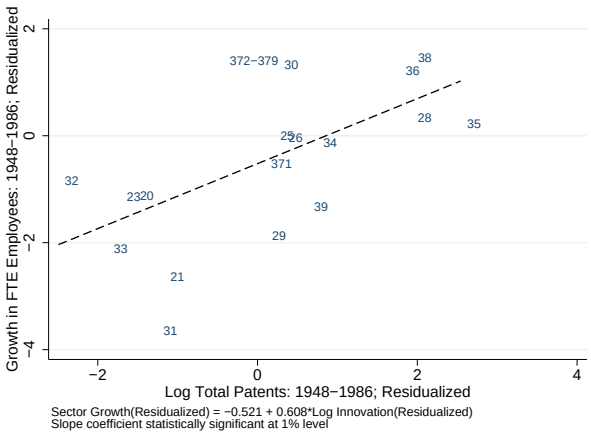
The positive relationship between innovation and output growth persists at the sector-level, as shown in Figure 8.[10]

Figure 8: Innovation and Long-run Growth: 3-digit Sectors between 1948-1986

Panel A: Growth in Value Added



Panel B: Growth in FTE Employees



Notes: Figure plots industry-level annualized growth in value added (Panel A) and full-time-equivalent employees (Panel B) against the log total patents used by the industry between 1948-1986, before the change of SIC code definitions in 1987. Both horizontal and vertical axes are residualized against 1948 value added (Panel A) or full-time-equivalent employees (Panel B). Each point represents a 2-digit SIC code, before the codes were changed in 1987. Patent classes are matched to sectors using data provided by William Kerr [3-digit version comes from Kerr (2008) and 4-digit comes from Acemoglu et al. (2016b)]. A patent class k is matched to an industry s if s is the modal user of patents from k. Industry data provided by the Bureau of Economic Analysis.

[10]For details on the match between patent classes and industries, see Appendix B.2.

Figure 8 plots industry-level annualized growth in value added (Panel A) and full-time-equivalent employees (Panel B) against the log total patents used by the industry between 1948-1986, before the change of SIC code definitions in 1987. Both horizontal and vertical axes are residualized against 1948 value added (Panel A) or full-time-equivalent employees (Panel B). Each point represents a 2-digit SIC code.

In Table 5, we provide the regression coefficients of Figure 8 and confirm the results using citation-weighted patent counts as our measure for innovation. The results highlight the strong positive association between innovation and economic growth at the sector level.

Table 5: Innovation and Sectoral Growth

Dependent Variable:	Value-Added Growth (1948-1986)		FTE Employee Growth (1948-1986)	
	(1)	(2)	(3)	(4)
Log Patents (1948-1986)	0.679***		0.609***	
	(0.191)		(0.159)	
Log Citations (1948-1986)		0.677***		0.617***
		(0.179)		(0.149)
1948 Dependent Variable Value (1000s)	-0.152	-0.142	-0.627	-0.595
	(0.102)	(0.095)	(0.457)	(0.420)
Observations	18	18	18	18
Mean of Dep. Var.	6.44	6.44	0.39	0.39
S.D. of Dep. Var.	1.61	1.61	1.45	1.45

Notes: Table reports estimated coefficients from a regression in which the dependent variable is the sector-level annualized growth rate in value added (columns 1 and 2) and full-time-equivalent employees (columns 3 and 4). Patent classes are matched to sectors using data provided by William Kerr [3-digit version comes from Kerr (2008) and 4-digit comes from Acemoglu et al. (2016b)]. A patent class k is matched to an industry s if s is the modal user of patents from k. Industry data provided by the Bureau of Economic Analysis. White heteroskedasticity robust standard errors reported in parentheses. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level.

Instrumental Variables

We now attempt to identify a causal effect in our growth regressions using OSRD contracts for wartime technological development as an instrument for innovation. Table 6 reports coefficients from a regression of post-war state-level growth in GDP per capita for a four decade time horizon (1947-1987) on state innovation levels immediately following World War II (1945-1950).[11] We report OLS estimates from this regression (columns 1, 2, 4, and 5) the corresponding IV estimates using OSRD contracts as an instrument (columns 3 and 6) and the first stage regression of log patents on the number of OSRD contracts (column 7).

The first stage relationship is strongly positive and interesting in its own right. Several studies including Barro (1981) and Field (2008) show that general wartime spending had little impact on economic growth and may have even crowded out private sector investment. Fishback and Cullen (2013) find that "growth in per capita measures of economic activity [to 1958] showed little relationship with per capita war spending" and Jaworski (2015) finds little effect of general

[11]The results are similar if we instead study the effect of patenting during the war, between 1940 and 1945.

wartime spending on subsequent growth rates in the US South. These studies suggest that our use of OSRD contracts as an instrument will not be invalidated by any correlated contemporaneous response of GDP per capita to other forms of government-spending. Furthermore, OSRD contracts were targeted towards technological development, whereas more general government contract spending on combat-related equipment like aeroplanes and tanks or incidentals such as clothing was not.

Our instrumental variables strategy requires that these contracts were correlated with innovation, uncorrelated with omitted determinants, and only influenced state growth rates through their effect on innovation. Note that if the OSRD contracted with only the best firms or academic institutions (which it did not), this would not be a violation of the exclusion restriction, so long as initial location decisions were orthogonal to a state's future growth rate. A brief survey of the institutional setting for the OSRD along with quantitative tests lends support to the credibility of our instrumentation approach.

Table 6: Innovation and Long Run Growth: OLS and IV between 1947-1987

	Annualized Growth Rate			DHS Growth Rate			1^{st} Stage
	OLS	OLS	IV	OLS	OLS	IV	OLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Log Patents (1945-1950)	0.139***	0.113***	0.135***	0.044***	0.036***	0.044***	
	(0.039)	(0.041)	(0.051)	(0.012)	(0.013)	(0.016)	
OSRD Contracts							0.734***
							(0.076)
Initial Log GDP per Capita	-1.678***	-1.777***	-1.836***	-0.512***	-0.543***	-0.563***	-0.114
	(0.229)	(0.232)	(0.247)	(0.069)	(0.070)	(0.076)	(0.612)
Population Density		1.397**	1.235**		0.435**	0.381**	-0.339
		(0.646)	(0.583)		(0.197)	(0.176)	(2.273)
Observations	48	48	48	48	48	48	48
Mean of Dep. Var.	2.501	2.501	2.501	0.909	0.909	0.909	7.257
Std. Dev. of Dep. Var.	0.439	0.439	0.439	0.134	0.134	0.134	1.413
F-Statistic							65.163

Notes: Table reports estimated coefficients from a regression in which the dependent variable is the state-level annualized growth rate in GDP per capita from 1947-1987. White heteroskedasticity robust standard errors reported in parentheses. DHS growth rate refers to the growth rate measure as proposed by Davis, Haltiwanger, and Schuh. The IV estimates are two-stage least squares estimates using the number of OSRD contracts in each state during World War II. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level.

The OSRD was established under President Roosevelt's Executive Order in June 1941 and operated until it was terminated in December 1947. It was headed by Vannevar Bush at the Carnegie Institution of Washington. The OSRD was responsible for major innovations that had an impact in wartime and beyond, including miniature electronics like the proximity fuse, navigation systems, solid fuel rockets, detonators and most famously the basic science used in the Manhattan Project (the Manhattan project was later transferred to the Manhattan District of the Army Engineers). Because of its significant impact, the OSRD spurred federal involvement in the development of US science and technology in the postwar years (Stephan (2014)).

The OSRD did not operate laboratories of its own; rather it contracted out the development of inventions. This reflected a new way of mobilizing public funding for the development of scientific

resources. During World War I scientists had worked at rudimentary laboratories established by the government on an *ad hoc* basis, and there was a long-standing concern among scientists that federal involvement in their activities would threaten creativity and intellectual independence. As Mowery (2010, p.1227) comments, "the contractual arrangements developed by the OSRD during World War II allowed the office to tap the expanded range of private sector and university scientific and engineering capabilities that had developed during the interwar period."

However, the OSRD did not know *ex ante* which firms or academic institutions would be successful because "the OSRD had long insisted that it was not working on materials or methods of wide use in industry" (NAS, 1964, p.28). In fact, due to uncertainty, the OSRD sometimes contracted with multiple entities to solve the same problem. The OSRD spent $450 million in total, about six and a half times the federal budget for science in 1940. Around this time universities had been spending about $50 million on research of which around $6 million was funded by the federal government to support mostly agriculture-related research (Payne, 1992, p.145). The OSRD created a large boost to firm-level R&D. For example, Radio Corporation of America invested heavily at its plants in Indiana and New Jersey (Chandler, 2001 p.27-28).[12]

We collected data on all contracts granted by the OSRD. We observe 1,717 contracts across 39 US States. The coverage of the OSRD contracts is wide. For example, Iowa State College received 10 contracts and the University of New Mexico received 7 contracts. Firms and academic institutions in the state of New York accounted for 30 percent of the total with the next largest concentrations of contracts being in Massachusetts (13 percent) and Pennsylvania (11 percent). The mean number of contracts per firm/academic institution is 4.3 and the median is 1. The most prolific private firm in terms of contracting is the Western Electric Company with 107 contracts. The most prolific university is MIT which was granted 89 contracts.

Using these data, columns 3 and 6 of Table 6 reveal that the OLS coefficients are confirmed by the IV estimates. To evaluate the validity of the exclusion restriction, we provide quantitative tests of the instrument in Table 7. Specifically, we check if contract allocation is correlated with pre-trend growth. A main concern would be that OSRD contracts simply proxy for states that were already growing relatively rapidly. We do not find a statistically significant effect of pre-period growth rates on contract allocation for a range of different time periods.

Finally, it is worth noting that the regression coefficients we have estimated in this section are informative about time period effects and the nature of the channel through which innovation influences economic growth. The effect we estimate at the state level (0.139 in Table 6) for the period 1947-1987 is double the size of the effect we estimate at the state level (0.066 in Table 4) for the period 1900-2000. This difference is consistent with US technological leadership being tightly linked to economic growth in the post World War II years (Nelson and Wright (1992)). In turn, both coefficients in Table 6 and 4 are substantially smaller than the estimated aggregate relationship between log GDP per capita and log patents from Figure 1 (0.64). When a new

[12]Procurement related contracts generally "allowed for work on a fixed price plus a reasonable profit for the contractor" ((Payne, 1992)). In terms of rights to patents, the contractor generally retained these, though the government was permitted to a royalty-free compulsory license from the contractor or to a patent buyout at a reasonable price. If the contractor chose not to patent, the government retained the right to do so, and would in turn grant the contractor a non-exclusive royalty free license to use the invention (Wellerstein, 2008).

product or process is invented, the knowledge embodied in the innovation diffuses across the economy, and is not confined by state borders. Because we do not take into account the positive cross-state spillovers from successful innovations our state-level estimates (and IV estimates) will likely understate the aggregate relationship between innovation and economic growth. Equally, this spillover argument may also contribute to the larger coefficient on innovation at the sector level in Table 5 than is found at the state level. While knowledge may easily diffuse across geographies and generate important spillover effects across states, knowledge at the sector-level may be more specialized, and less likely to influence other sectors directly.

Table 7: Testing for Selection Effects in OSRD Contracts

	$t = 1935 - 1940$		$t = 1930 - 1935$	
Growth Rate	Annual	DHS	Annual	DHS
	(1)	(2)	(3)	(4)
Real GDP Growth (t)	0.085	1.693	0.017	0.723
	(0.172)	(3.388)	(0.100)	(2.471)
Real GDP Growth ($t-1$)	-0.091	-2.119	0.118	2.592
	(0.060)	(1.489)	(0.150)	(3.182)
GDP per Capita ($t-1$)	2.212*	2.259**	2.179*	2.283**
	(1.129)	(1.120)	(1.140)	(1.126)
Population Density ($t-1$)	10.794*	10.783*	11.546**	11.688**
	(5.753)	(5.794)	(5.120)	(5.141)
Observations	48	48	48	48

Notes: Table reports coefficients from a regression in which the dependent variable is the dependent variable is the the number of OSRD contracts in each state during World War II and the independent variables are pre-trend growth rates, population density, and beginning of period GDP per capita. We consider growth rates from 1935-1940 (t) and 1930-1935 ($t-1$) in columns 1 and 2, while in columns 3 and 4 we consider growth rates from 1930-1935 (t) and 1925-1930 ($t-1$). White heteroskedasticity robust standard errors reported in parentheses. DHS growth rate refers to the growth rate measure as proposed by Davis, Haltiwanger, and Schuh. *,**,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: Bureau of Economic Analysis, USPTO patent records.

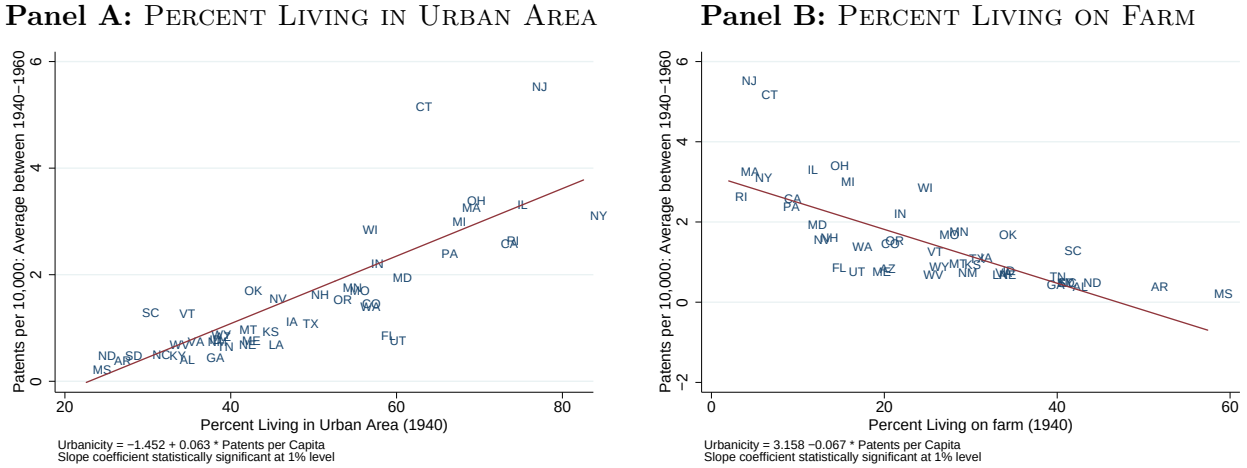
Fact 2. ***Densely-populated states were more inventive.***

We now seek to understand the environmental factors that determine whether a state is particularly inventive. We first consider the relationship between urbanization and state-level innovation. Table 1 shows population density was much higher in the most inventive states. This finding is in line with two parallel literatures: First, a growing theoretical literature arguing that human interaction is key for human capital accumulation and economic growth (e.g., Lucas (2009), Alvarez et al. (2013), Lucas and Moll (2014), Perla and Tonetti (2014)). Second, the agglomeration literature has long argued that physical proximity promotes creativity, the exchange of ideas and spillovers of knowledge capital among inventors (see Carlino and Kerr (2015) for a survey). We also find that the top inventive states were associated with higher levels of education and higher skilled occupations. Although inventor expropriation risk can be higher in cities where competing inventors can more easily learn about ideas, these results

are consistent with denser places being more likely to create positive externalities that lead to sustained economic growth.

Figure 9 confirms the relationship between urbanization and innovation. Panel A plots the percent of a state's population that lives in an urban area in the 1940 census against the average number of patents per capita granted in that state between 1940 and 2004. We see a robust positive relationship between a state's degree of urbanization and future inventive activity. This relationship is significantly different from zero at the 1% level. A one standard deviation increase in the percent of a population living in an urban area is associated with an increase in innovation that is 41.5% of its mean. Because the Census adopts a low threshold for urbanization as places that encompass at least 2,500 people, we repeated the analysis at different thresholds with substantively the same result.[13] Panel B repeats the same analysis using the fraction of population living on a farm. It shows that average patents per 10,000 people between 1940 and 1960 was decreasing in the percent of the population living on a farm in 1940.

Figure 9: Population Density and Innovation



Notes: Figure shows the relationship between the percent of individuals living in an urban area (Panel A) or on a farm (Panel B) in the 1940 census and the average patents per capita between 1940 and 1960. Delaware excluded as an outlier for visibility. Delaware excluded as an outlier for visibility. Source: 1940 Historical Census Data, USPTO patent records.

Robustness. We now show that the positive link between population density and innovation shown in Figure 9 holds at a finer level of aggregation. Table 8 report OLS estimates confirming a statistically significant relationship between innovation and population density at the county-level. To aid interpretation, we standardize the independent variables to have a zero mean and unit standard deviation. Table 8 also shows that this estimated relationship is not entirely driven by the sectoral composition of urban counties. Columns 2 and 4 control for the fraction of the population working in agriculture or manufacturing and condition on state fixed effects. More densely populated counties were also more inventive, even conditional on the county's industry mix and characteristics common to states. Conditional on our set of controls, a one standard deviation in the percent of a county's residents living in an urban area is associated with an

[13] In Figure 9 Panel A, the slope coefficient is 0.063. When we use a threshold of 5,000 people the coefficient is 0.061 and at 10,000 people the coefficient is 0.056. All are significantly significant at the 1% level.

increase of 0.414 in patents per 10,000 residents.

Table 8: Population Density and Innovation: County-Level Results

	(1)	(2)	(3)	(4)
% Urban	0.817***	0.414**		
	(0.139)	(0.176)		
% Living on Farm			-0.858***	-0.484**
			(0.096)	(0.242)
% Agricultural Occupation		-0.426***		-0.391**
		(0.112)		(0.163)
% Manufacturing Occupation		-0.021		-0.142
		(0.093)		(0.107)
State Fixed Effects	N	Y	N	Y
Observations	3087	3062	3087	3062

Notes: Table reports coefficients from OLS regression of average patents per 10,000 between 1940 and 1960 on various measures of a county's urbanicity. All independent variables measured as of 1940 and standardized to have 0 mean and unit standard deviation. White heteroskedasticity robust standard errors reported in parentheses. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: 1940 Historical Census Data, USPTO patent records.

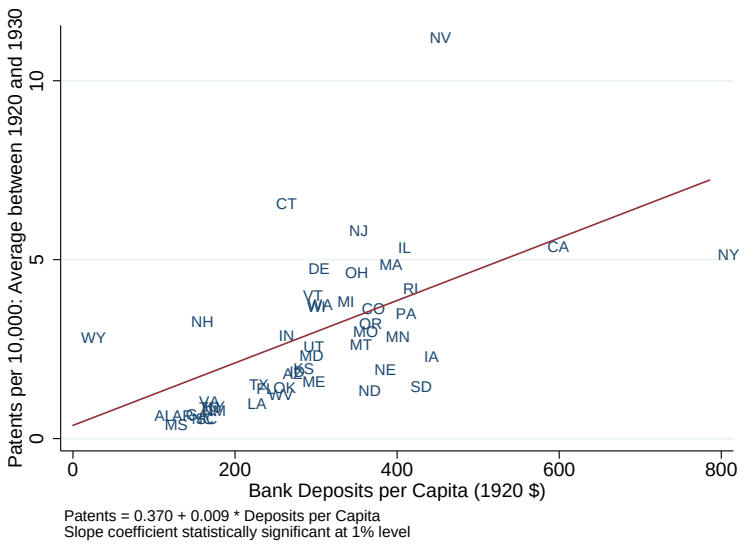
Fact 3. ***Financially-developed states were more inventive.***

There is a vast literature relating access to capital and innovation. Cross-country growth regressions have shown that higher levels of financial development are associated with faster rates of economic growth (e.g., King and Levine (1993), Rajan and Zingales (1998)). Within the US a range of evidence indicates capital availability mattered for innovation. For example, Lamoreaux et al. (2004) find that venture-style provision of capital dramatically reduced financing constraints for inventors in Cleveland, an important Second Industrial Revolution city. According to Kortum and Lerner (2000) venture capital had a strong causal impact on patenting rates in the US in the late twentieth century.

Modern financial markets are largely national in scope but were more local in the early twentieth century. Because of legal constraints on the functioning of the banking sector, banks were limited in their ability to operate across state lines and were often made up of unit banks that serviced local communities. Although California had extensive branching outside the bank's home office city this was not the norm (Carlson and Mitchener (2009)). From individual inventors up to large, publicly-traded corporations, an important component of finance could be sourced locally (Nanda and Nicholas (2009)). As such, one might expect a positive relationship between the health of a state's financial market and its propensity to innovate.

Measuring the level of financial development is fraught with difficulties. Private transactions between investors and inventors are not observable systematically and most later stage R&D is financed by firms internally. However, we can gain useful insights using FDIC data, which provides broad indicators of financial market development. We measure the health of a state's financial sector by bank deposits per capita in 1920 in order to proxy for the size of available funds in the state. We choose 1920 in order to avoid the pre-Depression inflation in stock prices and the market's subsequent collapse. Figure 10 plots the relationship, showing that a healthier financial sector is correlated with greater innovation levels, in line with expectations.

Figure 10: 1920 Deposits per capita and Innovation



Notes: Figure plots the relationship between 1920 financial health, and the average patents per capita between 1920 and 1940. Our measure of financial health is the amount of deposits per capita in 1920 dollars. Banking data originate from the FDIC dataset, downloaded from the University of Michigan's ICPSR repository (number 0007). We use 1920 data to remove the influence of the Great Depression from our data. Delaware excluded as an outlier for visibility. Source: FDIC, USPTO patent records.

Robustness. We now repeat and extend the analysis at the county-level. For this purpose we exploit the fact that dependence on external financial might vary across different types of inventors: those operating outside and inside the boundaries of corporations. Specifically, we would expect independent inventors to be more responsive to changes in the external financing environment than firms, where internal funds would be a more important driver of R&D investment. Empirical studies have shown that larger and more established firms exhibit a strong preference for financing innovation internally, to the point where they manage their cash flows in order to be able to do so (e.g., Hall and Lerner (2009)). With this fact in mind, Table 9 correlates individual- and firm-level patenting with the measure of financial development. Columns 1 and 4 reproduce at the county-level the basic correlation shown at the state-level in Figure 10. Columns 2 and 5 introduce sectoral controls as well as state fixed effects.

The results show a clear relationship between innovation and the level of financial development, and in the direction that we would expect based on the groups of inventors on which these specifications are run. The coefficients are more precisely estimated for non-corporate versus corporate patenting and their economic magnitude is also larger. Moreover, the results are robust to the possibility that even though the availability of finance may have been local at this time, inventors may have crossed state borders to access capital. When we run regressions at the state level in columns 3 and 6 we find a strong and statistically significant effect of finance on innovation for non-corporate patentees for whom we would expect the effect to be largest.

Fact 4. ***Geographically-connected states were more inventive.***

Another important dimension for innovation is access to other geographical regions. This could increase both the market size for innovation and the potential flow of knowledge spillovers. Both

Table 9: Financial Development and Innovation: County-Level Results

	Non-Corporate Patents			Corporate Patents		
	County	County	State	County	County	State
	(1)	(2)	(3)	(4)	(5)	(6)
Deposits per Capita	0.300***	0.185*	0.400***	0.164***	0.034	0.125
	(0.114)	(0.103)	(0.139)	(0.062)	(0.035)	(0.146)
% Agricultural Occupation		-0.383***	-0.184		-0.400***	-0.614
		(0.085)	(0.308)		(0.057)	(0.681)
% Manufacturing Occupation		-0.027	0.142		0.116**	0.244
		(0.067)	(0.314)		(0.059)	(0.574)
State Fixed Effects	N	Y	Y	N	Y	Y
Observations	3013	2279	48	3013	2279	48

Notes: Table reports coefficients from OLS regression of average patents per capita between 1920 and 1940 on the local amount of deposits per capita in 1920. Columns 1 through 3 consider only patents not assigned to corporations while columns 4 through 6 solely account for patents assigned to corporations. Columns 1, 2, 4, and 5 report estimates from county-level regressions, while columns 3 and 6 report analogous estimates from state-level regressions. All independent variables measured as of 1920 and standardized to have 0 mean and unit standard deviation. White heteroskedasticity robust standard errors reported in parentheses. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: FDIC, USPTO patent records.

mechanisms receive support in the literature. Acemoglu and Linn (2004) and Aghion et al. (2016) have shown that market size is a crucial factor for innovation in the pharmaceutical and auto industries, respectively. Sokoloff (1988) found that inventive activity in the early nineteenth century accelerated in locations that were proximate to navigable waterways, while Perlman (2016) finds strong effects on invention and agglomeration from the nineteenth century development of railroads. Donaldson and Hornbeck (2016) measure the increased level of market access caused by an expansion of the US railroad network between 1870 to 1890, finding the aggregate impact on growth to be large.

We use data provided by Donaldson and Hornbeck (2016) to investigate the relationship between a state's geographic connectivity and its level of patenting. By observing the development of roads, railroad, and waterways over time, and assigning different per-mile costs to each mode of transport, they construct the cost of shipping one ton of goods between every county pair in the United States every 10 years from 1830 through 1920. A high cost of shipping goods out of a state indicates that the state is geographically isolated.

Geographic connectivity may increase innovation both by allowing inventors to sell their inventions to a larger market, and by encouraging the free exchange of ideas across geographies. The average cost to transport goods out of a state is an imperfect proxy of these two measures. We therefore consider alternative measures of geographic connectivity derived from these transport cost data. Additionally, we ask whether the observed positive relationship between geographic connectivity and innovation persists at the more granular county level.

We construct two measures of a county's connectivity. Define $\kappa_{c,c'}$ to be the cost to ship one ton of goods from county c to county c' in 1880, and let $\bar{\kappa}_c$ be the weighted average outgoing transport cost for county c:

$$\bar{\kappa}_c = \frac{1}{N} \sum_{c'} \omega_{c,c'} \kappa_{c,c'}$$

for N the number of distinct c'. To account for the fact that connection to markets with more economic activity may increase the reward to innovating more than connection to relatively poor areas, we weight these averages by the mean wage income in county c' in our principal measures.[14] Let μ and σ be the mean and standard deviation, respectively, of $\bar{\kappa}_c$ in the sample. We call our first measure of county connectivity its "cost advantage" and compute it as

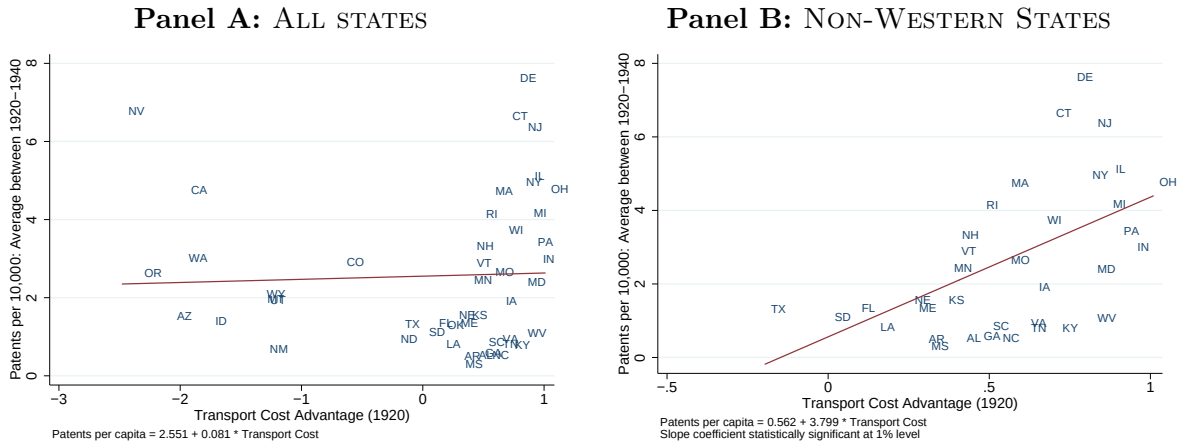
$$\text{Cost Advantage}_c = \frac{\mu - \bar{\kappa}_c}{\sigma}$$

A one unit increase in a county c's cost advantage has the interpretation that the average outbound transport cost for county c is one standard deviation below the mean outgoing transport cost. Our second measure of connectivity measures the number of people who live in counties which lay within the median county-to-county transport costs radius of c. Specifically, letting π^{50} be the median value of $\kappa_{c,c'}$, define $M(c) = \{c' : \kappa_{c,c'} \leq \pi^{50}\}$ to be the set of counties within median county-to-county transport costs of c. For P_c to the population living in county c in 1880, we define a county's "market size" to be

$$\text{Market Size}_c = \sum_{c' \in M(c)} P_c$$

The agglomeration literature would predict a positive relationship between transport cost advantage and innovation. Figure 11 tests this hypothesis.

Figure 11: Transport Cost State Scatters



Notes: Figure plots the relationship between outgoing shipment costs and innovation. The horizontal axis plots the average cost to transport one ton of goods from a county in state s to counties in state d different from s. For details on the construction of the county-to-county transportation cost measures, see Donaldson and Hornbeck (2016). Panel A plots the relationship for all states, while Panel B plots the relationship only for non-Western states. Non-Western states defined as those states with average outbound transport cost under \$18. Source: USPTO patent records, Donaldson and Hornbeck (2016).

The horizontal axis plots the average cost advantage and the vertical axis plots the average number of patents per 10,000 state residents over the period spanning 1920 to 1940. Panel A shows only a weak positive relationship between transport cost advantage and innovation across

[14] Using unweighted measures, or weighting by destination county population or patent counts does not qualitatively change the conclusions of this section.

all states. However, closer inspection reveals that the weak relationship between transport cost advantage and innovation is driven by a number of states with high transport costs. These states are all in the West of the United States. Although regional economies within the country integrated to form a national economy from the nineteenth to the early twentieth centuries (Kim (1995)), in such places the cost of transporting people, goods and information still represented a significant constraint.

Hence, the results are notably different when we consider the older, more integrated East Coast and Midwestern states where economic activity was concentrated. At the turn of the twentieth century the manufacturing belt contained almost three quarters of total manufacturing employment (Krugman (1991)). Panel B focuses only on the states with average shipping costs below $18, which includes most of this area. The nonlinearity is important. Among the states where the average outgoing transportation cost is below $18, the relationship between innovation and transportation cost advantage is strongly positive, suggesting that access to outside markets was an important component of innovation among traditionally developed states.

Robustness. These state-level patterns are confirmed through county-level analysis in Table 10. We report the coefficients from a county-level regression of patents per 10,000 residents on our two measures of county connectivity.

Table 10: Transport Costs, Market Size and Innovation: County-Level Results

	(1)	(2)	(3)	(4)
Transport Cost Advantage	0.096	0.583***		
	(0.080)	(0.199)		
Market size			0.040	0.713**
			(0.081)	(0.281)
% Agricultural Occupation		-0.569***		-0.573***
		(0.158)		(0.158)
% Manufacturing Occupation		-0.129		-0.137
		(0.169)		(0.168)
State Fixed Effects	N	Y	N	Y
Observations	2757	2073	2757	2073

Notes: Table presents estimates from OLS regressions of average patents per 10,000 between 1920 and 1940 on the degree of state geographic connectivity in 1920, as measured by a state's cost advantage or the population living within median transport cost of the county. A one unit increase in a state's cost advantage has the interpretation that the average outbound transport cost is one standard deviation below the mean outgoing transport cost in that state. All independent variables measured as of 1920 and standardized to have zero mean and unit standard deviation. Standard errors clustered at state level reported in parentheses. Observation counts drop in columns 2 and 4 due to missing occupation data in the county. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: USPTO patent records, Donaldson and Hornbeck (2016).

Columns 1 and 3 show little evidence that more connected counties were more innovative, unconditional on any covariates. However, this is subject to the same concerns as the state-level scatter plots presented in Figure 11: grouping states further westward with established eastern states masks the economic mechanisms at play during this time period. Once we control for state fixed effects and the county-level sectoral mix in columns 2 and 4, we find a strong positive relationship between a county's geographic connectivity and its innovation intensity. A

one standard deviation increase in a county's cost advantage (market size) is associated with an increase in patents per 10,000 of 0.583 (0.713) between 1920 and 1940, conditional on state fixed effects.

Fact 5. ***States associated with slavery were less inventive but religiosity is not robustly correlated with inventiveness.***

One potential explanation for these underlying state-level and county-level differences is that innovative places are relatively more open to unconventional and disruptive technological ideas. The recent literature has shown that the opening of the labor force to women and minorities has important consequences for growth. Viewing the changes in labor market outcomes for women and minorities through the lens of a Roy (1951) model, Hsieh et al. (2013) show that the convergence in the occupational distribution between white men, women, and black Americans can account for 15 to 20% of growth in aggregate output per worker between 1960 and 2010. The effect of this convergence on innovation, however, remains an open question.

Cultural differences may be an important determinant of a region's innovative activity and growth ((Gorodnichenko and Roland, 2011)). Yet, culture can have a positive or a negative effect. For example, productive aspects of social capital have been linked to financial development (Guiso et al. (2004)), whereas negative aspects of social capital have been linked to the rise of fascism (Satyanath et al. (2016)). Insofar as the makeup of society influences the incentives for creative invention (e.g., Florida (2002), Acemoglu et al. (2014)), we would expect to find correlations between a state's level of innovation and its demographic characteristics.[15]

An important aspect of openness of a society to innovation and economic growth could be seen from its approach towards slavery. Wright (1986) argues that the southern economy of the US was constrained by a lack of technological innovation in agriculture and manufacturing because of slavery. Slavery can undermine trust, having a persistent effect on beliefs and behavior (Nunn and Wantchekon, 2011). A lack of cultural freedom to deviate from established norms can strongly inhibit innovation. Cook (2011) finds that while African American inventors often made important technological discoveries during the nineteenth and early twentieth centuries, they were much less likely to do so in closed environments such as places that implemented segregation laws. Logan and Parman (2015) link racial sorting to a doubling in a newly constructed index of residential segregation in the US between 1880 and 1940.

Another important dimension through which societal attitudes towards innovation and economic growth can be shaped is through religion. Following Weber, "ascetic Protestantism"—the idea that idleness is sinful and rigorous self-discipline is necessary to serve God—is often hypothesized to have a positive effect on economic growth, though the relationship has proved difficult to verify empirically (e.g., Cantoni (2015)). A more recent set of studies has directly addressed the question of whether religiosity promotes or inhibits innovation. Benabou et al. (2013) and Benabou et al. (2015) conclude that the relationship is robust and strongly negative. While their

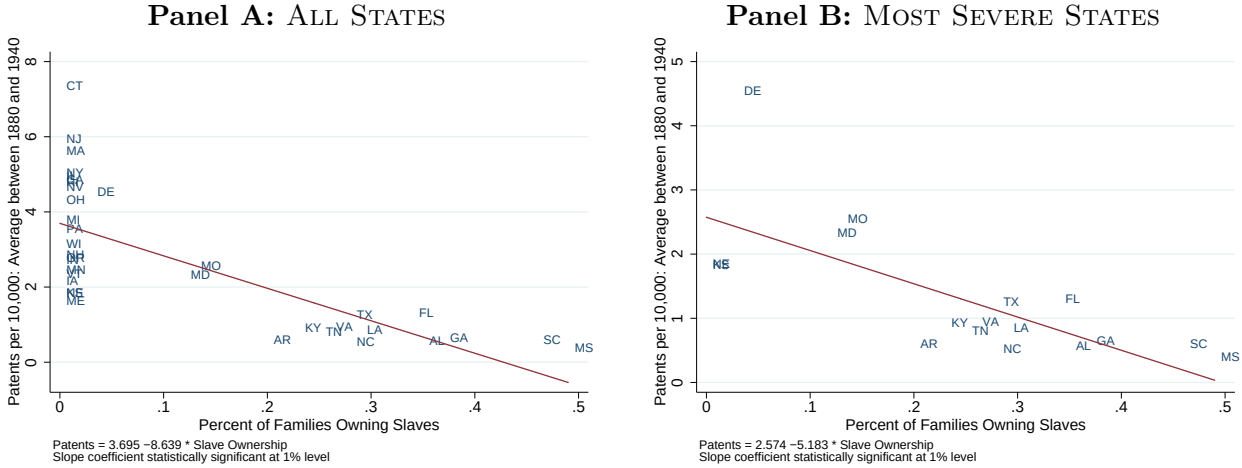
[15] Another project attempts to match patent data to the Censuses. Sarada et al. (2016) match patent data from the *Annual Report of the Commissioner of Patents* at decennial intervals from 1870 to 1940 to the US Federal Population Censuses in those years. Their study is complementary to ours as it focuses on the relationship between the demographics of patentees and economic and demographic characteristics at the county-level.

evidence is compelling for the United States in the modern era, we lack corresponding empirical evidence on the relationship between religion and innovation for the US in our time period.

We incorporate slavery and religion into our analysis. We use data from the 1860 Census to determine the percentage of slave-owning families.[16] If a state's slave ownership rate captures a society's openness to change, one would expect that innovation would correlate negatively with the slave ownership rate. Given the multidimensional nature of religion we use a variety of measures based on data in the Census of Religious Bodies. Specifically, we construct two measures: the share of the population belonging to any religion, and a Herfindahl Index of religious membership. The intuition behind the last measure is that it captures religious diversity and therefore the extent to which different beliefs are tolerated within broad communities. This variable proxies for the degree of openness to disruptive ideas, as a conduit to technological innovation (Acemoglu et al. (2014)).

Our results with respect to slavery have a natural interpretation. Figure 12 shows the relationship is unambiguously negative and statistically significant both when we correlate slave ownership with patenting in all US states and when we do this for only states in the southern part of the country, where slavery was most prevalent. The basic correlation can be taken alongside the descriptive statistics in Table 1 showing that slave states were disproportionately among the least inventive in the United States. Our findings are consistent with the view that slavery severely stunted economic growth (e.g., Wright (2006)).

Figure 12: Slave Ownership and Innovation



Notes: Figure plots the relationship between the percent of families which owned slaves in the 1860 census, and the average patents per capita between 1880 and 1940. Panel A plots the relationship for all states, while Panel B only includes states which have positive slave ownership rates. Many northern states did not have positive slave ownership in the 1860s. Source: 1860 Historical Census Data, USPTO patent records.

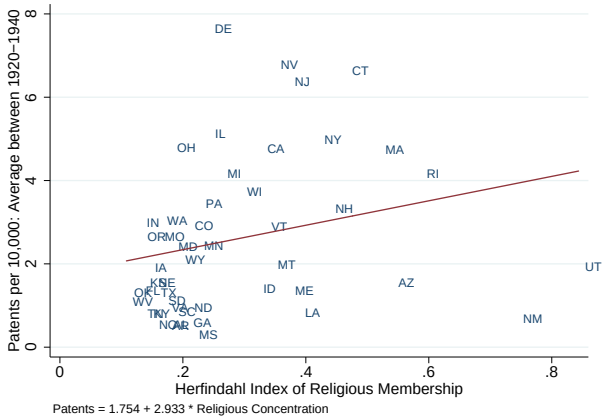
As shown in Figure 13, the relationship between religion and inventiveness is more complex. We find no strong correlation at the state level between the Herfindahl Index of religious openness measure (Panel A), or the share of the population of a state belonging to a religion (Panel B) and inventiveness. Although these findings contrast with Benabou et al. (2013) and Benabou et al. (2015) this difference could be explained by the fact that the effect of religion on innovation may

[16] County-level slavery statistics were calculated by Acharya et al. (2016).

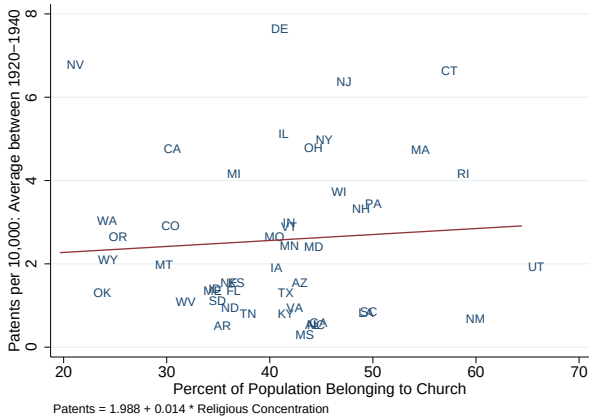
have varied over time. Historically—much more so than today—religion was associated with literacy and access to education (Goldin and Katz (2001)). The positive (though statistically insignificant) effect of religion that we find in Figure 13 may simply be capturing the underlying impact of human capital. The upshot of these results is that the effect of religion on innovation at the state-level is ambiguous.

Figure 13: Religion and Innovation

Panel A: Herfindahl Index of Religious Membership



Panel B: Share Belonging to any Religion



Notes: Figure plots the relationship between various measure of religion from the Census of Religious Bodies, and the average patents per capita between 1920 and 1940. Panel A plots the Herfindahl Index of religious membership, while Panel B plots the share of individuals in a state belonging to any religion. Source: 1916 Census of Religious Bodies, USPTO patent records.

Robustness. These contrasting findings for the nature of the relationship between social structure and innovation from the perspective of slavery and religion are confirmed at the county-level. Table 11 column 1 shows that the negative relationship illustrated in Figure 12 holds in the cross section of counties when we use the share of the population as slaves as an independent variable.

Table 11: Slave Ownership and Innovation: County-Level Results

	(1)	(2)
% of Population Slaves	-0.391***	0.274
	(0.117)	(0.232)
% Agricultural Occupation		-0.423***
		(0.095)
% Manufacturing Occupation		0.067
		(0.120)
State Fixed Effects	N	Y
Observations	2200	2186

Notes: Table reports coefficients from OLS regression of average patents per capita between 1920 and 1940 on county-level slavery. All independent variables measured as of 1920 and standardized to have 0 mean and unit standard deviation. Heteroskedasticity robust standard errors reported in parentheses. Source: 1860 Historical Census Data, USPTO patent records.

Although with additional controls for occupational mix the coefficient becomes statistically insignificant in column 2 and even reverses sign, this only tells us that differences in the share

of slave ownership across counties *but within states* had no effect on patenting. Variation in the cross-section of counties (and states) should be more important in this case given that unobservable traits like discrimination or institutional weaknesses that would be strongly correlated with slavery will tend to be be fixed within states.[17]

Table 12 reports corresponding estimates at the county-level for the relationship between religion and patenting. In columns 1 and 2 the coefficients on the Herfindahl Index of religious membership are economically small and statistically indistinguishable from zero. However, we do find a negative estimate in columns 3 and 4 on the broadest measure of religiosity, which is statistically significant in column 4. This finding is consistent with the negative effect found by Benabou et al. (2013) and Benabou et al. (2015) for the modern era. Overall, however, when taking the evidence in 13 and Table 12 together, it appears that religion was not a strong predictor of technological development during the US golden age.

Table 12: Religion and Innovation: County-Level Results

	(1)	(2)	(3)	(4)
Herfindahl Index of Church Membership	-0.045	-0.012		
	(0.040)	(0.060)		
% of Population Belonging to a Religion			-0.053	-0.135**
			(0.053)	(0.057)
% Agricultural Occupation		-0.396***		-0.413***
		(0.086)		(0.083)
% Manufacturing Occupation		0.028		-0.005
		(0.085)		(0.087)
State Fixed Effects	N	Y	N	Y
Observations	2732	2708	2715	2692

Notes: Table reports coefficients from OLS regression of average patents per capita between 1920 and 1940 on county-level measures of religiosity. All independent variables measured as of 1920 and standardized to have 0 mean and unit standard deviation. Heteroskedasticity robust standard errors reported in parentheses. Source: 1916 Census of Religious Bodies, USPTO patent records.

3.2 Personal Facts

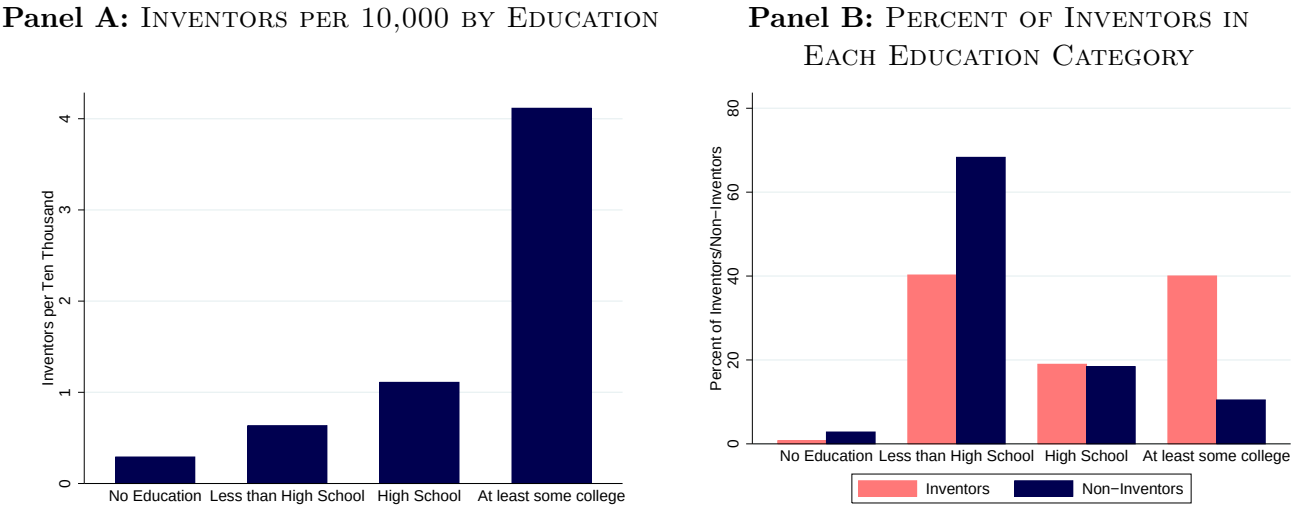
Moving on from environmental factors, we now make use of our microdata to present facts about the inventors of the golden age. While the macroeconomic facts we presented above are informed by linking our patent data to regional aggregates, this section relies largely on our dataset of patents matched to the Censuses. We examine the personal background of inventors, paying special attention to their educational attainment and age, entry and exit over the life cycle of inventiveness, lifetime migration decisions and marriage patterns.

[17] In Table 11 when we run the regression in column 2 without state fixed effects and just with controls the coefficient is still negative (-0.169, s.e. 0.232). Alternatively, if we include just state fixed effects but no controls, the sign on the coefficient reverses (0.179, s.e. 0.225).

Fact 6. ***Inventors were more educated on average and were most productive between the age of 36 and 55.***

A large literature suggests that education is a key factor for economic growth (see, for instance, Lucas (1998), Benhabib and Spiegel (1994), Bils and Klenow (2000), Goldin and Katz (2009), Barro (2001), Vandenbussche et al. (2006), Stokey (1991)). One of the main channels through which education affects economic growth may be its impact on innovation. Figure 14 shows the number of inventors per 10,000 people within each education group. While education seems to be an important determinant of becoming an inventor, the effect is particularly strong at the college degree level. Although the 1940 Census tended to overstate education levels (Goldin (1998)) the differences we see between categories are large. For example, an individual with at least a college degree is four times more likely to become an inventor than an individual with just a high-school diploma. Indeed, 40% of inventors had a college degree in 1940, compared with just 10% of the non-inventor population.

Figure 14: Education and Probability of Becoming an Inventor



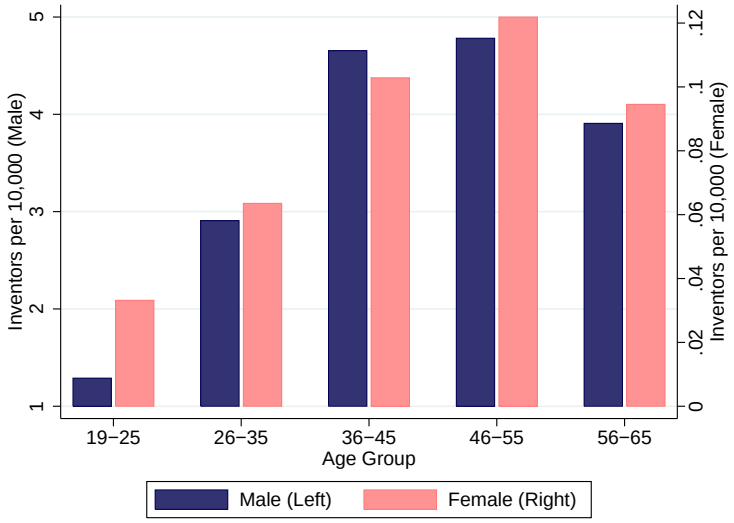
Notes: Figure plots the education of inventors and non-inventors in the 1940 census, the only census in our sample to provide sufficiently granular education information. Panel A plots the inventors per 10,000 people by education category. Panel B plots the percent of inventors and non-inventors that fall into each educational category. Source: 1940 Historical Census Data, USPTO patent records.

The opportunity cost of education is time spent in an active career. In theoretical models of education, individuals face a tradeoff between the benefits of higher education, which accumulate over the life cycle, and the costs which are incurred early on (Becker (1967), Card (2001)). By extension economic growth can be affected by the tradeoff inventors face between acquiring human capital to innovate and the potential delays this creates in the production of new technological discoveries that, in turn, benefit society. Jones (2010) argues that if true breakthroughs are developed by younger cohorts of individuals, the growth-slowing delay effect can be pronounced, especially if more human capital is required for the production of creative ideas as the demands of developing novel innovations increases over time. He finds that the age of great invention shifted upwards by about half a decade over the twentieth century.

We find that inventors were most productive between ages 35-55 as illustrated in Figure 15.

This is true for males and females, although we observe few female inventors. Interestingly, as shown by Sarada et al. (2016) the average age of invention in 1900 was approximately 40 years old, about what it is converging to today. Our data indicate that inventors had a reasonably long productive career-length. Long career-length is consistent with Khan and Sokoloff (2004)'s data on superstar inventors. They found that while 37 percent born prior to 1820 had careers over 30 years, 57 percent did in their post-1820 birth cohorts. A broad inventor life cycle, like we observe, tends to maximize creative output (Galenson, 2016).

Figure 15: Probability of Innovation by Age



Notes: Figure shows the average life cycle of inventiveness over the years 1880 to 1940. It plots the number of inventors per 10,000 individuals by gender. The dark blue bars plot the number of male inventors per 10,000 males against the left axis, while the bright red bars plot the number of female inventors per 10,000 females against the right axis. Source: 1880-1940 Historical Census Data, USPTO patent records.

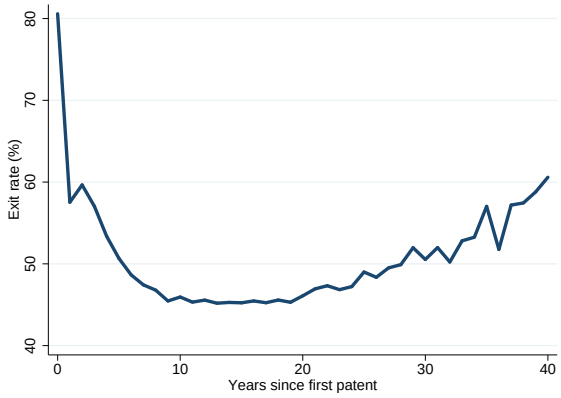
Fact 7. ***Inventors were positively selected through exit early in their careers, but were less productive and more likely to exit late in their careers.***

We extend the analysis to study career dynamics. If inventors approximate the life cycle of firms, some should enter, develop and succeed whereas others should fail and exit—entrepreneurial churn is an essential feature of a well-functioning innovation sector (Haltiwanger (2012)). There is a large growth literature examining firm selection and its mechanism. Acemoglu et al. (2015) examine potential misallocation of R&D inputs using a model in which old firms may be positively selected through endogenous exit of low-quality firms (as in Jovanovic (1982) and Hopenhayn (1992)), or negatively selected through obsolescence.

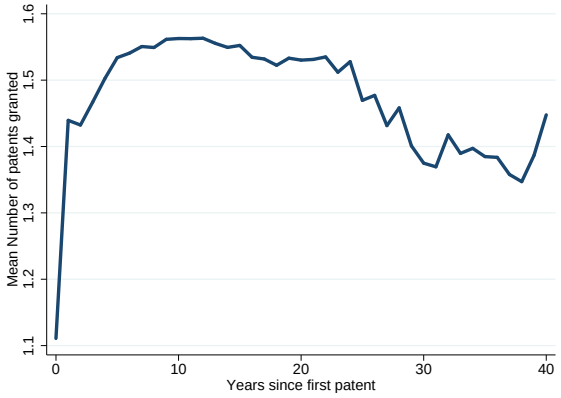
Figure 16 plots the career cycle of inventors using the universe of inventor data, as opposed to just the inventor data matched to the Census. Panel A plots the exit rate for inventors over their life cycle, where an inventor is said to have exited in period t if they file no successful patent applications in every period $t' > t$. Panel B plots the average number of patents conditional on survival for inventors over their tenure in the data. In both panels, the horizontal axis plots the number of years since the inventor filed his first successful patent application.

Figure 16: Inventors' Career Dynamics

Panel A: Exit Rates



Panel B: Productivity Cond. on Survival



Notes: In each panel, the horizontal axis plots the number of years since the inventor's first patent application. Panel A plots the exit rate for inventors over their life cycle, where an inventor is said to have exited in period t if they file no patent applications in every period $t' > t$. Panel B plots the average number of patents conditional on survival for inventors over their tenure in the data. Source: USPTO disambiguated inventor data 1920-2006, constructed by the authors using the algorithm of Li et al. (2014). Source: USPTO patent records.

The figure reveals both similarities and differences with the life cycle dynamics of firms. We find evidence for both positive selection through exit, and eventual obsolescence of inventors. Panel A of Figure 16 shows that inventor exit rates exhibit a U-shape, while Panel B shows that the number of patents conditional on survival has an inverted-U shape over the life cycle. Positive selection occurs early in an inventor's career, where low productivity inventors stop applying for patents. This yields a decreasing exit rate and increasing average productivity over the average inventor's life cycle. In later years of life, however, skill obsolescence and old age set in, reducing inventor productivity, and increasing exit rates. In the limit, biological constraints ensure that the inventor exit rate converges to one.

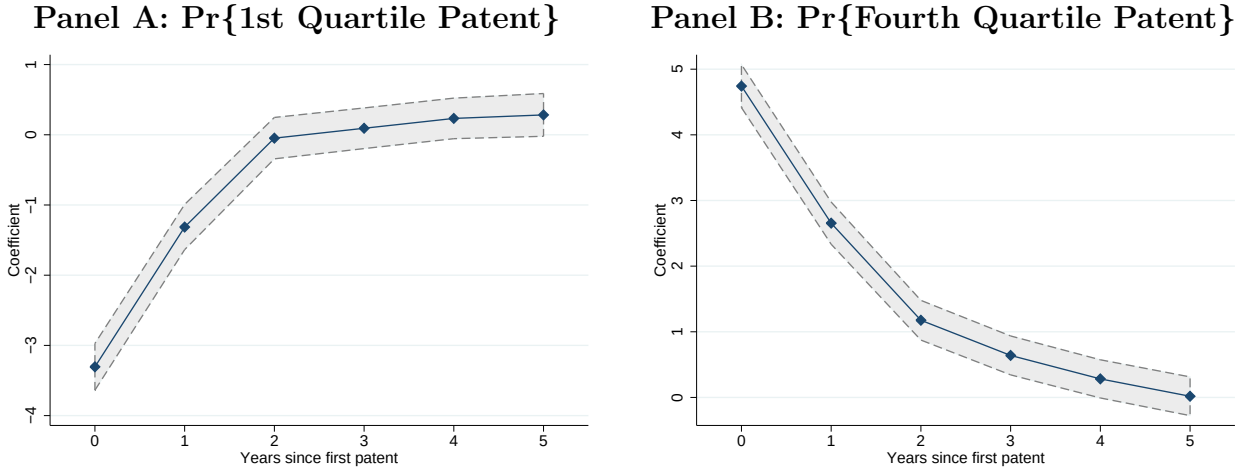
Fact 8. ***The patents of new inventors received more citations on average, and were more likely to be in the top decile of the citation distribution.***

Young firms are an important source of employment creation (Haltiwanger, 2012), and the entry of young firms may yield particularly radical innovations. Akcigit and Kerr (2016) estimate a model with heterogeneous innovations and find that innovative young firms do indeed invest more heavily, relative to their size, in products outside their normal span of expertise. Bernstein (2015) shows that publicly listed firms have less novel innovations and instead advance internal projects. In a recent survey, Lerner (2012) examines the advantages and liabilities of large companies for pursuing new innovation areas compared to start-ups.

There are reasons to suspect that the economics governing the differences between young and old firms might not carry over to the dynamics of a human inventor's career. For instance, the intuition laid out in the Lucas (1978) span-of-control model suggests that large firms might curtail innovation due to limits on its managers' time. On the other hand, there might exist innovation benefits for old, large firms if idea circulation is key to the innovative process (Hellman and Perotti, 2011). These mechanisms rely on the organizational structure of a firm, which has no clear counterpoint in an individual inventor's life cycle.

We therefore investigate the quality of the patents granted to inventors over their career path. We proxy a patent's quality and influence by considering its citation count, adjusted following the methodology of (Hall et al. (2001)). Figure 17 plots various moments of the patent quality distribution measured each year of an inventor's career, conditional on survival. Panel A plots the probability that a patent granted t years after the inventor's first successful patent application lies in the bottom quartile and Panel B repeats the same exercise with the top quartile of citations received. They show that patents granted by new inventors are more likely to be highly cited than patents granted by inventors with a long record of patenting, mirroring the dynamics of innovative firms found in the previous literature. Indeed, patents in the first year of an inventor's inventive tenure are 4.74 percentage points more likely to lie in the top quartile of patent citations, and 3.3 percentage points less likely to be in the bottom quartile than are patents granted 6 or more years after the inventor's first patent, conditional on individual and technology-year fixed effects. These plots are especially striking since they are conditional on survival, given that Figure 16 shows positive selection, measured by raw patent counts, among inventors who continue to innovate over a long career.

Figure 17: Patent Quality over an Inventor's Life Cycle



Notes: Figure plots regression coefficients from an OLS regression of the panel title on indicators for whether a patent was granted t years after the inventor's first appearance in the patent data. All regressions include individual and technology-year fixed effects. Grey bands indicate 95% confidence interval around point estimates, using standard errors which are clustered at the technology class-year level. Source: USPTO patent records.

Table 13 repeats the same analysis in a regression framework. Column 1 regresses the log of citations on an an indicator that is equal to 1 if the patent is granted within the first two years of inventor career and 0 otherwise. It shows clearly that patents obtained early in the career are of higher quality on average. Columns 2 to 5 replace the dependent variable with an indicator showing if the patent belongs to the relevant citation quartile. Again, on average, we see that inventors produce more influential work early in their career.

Fact 9. ***Inventors delayed marriage and had fewer children.***

To the extent that this life cycle dynamic created tradeoffs with respect to time allocation, one area in which we would expect to observe this is through marriage. Theory models specify

Table 13: Panel Relationship between Entry and Patent Quality

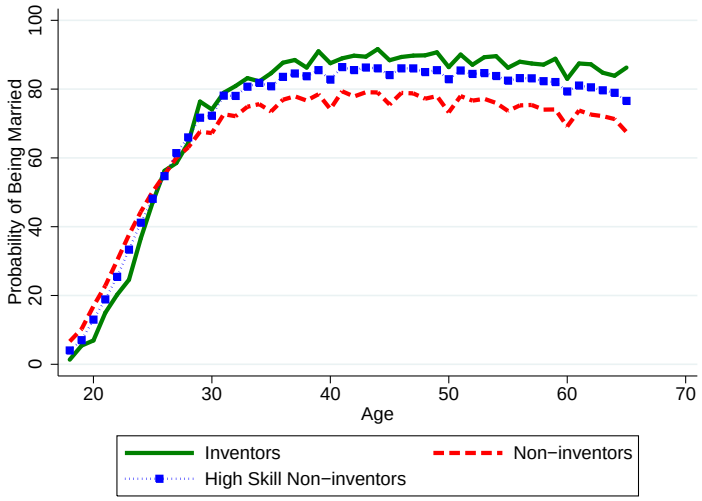
	Log Citations	Patent in quartile (coefficients sum to 0):			
		First	Second	Third	Fourth
	(1)	(2)	(3)	(4)	(5)
Patent granted in first	0.077***	-1.791***	-1.087***	0.350***	2.528***
two years of inventor career	(0.002)	(0.093)	(0.092)	(0.096)	(0.091)
Inventor Fixed Effects	Y	Y	Y	Y	Y
Class × Year Effects	Y	Y	Y	Y	Y
Observations	4290376	4765684	4765684	4765684	4765684

Notes: Table reports regression coefficients from an OLS regression of log citations and whether a patent was in a particular quartile of the citation distribution on an indicator for whether a patent was granted in the first two years of a career. All regressions include individual and technology-year fixed effects. Standard errors are clustered at the technology class-year level. Source: USPTO patent records.

that commitment to a spouse soaks up time and effort, and that if married partners did not gain from a union then they would remain single (e.g., Becker (1974)).

Anecdotally, some of the most prolific inventors did not believe in marriage. Nikola Tesla thought that marriage was inconsistent with great invention. He commented in the *New York Herald* in 1897 "I do not believe an inventor should marry, because he has so intense a nature, with so much in it of wild, passionate quality, that in giving himself to a woman he might love, he would give everything, and so take everything from his chosen field." Tesla went on to argue that "I do not think you can name many great inventions that have been made by married men." However, other great inventors did marry. Elias Howe (1819-1867), the sewing machine inventor, married when he was 21 years of age. Thomas Edison married first at age 24 and within a year had developed the revolutionary quadruplex telegraph for sending messages simultaneously over a single wire. Following the death of his first wife, Edison married again at age 39.

Figure 18: Family Decisions: Probability of Being Married



Notes: This figure plots the probability that an individual is married over their life cycle using data averaged across our six census years (1880, 1900, 1910, 1920, 1930, 1940). The data are constructed from this repeated cross section. The solid green line shows the marriage rate for inventors, while the red dashed line shows the rate for the universe of non-inventors. For comparison, the marriage rate for non-inventor high-skill occupations is plotted in the blue dotted line. Source: 1880-1940 Historical Census Data, USPTO patent records.

Figure 18 shows that inventors delayed marriage substantially relative to the population as a whole although inventors did indeed marry (or stay married) at a higher rate than their non-inventor counterparts at older ages. Inventors also tended to have fewer children. Table 3 above shows that inventors delayed fertility relative to the average American: 72.9% of inventors had a child before the age of 35, while 80% of non-inventors had a child by this time. Of course the relationship between delayed marriage and fewer children is mechanical. As Becker (1974, p.22) points out, "the age of entry [into marriage] would be earlier the larger the number of children desired."

Figure 18 also plots the probability of marriage for those working in a high-skill occupation, such as doctors and lawyers. The figure shows that inventors' marriage decision mirrors that of this group almost one-for-one. This comparison suggests that inventors' difference from the rest of the population is driven by underlying skill differences and human capital investment choices. This similarity in observable marriage patterns with high-skill workers can be reconciled with theoretical models of marriage markets like Bergstrom and Bagnoli (1993), where high-wage men gain by delaying marriage relative to low-wage men because accumulated income is a signal of quality when searching for the best partner.

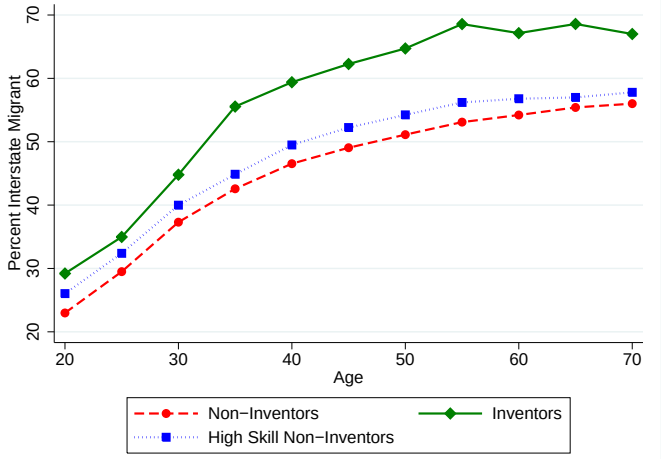
Fact 10. ***Inventors were more likely to have migrated from their state of birth. They moved to states that were more conducive to innovation.***

Individuals migrate in order to seek better job prospects in their destination state. This argument may apply particularly strongly for inventors, since, as shown above, environmental factors shift both the costs and benefits of innovation. The example of Thomas Edison illustrates this point. Not only did he stand to gain more from marketing his inventions in the larger market of New Jersey and New York but he also benefitted from the larger supply of skilled labor and financial development there. When inventors systematically move to such places, this generates spatial concentration giving rise to agglomeration externalities (Carlino and Kerr (2015)).

Figure 19 confirms that Edison's example is representative of the broader inventor population. The figure shows that inventors were most likely to move after the age of 35: the beginning of their most innovative period according to Figure 15. The high migration rate for inventors does not simply reflect their higher average skill level. Indeed, we see that highly skilled individuals in non-inventor occupations migrate significantly less than do inventors.

Migration can boost innovation at a more aggregate level as well. Immigrants can bring new ideas, expertise, and specialized labor to an area, all of which facilitate the production of patented innovations. Although modern studies have produced opposing conclusions on the role played by immigrants in US knowledge production (e.g., Kerr and Lincoln (2010); Hunt and Gauthier-Loiselle (2010); Borjas and Doran (2012)), historical evidence is more unequivocal. Moser et al. (2014) estimates that German emigres who fled the Nazi regime provided a significant boost to US invention during the twentieth century. In our data, Table 1 shows higher levels of interstate and international migration in inventive states. The higher share of international migrants we see in the top 10 inventive states is in line with Akcigit et al. (2016b)'s finding that inventors are internationally highly mobile. In the top 10 most inventive states, 20.6% of the population

Figure 19: Interstate Migration Rates by Age



Notes: Figure plots interstate migration rates by age of individual for the population of high skill individuals. An individual is defined to be an interstate migrant if their birth state is different from their current state of residence. Each point represents a 5-year forward-looking bin. For example, the point at age 20 measures the average migration rate for 20 to 25 year-olds. Figure uses data averaged across the four census years for which we have occupation information: 1880, 1920, 1930, and 1940. Source: 1880, 1920-1940 Historical Census Data, USPTO patent records.

were international migrants, while just 1.7% of the population of the least inventive states were born abroad.

Conditional on moving to a new location, where did inventors go? To answer this question, Figure 20 plots the characteristics of geographic origin and destination amongst inventors who move across state lines in our matched dataset of inventors to the Census. To facilitate the exposition, a 45-degree line is also plotted which denotes no change at all. To understand the plot, consider a point with 20 on the horizontal axis, and 60 on the vertical axis in Panel A. This point shows that an inventor migrant from a state in which 20% of the population lives in an urban area moves to a state with 60% of the population living in an urban area on average.

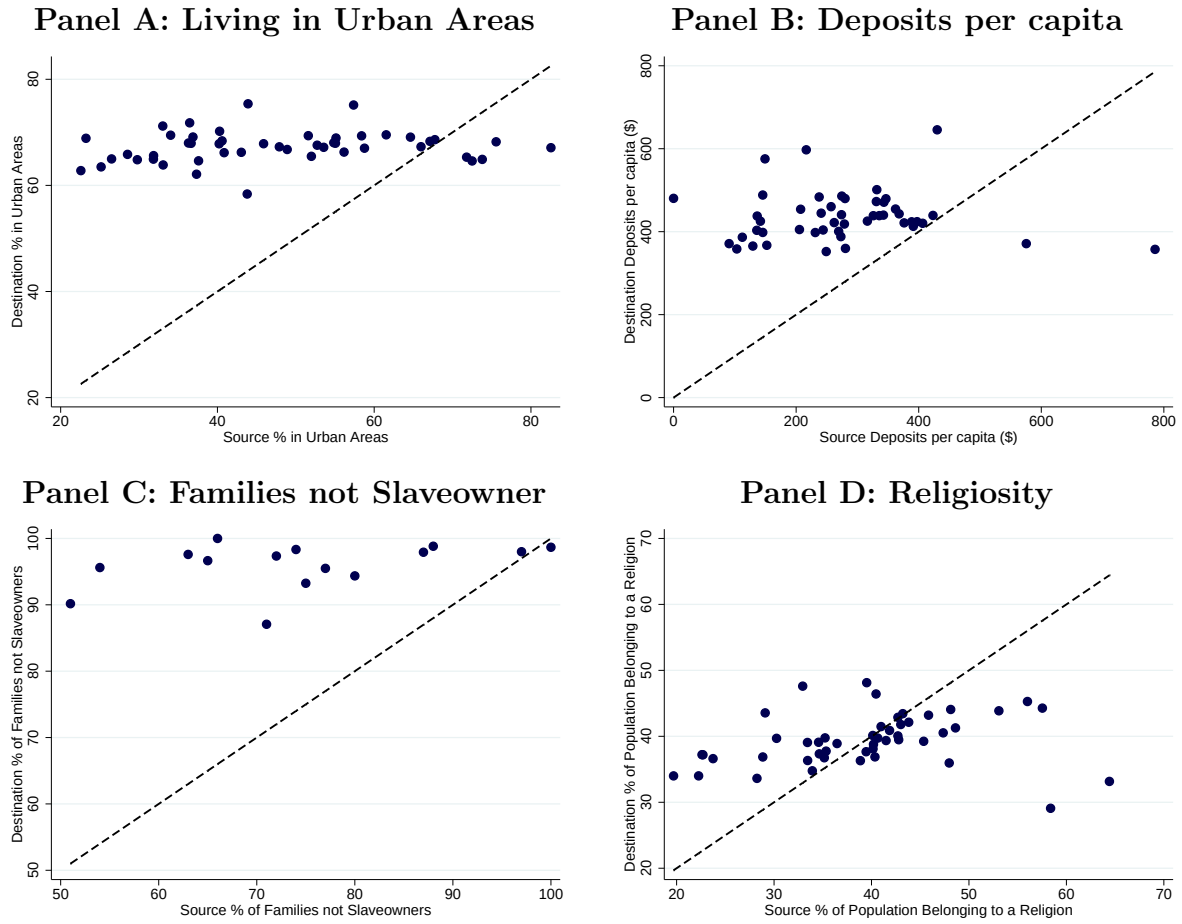
Most of the observations are clustered above the 45-degree line in Panel A, implying that inventors generally moved from less to more urbanized regions. Likewise, Panel B shows that inventors moved toward regions where deposit ratios were higher, suggesting that access to finance could have played a role in their migration decisions. Panel C shows that inventors moved toward regions where slave-ownership had been lower. Panel D shows little movement to places with a strong religious presence. Overall, inventors generally migrated to regions whose characteristics were well-suited to innovation.[18]

3.3 Family Background: Who Became an Inventor?

Having documented the basic facts about inventor careers, demographics and education, we now examine family backgrounds. Does parental affluence matter for the propensity to become an

[18]Equally these were places where the population was moving as well. Thus while Figure 19 shows inventors were much more likely to migrate than were non-inventors, *conditional* on migrating, inventors moved largely in lock-step with non-inventors. This is how clusters of economic activity start to emerge: the movement of people to centers of innovation becomes cumulative as it reduces the cost of moving products, accessing labor and diffusing new ideas (Ellison et al. (2010)).

Figure 20: To Where did Inventors Move?



Notes: In each panel, the horizontal axis plots the average value of the variable of interest in the source state for migrating inventors. The vertical axis plots the average of this variable in the destination state, weighted by the number of inventors who move there from the given source. The dashed line plots the 45^o line. Panel C plots only source states in which slavery was legal in 1860, and thus has fewer data points than Panels A, B and D. Source: 1860, 1940 Historical Census Data, 1916 Census of Religious Bodies, FDIC, USPTO patent records.

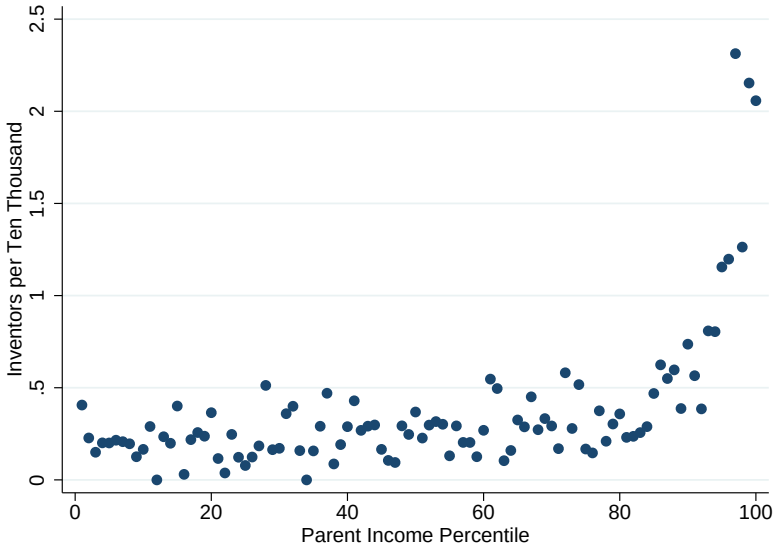
inventor? If so, through what mechanisms might this operate? Throughout this section we rely heavily on our parent-child matched dataset.[19] Because this covers individuals residing in the same household, we are capturing inventors early in their career. Home-leaving ages increased noticeably during the early twentieth century only starting to decline after World War II. Using Census data Gutmann et al. (2002) find that in 1940 the median home-leaving age for white males was 24 whereas 85% of unmarried white males lived at home between ages 15 and 29.

Fact 11. ***Father's income was positively correlated with becoming an inventor. This effect disappears once child's education is controlled for.***

Figure 21 illustrates the relationship between father's income and the probability of becoming an inventor. We find a strong association between the two, especially for the highest-income fathers. The convex relationship between parental income and the propensity to become an inventor is striking in its ubiquity. Aghion et al. (2015b) and Bell et al. (2015) document remarkably similar patterns in modern administrative data from Finland and the United States,

[19]For details of its construction, see appendix A.2.

Figure 21: The Relationship between Father's Income and Becoming an Inventor



Notes: Figure plots the number of inventors per 10,000 people by their father's percentile of wage income in the 1940 census. Only individuals successfully matched to their fathers are included in this plot. Wage income percentiles are calculated using the full sample of matched fathers in the U.S. Source: 1940 Historical Census Data, USPTO patent records.

respectively. The persistence of this relationship across time periods, geographies, and institutions is among the most noteworthy facts in this new literature on the backgrounds of inventors.

Several mechanisms can plausibly drive the patterns illustrated in Figure 21. If education was an important determinant of innovation, then the fact that only wealthy individuals had access to education could imply that credit constraints were binding for low-income families (e.g., Celik (2015)). Furthermore, credit constraints may inhibit the ability of prospective inventors to raise starting capital to develop their ideas. Alternatively, it is possible that high income parents interact in better-connected social circles, permitting their children to access high-quality funding, labor, and marketing resources. Finally, high income parents may have useful skills, knowledge, or genes which they pass on to their children.

We provide insight into some of these potential mechanisms through Table 14, which examines the relationship between fathers and sons using linear probability regressions. The dependent variable is an indicator for being granted at least one patent, scaled by a factor of 100 for legibility. Column 1 establishes a positive correlation between the father being an inventor and the child being an inventor. Column 2 introduces parental income instead and column 3 includes both measures. Having a father with income in the right tail of the distribution is very strongly correlated with the child becoming an inventor, even conditional on basic covariates.

Of course, a potentially confounding effect is that high-income parents could themselves be highly-educated, and so invest more in their children's development. To address this, column 4 adds parental education. Interestingly, parental income still matters. Finally, in column 5 we include the child's own education. The effect of parental income disappears, which suggests that parental income only positively affects the probability of becoming an inventor through its effect on children's access to education. In column 6 we show that this finding is robust to the

inclusion of occupation fixed effects, so it does not reflect occupational skill differences.

Table 14: Who Became an Inventor?

	(1)	(2)	(3)	(4)	(5)
Father Inventor	0.161**	0.159**	0.157**	0.155**	0.154**
	(0.075)	(0.076)	(0.075)	(0.075)	(0.075)
Father Income $90^{th} - 95^{th}$ %ile		0.003**	0.002	-0.001	-0.001
		(0.001)	(0.001)	(0.001)	(0.001)
Father Income 95^{th} %ile and above		0.008***	0.005***	0.000	-0.000
		(0.002)	(0.002)	(0.002)	(0.002)
Father: High School Graduate			0.004**	-0.001	-0.001
			(0.001)	(0.001)	(0.001)
Father: At least Some College			0.007***	-0.002*	-0.001
			(0.001)	(0.001)	(0.001)
Self: High School Graduate				0.006***	0.005***
				(0.001)	(0.001)
Self: At least Some College				0.029***	0.023***
				(0.004)	(0.004)
Occupation FE	N	N	N	N	Y
Observations	82810258	82810258	82810258	82810258	82810258
Mean of Dep. Var.	0.011	0.011	0.011	0.011	0.011

Notes: Standard errors clustered at the state-level reported in parentheses. All regressions include state fixed effects, and controls for race, sex, migration status, a quadratic in age, and father's age. Columns (2) through (5) include indicators for father being between the 50^{th} and 75^{th} percentile of income, and between the 75^{th} and 90^{th} percentile of income as independent variables. The omitted categories are below median income and less than high school eduction. Column (5) includes fixed effect controls for Census-defined occupation categories, including those with missing occupation data as a separate category. Source: 1940 Historical Census Data, USPTO patent records.

While Table 14 focuses on the extensive margin—the characteristics of those becoming inventors—Table 15 considers the relationship between an inventor's background and his productivity on the intensive margin, measured by the number of career patents and citations he generates. In column 1 we find a weak positive effect of the father being an inventor. In columns 2 through 4 and 6 through 8, we do not detect a strong effect of father's income, or father's education. In column 5 we introduce the child's own education and this is strongly correlated with long run inventiveness. In other words, the most highly educated inventors tended to be the most productive.

Two findings emerge when taking Table 14 and Table 15 together. First, the importance of education holds both at the extensive and intensive margins, which is consistent with a human capital explanation of invention. Second, both father inventor status and parental income matter on the extensive margin but not on the intensive margin, which suggests that the existence of credit constraints might have undermined inventiveness. This second finding is related to a long line of research in the family firm and management practice literatures, showing that privileged access to career paths (e.g., inherited CEO roles) is associated with under performance (e.g., Perez-Gonzalez (2006); Bloom and Van Reenen (2007) Caselli and Gennaioli (2013)).

Table 15: Individual Background and Long Run Inventiveness

Dependent Variable:	Log Career Patents				Log Career Citations			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Father Inventor	0.140	0.164	0.120	-0.019	0.693	0.693	0.779*	0.633
	(0.692)	(0.643)	(0.571)	(0.578)	(0.508)	(0.454)	(0.405)	(0.429)
Father Income $90^{th} - 95^{th}$ %ile		-0.211	-0.170	-0.180		-0.164	-0.142	-0.152
		(0.241)	(0.252)	(0.252)		(0.332)	(0.328)	(0.331)
Father Income 95^{th} %ile and above		0.185	0.077	0.062		0.370	0.294	0.277
		(0.208)	(0.203)	(0.195)		(0.357)	(0.395)	(0.384)
Father: High School Graduate			0.066	-0.032			0.248	0.145
			(0.118)	(0.122)			(0.232)	(0.229)
Father: At least Some College			0.234*	0.121			0.224	0.104
			(0.134)	(0.126)			(0.175)	(0.176)
Self: High School Graduate				0.065				0.042
				(0.042)				(0.063)
Self: At least Some College				0.297***				0.310***
				(0.052)				(0.054)
Observations	9032	9032	9032	9032	9032	9032	9032	9032
Mean of Dep. Var.	1.581	1.581	1.581	1.581	3.205	3.205	3.205	3.205
S.D. of Dep. Var.	1.365	1.365	1.365	1.365	1.964	1.964	1.964	1.964

Notes: Standard errors clustered at the state-level reported in parentheses. All regressions include state fixed effects, and controls for race, sex, migration status, a quadratic in age, and father's age. Columns (2) through (4) and (6) through (8) include indicators for father being between the 50^{th} and 75^{th} percentile of income, and between the 75^{th} and 90^{th} percentile of income as independent variables. The omitted income category is below median income, and we omit an indicator for the individual having less than a high school education. *,**,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: 1940 Historical Census Data, USPTO patent records.

3.4 Return to Innovation

With our dataset of inventors matched to the 1940 Census we can examine the private returns to innovation, with two caveats in mind. First, information on labor income is not recorded for all observations in the Census; and second labor income itself provides only a partial measure of the total financial returns to innovation. The discovery of new inventions may permit individuals to start their own business and earn a return on new capital assets. Hurst and Pugsley (2011) show that non-wage factors may be an important benefit of self-employment.

What we can do is examine the distribution and life cycle of wage earnings for inventors, both unconditionally and conditional on observable characteristics. Insights from inventor earnings profiles are important because the literature in this area is quite sparse due to data constraints. There is no systematic empirical evidence on inventor earnings during the late nineteenth and early twentieth centuries, even though anecdotally this was seen to be a key determinant of inventor behavior. Schmookler (1966) argued that the expectation of pecuniary gain was implicit to most inventors' careers, citing Thomas Edison as an inventor whose motivations were largely commercial and demand-driven by anticipated market size.

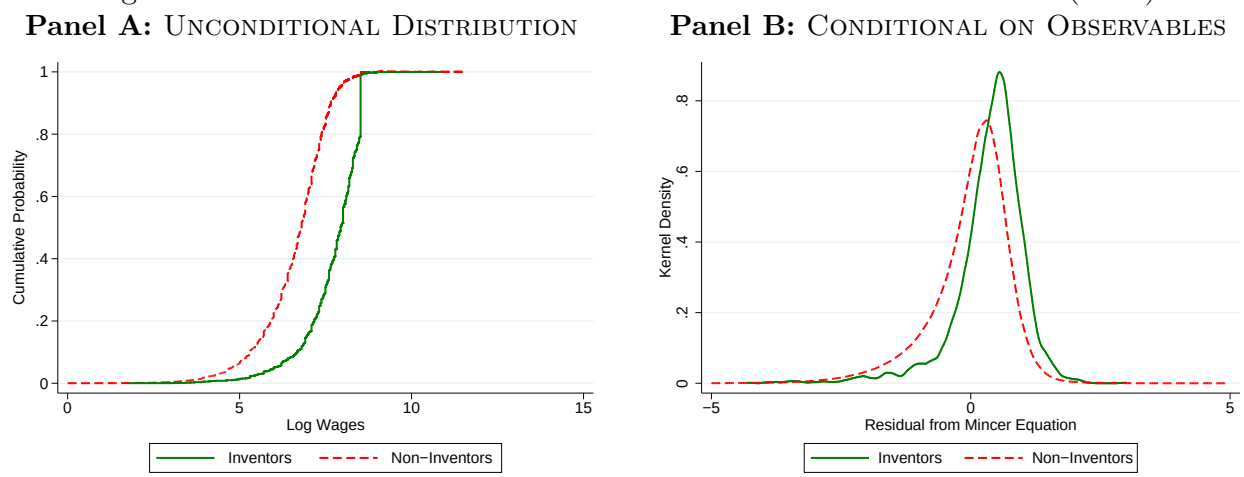
Even for modern periods, few studies on the profile of inventor earnings exist. There are some studies on the topic, however. For example, Astebro (2003) characterises the distribution of returns to independent invention for Canadian inventors between 1976 and 1993 as being highly

skewed, with the average inventor earning less than an equivalent investment in a portfolio of high-risk securities. Toivanen and Vaananen (2012) estimate a varying premium across inventions for a sample of Finnish inventors who patented in the US between 1991 and 1999, although they show the returns to the quality of technological development are particularly high. Depalo and Di Addario (2014) also find that superstar Italian inventors patenting between 1987 and 2006 earned the most.

Fact 12. ***Successful patentees had substantially higher labor income, even controlling for demographics, occupation, and education.***

Figure 22 plots the distribution of wage income for inventors and non-inventors in the 1940 Census. Panel A plots the unconditional CDF of log wage income for both groups. The solid green line represents the distribution of inventors' income, while the dashed red line shows the distribution for non-inventors. Unsurprisingly, inventors have relatively high incomes. Indeed, the inventors' income distribution first order stochastically dominates that of non-inventors.

Figure 22: The Distribution of Labor Income by Inventor Status (1940)



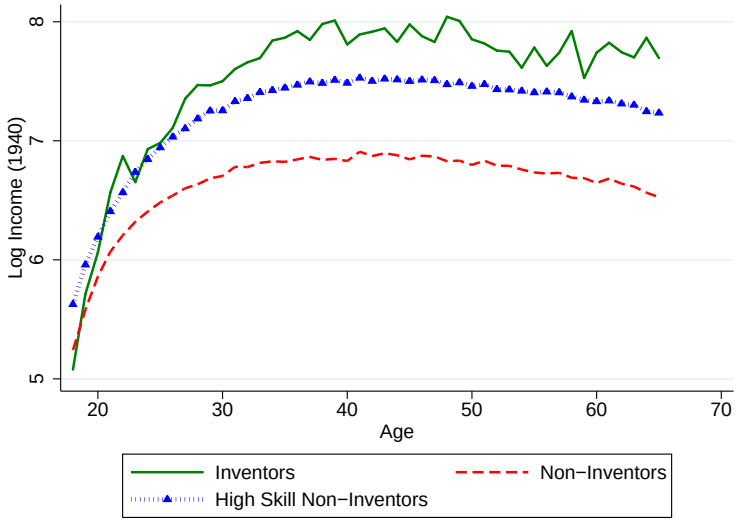
Notes: Figure plots the distribution of the natural log of wage income for inventors and non-inventors, as reported in the 1940 census. Many individuals report 0 wages, and are excluded from this plot. Solid green lines plot the distribution of inventors' wages, while dashed red lines plot the distribution of non-inventors' wages. Panel A plots the unconditional CDF of log wages. Panel B plots the density of log wages residualized against observable characteristics. Specifically, it plots the distribution of residuals from a regression in which the dependent variable is log wages, and includes controls for race, education, sex, international migrant status, residence state fixed effects, occupation fixed effects, and a quadratic in age. Source: 1940 Historical Census Data, USPTO patent records.

Of course, this result is expected given that inventors are better-educated, higher-skilled, and live in more urban states than non-inventors. Panel B therefore plots the distribution of wages for inventors and non-inventors after conditioning on observables. Specifically, we regress an individual's log wages on race, education, sex, international migrant status, residence state fixed effects, occupation fixed effects, and a quadratic in age. We then plot the distribution of residuals from this regression for inventors and non-inventors. Even after controlling for all observable characteristics, inventors have higher wage incomes throughout the distribution.

Inventors also have a steeper life cycle profile of wages. Figure 23 plots the average life cycle of log earnings for inventors, non-inventors, and non-inventors in high skill occupations. This

figure is constructed from the cross-section of individuals at each age. We see that inventors have higher earnings throughout their life cycle than non-inventors and high-skilled individuals. Indeed, Table A-5 in the Appendix shows that the difference between the wages of inventors and high-skill non-inventors is statistically significant at the 1% level from the age of 19 onwards. These figures provide suggestive evidence that invention was a key labor income differentiator.

Figure 23: The Life Cycle of Earnings by Inventor Status



Notes: Figure plots the evolution of log average wage income over the life cycle. The solid green line plots the evolution of inventors' wage income, while the dashed red line plots the wage evolution of the universe of non-inventors. The dotted blue line plots the life cycle of wages for doctors and lawyers. All plots use 1940 census data only. Source: 1940 Historical Census Data, USPTO patent records.

Fact 13. ***Inventor's income was highly correlated with the quality of invention, as measured by patent citations.***

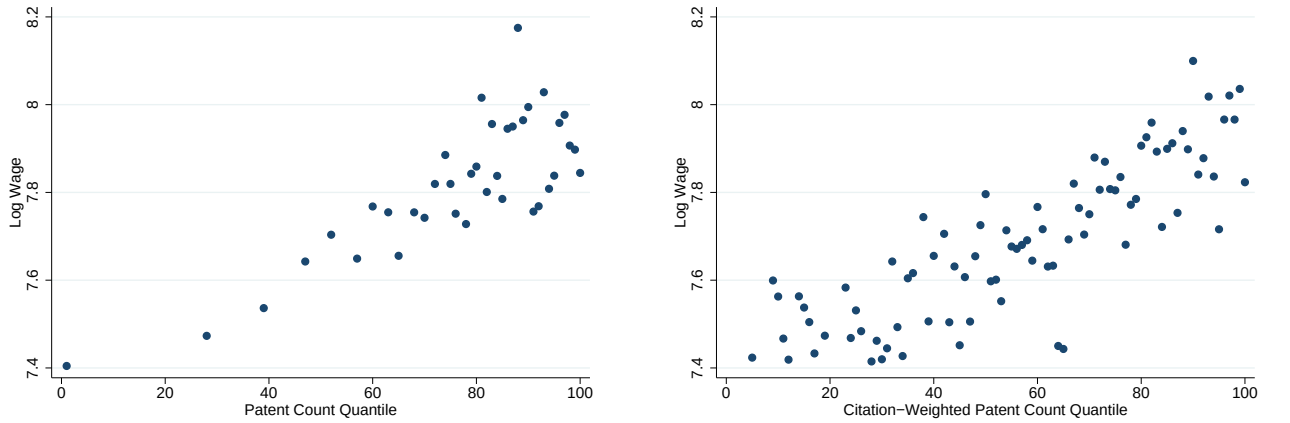
If the returns to invention reflected pecuniary gains from technological development we would also expect to observe a correlation between labor income and the quality of patents. Our data shows a strong correlation between the the quality of an inventor's patent portfolio and log wages. Figure 24 Panel A plots the relationship between the number of patents an inventor files over his life time and log average wages. Panel B mirrors Panel A, except the horizontal axis now weights each patent in an inventors portfolio by the number of citations the patent receives. Both panels exhibit a robust positive relationship between inventor productivity and log wages, suggesting that the higher-quality inventors were being compensated for their inventions.

There are two possible explanations for this positive relationship between wages and inventor productivity. First, an inventor may simply be more productive as a result of his past inventions. Alternatively, if invention is a signal of underlying worker type, an employer may pay an inventor more of a financial premium in anticipation of future productivity.

To disentangle these two effects, we regress log wages, measured in 1940, on an inventor's innovative activity both before and after 1940. If the current productivity effect dominates, we would expect pre-1940 innovation to have a strong effect on wages. However, if the anticipation effect dominates, the forward-looking innovative activity should predict an inventor's wages,

Figure 24: The Relationship between Innovative Productivity and Wages

Panel A: Unweighted Patent Counts **Panel B:** Citation-Weighted Patent Counts



Notes: Figure plots the relationship between log average wages and the quantile of inventive activity, conditional on being granted at least one patent. The median inventor is granted only 3 patents. Thus the unweighted patent count has relatively few percentile points at the low end of the distribution: the first 24% of the distribution is contained in the first percentile data point. Source: 1940 Historical Census Data, USPTO patent records.

so long as employers correctly anticipate an employee's future productivity. The results are reported in Table 16. Each regression controls for inventor demographics, education, and state. Whether we measure innovative activity as the log number of patents or as the log number of citations received, we find the anticipation effect to be a stronger determinant of wages.

One might expect the anticipation effect to be stronger for young inventors who have a longer career ahead of them at the point in time that they enter the most productive part of their careers (see Figure 15). In columns 2 and 4 of Table 16, we interact our forward- and backward-looking innovation measures with an indicator for whether the individual is over 35 years old in 1940. In line with theory, past productivity is a stronger predictor of wages for those over 35, while the reverse is true for those under age 35.

3.5 Income Inequality and Social Mobility

In this section we use the income data in the 1940 Census to study the effect of innovation on the make up of local societies. Do more inventive states have higher or lower wage inequality and social mobility? The nature of the relationship between innovation and income inequality is *ex ante* ambiguous. If innovation displaces incumbent firms and creates new wealth for competing entrants, more innovative societies are more likely to have lower income inequality. However, if innovation primarily strengthens incumbent firms, allowing them to increase markups and constrain output, income inequality in a society may rise with its level of innovation.

This issue is especially relevant given the large literature on income inequality and recent attempts to analyze the relationship between the top income share and patenting. The existing empirical literature is divided on the topic. Aghion et al. (2015a) examine modern US data finding a positive causal effect of innovation-led growth on top incomes shares at the state-level. However, they also find some sensitivity to measurement. The relationship between inequality and patenting becomes much weaker at different thresholds like the top 10% share, and they find

Table 16: What Determined Income?

	(1)	(2)	(3)	(4)
Log Patents Pre-1940	0.028**	-0.051***		
	(0.012)	(0.018)		
Log Patents Post-1940	0.064***	0.125***		
	(0.008)	(0.015)		
Log Patents Pre-1940 × Over 35		0.109***		
		(0.018)		
Log Patents Post-1940 × Over 35		-0.090***		
		(0.016)		
Log Citations Pre-1940			0.017**	-0.013
			(0.006)	(0.009)
Log Citations Post-1940			0.038***	0.067***
			(0.006)	(0.009)
Log Citations Pre-1940 × Over 35				0.041***
				(0.010)
Log Citations Post-1940 × Over 35				-0.042***
				(0.010)
Observations	6059	6059	6059	6059
Mean of Dep. Var.	7.636	7.636	7.636	7.636
S.D. of Dep. Var.	0.850	0.850	0.850	0.850

Notes: Table presents estimated coefficients from a regression of log wages on innovation measures. We restrict our attention to the sample of inventors matched to the 1940 census. Standard errors clustered at the state-level reported in parentheses. All regressions include state fixed effects, and controls for race, sex, migration status, occupation skill level, education and a quadratic in age. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: 1940 Historical Census Data, USPTO patent records.

a negative relationship when using the Gini coefficient, which considers all parts of the income distribution not just the top share. By contrast Jones and Kim (2014) shows theoretically that if innovations come from new entrants, the relationship between inequality and innovation could be negative. Given the lack of consensus on the topic, it is important to place these modern studies in a historical context.

Fact 14. ***Broad measures of income inequality (90/10, Gini) were negatively correlated with innovation.***

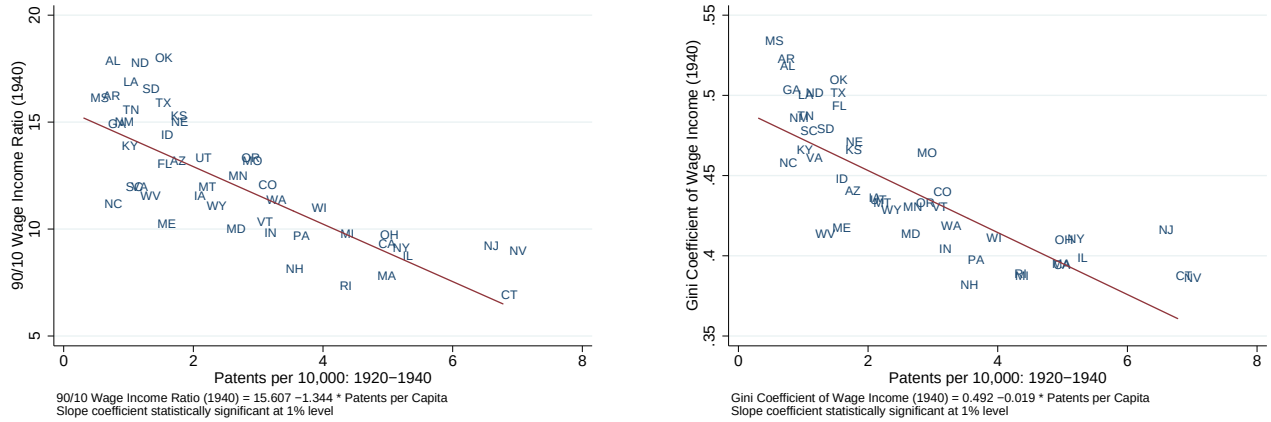
Our results in Figure 25 generally point to a negative relationship between income inequality and inventiveness.

The vertical axis plots the state-level 90/10 ratio and Gini coefficient as measured in the 1940 Census, while the horizontal axis plots backward-looking average patents per capita between 1920 and 1940. Both of these measures of inequality are strongly negatively associated with regional inventiveness.

Table 17 reports the results from a state-level regression of 1940 wage income inequality on average patents per capita between 1920 and 1940, and the state's occupation mix. All independent variables in the regression are standardized to have zero mean and unit standard deviation. Column 2 shows that increasing the number of patents per capita by one standard deviation is associated with a decline in the 90/10 ratio of 0.28 (= 0.828/2.98) standard deviations, conditional

Figure 25: Relationship between Wage Income Inequality and Inventiveness

Panel A: Ratio of 90^{th} to 10^{th} Percentile **Panel B:** Gini Coefficient



Notes: Figure plots the relationship between average patents per 10,000 residents between 1920 and 1940, and the state-level wage income inequality observed in the 1940 census. Panel A measures income inequality with the ratio of the 90^{th} percentile to the 10^{th} percentile of income, while panel B uses the Gini coefficient as its measure. Source: 1940 Historical Census Data, USPTO patent records.

on the state's occupation mix.[20]

Table 17: Wage Income Inequality and Innovation

Dependent Variable:	90/10 Ratio		Gini Coefficient	
	(1)	(2)	(3)	(4)
Av. Patents per Capita 1920-1940	-2.210***	-0.828**	-0.030***	-0.010
	(0.358)	(0.343)	(0.006)	(0.007)
% Agricultural Occupation (1940)		1.777***		0.020***
		(0.343)		(0.006)
% Manufacturing Occupation (1940)		-0.086		-0.012***
		(0.216)		(0.003)
Observations	48	48	48	48
R-squared	0.5545	0.7150	0.5239	0.7509
Mean of Dep. Var.	12.30	12.30	0.44	0.44
Std. Dev. of Dep. Var.	2.98	2.98	0.04	0.04

Notes: Table reports estimated coefficients from a regression of 1940 income inequality, measured by the ratio of the 90^{th} to the 10^{th} percentile of wage income (columns 1 and 2) and the Gini coefficient (columns 3 and 4), on the average patents per 10,000 residents between 1920 and 1940. Independent variables standardized to have zero mean and unit standard deviation. White heteroskedasticity robust standard errors reported in parentheses. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: 1940 Historical Census Data, USPTO patent records.

Fact 15. ***Top-1% labor-income share had a U-shaped relationship with innovation.***

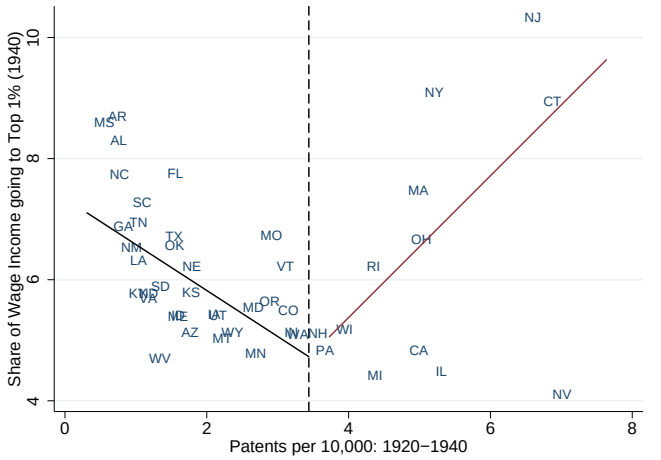
The estimated relationship between innovation and income inequality is sensitive to measurement.

Contrary to the findings of the previous section, the top 1% income share exhibits a non-linear, U-shaped relationship with patenting. In the least innovative states we find a negative

[20] We do not present evidence on the relationship between county innovation and county-level inequality, as such an analysis would miss the important cross-county inequality margin.

relationship. However, in the most innovative states such as New York, New Jersey and Massachusetts we find that more patenting was associated with more income held in by the top 1%. One potentially confounding effect is the different mixes of occupations in these innovative states. For example, Philippon and Reshef (2012) show that between 1909 and 1933 skill-based wage-compensation in finance was high. Addressing this concern, Appendix Figure A-8 shows that these patterns are robust to excluding individuals who work in the financial sector.

Figure 26: Relationship between Top-1% Labor-income Share and Inventiveness



Notes: Figure plots the relationship between average patents per 10,000 residents between 1920 and 1940, and 1940 state-level wage income inequality, measured by the share of income controlled by the top 1% of the state's wage earners. Source: 1940 Historical Census Data, USPTO patent records.

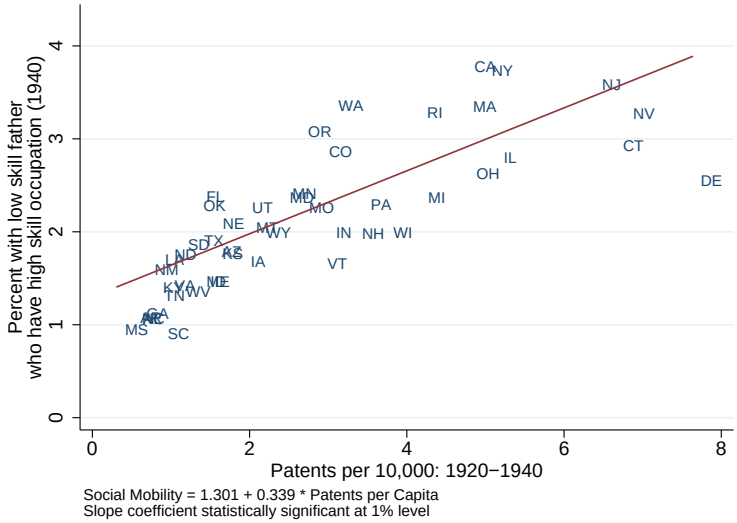
Although we do not claim to show a causal link between innovation and income inequality, our analysis yields a number of important insights. First, alternative measures of inequality may yield startlingly different results. The literature has not yet reached a consensus on the economics behind these various measures. Second, the correlations presented here suggest that innovation may indeed have an important effect on income inequality. Our time period covers a period in US history when income inequality was high (Goldin and Katz (1999), Piketty and Saez (2003)), and it was also associated with the diffusion of electricity as a major twentieth century general purpose technology(Jovanovic and Rousseau (2005)). Our findings are therefore consistent with the idea that technological change may drive a wedge between the relative earnings of the skilled and unskilled.

Fact 16. ***Innovation was strongly positively correlated with social mobility.***

While places with high income inequality tend to exhibit low levels of social mobility (e.g. Chetty et al. (2014)), innovation might actually decrease inequality if it acts as a social elevator. The Schumpeterian paradigm suggests that innovation allows a new entrant to capture markets from old incumbents. This process of creative destruction creates churn in the economy, allowing individuals and firms with limited market shares to grow. This mechanism lies at the heart of a large class of economic models, such as Aghion and Howitt (1992), Klette and Kortum (2004), and Akcigit and Kerr (2016). Yet, data constraints have prohibited direct evidence relating innovation to social mobility at the societal level.

We examine the relationship between innovation and social mobility directly using our occupation data. Figure 27 shows how a state's level of social mobility in 1940 correlates with the number of patents per capita granted between 1920 and 1940. Social mobility is measured as the fraction of individuals with a low skill father, who themselves have a high skill occupation. We see that more innovative regions feature more social mobility.

Figure 27: The Relationship between Inventiveness and Social Mobility



Notes: Figure plots the relationship between average patents per 10,000 residents between 1920 and 1940, and 1940 social mobility, measured by the share of those with a low-skill father who themselves have a high skill occupation. Source: 1940 Historical Census Data, USPTO patent records.

Of course, more innovative states may have higher social mobility for a number of reasons. For example, we have shown that innovative states tend to be more urban which may be correlated with social mobility while states with different types of economic activity (agricultural versus manufacturing) may also vary in their degree of change. To counter this concern, we regress our measure of social mobility in 1940 on average patents granted per capita between 1920 and 1940, and controls for a state's occupation mix. The percent of people in a state who work in the agricultural sector serves as a proxy for the state's degree of urbanization, while the states least open to technological disruption tended to be agricultural and in the south. Table 18 shows that the relationship between patents per capita and social mobility is positive and statistically significant, even after controlling for occupational variation.

These results are informative from the standpoint of innovation as a driver of social mobility given that Long and Ferrie (2013) find that around the turn of the twentieth century America was (from its mid-nineteenth century high-point) generally becoming a less socially mobile place. While we do not measure changing mobility levels over time, our results do indicate that innovative places were also socially mobile places. As we showed in Table 3 inventors were a small share of the total population who had a large effect on US development. Our findings underscore the need to study social movement within this sub-group of the population given the implications for the relationship between social mobility and economic growth.

Table 18: % of High-skill Child given Low-skill Father

	(1)	(2)
Av. Patents per Capita 1920-1940	0.746***	0.484***
	(0.116)	(0.149)
% Agricultural Occupation (1940)		-0.031***
		(0.011)
% Manufacturing Occupation (1940)		-0.016
		(0.019)
Observations	49	48
R-squared	0.5924	0.6844

Notes: Table reports estimated coefficients from a regression of 1940 social mobility, measured by the share of those with a low-skill father who themselves have a high skill occupation, on the average patents per 10,000 residents between 1920 and 1940. Both dependent and independent variables standardized to have zero mean and unit standard deviation. White heteroskedasticity robust standard errors reported in parentheses below coefficient. *,** ,*** represent that coefficients statistically differ from 0 at the 10%, 5%, and 1% level. Source: 1940 Historical Census Data, USPTO patent records.

4 Relation to Previous Studies

Finally, we attempt to overview our findings by codifying the facts we have presented throughout the paper in relation to the existing literature. We frame the discussion using Table 19, which summarizes areas in which our findings tend to agree with previous studies or disagree. In some cases our findings have no real counterpart in the literature. We highlight these as representing new facts about the dynamics of US inventive activity.

Our regional-level findings agree with the basic arguments underpinning four main literatures. The long-standing endogenous growth literature is framed by the idea that innovation and technological progress are key determinants of sustained development (e.g., Romer (1990), Aghion and Howitt (1992)). Our results (**Fact 1**) have verified that basic premise holds true in the data and in Section 3 we also attempted to show that the relationship between innovation and growth is likely to be causal. We have also been able to identify key mechanisms. Consistent with the large literature on agglomeration (e.g., Glaeser (2011)), we found that densely populated places were more innovative (**Fact 2**). Our findings (**Facts 3 and 4**, respectively) also emphasise the importance for innovation of access to finance (e.g., King and Levine (1993)) and spatial links that would increase the extent of the market (e.g., Donaldson and Hornbeck (2016)).

Equally, some of our findings create a degree of ambiguity in relation to the current literature. Given the frequently postulated relationship between the degree of openness of a region and its inventiveness we find some contrasting results (**Fact 5**). The experience of US slave states is perhaps the exemplar case of being closed to new ideas and we do find an empirically strong negative relationship between slavery and innovation. On the other hand, religiosity is a weak predictor of innovation in our data and time period. In the modern era the relationship is robustly negative (Benabou et al. (2013) and Benabou et al. (2015)). We consider these types of contrasting results to be equally informative because they reveal what has changed over time in the types of social structures that affect a society's propensity to innovate.

Given the large literature on the profiles of inventors historically (e.g., Lamoreaux and Sokoloff

Table 19: Codifying the Facts

		Agree	Disagree	New Fact
Fact 1	More inventive states and sectors **grew faster** on average.	✓		
Fact 2	**Densely-populated** states were more inventive.	✓		
Fact 3	**Financially-developed** states were more inventive.	✓		
Fact 4	**Geographically-connected** states were more inventive.	✓		
Fact 5	States associated with **slavery** were less inventive but **religiosity** is not robustly correlated with inventiveness.	✓	✓	
Fact 6	Inventors were **more educated** on average and were most productive between the age of **36 and 55**.	✓		
Fact 7	Inventors were **positively selected** through exit early in their careers, but were less productive and more likely to exit late in their careers.			✓
Fact 8	The patents of **new inventors received more citations** on average, and were more likely to be in the top decile of the citations distribution.			✓
Fact 9	Inventors **delayed marriage** and had **fewer children**.			✓
Fact 10	Inventors were more likely to **have migrated** from their state of birth. They moved to states that were more conducive to innovation.			✓
Fact 11	**Father's income** was correlated with becoming an inventor. This effect disappears once **child's education** is controlled for.			✓
Fact 12	Successful patentees had substantially higher **labor income**, even controlling for demographics and education.	✓		
Fact 13	Inventor's income was highly correlated with the **quality of invention** measured by patent citations.	✓		
Fact 14	Broad measures of **income inequality**, such as 90/10 ratio, Gini coefficient, were negatively correlated with innovation.	✓	✓	
Fact 15	However, the **top-1% income share** has a U-shaped relationship with innovation.	✓	✓	
Fact 16	Innovation was strongly positively correlated with **social mobility**.	✓		

(1999) and Khan and Sokoloff (2004)), our data help to confirm major findings about age profiles and education characteristics using a large sample of inventors (**Fact 6**). It is also reassuring that our results on the age distribution mirror those of Sarada et al. (2016) who match inventors to Census records for the same time period as us using different data and methods. These types of data are crucial to understanding the relationship between human capital and economic growth (e.g., Lucas (2009), Alvarez et al. (2013), Lucas and Moll (2014), Perla and Tonetti (2014)). Our results pinpoint productivity variation over the life cycle and continue to dispel the traditional argument that US inventors during the golden age were uneducated back-yard tinkerers.

We have also attempted to extend the analysis in a new direction by examining the life cycle of invention through the lens of career dynamics. Following the literature on firm dynamics (e.g., Haltiwanger (2012)) and the economic growth literature on entry, exit and reallocation (e.g., Acemoglu et al. (2015) and Akcigit and Kerr (2016)) we find that inventors were positively selected early in their careers on the quality of their inventions but their productivity dropped sharply later in their careers, presumably as new entrants disrupted existing ideas (**Facts 7 and 8**). These findings help to reconcile why the age distribution is heaped around middle-age (**Fact 6**) as does our evidence on marriage patterns, child rearing and time allocation (**Fact 9**). Finally, while it is known that inventors were susceptible to migration (Sarada et al. (2016)) we have presented further evidence on the nature of re-location decisions (**Fact 10**). These findings, in turn, relate to our regional results showing that inventors tended to move to urban and financially-developed places that were more likely to foster innovation (**Facts 2 and 3**).

We are also able to present new evidence on the family background of inventors (**Fact 11**). On the extensive margin our finding that the entry of inventors was increasing in father's income, but the mechanism appears to be operating through better off families providing access to education, is consistent with a human capital explanation of invention (**Fact 6**). Furthermore, our finding that father's income had no effect on the intensive margin of productivity is consistent with the broader hypothesis that US technological development was characterized by the democratization of invention (Khan and Sokoloff (2004)).

A large literature going back to Schmookler (1966) suggests that financial incentives drive innovation and our findings confirm that rewards did accrue to individuals developing new technological ideas. Most of the literature in this area is confined to modern periods where large scale administrative records are available.(e.g., Bell et al. (2015) and Aghion et al. (2015b)). While our results are novel given the lack of evidence on the returns to income for inventors in historical time periods, they confirm the hypothesis that incentives matter. Inventors earned more, especially if they produced high quality innovations (**Facts 12 and 13**).

Our results pertaining to the relationship between innovation, income inequality and social mobility are controversial relative to the recent literature (e.g., Aghion et al. (2015b) and Bell et al. (2015)) and they also indicate the need for more research. Much depends on the approach to measuring inequality: the relationship between inequality and innovation is different if one considers society as a whole (**Fact 14** on the 90/10 ratio and Gini coefficient) or only the upper echelons (**Fact 15** on the top-1%). Our final finding, that innovation appears to be a major social elevator (**Fact 16**) has important implications for how mobility connects to economic

growth. Inventors are a small sub-group of society who have a large influence on everyday life because of their technological ingenuity. We find that invention could be an important social elevator even though the US population as a whole was becoming less socially mobile over time (Long and Ferrie (2013)).

5 Conclusion

This paper presents a series of facts emerging from a major data collection and matching exercise combining information from US patent records with state and county-level aggregates and data on individuals from Federal Censuses between 1880 and 1940. The new data provide a comprehensive profile of inventions and their creators during the golden age of US invention. Our analysis begins with an attempt to identify a causal relationship between innovation and long run economic growth. We proceed to explore some of the main mechanisms driving this relationship using a framework for establishing macro and micro-level facts about the innovation process.

Examining the drivers of innovation during this historical time period is critical and sheds light on numerous key debates on innovation and long-run economic growth. Our data can also complement modern studies such as Aghion et al. (2015b) and Bell et al. (2015) to provide a more complete picture of inventor profiles over time and space. These new historical data exhibit a positive relationship between state-level innovation and regional growth, population density, financial development, geographic connectedness and social mobility. We further find that inventors during this period were largely middle-aged white males, with above-average education. Inventors in 1940 were positively selected through early exit, had steeper life cycle wage earnings profiles and right-shifted earnings distributions compared with similarly-skilled individuals; this was especially true of the period's most prolific inventors. Finally, our data suggest a non-monotone relationship between a state's level of innovation and its income inequality. These areas have allowed us to gain preliminary insights into the birth of technological ingenuity during one of the most important eras of American economic development.

A number of opportunities exist for further research. We have provided a range of interesting correlations in this paper, and establishing causal links between our variables will help to shed light on the nature of these facts and the various mechanisms at work. Yet taken at face value, our study has important implications for the design of innovation policy. Establishing the background of the most effective inventors informs well-targeted innovation interventions. Furthermore, while regional innovation yields growth, it is also related to inequality and social mobility. The extent to which innovation contributes to growth, inequality, and mobility is central to determining the societal costs and benefits of technological advance.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson, "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of*

Political Economy, 2014, *122* (3), 467–506.

Acemoglu, Daron and Joshua Linn, "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *Quarterly Journal of Economics*, 2004, *119* (3), 1049–1090.

_ , Jacob Moscona, and James A. Robinson, "State Capacity and American Technology: Evidence from the 19th Century," Working Paper 21932, National Bureau of Economic Research January 2016.

_ , Ufuk Akcigit, and Murat Alp Celik, "Young, Restless and Creative: Openness to Disruption and Creative Innovations," 2014. National Bureau of Economic Research WP # 19894.

_ , _ , and William R Kerr, "Innovation Network," *Proceedings of the National Academy of Sciences*, 2016, *113* (41), 11483–11488.

_ , _ , Nicholas Bloom, and William R Kerr, "Innovation, Reallocation, and Growth," 2015. NBER Working Paper #18993.

Acharya, Avidit, Matthew Blackwell, and Maya Sen, "The Political Legacy of American Slavery," *Journal of Politics*, 2016, *78* (3), 621–641.

Aghion, Philippe and Peter Howitt, "A Model of Growth through Creative Destruction," *Econometrica*, 1992, *60* (2).

_ , Antoine Dechezleprêtre, David Hémous, Ralf Martin, and John Van Reenen, "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry," *Journal of Political Economy*, 2016, *124* (1), 1–51.

_ , Ufuk Akcigit, Antonin Bergeaud, Richard Blundell, and David Hémous, "Innovation and Top Income Inequality," 2015. NBER Working Paper # 21247.

_ , _ , Ari Hyytinen, and Otto Toivanen, "Living the American Dream in Finland: The Social Mobility of Innovators," 2015. University of Chicago mimeo.

Akcigit, Ufuk and William R. Kerr, "Growth Through Heterogeneous Innovations," 2016. NBER Working Paper # 16443.

_ , Murat Alp Celik, and Jeremy Greenwood, "Buy, Keep or Sell: Economic Growth and the Market for Ideas," *Econometrica*, 2016, *84* (3), 943–984.

_ , Salome Baslandze, and Stefanie Stantcheva, "Taxation and the International Migration of Inventors," *American Economic Review*, 2016, *106* (10), 2930–2981.

Alvarez, Fernando E, Francisco J Buera, and Robert E Lucas, "Idea flows, economic growth, and trade," 2013. National Bureau of Economic Research WP #19667.

Astebro, Thomas, "The Return to Independent Invention: Evidence of Unrealistic Optimism, Risk Seeking or Skewness Loving?," *Economic Journal*, 2003, *113* (484), 226–239.

Barro, Robert, "Output Effects of Government Purchases," *Journal of Political Economy*, 1981, *89* (6), 1086–1121.

Barro, Robert J, "Economic Growth in a Cross Section of Countries," *Quarterly Journal of Economics*, 1991, *106* (2), 407–443.

_ , "Human Capital and Growth," *American Economic Review*, 2001, *91* (2), 12–17.

Becker, Gary, *Human Capital and the Personal Distribution of Income: An Analytical Approach*, Ann Arbor: University of Michigan Press, 1967.

Becker, Gary S, "A Theory of Marriage: Part II," *Journal of Political Economy*, Part II, 1974, *82* (2), S11–S26.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, "The Lifecycle of Inventors," 2015. Harvard mimeo.

Benabou, Roland, Davide Ticchi, and Andrea Vindigni, "Forbidden Fruits: The Political Economy of Science, Religion, and Growth," Working Paper, Princeton University April 2013.

_ , _ , and _ , "Religion and Innovation," 2015. National Bureau of Economic Research WP #21052.

Benhabib, Jess and Mark M Spiegel, "The Role of Human Capital in Economic Development: Evidence from Aggregate Cross-country Data," *Journal of Monetary Economics*, 1994, *34* (2), 143–173.

Bergstrom, Theodore C and Mark Bagnoli, "Courtship as a Waiting Game," *Journal of Political Economy*, February 1993, *101* (1), 185–202.

Bernstein, Shai, "Does Going Public Affect Innovation?," *Journal of Finance*, 2015, *70* (4), 1365–1403.

Bils, Mark and Peter J Klenow, "Does Schooling Cause Growth?," *American Economic Review*, 2000, *90* (5), 1160–1183.

Bloom, Nicholas and John Van Reenen, "Measuring and Explaining Management Practices Across Firms and Countries," *Quarterly Journal of Economics*, 2007, *122* (4), 1351–1408.

Borjas, George J. and Kirk B. Doran, "The Collapse of the Soviet Union and the Productivity of American Mathematicians," *Quarterly Journal of Economics*, 2012, *127* (3), 1143–1203.

Brunner, Robert F. and Sean D. Carr, *The Panic of 1907: Lessons Learned from the Market's Perfect Storm*, Hoboken: Whiley, 2007.

Cantoni, Davide, "The Economic Effects of the Protestant Reformation: Testing the Weber Hypothesis in the German Lands," *Journal of the European Economic Association*, 2015, *13* (4), 561–598.

Card, David, "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 2001, *69* (5), 1127–1160.

Carlino, Gerald and William R Kerr, "Agglomeration and Innovation," in "Handbook of Regional and Urban Economics," Vol. 5, Elsevier, 2015, chapter 6, pp. 349–404.

Carlson, Mark and Kris James Mitchener, "Branch Banking as a Device for Discipline: Competition and Bank Survivorship during the Great Depression," *Journal of Political Economy*, 2009, *117* (2), 165–210.

Carlson, W. Bernard, *Tesla: Inventor of the Electrical Age*, New Jersey: Princeton University Press, 2013.

Caselli, Francesco and Nicola Gennaioli, "Dynastic Management," *Economic Inquiry*, 2013, *51* (1), 971–996.

Celik, Murat Alp, "Does the Cream Always Rise to the Top? The Misallocation of Talent and Innovation," 2015. University of Pennsylvania Mimeo.

Chandler, Alfred D., *Inventing the Electronic Century: The Epic Story of the Consumer Electronics and Computer Industries*, Cambridge, Mass.,: Harvard University Press, 2001.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez, "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623.

Comin, Diego and Bart Hobijn, "An Exploration of Technology Diffusion," *American Economic Review*, December 2010, *100* (5), 2031–59.

Cook, Lisa D., "Inventing Social Capital: Evidence from African American inventors, 1843 – 1930," *Explorations in Economic History*, 2011, *48* (4), 507–518.

Davis, Steven, John Haltiwanger, and Scott Schuh, "Small Business and Job Creation: Dissecting the Myth and Reassessing the Facts," 1996.

Depalo, Domenico and Sabrina Di Addario, "Shedding Light on Inventors Returns to Patents," Working Paper, Bank of Italy November 2014.

Donaldson, Dave and Richard Hornbeck, "Railroads and American Economic Growth: A Market Access Approach," *Quarterly Journal of Economics*, 2016, *31* (2), 799–858.

Dorn, Harold, "The Possibility of Underenumeration in the 1920 Census. Part B," Technical Report, National Resources Committee. Population Statistics 1937.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr, "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *American Economic Review*, June 2010, *100* (3).

Feigenbaum, James J, "A New Old Measure of Intergenerational Mobility: Iowa 1915 to 1940," 2015. Harvard mimeo.

Field, Alexander J., "The Most Technologically Progressive Decade of the Century," *American Economic Review*, 2003, *93* (4), 1399–1413.

_ , "The Impact of the Second World War on US Productivity Growth," *Economic History Review*, 2008, *61* (3), 672–694.

Fishback, Price and Joseph A. Cullen, "Second World War spending and local economic activity in US counties, 193958," *Economic History Review*, 2013, *66* (4), 975–992.

Fisk, Catherine L., *Working Knowledge: Employee Innovation and the Rise of Corporate Intellectual Property, 1800-1930*, University of North Carolina Press, 2009.

Florida, Richard, "The Economic Geography of Talent," *Annals of the Association of American Geographers*, 2002, *92* (4), 743–755.

Galenson, David, "Creative Life Cycles: Three Myths," 2016. Becker Friedman Institute Working Papers Series # 2016–28.

Glaeser, Edward, *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*, Penguin, 2011.

Goldin, Claudia, "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century," *Journal of Economic History*, 1998, *58* (02), 345–374.

_ , "The Quiet Revolution That Transformed Women's Employment, Education, and Family," *American Economic Review*, 2006, *96* (2), 1–21.

_ **and Lawrence F. Katz**, "The Returns to Skill in the United States across the Twentieth Century," May 1999. National Bureau of Economic Research WP #7126.

_ **and** _ , "The Legacy of U.S. Educational Leadership: Notes on Distribution and Economic Growth in the 20th Century," *American Economic Review*, May 2001, *91* (2), 18–23.

Goldin, Claudia Dale and Lawrence F Katz, *The Race between Education and Technology*, Harvard University Press, 2009.

Gordon, Robert J, *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*, Princeton University Press, 2016.

Gorodnichenko, Yuriy and Gerard Roland, "Which Dimensions of Culture Matter for Long Run Growth?," *American Economic Review*, 2011, *101* (3), 492–498.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, "The Role of Social Capital in Financial Development," *American Economic Review*, June 2004, *94* (3), 526–556.

Gutmann, Myron P., Sara M. Pullum-Pinon, and Thomas W. Pullum, "Three Eras of Young Adult Home Leaving in Twentieth-Century America," *Journal of Social History*, 2002, *35* (3), 533–576.

Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg, "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools," 2001. NBER Working Paper # 8498.

_ **and Josh Lerner**, "The Financing of R&D and Innovation," Working Paper 15325, National Bureau of Economic Research September 2009.

Haltiwanger, John, "Job Creation and Firm Dynamics in the United States," in "Innovation Policy and the Economy, Volume 12" NBER Chapters, National Bureau of Economic Research, Inc, 07 2012, pp. 17–38.

Hellman, Thomas F. and Enrico C. Perotti, "The Circulation of Ideas in Firms and Markets," *Management Science*, 2011, *57* (10), 1813–1826.

Hopenhayn, Hugo, "Entry, Exit, and Firm Dynamics in Long Run Equilibrium," *Econometrica*, 1992, *60*, 1127–1150.

Hsieh, Chang-Tai and Enrico Moretti, "Why Do Cities Matter? Local Growth and Aggregate Growth," Working Paper 21154, National Bureau of Economic Research May 2015.

_ **, Erik Hurst, Charles I. Jones, and Peter J. Klenow**, "The Allocation Of Talent And U.S. Economic Growth," 2013. NBER Working Paper # 18693.

Hunt, Jennifer and Marjolaine Gauthier-Loiselle, "How Much Does Immigration Boost Innovation?," *American Economic Journal: Macroeconomics*, 2010, *2* (2), 31–56.

Hurst, Erik and Benjamin Wild Pugsley, "What Do Small Businesses Do?," *Brookings Papers on Economic Activity*, 2011, pp. 73–142. Fall.

Jaworski, Taylor, "World War II and the Industrialization of the American South," Working Paper, Queen's University 2015.

Jones, Benjamin F, "Age and Great Invention," *Review of Economics and Statistics*, 2010, *92* (1), 1–14.

Jones, Charles I and Jihee Kim, "A Schumpeterian Model of Top Income Inequality," 2014. NBER Working Paper # 20637.

_ **and Paul M Romer**, "The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital," *American Economic Journal: Macroeconomics*, 2010, *2* (1), 224–245.

Jovanovic, Boyan, "Selection and the Evolution of Industry," *Econometrica*, 1982, *50*, 649–670.

_ **and Peter Rousseau**, "General Purpose Technologies," in Philippe Aghion and Steven Durlauf, eds., *Handbook of Economic Growth*, 1 ed., Vol. 1, Part B, Elsevier, 2005, chapter 18, pp. 1181–1224.

Kaldor, Nicholas, *Capital Accumulation and Economic Growth*, Springer, 1961.

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

_ **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.

Khan, B Zorina, *The Democratization of Invention: Patents and Copyrights in American Economic Development, 1790-1920*, Cambridge University Press, 2009.

_ **and Kenneth L Sokoloff**, "Institutions and Democratic Invention in 19th-Century America: Evidence from" Great Inventors," 1790-1930," *American Economic Review*, 2004, *94* (2), 395–401.

Kim, Sukkoo, "Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U.S. Regional Manufacturing Structure, 18601987," *Quarterly Journal of Economics*, 1995, *110* (4), 881–908.

King, Robert G. and Ross Levine, "Finance and Growth: Schumpeter Might Be Right," *Quarterly Journal of Economics*, 1993, *108* (3), 717–737.

Klein, Alexander, "New State-Level Estimates of Personal Income in the United States, 1880-1910," in Christopher Hanes and Susan Wolcott, eds., *Research in Economic History (Volume 29)*, Emerald Group Publishing Limited, 2013, pp. 191–255.

Klette, Tor Jakob and Samuel Kortum, "Innovating Firms and Aggregate Innovation," *Journal of Political Economy*, 2004, *112* (5), 986–1018.

Kortum, Samuel and Josh Lerner, "Assessing The Contribution Of Venture Capital To Innovation," *Rand Journal of Economics*, 2000, *31* (4), 674–692.

Krugman, Paul, "History and Industry Location: The Case of the Manufacturing Belt," *American Economic Review Papers and Proceedings*, May 1991, *81* (2), 80–83.

Lamoreaux, Naomi R and Kenneth L Sokoloff, "Inventive activity and the market for technology in the United States, 1840-1920," 1999. National Bureau of Economic Research WP # 7107.

Lamoreaux, Naomi R., Margaret Levenstein, and Kenneth L. Sokoloff, "Financing Invention During the Second Industrial Revolution: Cleveland, Ohio, 1870-1920," 2004. NBER Working Paper # 10923.

Lerner, Josh, "150 Years of Patent Protection," *American Economic Review*, 2002, *92* (2), 221–225.

_ , *The Architecture of Innovation: The Economics of Creative Organizations*, Boston, MA: Harvard Business School Press, 2012.

Li, Guan-Cheng, Ronald Lai, Alexander D'Amour, David M. Doolin, Ye Sun, Vetle I. Torvik, Amy Z. Yu, and Lee Fleming, "Disambiguation and Co-authorship Networks of the U.S. Patent Inventor Database (1975–2010)," *Research Policy*, 2014, *43* (6), 941 – 955.

Logan, Trevon and John Parman, "The National Rise in Residential Segregation," 2015.

Long, Jason and Joseph Ferrie, "Intergenerational Occupational Mobility in Great Britain and the United States since 1850," *American Economic Review*, 2013, *103* (4), 1109–1137.

Lucas, Robert E., "On the Size Distribution of Business Firms," *Bell Journal of Economics*, 1978, *9* (2), 508–523.

Lucas, Robert E, "On the Mechanics of Economic Development," *Econometric Society Monographs*, 1998, *29*, 61–70.

_ , "Ideas and Growth," *Economica*, 2009, *76* (301), 1–19.

_ **and Benjamin Moll**, "Knowledge Growth and the Allocation of Time," *Journal of Political Economy*, 2014, *122* (1).

Martin, Robert F., "National Income and its Geographic Distribution," Technical Report, Conference Board Studies in Enterprise and Social Progress 1939.

Mokyr, Joel and Hans-Joachim Voth, "Understanding Growth in Europe, 17001870: Theory and Evidence," *Journal of Economic Sociology*, 2012, *13* (5), 57–102.

Moser, Petra, Alessandra Voena, and Fabian Waldinger, "German-Jewish Emigres and US Invention," *American Economic Review*, 2014, *104* (10), 3222–55.

Mowery, David C., "Military R&D and Innovation," in "Handbook of the Economics of Innovation," Elsevier, 2010, chapter 29, pp. 1219–1256.

Mowery, David C and Nathan Rosenberg, *Paths of innovation: Technological Change in 20th-century America*, Cambridge University Press, 1999.

Nanda, Ramana and Tom Nicholas, "Did Bank Distress Stifle Innovation During the Great Depression?," *Journal of Financial Economics*, 2009, *112* (2), 273–292.

National Academy of Sciences, *Federal Support of Basic Research in Institutions of Higher Learning*, Washington D.C.: National Research Council, 1964.

Nelson, Richard and Gavin Wright, "The Rise and Fall of American Technological Leadership: The Postwar Era in Historical Perspective," *Journal of Economic Literature*, 1992, *30* (4), 1931–64.

Nunn, Nathan and Leonard Wantchekon, "The Slave Trade and the Origins of Mistrust in Africa," *American Economic Review*, December 2011, *101* (7), 3221–52.

Payne, Carolyn S., *The Ames Project: Administering Classified Research as a Part of the Manhattan Project*, Iowa: Iowa State University PhD Thesis, 1992.

Perez-Gonzalez, Francisco, "Inherited Control and Firm Performance," *American Economic Review*, December 2006, *96* (5), 1559–1588.

Perla, Jesse and Christopher Tonetti, "Equilibrium Imitation and Growth," *Journal of Political Economy*, 2014, *122* (1).

Perlman, Elisabeth R., "Dense Enough To Be Brilliant: Patents, Urbanization, and Transportation in Nineteenth Century America," Working Paper, Boston University January 2016.

Petro, Diane, "The 1940 Census: Employment and Income," *Prologue Magazine*, 2016, *44* (1).

Philippon, Thomas and Ariell Reshef, "Wages and Human Capital in the U.S. Finance Industry: 19092006," *Quarterly Journal of Economics*, 2012, *12* (4), 15511609.

Piketty, Thomas and Emmanuel Saez, "Income Inequality in the United States, 19131998," *Quarterly Journal of Economics*, 2003, *118* (1), 1–41.

Rajan, Raghuram and Luigi Zingales, "Financial Dependence and Growth," *American Economic Review*, 1998, *88* (3), 559–586.

Romer, Paul Michael, "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5), S71–102.

Roy, A. D., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 1951, *3* (2), 135–146.

Sarada, Michael J. Andrews, and Nicolas L. Ziebarth, "The Demographics of Inventors in the Historical United States," Working Paper, University of Iowa 2016.

Satyanath, Shanker, Nico Voigtlaender, and Hans-Joachim Voth, "Bowling for Fascism: Social Capital and the Rise of the Nazi Party," *Journal of Political Economy*, 2016, *1* (1), 1–1.

Schmookler, Jacob, *Invention and Economic Growth*, Harvard University Press, 1966.

Sokoloff, Kenneth L., "Inventive Activity in Early Industrial America: Evidence From Patent Records, 1790–1846," *Journal of Economic History*, December 1988, *48* (04), 813–850.

Squicciarini, Mara P. and Nico Voigtlaender, "Human Capital and Industrialization: Evidence from the Age of Enlightenment," *Quarterly Journal of Economics*, 2015, *130* (4), 1825–1883.

Stephan, Paula, "The Endless Frontier: Reaping What Bush Sowed?," in "The Changing Frontier: Rethinking Science and Innovation Policy" NBER Chapters, National Bureau of Economic Research, Inc, February 2014, pp. 321–366.

Stern, Scott, "Do Scientists Pay to Be Scientists?," *Management Science*, 2004, *50* (6), 835–853.

Stokey, Nancy L, "Human Capital, Product Quality, and Growth," *Quarterly Journal of Economics*, 1991, *106* (2), 587–616.

Toivanen, Otto and Lotta Vaananen, "Returns to Inventors," *Review of Economics and Statistics*, 2012, *94* (4), 1173–1190.

Vandenbussche, Jérôme, Philippe Aghion, and Costas Meghir, "Growth, Distance to Frontier and Composition of Human Capital," *Journal of Economic Growth*, 2006, *11* (2), 97–127.

Wellerstein, Alex, "Patenting the Bomb: Nuclear Weapons, Intellectual Property, and Technological Control," *Isis*, 2008, *99* (1), 57–87.

Wright, Gavin, *Old South, New South: Revolutions in the Southern Economy since the Civil War*, New York: Basic Books, 1986.

_ , *Slavery and American Economic Development*, Baton Rouge: Louisiana State University Press, 2006.

Appendix

A Census Data Description

In this section we detail our Census data and the methods we use to prepare it for analysis. We use IPUMS complete-count data from the decennial Censuses in 1880, 1900, 1910, 1920, 1930, and 1940. We are limited to this set of years because the complete Census records are released only with a 72 year lag. In addition, the 1890 Census was largely destroyed in a fire in 1921. We start in 1880 because the Census for this year is systematic and contains a set of family related variables we use in our analysis. As the codebook for the 1880 Census writes:

> *"The 1880 Census is in several critical respects the first "modern" Census; it broke new ground in its completeness of coverage, accuracy of enumeration, and range and detail of questions. The supervision of enumerators shifted from a part-time responsibility of regular U.S. marshals to 150 Census Supervisors specifically appointed for the purpose. To make a full, accurate, and speedy enumeration practical, the size of enumeration subdistricts was reduced from a maximum of 30,000 inhabitants in 1870 to a maximum of 2,500 in 1880 while the number of enumerators was increased from 6,530 to 31,282. A variety of new questions were added that greatly enhance the value of the 1880 Census compared to earlier years. It was the first federal Census to inquire about marital status ... Equally important, a question on relationship to head of family was added, which makes it possible to distinguish kin from secondary individuals and allows construction of a wide variety of variables on family structure."*

The set of variables contained in the Census varies greatly over time. In addition, the micro-data from the 1940 Census is continuing to be populated with additional variables. Table A-1 summarizes the information available in our six decennial Census years.

A.1 Cleaning the Census Data

The Census provides a unique identifier for each individual in its records. These person identifiers, or "PIDs," are unique within Censuses, but are not constant across each Census year: an individual with PID 1 in 1880 is not the same individual with the PID 1 in 1900. We are unable therefore to create a panel dataset using our six Census datasets. Although the PIDs are unique in the vast majority of states and years, there are occasions in which the same individual shows up twice in the same year. Supposing data entry errors, we drop these duplicate PIDs.[21]

We take steps to impute missing data where it is easy to do so; for instance, we fill in missing age data by calculating the difference between the observed Census year and the individual's reported birth year.[22]

Before 1940, many variables are coded in strings rather than as categorical variables. For instance, sex variables can take on values "MALE," "FEMALE," "M," "F," and additional

[21] One individual in Georgia (PID 559409) has consistently non-sensical data, and is thus dropped from the 1900 Census.

[22] A number of individuals in 1900 have negative ages, or some ages above 130 years old. We drop these individuals from our analysis, supposing data entry errors.

Table A-1: Variables in the Censuses

Census Year	1880	1900	1910	1920	1930	1940
Age	✓	✓	✓	✓	✓	✓
Race	✓	✓	✓	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓
Marital status	✓	✓	✓	✓	✓	✓
Years married		✓				
Times married						✓
Birth place	✓	✓	✓	✓	✓	✓
Arrival year (immigrants)		✓	✓	✓	✓	✓
Mother's birth place	✓	✓	✓	✓	✓	✓
Father's birth place	✓	✓	✓	✓	✓	✓
Head of household	✓	✓	✓	✓	✓	✓
Family number				✓	✓	✓
Children born		✓				
Children living		✓				
Speak English				✓	✓	✓
Read				✓	✓	✓
Write				✓	✓	✓
Attended school				✓	✓	✓
Highest grade schooling						✓
Own home or rent				✓	✓	✓
Home mortgage				✓	✓	✓
Value of home					✓	✓
Radio					✓	
Occupation	✓			✓	✓	✓
Industry				✓	✓	✓
Class of worker				✓	✓	✓
Income						✓

Notes: This list focuses on those variables we use in our analysis and for which a large number of records have non-missing information. Home ownership variables are populated only for select group of individuals, and cannot be robustly matched to patent data.

codes indicating unknown. In many cases, these are easy to categorize into numeric categories. However, in certain instances, additional categorization must be done by hand. For instance, the race variable often mixes race and nationality. We therefore must make some assumptions as to what nationality corresponds to which race. For example, we classify those reporting that they are "Asian," "Chinese," "Filipino," "Japanese," "Korean," "Mongolian," or "Siamese" as one category "ASIAN."

There are two additional places where such categorization plays an important role in our analysis. First, the occupation variables contain over twenty thousand unique values in 1880, 1920, and 1930. Many of these unique values are the result of misspellings – such as "FARMR" in place of "FARMER" – or due to differences between British and American English, such as "LABOURER" instead of "LABORER." In order to reduce the dimension of the occupation data, we collapse the raw occupation data into three skill groups – low, medium, or high – and

three occupation categories: doctors, lawyers, and farmers. The three skill groups are mutually exclusive, and account for 79.4% of individuals with non-missing occupation data. Doctors and lawyers are all high skill, while farmers can be any skill level so long as they appear to work in farm related activities.

The skill classification proceeds as follows. We first classify individuals into low skill occupations using a string match. Low skill individuals perform routine jobs, sell their labor as hired hands, or work as servants or maids. For instance, if an individual reports an occupation containing the string "ASSIST," "CLERK," "LAUNDR," or "FARM," they are initially classified as low skill.[23] This method will classify those who say they are a shop clerk, blacksmith's assistant, laundry girl, or farm hand as low skill individuals. However, it will also classify farm supervisors and legal clerks as low skill. To correct for this, we next begin the classification of middle skill occupations.

Middle skill individuals are 1) those with particular specialties, such as carpenters or blacksmiths 2) those who perform middle management roles such as supervisors, or foremen, and 3) those in the clergy or law enforcement. We replace those coded with low skill occupations with a middle skill code if the individual both reports a string associated with a medium skill job, and is not an assistant or apprentice. Therefore, those who report that they are a "Foreman on a farm" will initially be classified as low skill because their occupation includes the string "farm," but will be updated to medium skill due to the string "Foreman." On the other hand, a "Blacksmith's apprentice" will not be updated to medium skill, as the string "apprentice" disqualifies classification as middle skill, even though the individual works with a blacksmith.

A similar routine is carried out for high skill classifications. Individuals are classified as high skill if 1) their occupation requires higher cognitive thought, such as a scientist, lawyer, or financier, 2) they are owners, directors, or upper management of ventures, 3) they are highly skilled manual workers, such as jewellers, goldsmiths, or silversmiths, or 4) they are public officials such as members of congress, or politicians. In addition, they must not be assistants, apprentices or hired hands. Once again, therefore, one who "Owns a farm" will initially be a low skill individual, but will be updated to high skill as a result of the string "own." Finally, students and those retired have missing occupation skills.

The occupation categories are more straightforward to classify. Doctors are those who are both high skill and who report an occupation string containing "DOCTOR," "MEDIC," "MD," "PHARM," "DENT," "PSYCH," or "OPTOM," among others. Thus pharmacists, dentists, psychiatrists, and optometrists will all be classified as doctors. Lawyers are high skill individuals with an occupation string containing "LAW," "JUDG," "ATTORN," and a number of legislator strings such as "SENATE." Thus attorneys, lawyers, judges, and legislators all count as lawyers by our broad definition. Finally, farmers are any individual who have an occupation string containing broad categories and common misspellings like "FARM," "FRM," "FIELD," and "CROP," or more narrow strings such as "HUSKER," "COTTON," "PICK," or "CHICKEN."

[23] This example is far from the full set of strings used to classify individuals. A full list of terms is available from the authors upon request. The set of strings was chosen by hand after examining the most common occupations.

The second major instance in which careful classification is required is in determining the birthplace of individuals. The majority of individuals report their place of birth at the state or country level. However, many give more specific answers such as the city, county, or (if abroad) principality of birth. In order to calculate robust migrant flows, it is necessary to aggregate these more refined answers to a state or country level. While there are too many small cities listed to code each person by hand, we make substantial progress in matching individuals to their state of birth: 86.7% of Census records with non-missing birthplace information are successfully matched.

We begin this refinement process by standardizing place names to be upper case, with no spaces. Next, we assign the largest cities in each state to its logical destination. For instance, "MOBILE" and "BIRMINGHAM" are assigned to Alabama. Note that individuals who were born in the much smaller town of Birmingham, Connecticut, for instance, will be incorrectly matched to an Alabama birthplace. While we are comfortable with this small error in most cases, it can prove quite difficult to address for city names that are large in multiple places. For instance, many people live in both Kansas City, Kansas, and Kansas City, Missouri. In such cases, we assume that the individual did not migrate across state lines if possible. That is, we assign an individual's birthplace to be Kansas if they currently live in Kansas and to be Missouri if they currently live in Missouri. These large cities that appear in multiple states are, as far as we can tell, only cities that straddle state lines. Therefore this conservative approach to migration appears to be justified - even if an individual moves from Kansas City, MO to Kansas City, KS, he will still be living in the same metropolitan area. Since classifying this individual as a migrant is thus misleading, we believe this no-migration error is justified.

A similar routine is carried out for international migrants as well. In particular, many German migrants provided specific states of birth, such as Bavaria, Württemberg, or Hamburg. Again, we aggregate these to the country level. We then divide the reported countries into nine regions: Western Europe, Scandinavia, Eastern Europe, Oceania, Africa, the Middle East, Latin America, Canada, and East Asia.[24]

With the cleaned birthplaces, we can then define the migration status of individuals. An individual is said to be an international migrant if they were born in any of the nine global regions defined above. An individual is defined as an interstate migrant if their birth state is different to their state of residence in the Census. Although we cannot calculate year-on-year migration flows, we can ask whether an individual has moved out of his state of birth, and has yet to move back.

A.2 Father Match

In order to study social mobility and the role of parental affluence, we attempt to form a robust link between individuals and their parents. The 1940 Census provides an explicit match between individuals and their spouse and parents, so long as they live in the same household. Using a household identifier and a variable giving an individual's person number in the household (e.g.

[24] A full list of classifications at both the state and country level is available from the authors upon request.

household head is 1, spouse may be 2, father may be 3, etc.), and the person number of their relevant family, we can explicitly ascertain the PID of the individual's family members.

Before 1940, however, we use our own algorithm to determine the PIDs of individuals' family members. First, we generate a family identifier, given by a unique surname, household ID, and city. We then use the relationship to household head variable recorded by the Census to determine the PID of individuals' family members. Specifically, we first limit ourselves to families with only one household head. Then we consider those individuals who report being either the household head or his/her spouse. We extract the PID of these household heads, and assign them to individuals reporting to be either the son or the daughter of the household head. This builds a crosswalk dataset - every son/daughter of a household head is matched to the PID of his/her mother and father.

To test the validity of this matching routine, we use our algorithm on the 1940 data, checking that we match the correct father to the correct child using the person number variables provided by the Census. Our match is nearly perfect: among individuals for whom the Census provides a person number match, we correctly ascertain the PIDs of an individual's father in over 99.99% of every state's population.

While this high match rate is encouraging, it does not suggest that our algorithm is perfect. Instead, it suggests that we are able to successfully replicate the Census' own algorithm for matching parents to children. The match is still subject to two important caveats. First, we can only match individuals to their parents if they live in the same household, and therefore our match may be more successful for younger children or poorer families. Second, we only match parents if at least one of them is the head of the household. In principle, one could attempt to match other members of the family by considering, for instance, the brother and nephew of the household head. However, matching based on non-nuclear family members introduces additional noise and incorrect matches, particularly in cases in which large families reside in the same household. We therefore avoid these kinds of matches in the construction of our final dataset.

B Patent Data Description

Our analysis is based on three main patent datasets we assembled using a mixture of hand entry and optical character recognition (OCR) techniques based on the original patent documents, and information from existing databases. These data are summarized as follows:

1. **Patent Dataset A**. 6,675,311 patents. Consists of close to the universe of patents granted by the USPTO between 1836 and 2004 covering the location of the first named inventor listed on the original patent documents down to the city level.
2. **Patent Dataset B**. 60,594 patents. Consists of the universe of patents granted by the USPTO for the years 1880, 1900 and 1910 covering both the name and location of the first named inventor down to the city level.

3. **Patent Dataset C**. 5,739,225 patents. Consists of the universe of patents granted by the USPTO including the name of all inventors and assignees from 1920 to 2006.

Note that these datasets contain overlapping and sometimes complementary information—for instance, **Patent Dataset A** contains the location of inventors but not their names whereas **Patent Dataset C** contains names but not locations.

We obtained address information for the first inventor from the original patent documents using OCR and hand entry (**Patent Dataset A**). This dataset contains US patents that were granted to both US citizens and individuals living abroad, but in our analysis we obviously limit ourselves to patents granted to individuals and entities based in the United States. In 1880 94% and in 1940 86% of patents were granted to inventors located in the US.

For the years 1880, 1900, and 1910 (**Patent Dataset B**), we extracted the name and address of the first inventor listed on the patent document, under the assumption that this individual was the principal inventor of the art. Single inventors were the norm during this time period. In 1880, 1900 and 1910 approximately 92%, 90% and 91% of patents were granted to a single inventor respectively.

From 1920 through 2006, we retrieved the name of every inventor listed on every patent each year using data supplied to us by the European Patent Office (**Patent Dataset C**). Additionally we created a panel dataset with an inventor identifier through disambiguating the inventor data using the algorithm of Li et al. (2014).

B.1 Citation Adjustment

Our data includes the number of citations each patent receives from patents granted from September 1947, when the USPTO began to note citation data in a systematic way, to February 2008. Thus, we have the full universe of citations received by patents granted during this time period. Citations start in 1947 because a USPTO Notice was issued on December 19th, 1946, instructing examiners to add citations in the published format of the patent, a practice that was incorporated into the *Manual of Patenting Examining Procedure* (paragraph 1302.12).

For patents granted before 1947, the noted citation count is left censored: a patent granted in 1940 will only have citations from patents granted after 1947, but will not have citations from patents between 1941 and 1946. This artificially deflates the number of citations received by patents before 1947, confounding attempts to use citations as an objective measure of a patent's quality. Furthermore, aggregate citation trends may weaken the link between raw citation counts and patent quality. For instance, if patents granted in 1960 cite an average of 5 prior patents, but those granted in 1990 cite 20 patents, one might expect the average citation received from a 1960 patent to be more indicative of a high quality innovation than a citation received in 1990. We therefore adjust the number of citations received by each patent following the quasi-structural approach laid out in Hall et al. (2001).

This approach relies on two critical assumptions. First, we assume that the citation process is *stationary*. That is, we assume that the evolution of citation shares does not change over time: a patent will on average receive a share $\pi_{k\tau}$ of its citations τ years after it is granted, regardless

of the grant year. This allows us to project back our adjustment factors to patents filed before the citation data began in 1947. Second, we assume *proportionality*. That is, we assume that the shape of the citation evolution does not depend on the total number of citations received so that highly cited patents are more highly cited at all lags. This allows the application of the same adjustment factor to every patent in our data granted in a given period and belonging to a given patent class.

The adjustment proceeds as follows. We start with the full patent citation network data, keeping only those patents granted in the United States. Let C_{kst} be the total number of citations to patents in year s and technology category k coming from patents in year t.[25] Further, define P_{ks} to be the total number of citations received by patents granted in year s in technological category k. One can then define $\pi_{kst} = C_{kst}/P_{ks}$ to be the average share of citations received by patents in class k in year s from patents granted in year t We assume that π_{kst} is some multiplicatively separable function of grant year, patent category, and a citation lag. That is, we can write

$$\log[\pi_{kst}] = \alpha_0 + \alpha_s + \alpha_t + \alpha_k + f_k(L) \tag{A-1}$$

for $L = t - s$ the lag between cited and citing patent grant years, and $f_k(\cdot)$ some category-specific function of these lags. For our purposes, we define $f_k(L) = \tilde{\gamma}_{k,L}$. We may then estimate equation A-1 using OLS to recover estimates of $\alpha_0, \alpha_s, \alpha_t, \alpha_k$, and $\tilde{\gamma}_{k,L}$ for each value of s, t, k and L in our data.[26] Taking exponentials of equation A-1 yields

$$C_{kst}/P_{ks} = e^{\alpha_0} e^{\alpha_s} e^{\alpha_t} e^{\alpha_k} e^{\tilde{\gamma}_{k,(t-s)}} \tag{A-2}$$

This formulation allows us to standardize citation counts over time and across categories. Specifically, in order to adjust for patent class, cited year, and citing year effects, we weight each citation from a patent in year t to a patent in class k in year s by $\exp(-\hat{\alpha}_k - \hat{\alpha}_s - \hat{\alpha}_t)$. Each patent's citation counts are therefore reflective of the patent's quality relative to the average patent in some base year and category.[27]

While this procedure accounts for aggregate differences across patent classes and grant years, it does not yet correct for bias arising from the left truncation of citation records. To build intuition for the truncation correction, consider an example in which each of the estimated α coefficients were 0: the only bias in our citation data arises from the lag. In that case, the assumptions of proportionality and stationarity suggest a natural adjustment factor for a patent granted L years before the 1947 cutoff. Define $G_k(L)$ to be the CDF of the lag distribution: the share of an average patent's citations received within the first L years after its grant. The

[25] For the purposes of the adjustment, we use technological categories as defined by the NBER patent data. For a detailed description of these data, see Hall et al. (2001).

[26] It is rare for a patent to receive citations more than 30 years after its initial grant date, and thus we top-code the citation lag L to have a maximum value of 30. That is, we define $L = \min\{t - s, 30\}$.

[27] For our purposes, we choose each patent citation to be relative to a patent in the "Other" category granted in 1975, receiving citations from patents also granted in 1975. Mechanically, this corresponds to setting the omitted categories in estimation of equation A-1 to be $k =$ "Other", $s = t = 1975$.

adjustment factor is then given by

$$\sigma_{k,L} = \frac{1}{1 - G_k(L)}$$

We would then predict that a patent in category k granted in year $1947 - L$ and receiving c citations from patents granted after 1947 would have received $\sigma_{k,L}c$ citations had the USPTO kept track of citations before 1947.[28]

In order to incorporate the year and category fixed effects into this truncation adjustment framework, one must establish a notion of the CDF of the lag distribution conditional on year and category effects. To do so, we interpret the $\exp(\tilde{\gamma}_{k,L})$'s as weights for each patent in the citation data. For instance, if the estimated $\exp(\tilde{\gamma}_{k,L=2})$ is 2, then an average patent is twice as likely to receive a citation after 1 year than in the year of patent grant, conditional on year and category effects. To construct the CDF of citations by lag conditional on year and class effects, we can sum our estimates of $\exp(\tilde{\gamma}_{k,L})$, normalizing the estimated coefficients so that they sum to 1. This gives us our estimate of $G_k(L)$:

$$\hat{G}_k(L) = \frac{\sum_{l=1}^{L} \exp(\tilde{\gamma}_{k,l})}{\sum_{l=1}^{30} \exp(\tilde{\gamma}_{k,l})} \tag{A-3}$$

We can then calculate our truncation adjustment factor as before[29]

$$\hat{\sigma}_{k,L} = \frac{1}{1 - \hat{G}_k(L)}. \tag{A-4}$$

To summarize, the citation adjustment proceeds in four steps:

1. Estimate equation A-1 using OLS to recover $\alpha_0, \alpha_k, \alpha_t, \alpha_s$ and $\gamma_{k,L}$.

2. For each citation made from a patent p' granted in year t to a patent p in class k granted in year s is weighted by

   $$\omega_{k,s,t} = e^{-\alpha_k - \alpha_t - \alpha_s}$$

   Define, for each cited patent p, the year- and category-adjusted citation count c to be the sum of the $\omega_{k,s,t}$ it received.

3. Calculate $\hat{G}_k(L)$ according to equation A-3

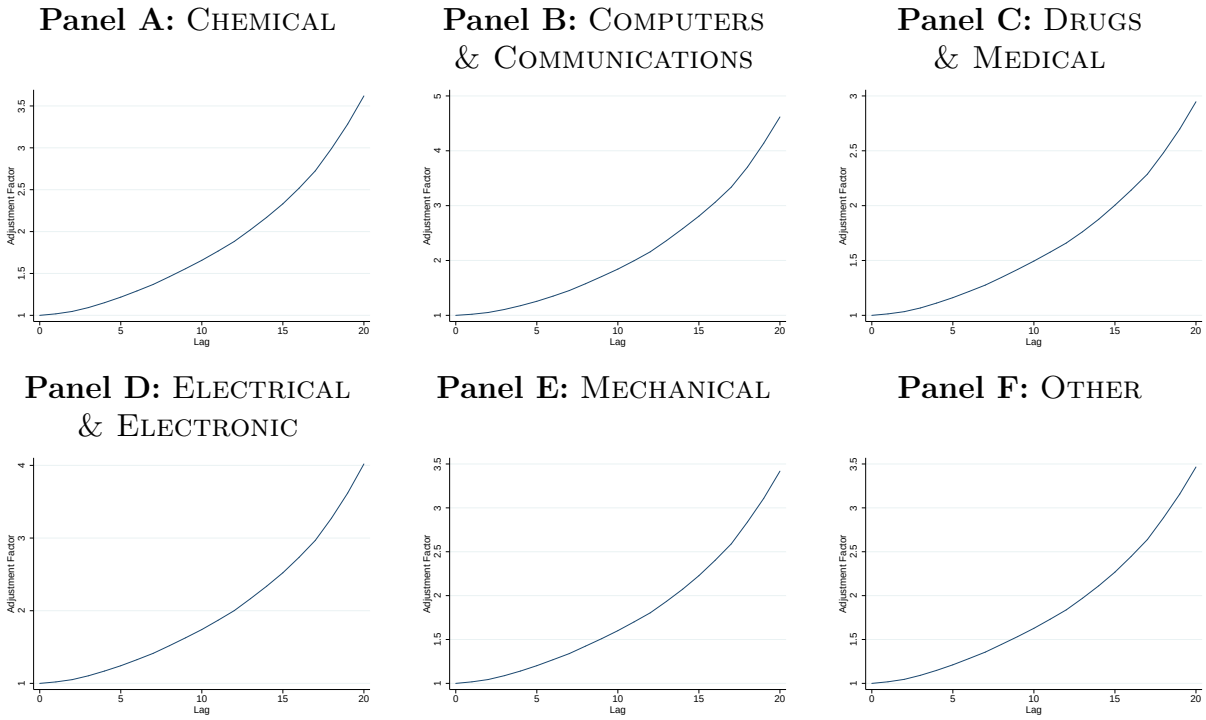
4. Using $\hat{G}_k(L)$, calculate the truncation adjustment factor $\hat{\sigma}_{k,L}$ according to A-4. Finally, define a patent p's adjusted citation count to be $\tilde{c} = c \cdot \sigma_{k,L}$ if p is in class k and was granted L years before 1947.

[28] Ignoring year and category effects and adjusting citations in this way does not significantly change the results presented in the main body of the paper.

[29] Note that we only calculate the truncation adjustment up to $L = 20$, despite estimating $\gamma_{k,L}$ for L as large as 30. This is to bound $\hat{G}_k(L)$ away from 1, so that we do not divide by 0 in the adjustment. For L larger than 20, we apply the adjustment factor for $L = 20$.

Figure A-1 plots the adjustment factors for truncation years for each of the six NBER patent categories. The multiplicative adjustment factors range from 1 to almost 5, and vary by NBER category. Meanwhile, Figure A-2 plots the distribution of log citations and the evolution of the average citation counts according to three adjustment regimes: no adjustment, full adjustment, and an adjustment in which we do not correct for truncation at 1947. We see that the fully adjusted citation counts have a much flatter time series relative to the unadjusted citation counts. This is by design: the purpose of the adjustment is to remove the aggregate fluctuations which do not accurately measure the relative quality of patents.

Figure A-1: Citation Adjustment: Adjustment Factors by Years Truncated



B.2 Patent Classes and Matching Patents to Sectors

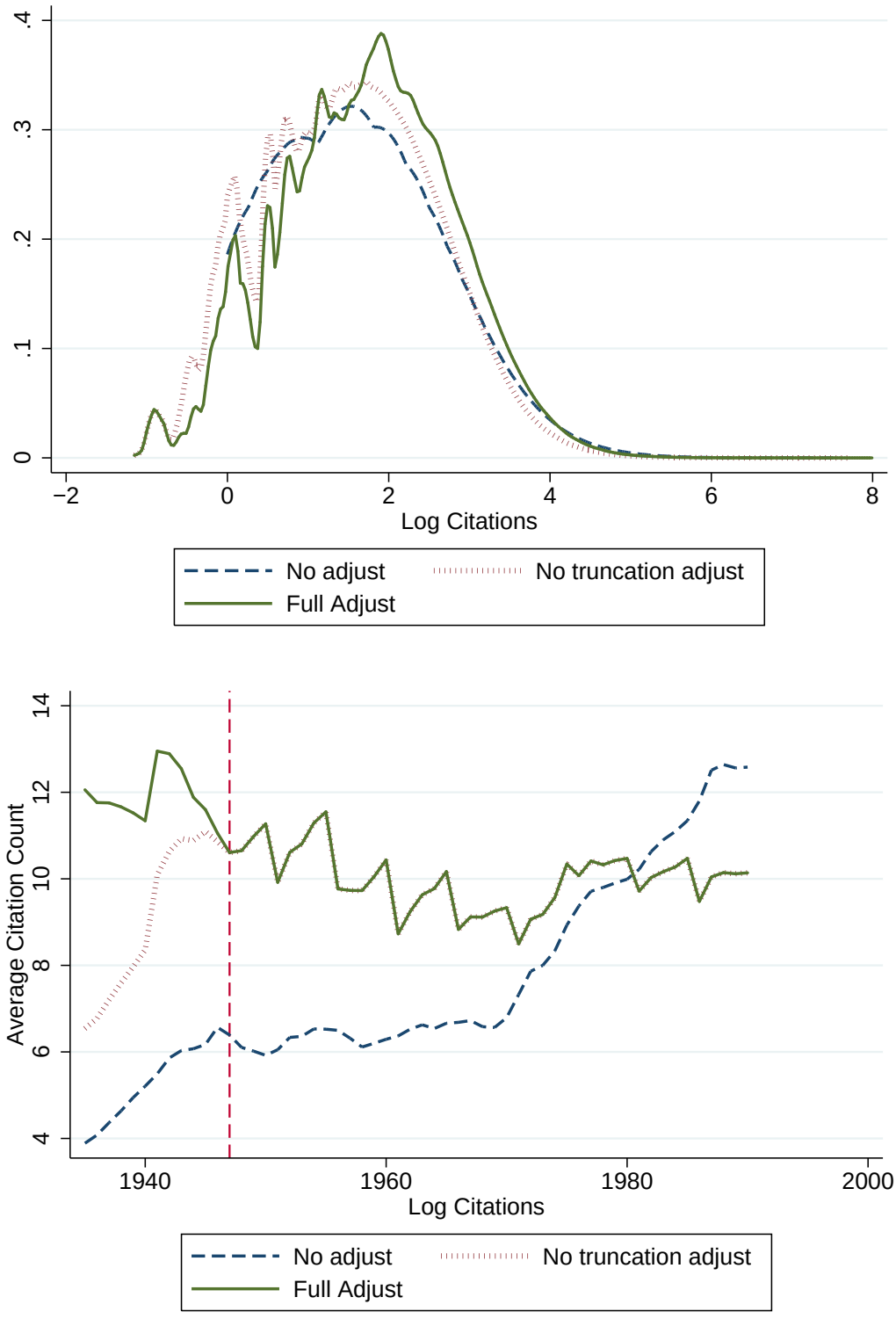
We obtain the main USPTO patent class for each patent and the NBER patent aggregations of these classes.[30] We match patents to sectors using the USPTO technology class of the patent.

We also use value added and full-time employment data by sector from 1947 through 1986, before the SIC was revised in 1987. These data are matched to data provided by Bill Kerr containing the fraction of patents in each class which were manufactured and used by every 3-digit SIC code (Kerr, 2008). We first aggregate these SIC codes into the same categories contained in the industry value added data from the BEA. Table A-2 shows this aggregation.

Once we know the fraction of patents in each class that are accounted for by the BEA-provided industries, we assign each class to an industry. We say a patent class c is affiliated with industry

[30] The USPTO occasionally reclassifies patents based on the emergence of new technologies. Throughout the paper, we use the 2006 classification.

Figure A-2: Citation Adjustment: Citation Distributions and Average Citation Counts over Time



j if industry j manufactures the highest share of patents in class c. We can then calculate the total number of patents for each BEA industry.

Table A-2: 2- and 3-Digit SIC Codes and Associated Industry Titles

01-02	Farms	23	Apparel and other textile products	61	Credit agencies other than banks
07-09	Agricultural services, forestry, and fishing	26	Paper and allied products	62	Security and commodity brokers
10	Metal mining	27	Printing and publishing	63	Insurance carriers
11-12	Coal mining	28	Chemicals and allied products	64	Insurance agents, brokers, and service
13	Oil and gas extraction	29	Petroleum and coal products	65-66	Real estate
14	Nonmetallic minerals, except fuels	30	Rubber and miscellaneous plastics products	67	Holding and other investment offices
15-17	Construction	31	Leather and leather products	70	Hotels and other lodging places
24	Lumber and wood products	40	Railroad transportation	72	Personal services
25	Furniture and fixtures	41	Local and interurban passenger transit	73	Business services
32	Stone, clay, and glass products	42	Trucking and warehousing	75	Auto repair, services, and parking
33	Primary metal industries	44	Water transportation	76	Miscellaneous repair services
34	Fabricated metal products	45	Transportation by air	78	Motion pictures
35	Machinery, except electrical	46	Pipelines, except natural gas	79	Amusement and recreation services
36	Electric and electronic equipment	47	Transportation services	80	Health services
371	Motor vehicles and equipment	48	Communications	81	Legal services
372-379	Other transportation equipment	481,482,489	Telephone and telegraph	82	Educational services
38	Instruments and related products	483	Radio and television	83	Social services
39	Miscellaneous manufacturing industries	49	Electric, gas, and sanitary services	86	Membership organizations
20	Food and kindred products	50-51	Wholesale trade	84,89	Miscellaneous professional services
21	Tobacco products	52-59	Retail trade	88	Private households
22	Textile mill products	60	Banking	43,91-97	Government

Notes: Sector codes retrieved from the documentation of value added statistics provided by the BEA: https://www.bea.gov/industry/xls/GDPbyInd_VA_SIC.xls on August 10, 2016.

C Merging Patent and Census Data

C.1 Data Preparation

We first standardize the names and places listed in the patent and Census data. We begin by ensuring that all names are fully capitalized, and remove all special characters (e.g. "."s) from the names.[31] In addition, we remove suffixes such as "JR," "Senior," and "III" from listed names. We next parse the names into different words. The surname is taken to be the last word of an individual's name, while an individual's first name is taken to be the first word. The first letter of the second word of an individual's name is taken to be their initial, so long as the name contains at least three words. For example, a name originally recorded as "Thomas Alva Edison," will return three pieces of information: the surname "EDISON," a first name "THOMAS," and an initial "A." Note that this procedure implies that those with multiple words in their surname are constrained to have a first name, single middle initial, and one-word surname. For example, Robert Van de Graaff, inventor of the Van de Graaff generator (a machine that generates static electricity), is eventually listed as "ROBERT V GRAAFF."[32]

Locations are likewise standardized. First, we capitalize all place names listed in the Census and on the patent records. We then ensure that the spelling of common pieces of the place name are constant across the two data sources. For instance, we enforce that the word "SAINT," as in "SAINT LOUIS," are all listed as "ST." In addition, we remove superfluous words such as "WARD," "DISTRICT" or "CITY;" for instance, "NEW YORK CITY" becomes simply "NEW YORK." Finally, we standardize a number of common place names by hand; for example, we impose that the five boroughs of New York City – Brooklyn, Manhattan, Queens, the Bronx, and Staten Island – are all coded as "NEW YORK."

C.2 Merging the Data

We next merge the patent data to the decennial Censuses. To do so, we first insist that records in the Census have the same first name, last name, county, and state as the inventor listed on the patent. In addition, the patent in question must have been granted in the same year as the Census was conducted. While we make a strong effort to clean our data before matching, there remain some cases that do not match even on these basic criteria. Of all patent-inventor instances in the patent data, 70.8% find a match in the census based on these criteria. The remaining 30% may not match either because their names were incorrectly entered in either the Census or patent data, or because they may have moved across state lines between the time the Census was conducted and the patent was granted. Predictably, this problem is particularly pronounced immediately following the end of the First World War: we match just 61.1% of patent-inventor observations in 1920.

[31] We drop Census records with first or last names longer than 40 characters. We do this because we suppose that such long names arise from input errors.

[32] The 1910 Census provides multiple name fields. We take the most well-populated field, and fill in missing values with the names contained in the other name variables. In the vast majority of cases, the names provided in the two variables are identical.

Naturally, there may remain multiple inventor matches if, for example, there are multiple Thomas Edison's living in Middlesex county, New Jersey in 1900. Indeed, 44.3% of our initially matched patents have multiple candidate inventors. We then refine the match further based on other information in the patent documents. First, for each patent, we look to see if one of the candidate inventors in the Census data has the same middle initial as listed on the patent document. If so, we only keep those that match. This removes 8.2% of our multiple matches.

At this stage, the multiplicity concern arises from the possibility of multiple Thomas A Edison's living in Middlesex, NJ in 1900. Thus we refine to a more granular geography. Our second refinement asks whether there are any candidate inventors living in the same city or township as was listed on the patent document. We only keep those who match on this criterion, so long as the patent has at least one matched candidate. Thus we limit ourselves to Thomas A Edison's living in Menlo Park, Middlesex County, New Jersey in 1900. The refinement based on cities removes 7.3% of the duplicate inventors, who survived the refinement based on middle initials.

Multiplicity can still persist, however, and may be particularly common within family units if a son is named after his father. At this stage, both John J Smith Jr and his father John J Smith Sr, living in the same household, would be matched to the same patent. To combat this, we finally refine the match based on an age criterion. For a given patent, we ask if there is a candidate inventor between the ages of 15 and 85 in the Census. If so, we keep that candidate inventor, and discard the candidate children under 15 years old and the elderly above 85. This age refinement removes 5.5% of the multiple inventors present at this stage. We next repeat this refinement with a sharper age criterion, keeping those between 18 and 65 years of age if such a match exists.

Finally, if there are still multiple matches for a given patent, then we exclude the individual and patent from the sample altogether and they are counted neither as an inventor nor as a non-inventor. This is done to be conservative about our match rate, and to avoid inducing spurious correlations from incorrect matches. As a robustness check, we also run our analysis on a sample in which we keep a random inventor for each patent with a multiple match. The results are qualitatively similar, and are available from the authors upon request.[33]

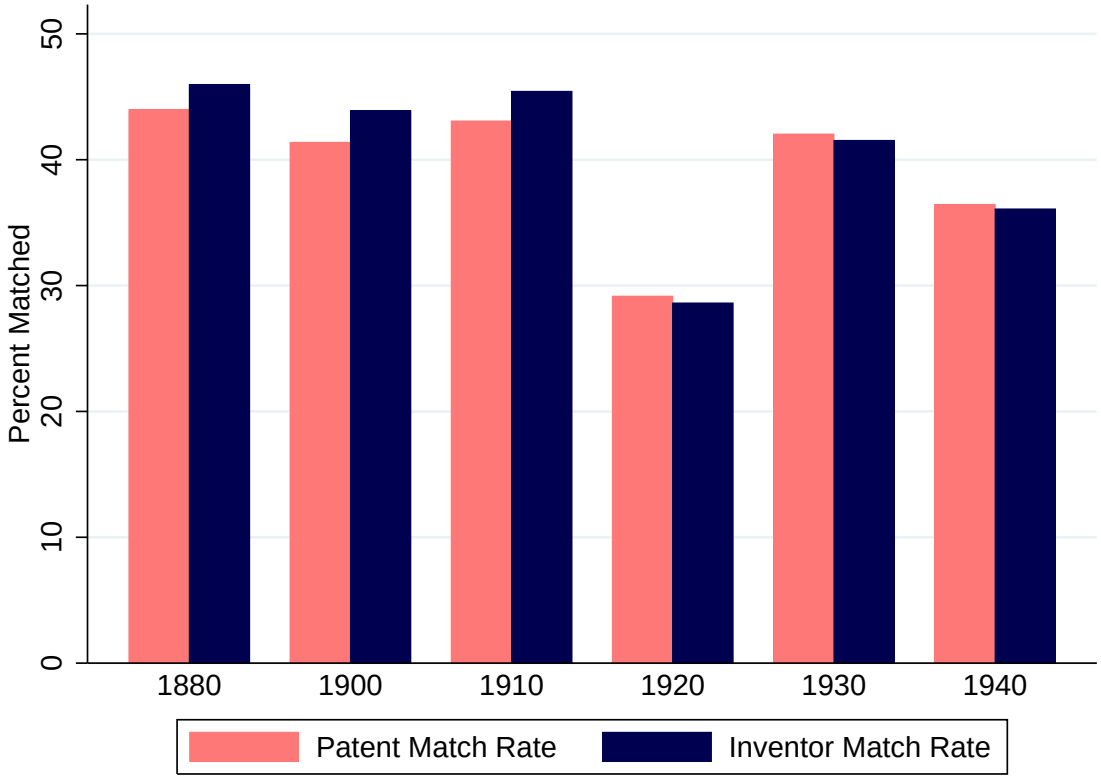
We then merge into our data every patent ever granted to each inventor we have successfully matched. Thus, while we only match inventors to the Census if they are granted a patent in a decennial Census year, our matched data contain patents granted to inventors in every year from 1920 through 2006, as well as patents granted to inventors in 1880, 1900, and 1910.

[33]There is one exception to this similarity in headline results. Table 14 in the main body of the paper show a weak correlation between the probability that an individual becomes an inventor and the inventor status of one's father. When we keep a random matched inventor, this correlation becomes large, significant, and positive. This change can be best understood with an example. Suppose that John J Smith Jr is 25 years old and cohabits with his father, the 50 year old John J Smith Sr. The younger John Smith is an inventor of two patents, but his father is not. Because they are both between the age of 15 and 85 with the same first name, last name and middle initial, and live in the same city, we must keep a random John J Smith for each of the two patents. For the first patent, suppose we kept the younger John Smith, while the other patent is assigned to his father. This generates a spurious correlation between an individual's inventor status and that of his father: even though John Smith Sr was never granted a patent, it appears as though he was in our data. These family relationships might be a persistent source of multiplicity, and thus likely drives this particular difference in our results. We therefore favor the more robust results presented in the main body of the paper.

C.3 Match Success

Figure A-3 shows the match rate by decennial Census year. Our most successful match year is 1880, in which we match 44.0% of all patents and 46.0% of all inventors in the patent data to a unique individual in the Census. While the match rate hovers around 40% for most years, in 1920 we match 29.2% of US patents and 28.7% of inventors. The relatively low match rate we observe for 1920 may simply be idiosyncratic. As we point out in Section 2 the 1920 Census was conducted in the winter which led to some level of underenumeration, though not on a scale to bias our results. The effects of World War I demobilization on the movement of ex-servicemen in the population who were also inventors may also have had an effect.

Figure A-3: Match Rate by Decennial Census Year - All States



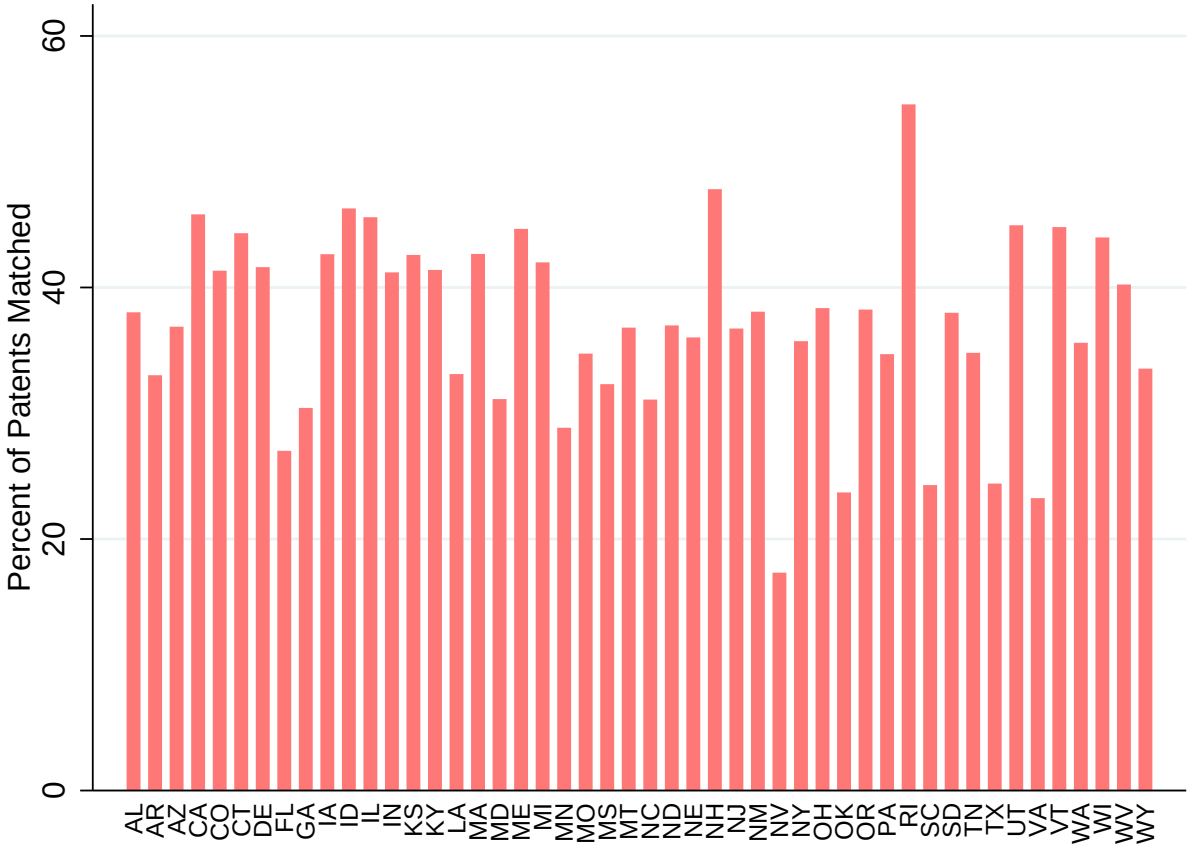
Notes: Figure shows the percent of inventors (solid red bars) and patents (dashed blue bars) present in the patent data who successfully match to the Census data by year. All states are aggregated together to produce this plot.

Figure A-4 shows the match rate by state, pooling all years together. Panel A shows the match rate for patents, while Panel B shows the match rate for inventors. There is heterogeneity in the match success across states. While Rhode Island enjoys a successful match rate of 54.5% for patents and 55.9% for inventors, we only match 17.3% of patents and 21.0% of inventors in Nevada. Part of this difference may be attributable to the changing county (and even state) boundaries in the early part of our sample, as frontier states saw rapid increases in population.

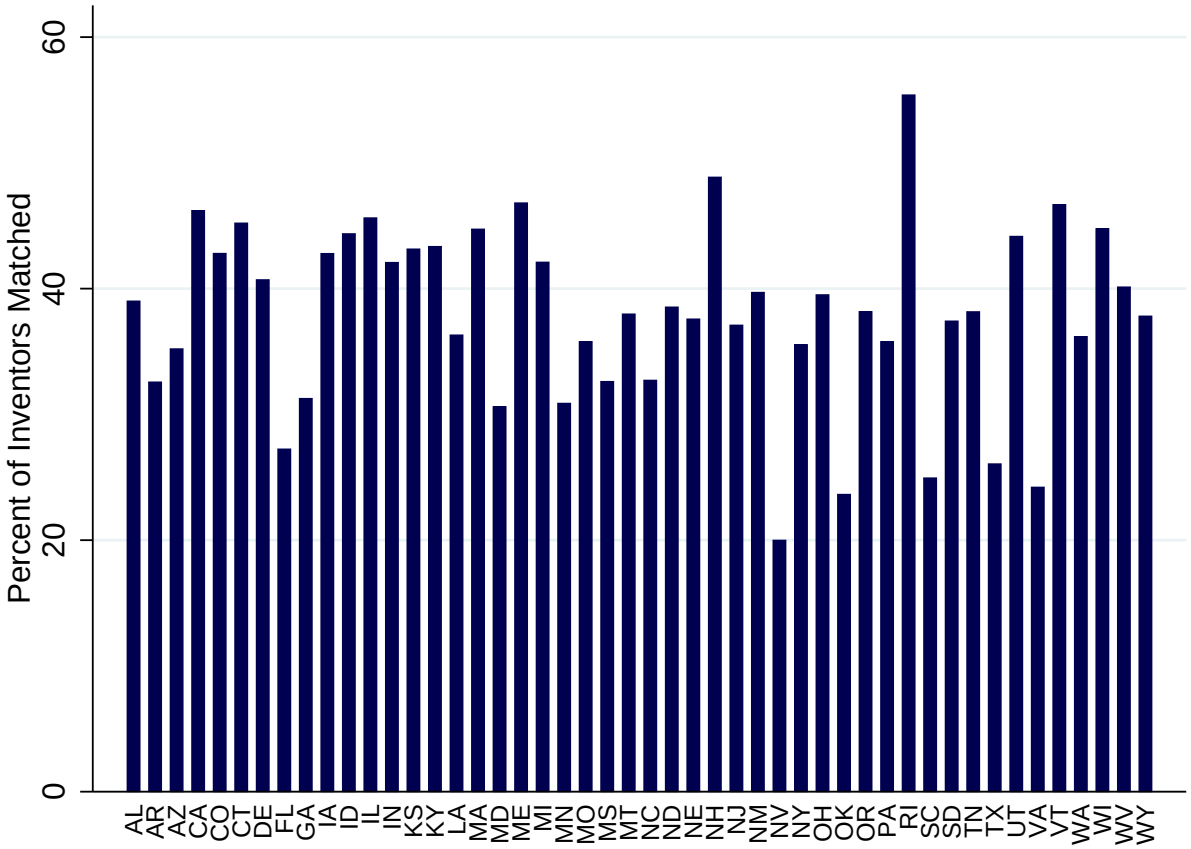
A potential concern with our results is that they may be driven by systematic match errors, rather than the unique characteristics for inventors. For instance, if name disambiguation proves especially difficult for common names, our match success will reflect only rare names, which may disproportionately represent a particular race, sex, or age profile. Alternatively, if data input

Figure A-4: Match Rate by State - All Years

Panel A: Patent Match Rate



Panel B: Inventor Match Rate



Notes: Figure shows the percent of inventors (Panel B: dark blue bars) and patents (Panel A: bright red bars) present in the patent data who successfully match to the Census data by the state listed on the patent application. All years are aggregated together to produce this plot.

errors are common within the Census, especially uncommon or foreign-sounding names may be matched at a lower rate than traditional American names.

To test for any biases, we ask whether the patents and inventors that are successfully matched to the Census are observably different from those that are not matched. For this exercise, we consider the universe of patents granted in each of our decennial Census years in the 48 mainland states used in our analysis. We then generate a binary variable equal to 1 if that patent and inventor were successfully matched to our Census data and survived the refinements detailed above. We then regress this indicator on characteristics of the patent and inventor. One might be particularly concerned that we have more success matching common, traditional American names, or particularly prominent inventors. We measure inventor and patent prominence by the number of citations received between 1947 and 2008. We use two indicators for the rarity of the inventor's name. First, we construct the share of the population with each first name using Census data. Second, we include the string length of the inventor's surname.

The results of this regression exercise are displayed in Table A-3. Column 1 show that those with longer names are less likely to be matched, and those with common first names are slightly more likely to be matched. However, we do not disproportionately match patents or inventors of a higher quality. These effects are small: increasing name prevalence by 100 (approximately 1 standard deviation) is associated with just a 1.3 percentage point increase in the match rate, roughly 3% of its mean. Meanwhile, a one standard deviation (1.75) increase in an individual's surname length reduces the match rate by 0.6 percentage points.

To test for disproportionate matching of particular population groups, we again use the Census to construct our variables of interest at the first name level. We thus include the percent of individuals with the inventor's first name who were international migrants and the average age of those with the inventor's name in the Census as dependent variables in columns 2 through 4. Column 2 includes no fixed effects. Column 2 would suggest that, we are less likely to match those with names commonly associated with international migrants, while there is hardly any bias in our age match. This implies that, although we find little difference in the *international* mobility between inventors and non-inventors using our matched data, it remains possible that inventors were more likely to be international migrants. Indeed, this observation might partially explain the difference in international migration rates between the most and least inventive states, as shown in Table 1. A one standard deviation increase in the percent of people with the inventor's first name who are international migrants (13 percentage points) is associated with a 2.64 percentage point reduction in the patent match rate.

Columns 3 and 4 control for state-year fixed effects in our selection regressions, while column 4 additionally controls for the patent's technology class, and the technology class of the inventor's first granted patent. Column 3 most closely matches that of our previous regression analyses, which use matched data for just one year of the census, and include state fixed effects (see, for example, Table 14). The inclusion of these fixed effects does not significantly change the patterns shown in columns 1 and 2.

Another concern would be that the substantial heterogeneity in state match rates is systematically correlated with key state variables of interest. Although we do not use the matched data

Table A-3: Selection into Matching: Regressions on Probability of Match

Panel A: Patent Match Probability				
	(1)	(2)	(3)	(4)
Surname Length	-0.346**	-0.219	-0.285*	-0.302**
	(0.154)	(0.154)	(0.153)	(0.149)
Name prevalence (per 10,000 people)	0.013***	0.013***	0.008***	0.008***
	(0.003)	(0.003)	(0.003)	(0.003)
Citations between 1947-2008	-0.011	-0.010	0.005	-0.004
	(0.011)	(0.011)	(0.007)	(0.007)
Percent First Name Int'l Migrant		-0.203***	-0.201***	-0.201***
		(0.025)	(0.018)	(0.017)
Average Age with First Name		-0.085*	-0.000	-0.012
		(0.051)	(0.036)	(0.036)
Fixed Effects	None	None	State× Year	State× Year Tech Class
Observations	175093	175093	175093	175093
Mean of Dep. Var.	38.65	38.65	38.65	38.65

Panel B: Inventor Match Probability				
	(1)	(2)	(3)	(4)
Surname Length	-0.675***	-0.535***	-0.592***	-0.598***
	(0.142)	(0.141)	(0.140)	(0.139)
Name prevalence (per 10,000 people)	0.022***	0.022***	0.015***	0.015***
	(0.002)	(0.003)	(0.002)	(0.002)
Citations between 1947-2008	-0.028***	-0.028***	-0.008	-0.009
	(0.010)	(0.010)	(0.007)	(0.007)
Percent First Name Int'l Migrant		-0.222***	-0.218***	-0.218***
		(0.024)	(0.016)	(0.015)
Average Age with First Name		-0.124***	0.014	0.000
		(0.047)	(0.030)	(0.029)
Fixed Effects	None	None	State× Year	State× Year Tech Class
Observations	122095	122095	122095	122095
Mean of Dep. Var.	39.12	39.12	39.12	39.12

Notes: Dependent variable is an indicator for an observation being matched to the census data, multiplied by 100 for legibility. White heteroskedasticity-robust standard errors reported in parentheses. *, **, and *** represent coefficient statistically different from 0 at the 10, 5, and 1% level respectively. Inventor technology class defined to be the technology class of his/her first patent.

for our state-level analysis, it is worth considering this claim. Table A-4 reports estimates from an OLS regression of a state's match rate on its observable characteristics. We see that none of our regional variables predict a state's match rate. Indeed, the full set of variables only explains approximately 12% of the variation in state match rates, as measured by the regression's R^2.

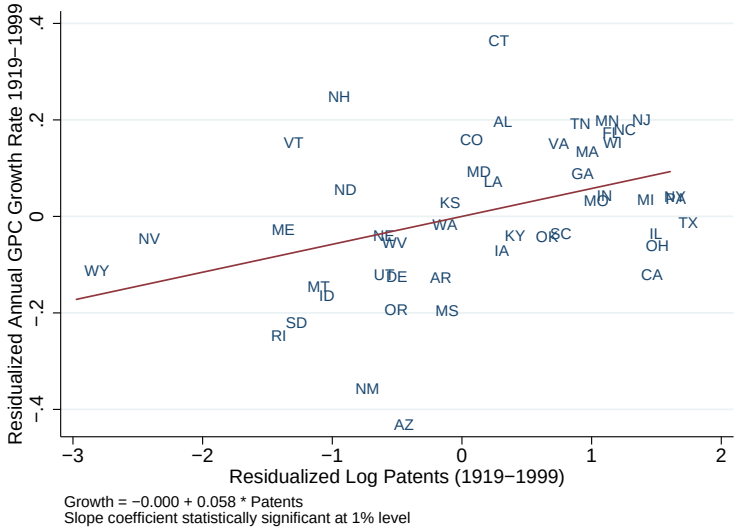
Table A-4: Selection into Matching: State Match Rate Regressions

	Panel A: Patent Match		**Panel B:** Inventor Match	
	(1)	(2)	(3)	(4)
90-10 Wage Income Ratio	-0.542	-0.364	-0.541	-0.321
	(0.627)	(0.719)	(0.620)	(0.709)
Average Income	-0.003	-0.006	-0.004	-0.007
	(0.012)	(0.017)	(0.012)	(0.017)
Population Density	0.030	0.021	0.034	0.022
	(0.034)	(0.040)	(0.034)	(0.039)
Deposits per capita	0.006	0.005	0.004	0.003
	(0.011)	(0.011)	(0.010)	(0.011)
Average outbound transport cost	0.101	0.134	0.055	0.101
	(0.281)	(0.311)	(0.278)	(0.307)
Percent of residents with college degree	0.083	0.107	0.187	0.209
	(0.721)	(0.773)	(0.713)	(0.762)
Percent employed in manufacturing		-0.132		-0.183
		(0.418)		(0.412)
Percent employed in agriculture		-0.154		-0.196
		(0.242)		(0.239)
Observations	47	47	47	47
R-squared	0.122	0.131	0.114	0.130
Mean of Dep. Var.	37.363	37.363	37.699	37.699

Notes: Dependent variable is the percent of a state's patents matched to the census in one of our six census years. White heteroskedasticity-robust standard errors reported in parentheses. *, **, and *** represent coefficient statistically different from 0 at the 10, 5, and 1% level respectively.

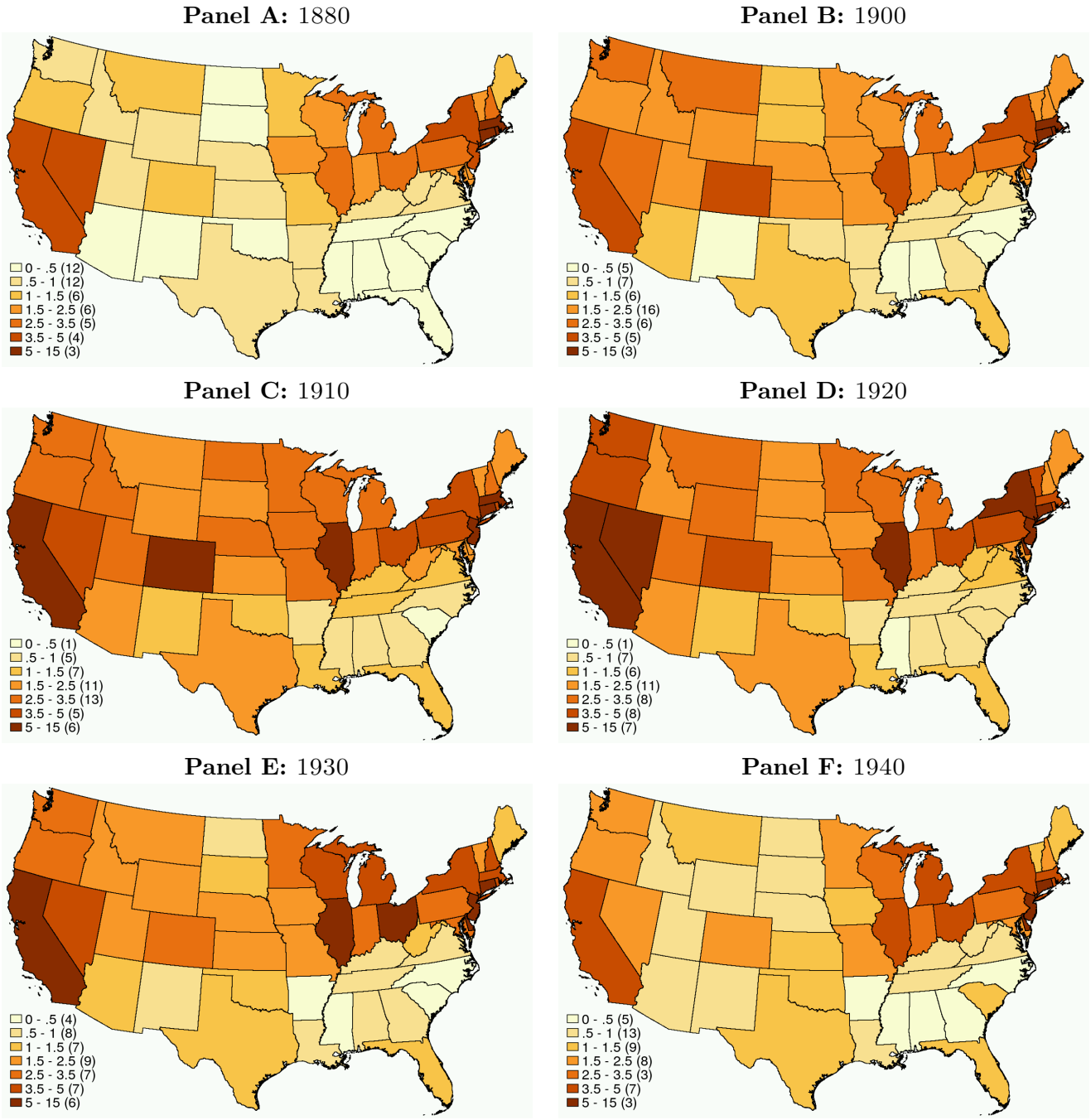
D Additional Robustness Checks

Figure A-5: Innovation and Long-run Growth: US States between 1919-1999



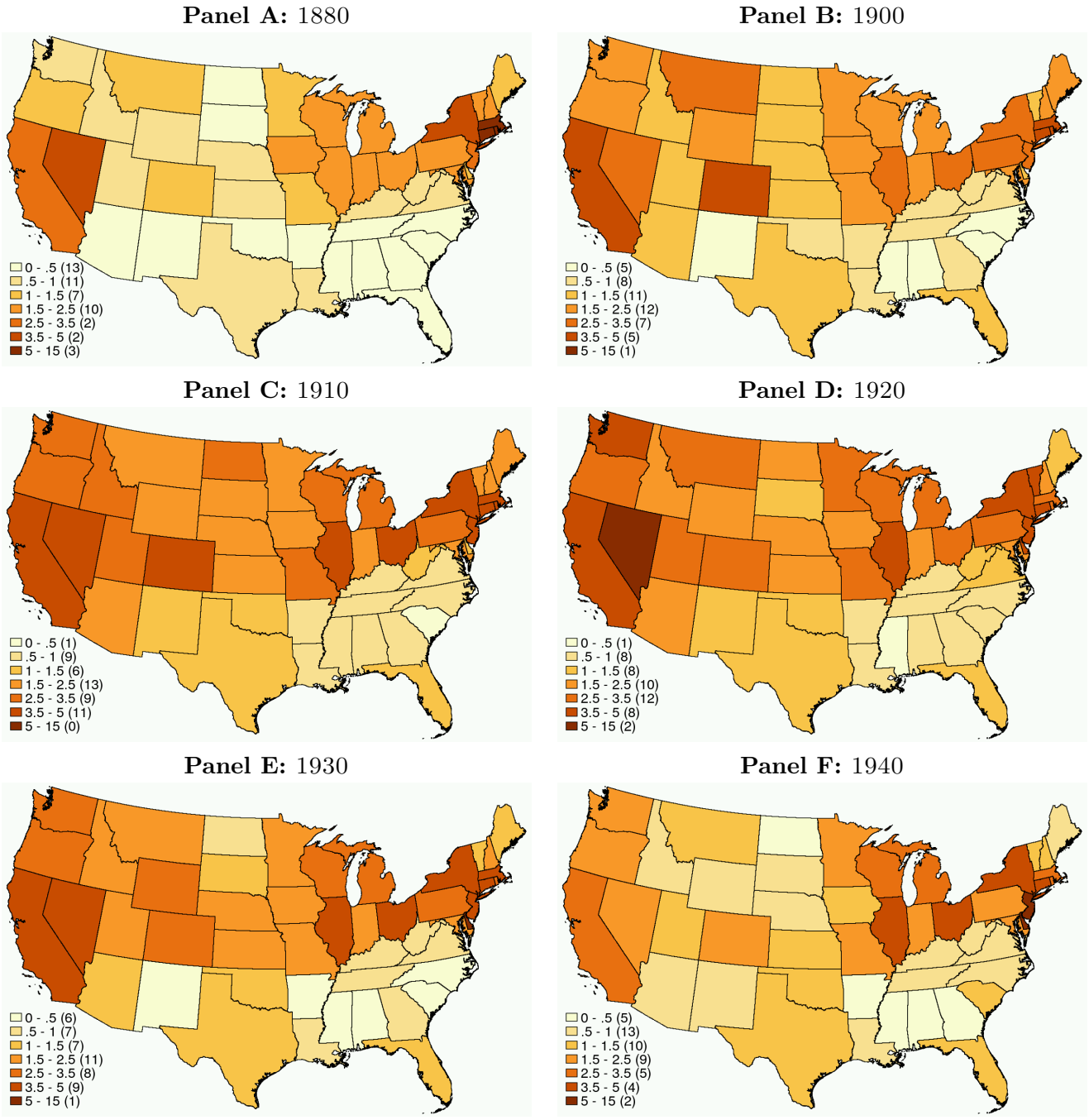
Notes: Figure plots the total number of patents granted to inventors in each state between 1919 and 1999 on the horizontal axis, and the annualized growth rate in state GDP per capita between 1919 and 1999 on the vertical axis. Both horizontal and vertical axes plot the variables of interest residualized against 1919 log GDP per capita, to account for conditional convergence. Source: BEA Historical Regional Economic Accounts, and data from (Martin (1939)) courtesy of Price Fishback.

Figure A-6: The Geography of Inventiveness over Time: Patents per 10,000



Notes: Figure maps the number of patents per 10,000 residents in each state of the mainland U.S. in each decennial census year of our data. Darker colors represent more inventive activity per resident. Patent data come from the USPTO's historical patent files, while population counts are calculated using the U.S. Census.

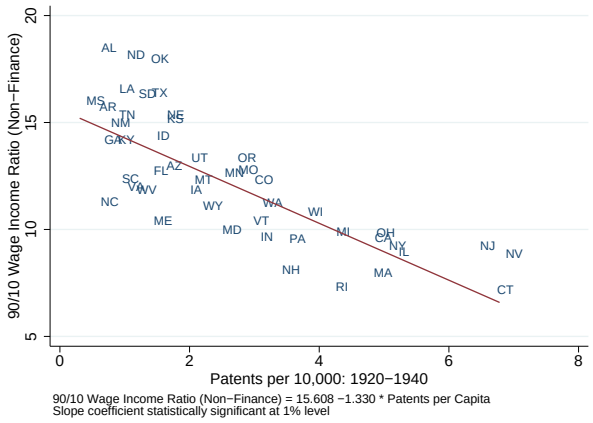
Figure A-7: The Geography of Inventiveness over Time: Inventors per 10,000



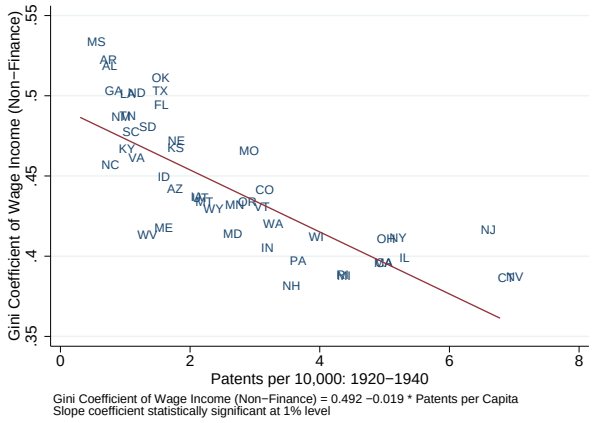
Notes: Figure maps the number of unique inventors per 10,000 residents in each state of the mainland U.S. in each decennial census year of our data. Darker colors represent more inventive activity per resident. Patent data come from the USPTO's historical patent files, while population counts are calculated using the U.S. Census. Source: Historical Census Data, USPTO patent records.

Figure A-8: Relationship between Wage Income Inequality and Inventiveness: Inequality Measures Excluding those Working in Financial Sector

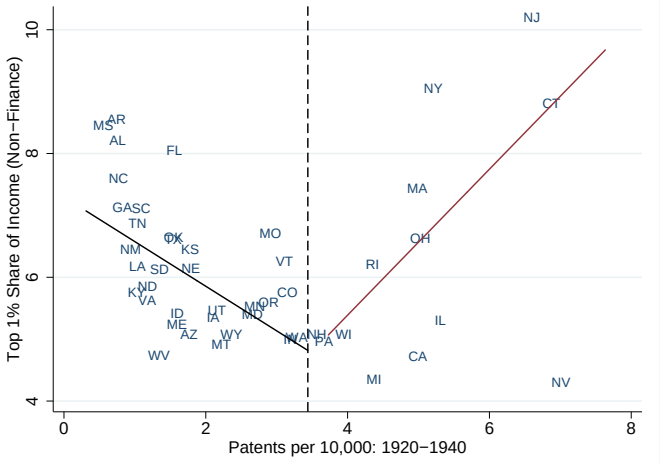
Panel A: Ratio of 90^{th} to 10^{th} Percentile of Income



Panel B: Gini Coefficient



Panel C: Share of Income held by Top 1%



Notes: Figure plots the relationship between average patents per 10,000 residents between 1920 and 1940, and the state-level wage income inequality observed in the 1940 census. All wage inequality measures exclude those who work in the financial sector. Panel A measures income inequality with the ratio of the 90^{th} percentile to the 10^{th} percentile of income, while panel B uses the Gini coefficient as its measure. Panel C measures inequality by the share of income controlled by the top 1% of the state's wage earners. Source: 1940 Historical Census Data, USPTO patent records.

Table A-5: T-tests of Difference between Inventor and High-Skill Non-Inventor Wages over the Life Cycle

Age Group	Inventor Mean Log Wage	High-Skill Mean Log Wage	p-value
19-25	6.621	6.141	0.000
	(1.003)	(0.922)	
26-35	7.615	6.665	0.000
	(0.708)	(0.856)	
36-45	7.913	6.789	0.000
	(0.721)	(0.908)	
46-55	7.845	6.741	0.000
	(0.805)	(0.942)	
56-65	7.735	6.588	0.000
	(0.922)	(0.997)	

Notes: Table reports average log wages for inventors and high-skill non-inventors within each age group. Wage income data taken from 1940 Census. Standard deviations reported in parentheses below means. Final column presents p-values from a two-sided t-test of means among inventor and high-skill non-inventor populations. Source: 1940 Historical Census Data, USPTO patent records.