# The Production of Knowledge and Culture: The US 1790-1870

Tancredi Rapone

September 7, 2023

**Abstract**

This paper uses new data on copyright registration title pages from the Library of Congress (LOC) to analyze the intellectual and cultural development of the United States over 1790-1870. I construct national time series of book production over this period which show an uptake in per-capita terms in 1830, well before the start of the Second Industrial Revolution and the era of "knowledge based progress" (Abramovitz & David, 1996). Matching authors to locations (at the county level) using declassified census data reveals that the spatial distribution of intellectual production in the early 19th century is strongly correlated with inventive activity over 1860-1940 and the evolution of the manufacturing sector. Identification is based on a shift-share type instrument exploiting the large internal migration patterns occurring in this time period. I then use topic modeling to classify books according to subject matter. Contrary to commonly held beliefs, scientific works are not the strongest predictors of the economic trajectories of US counties. Their correlation with manufacturing activity is relatively large in the short-run but disappears over a few decades whereas non-scientific works show an enduring relationship with economic development well into the 20th century. A theoretical model is briefly sketched which rationalizes these results.

# Contents

**1 Introduction** **3**

**2 Literature Review** **5**

**3 Historical Context** **8**

**4 Data Construction** **10**
4.1 Copyright Registration Data . . . . . 10
4.2 Instrument Construction . . . . . 14

**5 Descriptive Statistics** **16**

**6 Empirical Methods** **24**

**7 Theoretical Model** **39**
7.1 Motivation . . . . . 39
7.2 Household Problems . . . . . 40
7.3 Simulation . . . . . 42

**8 Conclusion** **43**

**References** **45**

**9 Appendix** **49**
9.1 Human Capital and Agricultural Diversity . . . . . 49
9.2 Monte Carlo Exercise: Difference between IV and OLS Estimates . . . . 51
9.3 Supplementary Results . . . . . 52
9.4 Data Appendix . . . . . 57

# 1 Introduction

According to a popular view in the literature, the principal reason as to why Western countries are richer today than they were two centuries ago is that these countries *know* more (Mokyr, 2005). The sustained accumulation of knowledge is what made technological progress and economic growth possible for today's developed countries and what separates us from the Malthusian (pre-industrial) world. However, while the proposition that modern economic growth is driven by research efforts leading to useful knowledge is widely accepted in mainstream macroeconomics (e.g. Jones, 2022), we still don't fully understand when this process began and why. Historians of technology and science emphasize the cultural and political changes which occurred between the 16th and 17th centuries and are sometimes referred to as the Scientific Revolution (Mokyr, 2009). However, the space of papers investigating the relationship between knowledge production and economic development is surprisingly limited (some examples include Baten & Van Zanden, 2008; Chaney, 2016; Dittmar, 2011; Squicciarini & Voigtländer, 2015). This is perhaps partially due to the fact that it is difficult to identify individuals driving knowledge production in historical data, and even if one succeeds in this task, understanding the causal processes at play is far from straightforward. For instance, it may seem evident that sustained innovation requires an expanding body of knowledge, but the amount of resources devoted to knowledge production are likely to be themselves a function of the economic environment and preferences.

In this paper, I use the universe copyright title pages over 1790-1870, obtained from the Library of Congress (LOC henceforth), to identify the individuals driving knowledge production in the early United States. These primary sources provide an extremely rich set of data concerning the intellectual and cultural history of a nation which, in this time time period, went from being a primarily agrarian country to the world's leading economy.[1] Using the declassified census, we can trace these individuals to find out where they lived and, potentially, how they contributed to the extraordinary economic development which took place over the 19th and early 20th centuries. While our data is limited to the time period of 1790-1870, the largest drive towards industrialization occurred in the US towards the end of the 19th century. This is functional to our purpose as it will limit the extent of reverse causality which could ensue from a more developed manufacturing sector stimulating the production of copyrighted materials. Care must be still exercised when interpreting OLS coefficients however as the production of knowledge is certainly correlated with other factors which could influence the development of a manufacturing sector. An advantage of focusing on the US as a

[1] By the end of the 19th century the US had surpassed Britain to become the world's largest, richest and most technologically advanced economy (Gallman, 2000).

case study, is of course the abundance of historical data which is easily available and will enable us to control for many possible channels in the baseline specifications.

To further assuage potential endogeneity concerns, I attempt to obtain exogenous variation in the spatial distribution of knowledge producing individuals, through a shift-share instrument which exploits the large amount of internal migration which resulted as a consequence of the Westward movement. Up to 65% of all native individuals who registered a work for copyright between 1840 and 1860 did not reside, in 1850, in the same state where they were born.[2] Around 15% of all authors were foreign born, which was considerably higher than the 10% national average[3]. Using county level shares of individuals born in different states I predict the number of authors residing in a county, exploiting the cross-state differences in copyrighting rates and the fact that in inter-state migration patterns individuals migrate to locations with higher shares of people native to their state of origin. The instrument is valid as long as the share of individuals native to each state does not vary systematically with county level unobservables. To mitigate concerns that this assumption may be violated, I control for a wide array of observable characteristics. Using this strategy, I show that the spatial distribution of innovation, as measured by patents, during the Second Industrial Revolution (1860-1940) is strongly related to the intensity of knoweldge production, measured by copyright registrations, over 1840-1860. This paper adds to the recent literature on the determinants of American long-run structural change and growth which finds that geographic endowments leading to higher agricultural diversity in production in 1860 contributed to the industrialization process (Fiszbein, 2022). I find that, controlling for the abundance of preindustrial human capital identified as the intensity of copyrighting at the county level, there is no effect of agricultural diversity due to better geographic endowments.[4] This potentially indicates that exogenous geographic characteristics were functional in promoting preindustrial economic development and human capital accumulation, but did not have an independent long-run impact on the structural change process.

I also contribute to the literature on the connection between culture and long-run economic performance by identifying a correlation between religious fractionalization and intellectual production, measured through copyright registrations, in this time period. Previous literature has focused on the negative consequences of ethnic and religious diversity in developing countries, highlighting their increased incidence on the severity and frequency of conflicts (Alesina & La Ferrara, 2005; Montalvo & Reynal-Querol, 2003; Montalvo & Reynal-Querol, 2005). In contrast, in the American context

[2] This is the definition of internal migration I use throughout the paper.
[3] See Appendix table 20 for more statistics comparing authors to the general population
[4] As the paper already has an abundance of Tables and Figures, this explicit comparison with Fiszbein (2022) is left to a dedicated Appendix.

increased ethnic diversity during the age of mass migration has been found to have a positive impact on long-run outcomes, potentially due to the importance of skill variety for industrial development (Ager & Brückner, 2013). I construct indices of religious fractionalization and polarization following the methodology of Montalvo and Reynal-Querol (2003) which exhibit, respectively, a positive and negative correlation with intellectual production over 1840-1860. I hypothesize that these correlations reflect an underlying relationship between cultural and religious diversity and the production of knowledge due to the positive effect that competition between different denominations had on the provisioning of religious education. A large portion of copyrighted works consist in textbooks and pedagogical materials for use in *sunday-schools* which were effectively the first schooling institutions in the United States. I am still exploring avenues to obtain an exogenous variation in religious diversity to identify whether this is a causal driver of intellectual production as measured through copyrighting, however this is challenging due to the lack of data on religious affiliations and establishments prior to 1850.

The rest of this paper is structured as follows. Section 2 provides a brief overview of the literature. Section 3 describes the historical context with particular attention to internal re-settlement patterns and human capital growth in this time period, Section 4 describes the data construction process (including the construction of the instrument) and discusses all primary data sources used therein. Section 5 presents some descriptive statistics using the novel data. Section 6 presents the main empirical results concerning the impact of human capital accumulation on regional growth and structural change. Section 7 describes a theoretical model which attempts to rationalize some of these results. Section 8 concludes.

## 2 Literature Review

A growing body of empirical literature identifies human capital accumulation as a driver of long-run economic outcomes. This literature is at least as old as the debate between Acemoglu et al. (2001) and Glaeser et al. (2004) over the "colonial origins" of comparative development. Glaeser et al. (2004) were the first to suggest that heterogeneity in cultural characteristics of colonial settlers could have long-run impacts on the accumulation of human capital and growth, independently of institutional arrangements set up in colonies as argued by Acemoglu et al. (2001). However, given the well-known problems involved in using cross-country regressions to identify causal relationships, these papers were not able to conclusively determine in what way the international migration associated with colonial settlements affected long-run outcomes. This is a challenging question to answer because, as argued by Glaeser et al. (2004), it remains fundamentally uncertain what Europeans brought with them when they settled to the new world: knowledge, culture or institutions? Most likely a mix of all

three and potentially other factors as well. Studies using within country spatial data, where certain features such as the institutional environment can be kept constant, have had more success in establishing a causal link between human capital and economic development. Dittmar (2011) uses city level data from Germany to investigate the effects of the introduction of the printing press, a radical innovation in the technology used to store and produce information, finding large and statistically significant effects. Also in the German context Becker and Woessmann (2009) and Cantoni (2015) find that the protestant reformation had large effects on the accumulation of human capital and, through this channel, city growth and urbanization. In the American context, a growing number of studies has focused on international migration and the skills brought by immigrants when resettling in the United States (Abramitzky et al., 2014; Ferrie, 1999; Sequeira et al., 2020). As noted by Smith (1925), at p.72 "with mercantile sagacity, England prohibited the export of [...] machinery, but she failed to prohibit travel. So one day, Samuel Slater arrived in Providence with a head full of knowledge". Referred to by Andrew Jackson as the "father of the American industrial revolution", Samuel Slater was the man responsible for importing British textile technology to the United States. Although Slater's contributions do not specifically show up in our database due to his lack of written works, it is worth noting that many other key figures in American industrialization, such as Oliver Evans, Thomas Blanchard and John W. Nystrom, who contributed respectively to the advancement of steam power, the assembly line and early computing technology, authored numerous published works, contributing to the dissemination of knowledge in their fields. Identifying the location of these and other individuals who constituted the knowledge elite of the time, as I do in this paper, is thus key to mapping the distribution of human capital and understanding its relationship with the industrialization process.

In a similar paper, Squicciarini and Voigtländer (2015) find that the location of individuals who subscribed to the Encyclopedia in preindustrial France significantly predicted the adoption of technologies and wage growth during the industrial transition of the late 19th century. These authors also found that, consistently with evidence from Britain, conventional indicators of human capital such as literacy rates are only related to economic development in the cross-section, whereas in the time series data only the density of the "upper-tail" of the human capital distribution, in their analysis the encyclopedia subscribers, predicts wage growth and industrialization. By shifting attention to the "upper-tail" of the human capital distribution, being the inventors and knowledge producers, Squicciarini and Voigtländer (2015) reconciled the robust positive relationship between human capital and economic development in contemporary data with the long-standing belief that conventional indicators such as literacy rates played little role in the industrial revolutions of continental Europe (Mitch, 1993). While this makes sense in most European countries where mass schooling, advanced through the promotion of formal reforms enacted by governments, took place only

decades after industrialization, the United States has had a long tradition of highly localized schooling provision which started since the founding of the country and is still a topic of debate among economists (Fernandez & Rogerson, 1996, 1998). The provision of schooling started in a grass-roots fashion at the district level, introduced by local communities of parents and, in some cases, religious organizations (Goldin, 2016). This local endogeneity of schooling provision could therefore provide a channel through which the presence of a knowledge elite, identified from the copyright registrations, stimulated the provisioning of schooling, which is indeed consistent with the findings provided in Section 6.

The interest in books and copyrighted materials as indicators of human capital is, therefore, not necessarily parasitic on the presumed relationship between the production of useful knowledge and productivity growth. Baten and Van Zanden (2008) postulate that book production indicates the presence of a "modern" culture.[5] Books are themselves more than merely normal goods: their consumption does not only vary with income but also with consumer preferences which are changing as a result of historical processes such as secularization and other cultural changes. The national time-series of book production which I show in Section 5, when compared to the data collected by Baten and Van Zanden (2008), shows the extent of the vertiginous rise of the United States as the human capital leader in the Western world: between 1790 and 1840 the US achieved the increase in per-capita book production attained by European countries over the two centuries leading up to the Industrial Revolution. As in the case of the European countries studied by Baten and Van Zanden (2008), it is important to note that the rise in book production occurred before industrialization began.

The analysis carried out here is not inconsistent with the literature arguing that innovation in the industrial revolution was driven by the incentive to replace expensive labor with inexpensive capital (Acemoglu, 2002, 2003; Allen, 2009). This body of literature has been extensively criticized by, among others, Kelly et al. (2014) who point out that endogeneity of factor prices makes these sorts of theories difficult to sustain. While I make no attempt to integrate factor prices in this paper, it is worth noting that this may be a necessary but insufficient condition to stimulate industrialization. It is far from certain that innovations which reduce labor input will be attainable in all circumstances, even if incentives for their development due to high labor costs are present. Individuals capable of inventions and an abundance of entrepreneurs willing to invest in them, along with some level of financial intermediation (Mao & Wang, 2022), must also exist. In the British case for instance, financial incentives alone were

[5] See also Mokyr (2016).

surely not the principal motivation behind the rise in inventive activity, given the dismal fortunes of most inventors (Clark, 2008). At a more basic level, the presence of a knowledge elite which advances the understanding of natural phenomena can be understood as another necessary, albeit perhaps insufficient, condition for industrialization to occur. Another condition which has been emphasized by Mokyr (2016) and Clark (2008) is that the pursuits of this knowledge elite must be sufficiently secular and interested in practical matters. Doepke and Zilibotti (2008) argue that a couple of the richest landowing families in England could have financed the entire capital accumulation necessary for the Industrial Revolution. The reason why the landowing class repeatedly resisted this change instead of capitalizing on it, in their view, was that investment in new technologies and ventures was simply too far removed from their preferences. Using the title pages of books produced by American authors over time period, can therefore help us understand whether the interests of the knowledge elite are becoming more secular and concerned with practical issues, which is indeed what we find over 1790-1840.

Finally, I contribute to the theoretical literature on the transition from the malthusian world to modern economic growth (Galor, 2011; Galor & Moav, 2002; Kremer, 1993). Most of the theoretical literature on this topic has been rooted in the endogenous growth tradition which focuses on population growth as the ultimate driver of the transition. I explore an alternative theory where the spreading of skills via cultural transmission of preferences and technology à la Bisin and Verdier (2001) is the driver. Unlike in the classical unified growth model by Galor (2011) agents' incentives for human capital accumulation are not homogeneous. The probability an individual becomes of the high skill type, and can hence profit from human capital investment, is endogenous to parental decisions as well as "neighborhood effects" which consist in the relative composition of the population at the county level. Individuals born to low-skill families living in high-skill areas have, ceteris paribus, a higher probability of becoming high-skilled in the model presented in Section 7. Using this framework, I show that variation in the initial proportion of the high-skill type (the model counterpart of the "knowledge elite") can have long-lasting persistent effects.

# 3 Historical Context

The United States in this time period saw some of the most important transformations in its history. Most notably, the physical size of the country was changing from year to year as a result of military conflicts such as the Mexican American war and numerous purchases from foreign countries (the largest of which was the Louisiana Purchase of 1803 which conferred to the US 23% of its current territory, previously belonging to France). This abundance of land created vast opportunities for internal migration. While there are many studies focusing on international migration into the US in the end

of the nineteenth century (e.g. Abramitzky et al., 2012, 2014; Ager & Brückner, 2013; Sequeira et al., 2020), there has been relatively little attention given to the importance of internal migration and the Westward movement in the economic literature. In two recent papers Bazzi et al. (2020) and Bazzi, Fiszbein, et al. (2021) show that the relative time a county spent on the frontier contributed to the development of a culture of "rugged individualism" which persists to the present day in the form of support for the Republican party and resistance to lockdowns during the Covid-19 pandemic. Their analysis focuses on the causal influence that frontier conditions, characterized by low population density and institutional presence, had on the development of cultural traits favoring individualism and discouraging collective action. While their research design is uniquely suited to the purportedly exogenous expansion of the country in this time period, the mechanism through which cultural traits developed and persisted to the present should apply to other settings as well. For instance, areas settled by individuals coming from Northeastern states, which had a high level of human capital and urbanization even before the beginning of America's industrial revolution, could exhibit different development trajectories than those settled by individuals coming from the Southern states, due to the same mechanism of intergenerational transmission of preferences and human capital. The heterogeneity of settlers in the Westward movement is therefore a potentially significant area of research which has previously received little attention. This is especially important for human capital accumulation as there was no federal schooling policy before 1870. As I mentioned in Section 2, the development of schooling prior to 1870 was entirely a grass-roots movement in which local communities of parents came together, often autonomously funding schools for their children (Goldin, 2016). While the culture of "rugged individualism" which discouraged collective action may have hampered this, other cultural traits may have been more favorable. It is therefore likely that the composition of settlers in the westward movement should have some significance on future economic development. Indeed, the empirical strategy used in this paper will leverage the fact that counties with higher shares of people coming from states with higher than average values of knowledge production have themselves higher copyrighting activity.

It is important to note that, although not all territory West of the Appalachians was uninhabited prior to 1790, the best estimates of population density were extremely low even as of the first decades of the nineteenth century (Porter et al., 1895), meaning that most of these areas, excepting geographic characteristics, were relatively similar prior to the Westward movement.[6] The most important factors contributing to the settlement of the Mid-West identified in the literature are the reduction in transportation costs as a result of technological improvements such as steam-boats and the construction of canals and, to a lesser extent, the supply push of international mi-

[6] This is of course abstracting from the pre-colombian settlements of native Americans which were the victim of massive destruction at the hands of the American government (Cozzens, 2016).

grants which put population pressure on the Eastern seaboard (Vandenbroucke, 2008a, 2008b). Land policy, which was since independence a major source of disagreement between the founding fathers, although less explored also played a major role (Atack et al., 2008). While Thomas Jefferson favored a policy of free land concessions to settlers, envisioning the United States as a country of independent small-holders whose relative equality would be an integral part of a political democracy, others lead by Alexander Hamilton beleived that such a policy would stifle an emerging manufacturing sector and significantly retard economic progress (Atack et al., 2008). For a long time the Hamiltonian faction had the upper hand and land prices, although varying, significantly biased the composition of westward migrants in favor of higher skill occupations and richer individuals. This is why we observe that while on average 35% of all Americans were internal migrants in 1850 (defined as not living in the state were they were born), this number was closer to 65% for authors who registered a work for copyright between 1840 and 1860. By way of comparison, the fraction of Americans currently living in a different state from the one where they were born is 38%, hence while these levels of internal migration are not necessarily unusual the fact that areas west of the Appalachians were largely unpopulated before the early 19th century gives us a unique opportunity to examine how the composition of settlers affected subsequent development patterns.

# 4 Data Construction

In this section I describe the process of constructing the data which I use in Sections 5 and 6 to study the relationship between knowledge production and economic development. I divide this Section in two Sub-Sections for ease of reference. First I describe in detail the process of constructing the data on copyright registrations based on the primary sources gathered from the Library of Congress. In the following subsection I describe the construction of the instrument used to obtain conditionally exogenous variation in knowledge production at the county level.

## 4.1 Copyright Registration Data

The data on copyright registrations I use throughout the paper is constructed from the collection of "Early Copyright Title Pages" available from the United States Library of Congress in digital form since early 2020. In order to register an item for copyright in the US from 1790 to 1870, the author had to submit a copy of the title page of the work along with a form and pay a registration fee at their local District Court. This was overhauled in 1870 at which point the responsibility for handling copyright requests was transferred to the Library of Congress along with all the extant copyright records. The nominal registration fee was set at around 60 cents throughout the period under study, which was arguably expensive enough that the author had to believe the

work was of some value but not so expensive as to be prohibitive to many authors.[7] Excepting patents, and a small fraction of records which never made it to the Library of Congress, this database arguably comprises the universe of the intellectual production in the US over 1790-1870, or at least the subset which has been preserved in writing. The LOC estimates this corpus to comprise approximately 50 thousand individual copyrighted works in over 90 thousand image records.[8] To the best of my knowledge, this paper represents the first attempt to classify these records by time, location and subject matter.

The data construction process is done as follows. First, I download the transcriptions of the title pages from the LOC's website using a webscraper.[9] I then proceed to process and clean the transcriptions. Ideally, the final product would be a dataset containing for each copyrighted work the year of registration, name of the author, place where the author lived and subject matter. Extracting the year is the easiest task which can be done by filtering the transcription for groups of numeric characters starting with "17" or "18", after converting all roman numerals to western digits. I am able to confidently assign a date to approximately 70 thousand individual entries which is higher than the LOC's estimate of the total number of works, most likely due to resubmissions and digitization errors by LOC clerks.[10] I then proceed to extract the name of the author. To do this I simply filter the transcription, removing all words before the first instance of the word "by", and then match individual words in the transcription to the list of first and last names contained in the 1850 census. A more sophisticated machine learning approach which uses the particular layout of book title pages (which usually contain title, author and publisher name, in that order), following the lines of Shen et al. (2021), would most likely perform more accurately than the procedure I have so far implemented. However, this would require the creation of a large training dataset. The gains to usign a layout parser would appear to be limited, I estimate 80% of correct matches, although for works with multiple authors I can only match the first author.

The next step is assigning an author to a location at the county level. To reduce false positives I first extract the name of the state where the book was published, which usually figures on the title page. To correct for misspellings, which are pervasive in historical censuses, I apply *soundex* to both the author names extracted from the primary sources and the individual census records. The soundex is a user written

[7] This fee was equivalent to about 20$ in today's money. See Khan (2005) for more information.
[8] The larger number of images is due to some entries being accompanied by forms and blank pages.
[9] This is perfectly legal as the title pages are in the public domain. All codes are provided upon request.
[10] Copyright protection expired after 14 years, at which point the author could register their work again for a one time extension. It is important to note that due to human errors there may be multiple entries, that is, the same record may have been scanned twice. Removing the duplicates is, however, not a straightforward process as it is difficult to distinguish them from resubmissions (which could potentially indicate books quality).

command available in Stata which assigns to each string a code based on its phonetic pronunciation in the English language, which helps deal with the misspellings which are common in historical census data. I then use the first and last name of the author of a work to search for the location based on the census records. In order to have an informative match, I am forced to restrict the pool of potential matches to certain high skilled occupational groups (the list of occupations is provided in Table 18). I further impose the restriction that the author must have been at least 25 years old at the time he/she[11] wrote the book. This allows me to exactly match approximately 70% of authors (see the Appendix, Table 17). An exact match here means that for one author I observe one person in the census living in the relevant state and satisfying the occupational/age criteria. Unmatched authors are matched in a second round search, which is generally less accurate, by searching for individuals in other states. It is important to note that there is, very likely, an important amount of measurement error involved in this process which will potentially lead to downward bias in parameter estimates. This is due both the inaccuracies involved in the matching and author extraction processes and the fact that copyright registrations is a proxy which likely understates and imperfectly captures the true extent of knowledge production in a county. Adding to the confusion, this measurement error is likely positively correlated to population, as finding a false positive is easier in a highly populated area with more potential matches. This generates the potential for spurious results via the construction of a population proxy. The fact that in most empirical models including population as a control does not make the variable of interest insignificant is therefore a reassuring test that we are not merely constructing a population proxy via spurious matching.[12]

I then proceed to assign topics to the copyrighted works. For now, the only classification I am attempting is academic/non-academic works. As with most topic classification tasks, there is some arbitrariness in the definition of the topics. I consider a work to have academic value as long as it fulfills two conditions: (i) its main purpose is the diffusion of some kind of knowledge and (ii) this knowledge is not mainly of a religious character.[13] An ideal topic classification strategy would use a large training dataset of labeled title pages to classify the corpus (Jelodar et al., 2019). As such a dataset does not exist, and constructing it would be too time intensive, I attempt a second best strategy using keywords. Importantly, this will only work for binary classifications (such as academic/non-academic) but not for a more find grained topic

[11] I estimate 1.32% of authors between 1790 and 1870 were women using the relative frequency of the string "Mrs." or "Miss". While this number is small, some of the most prolific authors such as Harriette Baker who wrote children's books under the pen name of Madeline Leslie, were women.

[12] I conduct placebo tests where I randomly select 4000 individuals from the census and pretend they are authors. In nearly all cases, the resulting indicator of county level copyright registrations is unrelated to all outcome variables when controlling for population.

[13] Although some academic works may be written by exponents of the clergy such as manuals for Sunday school teachers.

classification.[14] This is in principle more straightforward as I can rank works by the relative frequency of certain words which I associate with academic topics and classify according to a threshold rule (e.g. at least 10% of the words must be on this list). However, there is an important tradeoff here between large lists of keywords, which avoid Type 2 errors and smaller lists which avoid Type 1 errors. To solve this I opt for a small list (see Table 22) and use the top 10% of the works and bottom 10%, ranked according to the sum of the relative frequencies of the words in the keyword list, as a training dataset to classify the remaining 80% of the corpus. This classification is done using a logistical regression model of the following form:

$$y = X'\beta + \epsilon \tag{1}$$

Where $y$ takes on value 1 if the work is labeled as scientific and 0 otherwise (according to the keywords). The matrix $X$ is a matrix of all unique words appearing in the corpus and takes values 0 or $n$ if the word appears in the title page respectively 0 or $n$ times. The error term vector $\epsilon$ follows an i.i.d. extreme value distribution. Once we have estimated the coefficients $\hat{\beta}$ we can use this model to predict the topic (academic/non-academic) of the remaining data. As topic classification is an inherently subjective task, in the code I have included a section which extracts a random sample of scientific and non-scientific works after applying this procedure to classify the corpus and saves them to a csv file for the reader to inspect. Cursory checks of this classification procedure reveal that it is relatively successful and, hopefully, superior to a simple keyword strategy as used by Chaney (2016).[15]

The strategy described above, which involves a binary classifier (academic or non-academic), is however less well suited to identifying works which may transcend clear-cut topic boundaries. This will be the case for books which are both religious and philosophical, or textbooks which touch upon many different subjects. To divide the corpus into many different categories, unsupervised classification algorithms such as LDA and BERT will be more adequate. However, these algorithms work better with larger corpuses of texts, such as entire books. This is especially true of LDA which I have attempted to use with disappointing results.

[14] I also attempted using a Latent Dirichlet Allocation model (LDA) which would allow classification into more fine grained topics without a training dataset. However, the results are disappointing as LDA works best with large amounts of text.
[15] All codes are available upon request.

## 4.2 Instrument Construction

As discussed in Section 3, I will use an empirical strategy which leverages the large patterns of internal migration occurring in this time period. Table **??** shows the fraction of individuals by region who are classified as internal migrants in the 1850 census.[16] As we can see, this fraction is especially high for the Midwestern region and is always higher for authors, although it is still substantial even in the Northeastern region which was already largely settled by the end of the 18th century.

| Region | Mean | | Frequency |
|---|---|---|---|
| | Non-Author | Author | |
| Midwest | 0.45 | 0.78 | 4,584,383 |
| Northeast | 0.12 | 0.30 | 7,207,901 |
| South | 0.22 | 0.48 | 5,434,905 |
| Total | 0.24 | 0.36 | 17,227,189 |

**Table 1:** Summary Statistics of Internal Migration for natives by Region and Author Status. Source: US Census of Population 1850 and Copyright Registration Data from the Library of Congress.

The identification strategy exploits the fact that authors, just like other migrants, tend to choose locations where larger fractions of people from their home state settled. This is what is known in the migration literature as the "migrant network instrument" (Bartel, 1989; Munshi, 2003), with the exception that I have no time variation as data on state of birth only becomes available in the 1850 census.[17] Nonetheless, I argue that this instrument will alleviate selection bias concerns as factors which induce the knowledge elite to settle in particular areas are likely to cut across state origins: that is, they attract authors from Connecticut just as they do authors from New York. This could be important as, if authors across states tended to choose particular locations with better unobservable characteristics OLS coefficients may capture a selection effect, due to authors systematically choosing better locations, as opposed to an effect of knowledge production on subsequent economic development. Alternatively, measurement errors in knowledge production due to imperfect matching and failure in identifying author names from the primary sources will almost certainly lead to downward bias in OLS coefficients. The instrument will thus exploit the fact that authors tend to locate in regions with larger shares of individuals from their state, alleviating the selection concern, and that individuals from states with higher levels of copyrighting are more likely

[16] As discussed above, my definition of internal migrant is an individual who did not live in 1850 in the same state where they were born.

[17] The traditional "migrant network instrument" uses shares from the previous period (which are considered to be more plausibly exogenous) to predict current migration flows. In a traditional migration paper using the current shares would be inapproprate as there would be a, meaningless, mechanical relationship between the level of immigrants from a certain state and their share, however in our case this is not true as authors are a very small subset of the population (approximately 4000 out of 19 million).

themselves to produce more unobservable knowledge, alleviating the measurement error concern.[18] The identification assumption is that, aside from having higher levels of human capital, settlers from different states are otherwise similar and they do not select systematically different locations. While this assumption is admittedly strong, especially since other characteristics of Westward moving settlers have been studied in the literature (Bazzi et al., 2022; Bazzi, Ferrara, Fiszbein, et al., 2021), we can control for a wide variety of county level characteristics to identify other channels through which initial settlement patterns may affect subsequent development. Formally the instrument is constructed as follows:

$$\hat{books}_{i,s} = \sum_{l \neq s} \left( \frac{authors_l}{people_l} - \frac{authors_{us}}{people_{us}} \right) \pi_{i,l} \tag{2}$$

Where the predicted level of copyrighting in county $i$ of state $s$ is a weighted sum of the average deviations (from the national average) of copyrighting per capita in all other states $l$ ($authors_l$ is the number of people who registered a work for copyright between 1840-1860 and were born in state $l$ and $people_l$ is the number of people in the 1850 census born in state $l$) weighted by the relative fraction of individuals from that state in the county population (in 1850). I also add the largest foreign suppliers of international migrants to the list of US states. These countries are the UK, France and Germany, which are treated as if they were US states in 2.[19] This prediction is used as an instrument for the actual level of copyrighting at the county level to isolate the part of knowledge production which is due to the cultural composition of settlers and not based on selection mechanisms which would attract authors across states. To motivate this as a valid instrument, we can run a reduced form regression to see how the instrument correlates with the observable geographic characteristics of different counties. If there is selection on observables (meaning that people from states with high levels of copyrighting choose better locations), there may be also selection on unobservables which casts doubt on the validity of the instrument. To test this, consider the following model:

$$\hat{books}_{i,s} = \alpha_s + \delta \Gamma_{i,s} + \epsilon_{i,s} \tag{3}$$

Where $\alpha_s$ is a state fixed effect and $\Gamma$ is a vector of geographic controls including average temperature and precipitation, distance to lakes and oceans, potential yields and terrain ruggedness (measured as the average slope in a county). The results are shown in Table 2 which shows that, if anything there is negative selection. People

[18] In the Appendix I formalize this conjecture with a Monte Carlo experiment. See Figure 13.
[19] As we have no data on copyright registrations outside the US, the fraction $authors_l/people_l$ is the rate of copyrighting for individuals from those countries in the US.

from states with higher levels of copyrighting activity tend to resettle in areas which are farther from oceans and lakes, more rugged and colder. The association with land suitability and mean precipitations is insignificant.

**Table 2:** OLS Regressions: Selection on Observables

| | (1) | (2) | (3) |
|---|---|---|---|
| Terrain Ruggedness | 0.811*** | 0.848** | 0.581** |
| | (3.79) | (3.18) | (3.19) |
| Distance to the Ocean or the Great Lakes (in km) | -0.0367*** | -0.0333*** | -0.0102 |
| | (-9.03) | (-5.55) | (-1.43) |
| Temperature | -1.069** | 0.947 | 1.514 |
| | (-3.22) | (1.71) | (1.58) |
| Annual Rainfall | 0.000104 | -0.00642 | 0.00652 |
| | (0.02) | (-0.81) | (1.24) |
| Land Suitability | -1.744 | 4.738 | 4.760 |
| | (-0.50) | (1.73) | (1.89) |
| N | 1507 | 1507 | 1507 |
| R2 | 0.423 | 0.145 | 0.235 |
| State FE | no | yes | yes |
| Lat/Lon Polynomial | no | no | yes |

$t$ statistics in parentheses.
Dependent variable is the instrument constructed using 3. Conley standard errors used with a 100km cutoff. For variable definitions and sources, see the Data Appendix.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

# 5 Descriptive Statistics

This section presents some descriptive evidence relating to the original data constructed in this paper. We start off by examining the topic composition of the entire corpus of works registered between 1790 and 1870 using a simple keyword strategy. The lists of keywords are available upon request, they are not included in the appendix due to space limitations. At this stage we classify a work as belonging to a specific topic as long as it contains at least one of the keywords, which will potentially lead to many overlapping classifications, but is still useful to have a broad overview of the data.

As we can see from Table 3, the keyword strategy for classifying titles leaves a large number of works unclassified and shows substantial overlap between categories. Given the large amounts of overlap in this classification, in the following sections we will be only looking at binary distinctions between, for instance, science non-science performed using the logistical regression model described in Section 4 which is far more precise than a simple keyword strategy. One of the main contributions of this paper is the

**Table 3:** Descriptive Statistics of Work Categories

| Category | Non-Overlap | Overlap | % of Total |
|---|---|---|---|
| Academic | 5,389 | 23,717 | 9.2% |
| Textbook | n.a. | 9,865 | |
| Invention | n.a. | 3,280 | |
| Novel | 3,182 | 10,419 | 5.4% |
| Religious | 4,486 | 14,387 | 7.7% |
| Business | 4,642 | 16,305 | 7.9% |
| Not Classified | 26,000 | n.a. | 44.4% |
| Ambiguous | n.a. | 14,487 | 24.7% |
| Total | 58,498 | n.a. | 100% |

Descriptive statistics, showing only categories with more than 5000 works classified. The first column shows the number of works falling unambiguously in a category (e.g., scientific works only classified as such), the second shows the total number of works falling in a category allowing for overlap with other categories in the table. Total number of works is defined as the sum of (i) all non-overlapping works in each category, (ii) works falling in more than one category (ambiguous works), and (iii) non-classified works. The third column shows the percentage of each entry in the total. The lists of keywords are available in the Appendix. Textbook and Invention are treated as subcategories of Science and hence do not have a non-overlapping component.

construction of a national time-series of the number of books registered for copyright in the US over the 19th century, which is then broken down by location and topics. This is interesting even disregarding the potential to parse through the subject matter of individual publications, as it shows the evolution of the intellectual production in the US during its most important phase of economic development, or at least of what has been transmitted to us through written records. Figure 1 shows the number of works registered for copyright per year over 1790-1870. The large drop in the early 60s is, most likely due to the Civil War which ravaged the country between 1861-1865.

When normalized by population, we can see that a persistent uptake begins around 1830. This roughly coincides with the earliest growth of manufacturing activity in the US (Davis, 2004), although it is substantially before the introduction of formal schooling and the massive industrialization which began in the end of the 19th century. Figure 3 plots the number of copyrighted works and patents registered in each decade normalized by the size of the population. As we can see, the rise in copyrighting started around a decade prior to patenting which has been found in previous studies to be substantially related to the growth of markets and a manufacturing industry in the US (Khan & Sokoloff, 1993; Sokoloff, 1988). Regression models in Section 6 will show that the level of knowledge production over 1840-1860 is a significant predictor of patenting during the Second Industrial revolution even when including a wide array of controls and instrumenting for the level of copyrighting exploiting internal migration patterns. Turning to the spatial distribution of knowledge production, Figure 4 shows a map of the US where with state boundaries and copyrighting intensity at the county

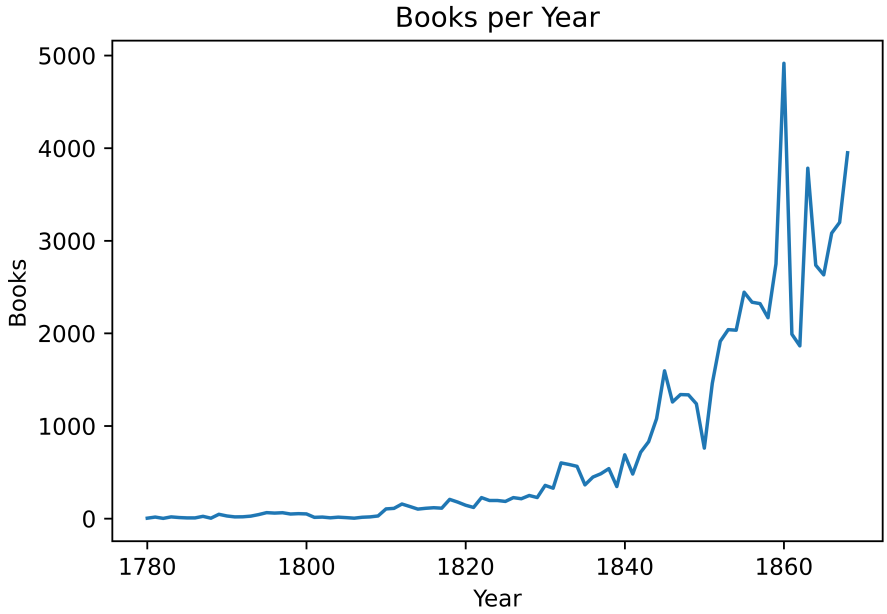
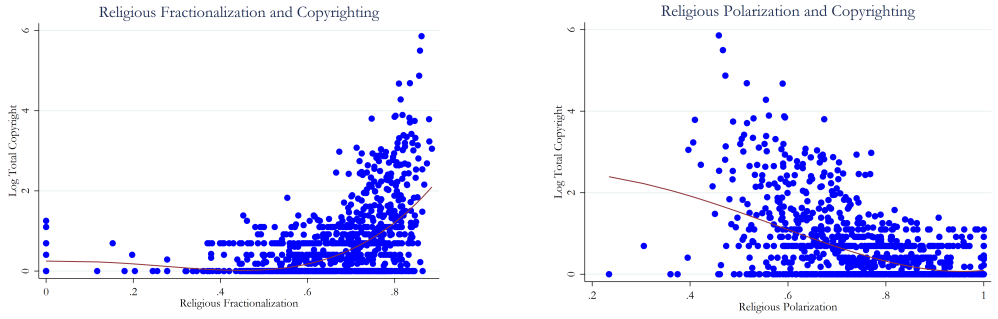



**Figure 1:** Number of copyrighted works by year over 1790-1870. Source: author's calculations using the Library of Congress (1868) data.

level. Academic works are shown in blue, whereas non-academic are in red, hence overlap appears as purple. As we can see, while the majority of registrations occur in the Northeast of the country (especially for academic works) there is substantial variation within states, even those in the Midwest and South of the country. While knowledge production is of course non-rival (and only with some difficulty excludable) as emphasized in the endogenous growth literature (Romer, 1990), we may still expect that regions which produce knowledge may vary in certain cultural traits. This can be especially relevant to those works which are not academic and thus speak to the interests and pursuits of the general consumer. In the following section we will turn to how these traits correlated with subsequent economic development. For now, Table 4 shows us the summary statistics of, respectively, above and below median counties ranked by knowledge production. Counties with higher levels of knowledge production exhibit larger populations, higher Solow residuals,[20] literacy and urbanization rates and more colleges and schools (in 1840). In contrast to some of the literature which studies European countries (e.g. Becker & Woessmann, 2009), no particular religious denomination appears to be positively related with copyright registrations. Notice that, while counties with above median levels of knowledge production do indeed have more protestant churches, the levels are exceedingly small (e.g. 3.5 % compared to 2% of all churches).

[20] The Solow residuals are calculated as OLS residuals from a regression of log output per worker in on capital per worker in the manufacturing sector.

Rather, as I show in Figure 2, indices of religious fractionalization and polarization constructed following Montalvo and Reynal-Querol (2003) respectively correlate positively and negatively with copyrighting activity. These relationships are especially strong (explaining alone around 23% of the variation in copyright registrations) and regression analysis shows that they are significant predictors of copyrighting even when conditioning on a wide variety of controls (see the Appendix Table 19). This is important as the production of knowledge is itself an endogenous product of some historical process and religious competition has been found to have had meaningful consequences for human capital accumulation in Europe by Cantoni et al. (2018). Nonetheless, for now this remains something which is left for future research as a deeper dive into the literature is needed to investigate this point.




**(a)** **(b)**

**Figure 2:** Religious Fractionalization (a) and Polarization (b) against total copyright registrations. For variable definitions and sources see Appendix 9.3 and 9.4.

| | Above Median | | Below Median | | Diff. |
|---|---|---|---|---|---|
| | mean | sd | mean | sd | t |
| Total population | 25797.150 | 33419.058 | 8367.849 | 5061.875 | (-12.923) |
| Solow Residual | 0.065 | 0.403 | -0.074 | 0.501 | (-5.168) |
| Literacy % | 0.876 | 0.114 | 0.792 | 0.151 | (-11.134) |
| Urbanization % | 0.021 | 0.120 | 0.000 | 0.000 | (-4.284) |
| #Slaves/#Population | 0.167 | 0.223 | 0.211 | 0.226 | (3.433) |
| # Colleges | 0.306 | 0.720 | 0.053 | 0.230 | (-8.390) |
| # Schools per child | 0.020 | 0.014 | 0.015 | 0.012 | (-5.605) |
| Protestant churches % | 3.520 | 8.308 | 2.056 | 6.292 | (-3.212) |
| Catholic churches % | 3.804 | 8.144 | 5.014 | 15.123 | (1.733) |
| German Reformed churches % | 1.349 | 4.055 | 0.395 | 1.982 | (-3.765) |
| Observations | 628 | | 629 | | 1257 |

**Table 4:** County-level summary statistics by intensity of copyrighting. Columns (1) and (2) show summary statistics for respectively above and below median counties in terms of total registered works for copyright. Sources: data on copyrighting is based on author's calculations using Library of Congress, 1868, all other data is from Haines et al., 2010.

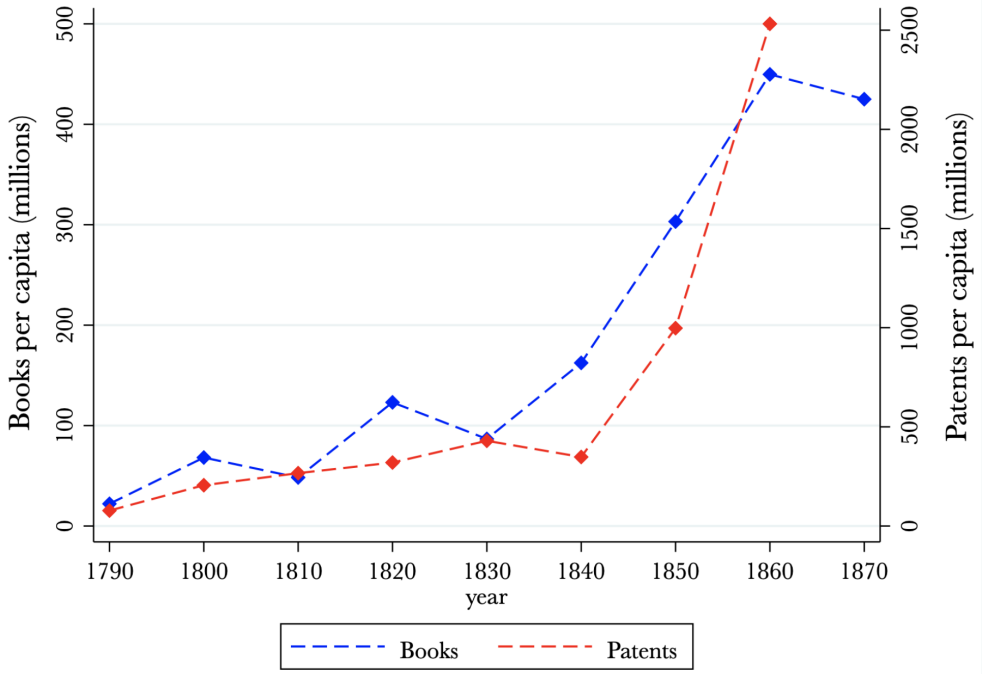



**Figure 3:** Copyrighted works and patents normalized by population. Source for copyright data: author's calculations using the Library of Congress (1868). Source for patents: USPTO (2022).

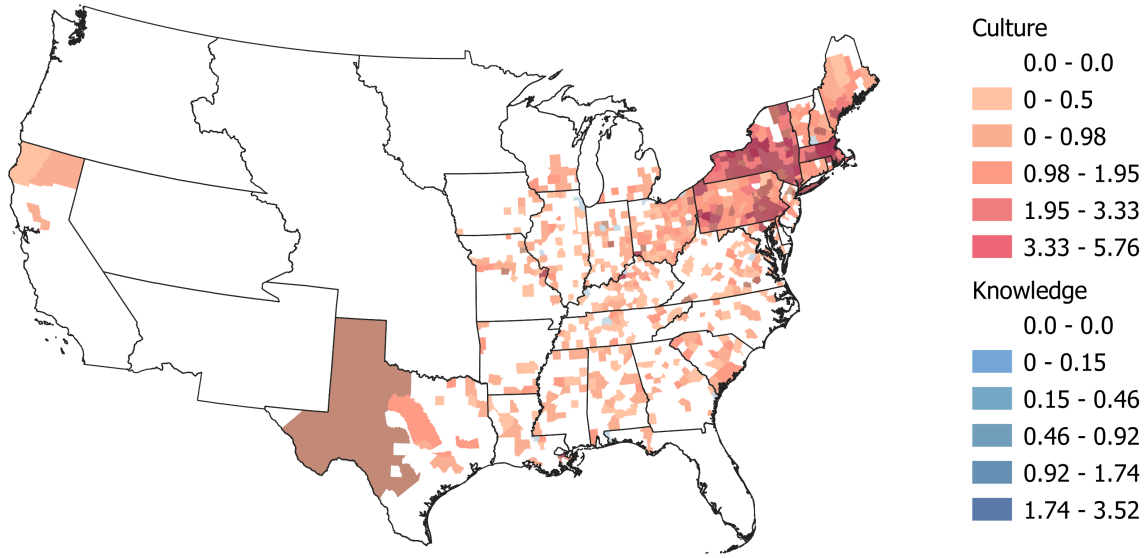



**Figure 4:** Log of copyrighted works over 1840-1860 by county. Academic and non-academic works are respectively in blue and red, hence overlap appears as purple. Source: author's calculations using Library of Congress (1868).

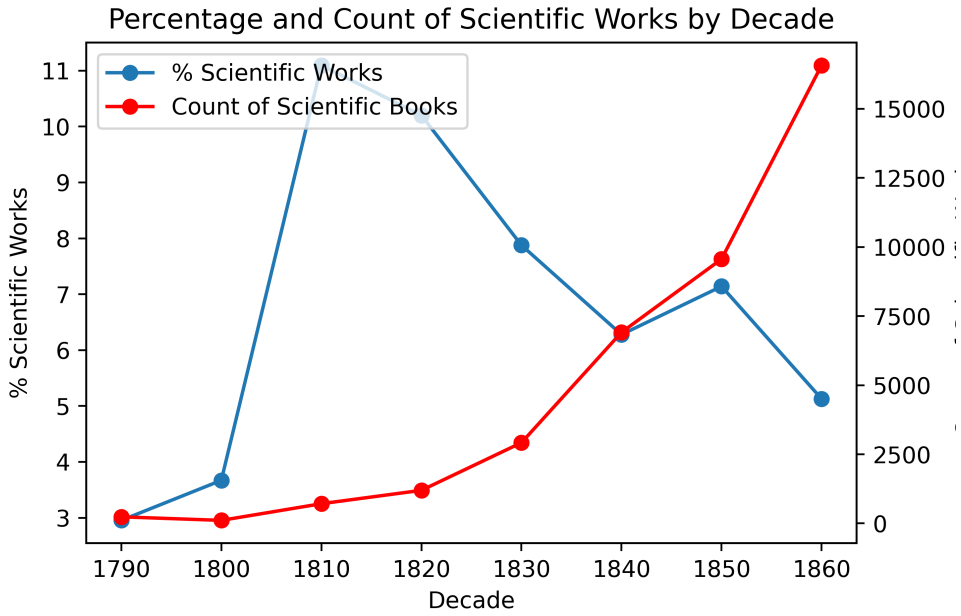



**Figure 5:** Total and relative frequency of works classified as "scientific" following the methodology laid out in Section 3. The code used to classify works is provided in the Appendix. Notice that these results need not be consistent with those presented in Table 3 as they come from a different classification procedure.

Turning to the topic composition of copyrighted works, as we can see in Figure 6, the total count and fraction of all works which can be classified as academic is increasing over time reaching a peak of 22%. This would seem to indicate substantial interest in the pursuit of secular knowledge starting long before the rapid industrialization experienced in the second half of the 19th century. If we restrict our attention to works which have scientific value, by which we mean works attempting to push the knowledge frontier forward, these appear to be declining in relative importance from 1810 onward.[21] This is not surprising as the growth of the reading public implies that more accessible publications are growing in demand faster than scientific ones.[22] To better understand the topic composition of the growing intellectual production we observe in this time period we can use a wordcloud to visualize the most frequent words by decade. Although purely descriptive in nature, such an exercise can be particularly useful when a time dimension is included as it can show whether the focus of the knowledge elite is changing or stable over time. Table 5 below shows wordclouds over 20 year intervals for the whole sample period.

[21] This classification is done following the same methodology outlined in Section 4 for the academic/non-academic classification.

[22] Being literate in 1790 is far less common than being literate in 1870. As the reading public grows it is natural to expect that the focus of publications should shift towards layman's topics.

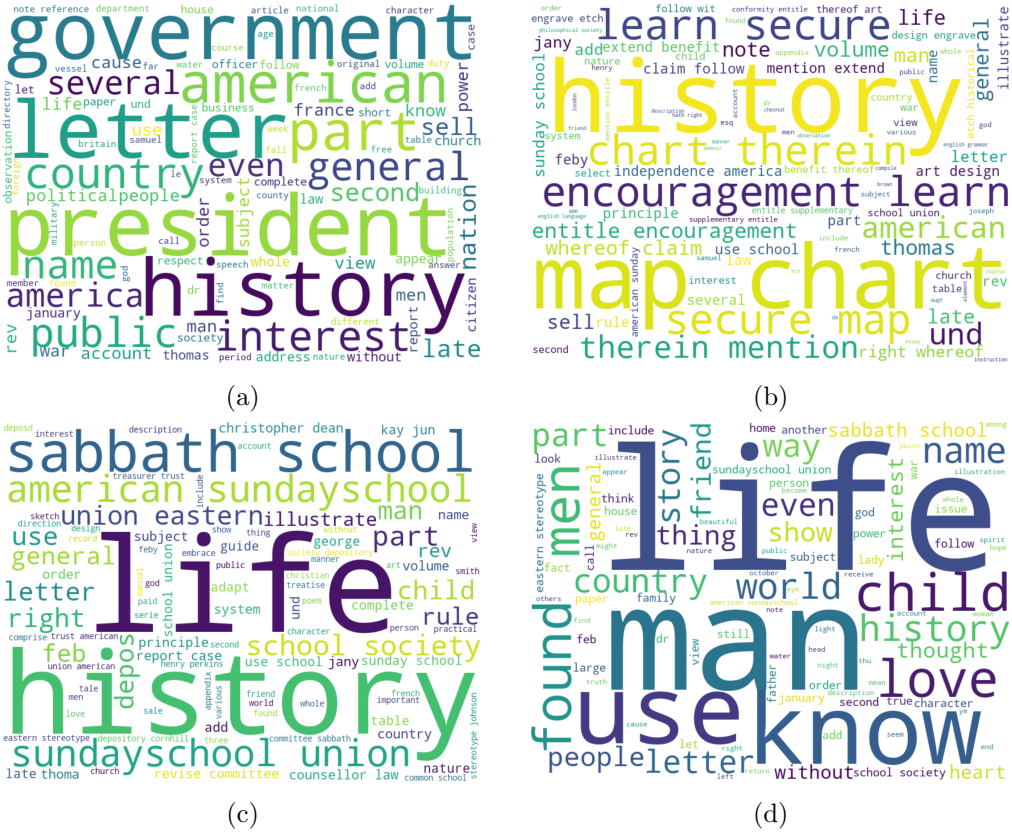



**Table 5:** Wordcloud plots of the corpus at different moments in time. Panels (a), (b), (c), and (d) refer respectively to time periods 1790-1810, 1810-1830, 1830-1850 and 1850-1870. The list of excluded stopwords is available in the code provided in the Appendix. The size of each word in the cloud is proportional to its frequency in the corpus of text.

The early focus of intellectual production in the US appears to be concerned with books relating to American history and accounts of the parliamentary discussions and political discourse of the time. However, as early as the period between 1810 and 1830, we can see that the subject matter changes noticeably. An increasing importance of the words "learn", "encouragement" and, in the following periods, "school" and "know" indicates that the spread of education in the US may have started well before the introduction of mass schooling towards the end of the 19th century. The appearance of the word "life" in the period after 1830 is itself interesting and calls for further investigation. While this could be related to the "life sciences", this is unlikely to be the case as the word "science" appears nowhere in these wordclouds. Most likely, the frequency of this word is the result of the growing number of people who feel the need to write their autobiography or the biography of key figures in American history. This is an interesting result as it indicates that people feel their lives, and those of the key figures in the history of their country, are worth recording and transmitting to future generations. This is highly significant as a marker of economic change as traditional societies where life is "nasty brutish and short" are, arguably, unlikely to be characterized by this kind of behavior.

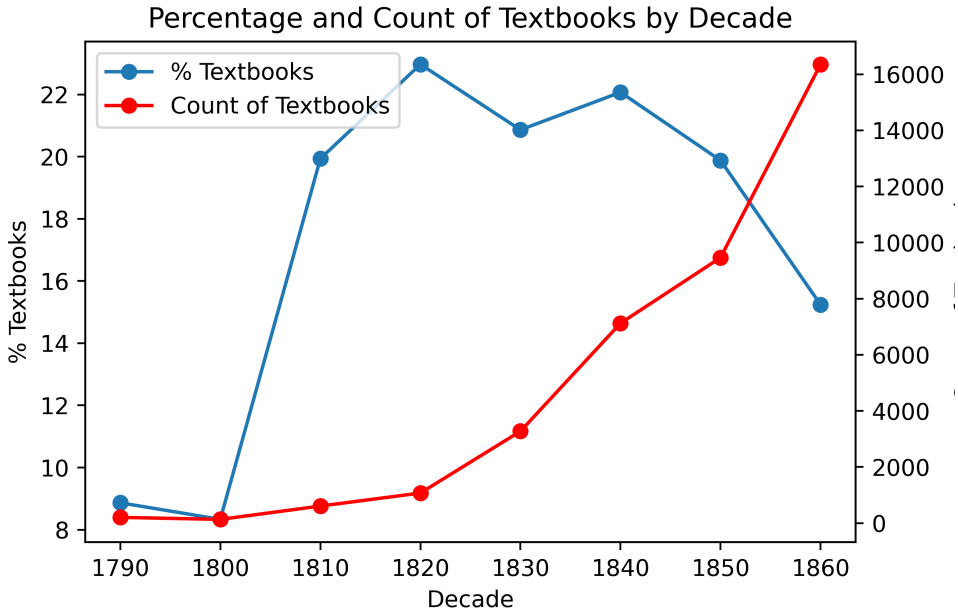



**Figure 6:** Total and relative frequency of works classified as "Textbooks" following the methodology laid out in Section 3. The code and list of keywords used to classify works is provided in the Appendix. Notice that these results need not be consistent with those presented in Table 3 as they come from a different classification procedure.

# 6 Empirical Methods

This section uses empirical models to quantify the relationship between pre-industrial copyright registrations, which I take as a proxy for the presence of a local knowledge elite, and economic development throughout the 19th and 20th centuries. We start by considering the following dynamic panel data model:

$$y_{i,t} = \alpha_i + \delta_t + \sum_{k=1}^{2} \rho_j y_{i,t-k} + \sum_{k=1}^{2} \beta_k \ln(Copyright)_{i,t-k} + \varepsilon_{i,t} \tag{4}$$

Where $y_{i,t}$ is an outcome variable measured at time $t$ in county $i$, $\alpha_i$ is a source of unobservable time-invariant heterogeneity, $\delta_t$ is a time fixed effect and $\ln(Copyright)_{i,t-k}$ is our variable of interest denoting human capital measured through copyright registrations, lagged up to $k$ periods to capture potential delays in the effect of the independent variable on the outcome. I use the log transformation throughout this Section as it helps deal with a highly skewed variable of interest and eases the interpretation of the coefficients. I also include $k$ lags of the dependent variable to absorb any persistence in the outcome variable over time. Although our data on copyright registrations has a yearly frequency, due to the fact that the census of population is taken at decade intervals, this means that 4 is estimated at the decade/county level.

The fact that the census questionnaire was changing over time also imposes some restrictions on the selection of outcome variables. The early censuses, up to and including 1810, contained very limited information. The only variables which are consistently available at the county level in all censuses are population and the number of slaves (for censuses prior to 1870). For censuses after 1820 this was complemented with the number people employed in manufacturing, the number of foreigners and, starting in 1840, with detailed information on schooling and other economic variables. Moreover, due to the fact that pre-1850 censuses are still not fully processed in IPUMS' data enclave, I am currently unable to match counties between 1850 and 1840 which means that we will need to restrict our attention to two sub-periods: 1790-1840 and 1850-1870. This is not ideal, but is easily remedied by adding county and state fips (or icpsr) codes to the pre 1850 data.[23] What is perhaps more frustrating is that, for the first sub-period, data on manufacturing, urbanization and population density is available starting in 1820 but cannot be accessed through the IPUMS data enclave directly.[24] Hence, for the period 1790 to 1840, I have no choice but to choose the log of population as the outcome variable. This is at least convenient as it allows me to

[23] IPUMS staff are working on this.
[24] This data needs to be uploaded separately, which requires approval from IPUMS staff.

exploit the full period as data on other outcome variables (such as urbanization) is generally not available before 1820.

We can estimate 4 using the Arellano-Bover/Blundell-Bond estimator including up to 3 lags $k = 3$ (Arellano & Bover, 1995; Blundell & Bond, 1998). In Table 6 we treat the log of copyright registrations as an endogenous variable, meaning that it is instrumented using its lagged levels and first differences. While this approach should deal with some of the endogeneity in the variable of interest, and the high p-values in the second order serial correlation tests are reassuring, we should still be guarded against a causal interpretation of these coefficients. This is especially the case as there is a clear relationship between the error term and the variable of interest given that potential spurious matching is more likely in highly populated areas. Hence, it is safest to stick to a descriptive interpretation which will tell us whether counties which are growing in population have higher or lower levels of human capital accumulation as measured by copyright registrations.

**Table 6:** Panel Data Regressions: Dependent variable is Log of Population

| | (1) | (2) | (3) |
|---|---|---|---|
| L1. Log Population | 0.672*** | 0.733*** | 0.541** |
| | (0.0423) | (0.169) | (0.236) |
| L2. Log Population | | 0.085 | 0.234 |
| | | (0.082) | (0.196) |
| L3. Log Population | | | -0.181* |
| | | | (0.110) |
| L0. Log Copyright | 0.357*** | 0.897*** | 0.135 |
| | (0.112) | (0.277) | (0.251) |
| L1. Log Copyright | 0.085 | -0.294 | 0.597 |
| | (0.152) | (0.437) | (0.547) |
| L2. Log Copyright | | -0.599** | -0.146 |
| | | (0.269) | (0.331) |
| L3. Log Copyright | | | 0.108 |
| | | | (0.296) |
| N Obs | 2264 | 1338 | 700 |
| N Groups | 917 | 638 | 373 |
| P > z (1) | 0.007 | 0.322 | 0.126 |
| P > z (2) | 0.328 | 0.228 | . |

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

**Note:** Dependent variable is log of population. Panel data regressions with county and time fixed effects estimated using the Arellano-Bover/Blundell-Bond estimator. Robust standard errors in parentheses. The reference period is 1790-1840 and observations are at the decade/county frequency. For variable definitions and sources see the Data Appendix.

As we can see from table 6, the model with 1 lag of the dependent variable already captures most of the autocorrelation in the residuals as can be seen from the p-value of the second order test of 0.328. This model, and those including more lags, show a robust positive association between copyright registrations and population over 1790-

1840. This is expected, but nonetheless reassuring. Turning to the second sample period 1850-1870, we now have a larger choice of dependent variables to understand the relationship between human capital accumulation and economic development. As we only have three periods, we will be able to include maximum one lag of the dependent variable. This is unfortunate, especially since three of our dependent variables, namely real property values, urbanization rates and population show coefficients of the lagged dependent variable which are larger than one when estimating models with only one lag. This is clearly evidence of misspecification, hence results from these models should be considered unreliable and are not shown. This leaves us with literacy rates and the labor share in non-farm activities.[25] Table 7 shows results estimating equation 4 with the literacy rate and labor share in non-farm activities as the dependent variables which can tell us, respectively, something about the relationship between knowledge production and broader measures of human capital accumulation and structural change.

**Table 7:** Panel Data Regressions: Literacy (1) and Non-Farm LS (2)

| | (1) | (2) |
|---|---|---|
| L1. Literacy | 0.0710 | |
| | (0.519) | |
| L1. Non-Farm Share | | 0.553 |
| | | (0.476) |
| L0. Log Copyright | 0.296** | 0.329 |
| | (0.118) | (0.312) |
| N Obs | 3216 | 3216 |
| N Groups | 1608 | 1608 |

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

**Note:** Dependent variable is the log of the population. Panel data regressions with county and time fixed effects estimated using the Arellano-Bover/Blundell-Bond estimator. Robust standard errors in parentheses. The reference period is 1850-1870, and observations are at the decade/county frequency. Notice that as $T = 3$, the Arellano-Bond test on the serial correlation of residuals (first-differenced) is unavailable here. For variable definitions and sources see the Data Appendix.

As we can see, copyright registrations are strongly correlated with the literacy rate over 1850-1870. The relationship with the labor-share of non-manufacturing activities appears to be insignificant over this time period. While somewhat surprising, the fact that the own lags of the dependent variables are insignificant may mean that the models in Table 7 are misspecified. Merging this time period with the pre-1850 data, where there is likely more variation in literacy and labor shares could perhaps lead to more reliable estimates and, with the availability of additional lags, the inclusion of more dependent variables.

We now turn to the relationship between pre-industrial human capital and economic development over the course of the Second Industrial Revolution and into the 20th

[25] This is not the same as the labor share of manufacturing, for which we would need to upload a separate data file to the IPUMS data enclave requiring administrative permission.

century. The basic empirical model we are interested is the same one Fiszbein (2022) which takes the following form:

$$y_{i,t} = \alpha_s + \beta \ln(Copyright_{1840-1860}) + \Gamma' X_{1850} + \varepsilon_{i,t} \tag{5}$$

Where $y_{i,t}$ is an outcome variable measured at time $t$, $\alpha_s$ is a state fixed effect, $\ln(Copyright_{1840-1860})$ is the variable of interest which is the logged number of copyright registrations between 1840 and 1860 in a given county[26] and $X$ is a vector of control variables. There are three types of controls considered: geographic controls which are unaffected by human interaction with the environment (these are potential yields, average rainfall and temperature, distance to oceans and the great lakes and terrain ruggedness), social controls which are potentially endogenous (these are log of population, urbanization, the fraction of foreigners and internal migrants and population density) and human capital controls which account for the potential correlation between knowledge production and other types of human capital which could have independent effects on economic development (these are literacy, and school enrollment rates). All social and human capital controls are measured in 1850. In addition to these controls, all regressions (unless stated otherwise) include a third degree polinomial of longitude and latitude to limit concerns relating to spatial autocorrelation (e.g. Kelly, 2020).[27] Furthermore, standard errors are calculated using the Conley correction (Conley, 1999) with a cutoff of 100km in all tables. This means we allow for dependence between observations within 100km for each other. Alternative estimates using state clustered standard errors are qualitatively equivalent. One advantage of this specification compared to 4 is that by keeping the variable of interest fixed in time, we can limit reverse causality concerns running from the dependent variable to the variable of interest. This is also a similar empirical setup to Squicciarini and Voigtländer (2015), Fiszbein (2022) and several other papers which study path dependence and persistence in economic outcomes.

The first outcome variable we consider is city growth in during the Second Industrial Revolution (1860-1940). Following a long tradition in the historical political economy literature, population growth - and urban population growth in particular - is often considered to be a fairly decent proxy of economic development (De Long & Shleifer, 1993). The presence of knowledge elites has already been found to predict city growth during industrialization in France by Squicciarini and Voigtländer (2015), who stress that while traditional measures of human capital such as literacy rates correlate with

[26] The period 1840-1860 is chosen as the only census which I have available to use outside of the Minnesota Population Center's remote desktop computers is the 1850 one. In future versions of this paper this variable should be the total over 1790-1850. However, given that copyright registrations only pick up speed after 1830 this should change the results by much.

[27] The results are not sensitive to changing the degree of the polynomial.

economic development in the cross-section, only upper-tail human capital predicts changes in the long-run. The results I show in Figure 7 and Tables 8 and 9 are consistent with these findings, showing that the intensity of copyrighting activity over 1840-1860 strongly predicts the growth in the share of people living in urban centers.

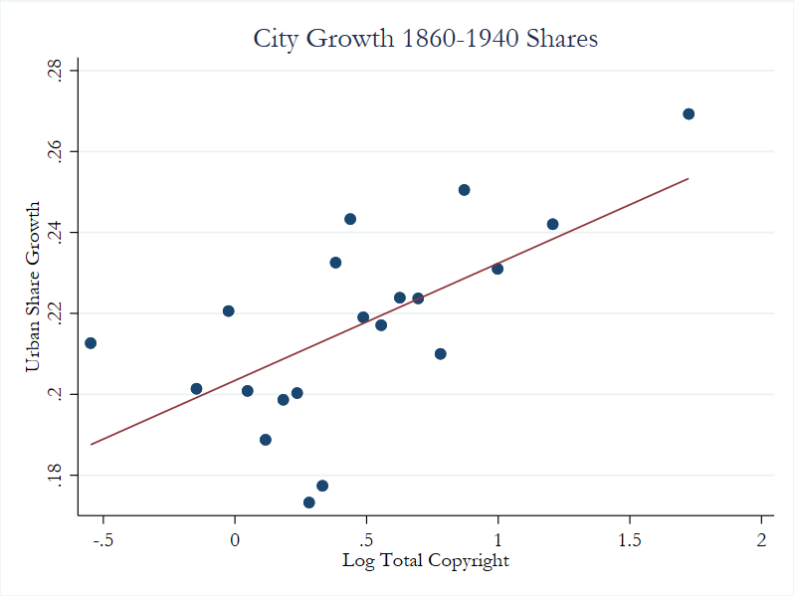



**Figure 7:** Binscatter plot of the growth of urban share population over 1860-1940 against the log of copyright activity between 1840-1860. Number of quantiles is 20 and all controls - social, human capital, geographic, the latitude-longitude polynomial and state fixed effects - are partialled out. For variable definitions and sources see the Data Appendix.

We now turn to the labor share of manufacturing and population density over 1850-2020, to understand the mechanisms through which pre-industrial knowledge elites influenced the development trajectories of counties. While population density is a general proxy for economic development both in pre-industrial and modern times, the labor share emphasizes the relationship between human capital and industrialization which is at the core the economic transformations occurring in the US and other Western countries in the end of the 19th century. To motivate this relationship I will show that copyrighting is a significant predictor of innovation as measured by per-capita patent rates over 1860-1940. Surprisingly, the variable of interest appears to be only weakly related to the Solow residual over 1850-1920.[28] This, however, may be due to heterogeneity in prices between counties as output per worker is measured in nominal terms. As data on county level prices is unavailable, I do not show these results here.

[28] The Solow residual is calculated as an OLS residual of a regression of log output per worker on log capital per worker in manufacturing.

**Table 8:** IV Regressions: Urbanization Share Growth 1860-1940

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Log Copyright | 0.328* | 0.365* | 0.358* | 0.358* | -2.832 |
| | (2.38) | (2.45) | (2.13) | (2.13) | (-0.05) |
| % Urbanized 1850 | -1.200*** | -1.279*** | -1.260** | -1.260** | 1.954 |
| | (-3.44) | (-3.62) | (-3.12) | (-3.12) | (0.04) |
| % In School 1850 | | | 0.00831 | 0.00831 | -0.127 |
| | | | (0.23) | (0.23) | (-0.04) |
| % Literate 1850 | | | 0.0678 | 0.0678 | -0.698 |
| | | | (0.26) | (0.26) | (-0.04) |
| Log Population 1850 | | | | | 0.561 |
| | | | | | (0.06) |
| N | 1472 | 1467 | 1466 | 1466 | 1465 |
| Kleibergen-Paap F | 16 | 17 | 13 | 13 | 1 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses

Dependent variable is city population growth in levels over 1860-1940. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As we can see in Table 10, even in the most conservative specifications OLS estimates show a significant relationship between knowledge production in the pre-industrial period and structural change. In all tables, the coefficients shown are a selection of the complete set of 20 control variables included in the fully specified models. The point estimates in Table 10 peak around 1920 and subsequently decline, following the hump-shaped pattern of manufacturing shares exhibited by most developed countries over this time period. Controls are often insignificant (e.g. urbanization) or enter the regression with the wrong sign (e.g. population density and schooling enrollment rates). No variable in 5 predicts the structural change pattern as well as the log of copyright registrations over 1840-1860. While these results suggest the presence of a relationship between pre-industrial knowledge elite presence and industrialization, we may still be capturing spurious selection effects. As we are controlling for a significant number of observable characteristics of counties, including several controls which may be potentially endogenous to the variable of interest, we can interpret the estimates in Table 10 as being fairly conservative. Nonetheless, counties with higher levels of knowledge production in pre-industrial periods may be different in terms of unobservable characteristics. The main concern is that, given the high rates of internal migration,

**Table 9:** OLS Regressions: Urbanization Share Growth 1860-1940.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Log Copyright | 0.0654*** | 0.0617*** | 0.0548*** | 0.0548*** | 0.0169 |
| | (4.45) | (4.36) | (3.92) | (3.92) | (1.11) |
| % Urbanized 1850 | -0.503*** | -0.504*** | -0.482*** | -0.482*** | -0.712*** |
| | (-7.98) | (-7.66) | (-7.51) | (-7.51) | (-8.02) |
| % In School 1850 | | | 0.0673** | 0.0673** | 0.0285 |
| | | | (2.98) | (2.98) | (1.31) |
| % Literate 1850 | | | 0.386* | 0.386* | 0.202 |
| | | | (2.20) | (2.20) | (1.16) |
| Log Population 1850 | | | | | 0.0431*** |
| | | | | | (3.37) |
| N | 1472 | 1467 | 1466 | 1466 | 1465 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses

Dependent variable is city population growth in levels over 1860-1940. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

authors and other knowledge producers select locations which have better chances of industrializing at the end of the 19th century. This is partially addressed by controlling for urbanization and population density in 1850, however, these variables may still imperfectly capture the factors driving selection and hence only partially address the problem.

An instrumental variable strategy is therefore needed to isolate the effect of knowledge production on industrialization which is not driven by pre-existing characteristics operating through selection mechanisms. Table 11 addresses this by using the shift-share level of predicted knowledge production described in Section 4. This instrument exploits the fact that when migrating authors tend to choose locations which have relatively larger settlements of individuals from their home state. Identification here comes from the fact that any factor which is potentially attractive to authors, and hence introduces selection bias in our coefficients, is likely to cut across state boundaries. It is challenging to think of a selection mechanism which operates only for individuals from Connecticut and not for those from New York and simultaneously affects long-term development trajectories. It may however be true that individuals from Connecticut are different from individuals from New York in dimensions which go beyond the relative abundance of authors in New York and relative scarcity in Con-

**Table 10:** OLS Regressions: Labor Share all Controls

| | (1) 1850 | (2) 1870 | (3) 1920 | (4) 1970 | (5) 2000 |
|---|---|---|---|---|---|
| Log Copyright | 0.00742*** | 0.00927*** | 0.0111*** | -0.00324 | -0.00239 |
| | (5.65) | (3.57) | (3.50) | (-1.06) | (-0.51) |
| % In School 1850 | -0.0104** | 0.00651 | 0.0160** | 0.0304*** | 0.00853 |
| | (-2.85) | (0.57) | (2.69) | (4.38) | (1.13) |
| % Literate | -0.0114 | 0.0556 | 0.0601 | 0.0460 | 0.0241 |
| | (-0.65) | (1.18) | (1.58) | (1.73) | (0.67) |
| % Urbanized | 0.0218 | -0.00409 | -0.0814*** | -0.0602** | -0.00836 |
| | (1.06) | (-0.18) | (-3.51) | (-2.63) | (-0.47) |
| Log Population 1850 | 0.000840 | -0.00419 | 0.00127 | 0.00103 | 0.00939** |
| | (0.34) | (-0.66) | (1.09) | (0.48) | (2.82) |
| N | 1114 | 1504 | 1504 | 1504 | 1504 |
| R2 | 0.271 | 0.193 | 0.204 | 0.221 | 0.129 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | yes | yes | yes | yes | yes |
| HC Controls | yes | yes | yes | yes | yes |
| Social Controls | yes | yes | yes | yes | yes |

$t$ statistics in parentheses

Dependent variable is the labor share of manufacturing 1850-2000. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

necticut. To address this we keep controlling for all the human capital controls and urbanization in 1850. We are however forced to drop the rest of the social controls as partialling them out in the first stage leads to a weak instrument problem which generates imprecise coefficients (although the sign of point estimates is stable). The controls which are driving this problem are the fraction of internal migrants and foreigners, which is unsurprising given that they have an explicit relationship with the instrument.

As we can see from Table 11, point estimates are larger in the IV specification and broadly follow the same pattern. The instrument is strongly relevant as can be seen from the large Kleibergen-Paap statistics which are generally above 25.[29] The human capital and urbanization controls enter the regression with the wrong sign and are generally not significant. It is important to comment on the relative magnitudes of the IV and OLS coefficients as large differences between these parameter estimates

[29] However, if we add the rest of the social controls (which are population and the share of foreigners and internal migrants) the association between the instrument and the variable of interest vanishes leading to large standard errors.

**Table 11:** IV Regressions: Labor Share all Controls

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | 1850 | 1870 | 1920 | 1970 | 2000 |
| Log Copyright | 0.0630*** | 0.0360 | 0.185*** | 0.102 | 0.0350 |
| | (16.47) | (0.44) | (3.88) | (1.96) | (0.81) |
| % In School 1850 | -0.0234*** | 0.00458 | -0.00765 | 0.0174 | 0.0101 |
| | (-5.53) | (0.19) | (-0.73) | (1.54) | (1.15) |
| % Literate | -0.0520** | 0.0328 | -0.102 | -0.0548 | -0.00929 |
| | (-3.05) | (0.28) | (-1.38) | (-0.70) | (-0.12) |
| % Urbanized | -0.0597* | -0.0119 | -0.401*** | -0.253 | -0.0781 |
| | (-2.15) | (-0.06) | (-3.42) | (-1.83) | (-0.68) |
| N | 1114 | 1505 | 1505 | 1505 | 1505 |
| Kleibergen-Paap | 23 | 13 | 13 | 13 | 13 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | yes | yes | yes | yes | yes |
| HC Controls | yes | yes | yes | yes | yes |
| Social Controls | no | no | no | no | no |

$t$ statistics in parentheses

Dependent variable is the labor share of manufacturing 1850-2000. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

can sometimes indicate issues of instrument endogeneity (Oster, 2019). Selection on unobservables would if anything lead to upward bias in the OLS coefficients, which is exactly the opposite of what we observe. Hence, the likely source of downward bias in the OLS coefficients is due to measurement error in the variable of interest. This measurement error has two sources. Firstly, the variable we are measuring (total copyrighting at the county level) is an imperfect proxy for the general level of knowledge production. While a large component of this is likely to be unobservable in copyright records, the instrument can potentially correct for this by considering the variation in copyrighting activity by states. As long as individuals from states which produce more copyrighting also produce, on average, more unobservable knowledge the instrument can address this issue. Secondly, the measurement of copyrighting activity by county is itself subject to error due to potential false matches and failed extraction of author names from the title pages. Monte-Carlo experiments can show us that these two sources of measurement error can generate substantial levels of downward bias in OLS coefficients (see Appendix Figures 13a and 13b).

To better visualize the strong relationship between pre-industrial human capital and structural change, Figure 8 shows a coefficient plot of the variable of interest over time using the preferred specification, which includes all except social controls (but including urbanization). We can see that, while there are large differences between the

OLS and IV coefficients, these estimates follow the same patterns and adding more years does not overturn the results in Tables 10 and 11. Figure 9 applies the same procedure to population density, showing a robust and increasing positive association with pre-industrial knowledge production. Agricultural diversity, which was argued to have exerted a determining role in the US' structural change in a recent paper (Fiszbein, 2022), appears to have no explanatory power once pre-industrial human capital measured through copyrighting activity is taken into account (see Figure 12 in the Appendix).

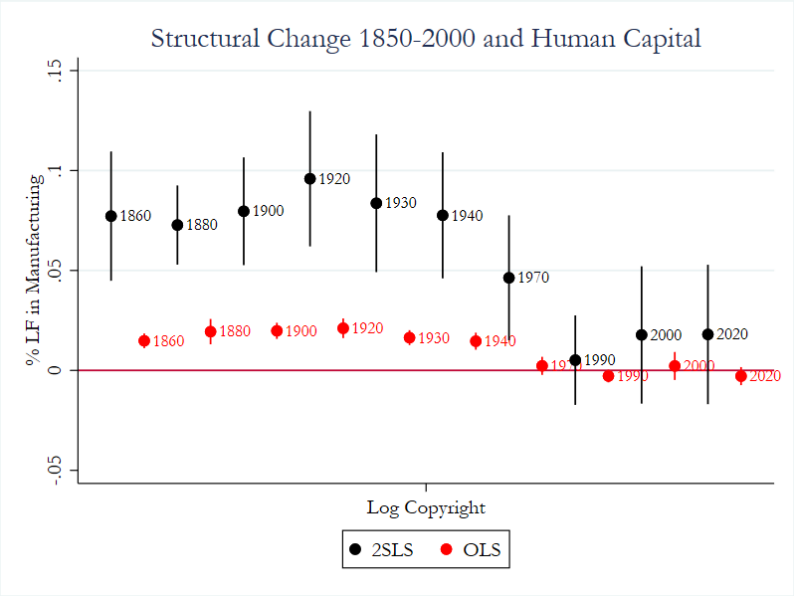



**Figure 8:** Point estimates with 95% confidence intervals of the effect of copyrighting in 1840-1860 on the labor share of manufacturing over 1850-2020.

As we can see from Figure 9, knowledge production in the pre-industrial period shows an enduring relationship with economic development. This is also evident in regressions where the outcome variable is income per-capita in 2000 as we can see in Table 12 and Figure 10. As before, when we include the full set of social controls the estimated effect of knowledge production on income per capita turns insignificant. This is due to the fact that the strength of the instrument in predicting the variable of interest is obliterated once the social controls (especially the share of foreigners and internal migrants) are introduced. While this is inconvenient, it does not necessarily invalidate our empirical strategy unless we have a strong reason to believe that these additional controls had their own independent impact on subsequent economic development. While this may be true, at least our OLS estimates always show that the variable of interest remains almost always significant even in the fully specified models.

An advantage of using title pages instead of other primary sources, such as encyclopedia subscriptions (Squicciarini & Voigtländer, 2015), to identify the knowledge elite,

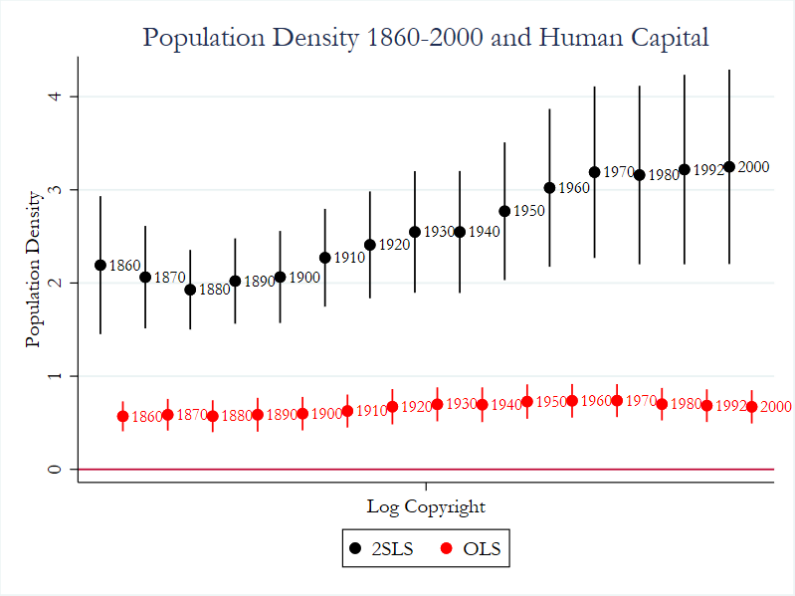



**Figure 9:** Point estimates with 95% confidence intervals of the effect of copyrighting in 1840-1860 on population density over 1860-2020.

is that the title page can give us information on the content of publications which provides insights on the types, and not only quantities, of knowledge being produced. The information available is of course limited to the title of the book, which provides a challenge when applying topic classification models as these generally perform poorly on short texts. This is why, for now, I have opted for a binary classifier separating between academic and non-academic publications. As discussed in Section 4, an Academic publication here comprises anything which serves the purpose of diffusing or advancing knowledge. As can be gleaned from Table 5 and a cursory glance at the classified primary data, a large portion of these works are Sunday School teaching manuals used by educators to teach English and other subjects to students. Another class of works consists in publications by actual academics and luminaries of the times relating to subject matters ranging from astronomy to classical literature and philosophy. Admittedly, the connection between these works and innovations fostering productivity is not always clear. While some publications in the academic category are certainly relevant to the growth of knowledge which spurred innovation and adaptation of foreign technology during the Second Industrial Revolution (see the discussion in Section 2), these are certainly a minority. However, this should not be taken to mean that the large uptake in copyrighting shown in Figures 1 and 3 is of no significance. As noted by Baten and Van Zanden (2008), the production of books, regardless of their content, can itself indicate changes in the tastes and attitudes toward knowledge of the average consumer.

**Table 12:** IV Regressions: Income p.c. in 2000

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Log Copyright | 0.417*** | 0.421*** | 0.406*** | 0.782** | 7.988 |
| | (6.12) | (5.82) | (5.14) | (3.24) | (0.12) |
| % In School 1850 | | | 0.0319 | -0.0372 | 0.597 |
| | | | (1.22) | (-0.79) | (0.14) |
| % Literate | | | 0.149 | -0.286 | 2.467 |
| | | | (1.16) | (-1.01) | (0.14) |
| % Urbanized | | | | -1.696** | -7.826 |
| | | | | (-3.19) | (-0.12) |
| Log Population 1850 | | | | | -1.432 |
| | | | | | (-0.12) |
| N | 1512 | 1507 | 1505 | 1505 | 1504 |
| Kleibergen-Paap F | 24 | 28 | 22 | 13 | 1 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses

Dependent variable is log income per capita in 2000. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This is supported by the results shown in Table 14 which shows that, in fully specified models, non-academic works are significant predictors of the growth in the manufacturing sector while having lower point estimates. A possible explanation for this is that, while academic publications are highly clustered in certain areas (especially the Northeastern states), the production of non-academic works in the rest of the country is a marker for the general attitudes towards human capital investments. This is consistent with the evidence that counties with higher levels of copyright registrations in the pre-industrial period show higher patenting rates throughout the wave of industrialization occurring at the end of the 19th and early 20th century. Tables 16 and 15 show that patenting rates over 1860-1940 are significantly higher in counties with more copyright registrations over 1840-1860.

**Table 13:** OLS Regressions: Income p.c. in 2000

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Log Copyright | 0.0861*** | 0.0704*** | 0.0629*** | 0.0500*** | 0.0135 |
| | (7.09) | (5.43) | (5.13) | (4.42) | (1.04) |
| % In School 1850 | | | 0.0930*** | 0.0955*** | 0.0783*** |
| | | | (3.44) | (3.45) | (3.66) |
| % Literate | | | 0.500*** | 0.518*** | 0.344*** |
| | | | (8.40) | (9.51) | (5.64) |
| % Urbanized | | | | 0.180* | -0.180* |
| | | | | (2.33) | (-2.12) |
| Log Population 1850 | | | | | 0.0122 |
| | | | | | (1.63) |
| N | 1512 | 1507 | 1505 | 1505 | 1504 |
| R2 | 0.0877 | 0.155 | 0.178 | 0.183 | 0.232 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses

Dependent variable is log income per capita in 2000. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 14:** OLS Regressions: Academic/Non-Academic

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | 1860 | 1870 | 1920 | 1970 | 2000 |
| Log Science | 0.00899*** | 0.00785** | -0.00307 | -0.00264 | 0.00708 |
| | (3.34) | (2.10) | (-0.43) | (-0.45) | (1.65) |
| Log Culture | 0.00670*** | 0.00873*** | 0.0105*** | -0.00254 | -0.00246 |
| | (4.97) | (3.50) | (3.25) | (-0.80) | (-0.54) |
| N | 1504 | 1504 | 1504 | 1504 | 1504 |
| R2 | 0.295 | 0.195 | 0.203 | 0.221 | 0.130 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | yes | yes | yes | yes | yes |
| HC Controls | yes | yes | yes | yes | yes |
| Social Controls | yes | yes | yes | yes | yes |

$t$ statistics in parentheses

Dependent variable is the labor share of manufacturing 1850-2000. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

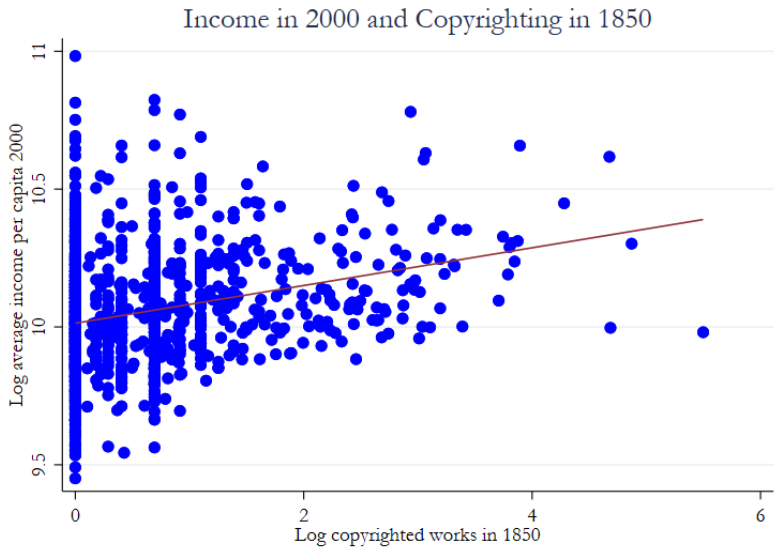



**Figure 10:** Scatterplot of log income per capita in 2000 and log copyright registrations in 1840-1860.

**Table 15:** IV Regressions: Patents 1860-1940

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Log Copyright | 1.448*** | 1.416*** | 1.344*** | 2.353*** | 53.89 |
| | (4.48) | (4.54) | (4.51) | (3.59) | (0.10) |
| % In School 1850 | | | 0.407** | 0.220 | 3.750 |
| | | | (3.00) | (1.52) | (0.11) |
| % Literate | | | 1.630* | 0.463 | 17.00 |
| | | | (2.35) | (0.42) | (0.11) |
| % Urbanized | | | | -4.550** | -52.05 |
| | | | | (-2.97) | (-0.10) |
| Log Population 1850 | | | | | -9.626 |
| | | | | | (-0.10) |
| N | 1506 | 1501 | 1499 | 1499 | 1498 |
| Kleibergen-Paap F | 24 | 28 | 22 | 14 | 1 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses

Dependent variable is the log of patents per-capita/decade over 1860-1940. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

**Table 16:** OLS Regressions: Patents 1860-1940

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Log Copyright | 0.486*** | 0.441*** | 0.400*** | 0.360*** | 0.195*** |
| | (9.30) | (12.28) | (10.55) | (7.47) | (3.40) |
| % In School 1850 | | | 0.578*** | 0.585*** | 0.398*** |
| | | | (4.29) | (4.33) | (3.39) |
| % Literate | | | 2.592*** | 2.650*** | 2.195*** |
| | | | (4.74) | (4.71) | (4.20) |
| % Urbanized | | | | 0.557 | -1.098 |
| | | | | (1.07) | (-1.78) |
| Log Population 1850 | | | | | 0.0856* |
| | | | | | (2.45) |
| N | 1506 | 1501 | 1499 | 1499 | 1498 |
| R2 | 0.104 | 0.147 | 0.179 | 0.181 | 0.223 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses

Dependent variable is the log of patents per-capita/decade over 1860-1940. Conley standard errors used with a 100km cutoff. For variable definitions and sources see the Data Appendix.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

# 7 Theoretical Model

## 7.1 Motivation

The reduced form results I have shown so far seem to suggest that there is a robust association between the intensity of copyright registrations in the pre-industrial period and subsequent economic development. Crucially, this relationship appears to reflect characteristics of the communities of people living in particular counties as opposed to an effect of the economic environment or geographic factors on them. Counties which received influxes of individuals from states which had higher levels of copyrighting activity in turn show higher levels of copyrighting and higher schooling provision. Furthermore, this finding is not specific to academic publications which purportedly advance the knowledge frontier and we would expect to raise the returns to schooling. This suggests that, contrary to endogenous growth models of the form of Galor (2011), Voigtländer and Voth (2006) and Jones (2001) which consider countries as a whole to be the relevant unit of analysis, the spatial distribution of knowledge production in pre-industrial times matters for subsequent economic development. As most of our empirical models in Section 6 include urbanization in 1850 as a control variable, we can hypothesize that these results do not merely capture agglomeration economies due to the fact that authors live in more urbanized areas.

To rationalize this result, a different kind of model of the transition from the Malthusian world to sustained economic growth (these models are also called "unified" growth models) is needed. As the model is not yet calibrated, we cannot use it to perform counterfactual analysis. However, the central intuitions are clear from observing two transition paths which will show that an initial difference in the composition of a county's settlers can have long lasting effects on productivity growth and the speed of the transition away from the Malthusian equilibrium. The key difference with other unified growth models along the lines of Galor (2011) is that in this model we have heterogeneous agents of two types (one "modern" type which possesses technology for skill accumulation and the other "traditional" type does not) and cultural dynamics between the types à la Bisin and Verdier (2001). The potential for oblique socialization will mean that low skill types in environments dominated by high skill types are more likely to transition to the high skill type. This yields the desired result that the initial composition of the population has a persistent effect on the speed of transition out of the Malthusian equilibrium, although initial conditions do not affect the long-run steady state. If agglomeration economies were introduced, this effect would be compounded and convergence between two regions with different initial compositions of population would, likely, not occur.

## 7.2 Household Problems

There are two cultural types (traditional and modern) which differ in their preferences over human capital and the technology they use in production. All agents live for one period and make choices relating to fertility, consumption, human capital investments and socialization investments.[30] As the model is already quite involved, the production side is bare bones. Each household produces as an independent unit with no outside hiring and the only difference between modern and traditional households is that the former can use skills in production whereas the latter only use labor and land. The traditional type population share is $q$ and they are characterized by the following representative household problem:

$$\begin{aligned} V_t(S,q) = \max_{\{c,\tau^q,e,n\}} \{ & u^t(c,n) \\ & + \beta[\pi_{mm}(\tau^q,q)V_t(S',q') + (1-\pi_{mm}(\tau^q,q))V_m(S',q')]\} \end{aligned} \tag{6}$$

such that:

$$\begin{aligned} c &= w_t l_t \\ l_t + n(\bar{\tau} + \tau^q + \tau^e(S)e) &= 1 \\ S' &= f(S,e) + \delta S \end{aligned}$$

Where the index $t$ refers to the "traditional" type and primes refer to the children's values of a given variable. The rest is standard: $c$ is consumption, $n$ is the number of children, $\beta$ is the discount factor, $\bar{\tau}$ is the fixed time cost associated with raising a child, $\tau^q$ is the socialization effort, $e$ is the time investment in education, $w_t$ is the wage of the traditional type and $\delta$ is the depreciation rate of skills.[31] Furthermore, assume that the skill accumulation function $f(e,S)$ is increasing in both arguments and strictly concave. Finally, $\pi_{ij}$ is the probability that parent of type $i$ socializes her child to type $j$. These probabilities follow the logic in Bisin and Verdier, 2001 with cultural substitution:

[30] Following the literature on the cultural transmission of preferences, socialization investments are done by controlling the probability that an agent's child(ren) remain of the same type as them.

[31] The wage can be though of as coming from a firm maximization problem with the following production function $F(L,X) = (AX)^{\alpha}L^{1-\alpha}$ where $L$ is the labor supplied by the traditional household.

$$\pi_{tt} = \tau^q + (1 - \tau_t^q)q$$
$$\pi_{tm} = (1 - \tau^q)(1 - q)$$

Where the parent controls the probability of direct socialization $\tau^q$ and the oblique socialization is linear in $q$. The cultural dynamics will satisfy:

$$q' = \frac{qn_t(\tau_t^q + (1 - \tau_t^q)q) + (1 - q)n_m(1 - \tau_m^q)q}{(1 - q)n_m + qn_t} \tag{7}$$

The proportion of $q$ types evolves according to a convex combination of fertility rates of the two types, weighted by the respective probabilities of transitioning to the traditional type.

Notice that while the traditional type does not gain any return from skill accumulation in production, there is still an incentive to invest in education due to the possibility of oblique socialization. The modern agent on the other hand, faces an income profile which is elastic to skills $S$. Their corresponding problem is:

$$\begin{aligned} V_m(S, q) = \max_{\{c, \tau^q, e, n\}} \{ & u^m(c, n, S) \\ & + \beta[\pi_{mm}(\tau^q, q)V_m(S', q') + (1 - \pi_{mm}(\tau^q, q))V_t(S', q')]\} \end{aligned} \tag{8}$$

such that:

$$\begin{aligned} c &= w_m(S)l_m \\ l_m + n(\bar{\tau} + \tau^q + \tau^e(S)e) &= 1 \\ S' &= f(S, e) + \delta S \end{aligned}$$

Where the index $m$ now refers to the "modern" type. The difference in the types is hence layered over preferences and technology with the modern type having a preference for skills $S$ in their utility function and being able to use them in production. Utilities are increasing and strictly concave in all arguments for both types. In order to numerically solve the model, I assume the following functional forms, where $\bar{c}$ is a subsistence constraint and $\nu$ controls the importance of parental skills in children's skill development.

$$
\begin{aligned}
u^m(c, n, S) &= \chi \ln(c - \bar{c}) + \ln(S) + \ln(n) \\
u^t(c, n) &= \chi \ln(c - \bar{c}) + \ln(n) \\
\tau^e(S) &= \frac{1}{S} \\
f(S, e) &= S^{\nu} e^{1-\nu}
\end{aligned}
$$

## 7.3 Simulation

The model is numerically solved by value function iteration using a grid search method. As there are four choice variables, this can be computationally expensive, which is why I choose small grids which generate the kinks in the transition path shown in Figure 11. To solve the model I take an arbitrary parametrization which delivers interior solutions. The solution to the household problems laid out in Section 7.2 is four policy functions, one for each of the choice variables, which only depend on the states (skills $S$ and the fraction of the traditional type $q$). Using these policy functions, we can simulate the transition paths for two economies which receive an influx of population. One of these is ex-post abundant in the modern type ($q = 0.1$), which has the ability to use skills in production, and the other is ex-post abundant in the traditional type ($q = 0.9$). Prior to the arrival of the modern type the economy is fully Malthusian, meaning that income per capita is constant and exogenous technological progress is fully offset by fertility decisions.[32]

[32] This can be easily shown although I refrain from doing so here.

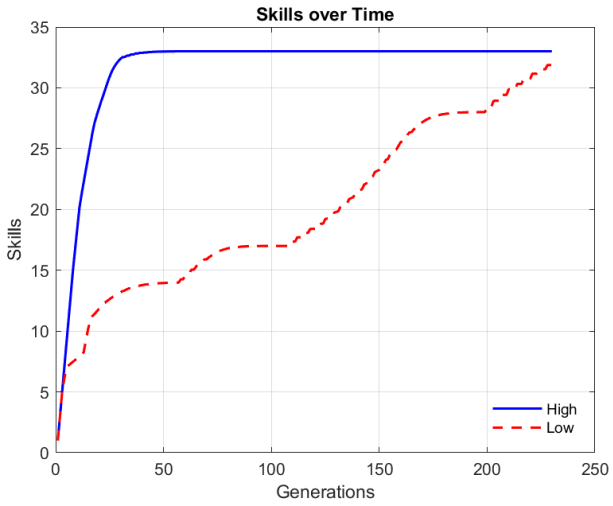



**Figure 11:** Transition path of skills for high ($q = 0.1$) and low ($q = 0.9$) influx of the modern type in two economies.

Once the modern type arrives the economy transitions to a new steady state with higher per-capita consumption and higher skills. However, we can see in Figure 11 that convergence to this new steady state can be extremely slow, especially for the economy which receives a smaller fraction of the modern type. The intuition behind this result is that the incentives for skill accumulation are held back by the abundance of the traditional type in the "low" economy. Modern parents face a more hostile environment which induces them to invest more in $\tau^q$ to ensure that their children do not lose their ability to use skills in production and less in $e$ which raises their skill level. The externality that comes from being surrounded by the modern type, which makes the probability that a child of a modern parent remains modern - exogenously - higher, can have a large long-lasting effects. In this parametrization, convergence only occurs in 200 generations (approximately 7000 years). Traditional unified growth models with representative agents will miss this effect that the environment has on optimal human capital investment decisions, which are generally an increasing function of the rate of technological progress which is an increasing function of population growth (Galor, 2011).

# 8 Conclusion

This paper is the first to empirically assess the early intellectual production of the United States using copyright registration data. Constructing time series of book production at the national level reveals that authorship of books and other copyrighted materials began a sustained upward trend in 1830, which is well before the start of rapid industrialization. Assigning copyright registrations to locations using the de-

classified 1850 census shows that the spatial distribution of knowledge production in pre-industrial times is strongly correlated with industrial development and population density over time. Exploiting the large internal migration which occurred at this time as a result of the Westward movement, I attempt to obtain exogenous variation of "knowledge production" at the county level. IV coefficients obtained using a shift-share type instrument are statistically significant and larger than their OLS counterparts. To rationalize these results I consider a variation of the standard unified growth models developed by Galor (2011) where cultural transmission of preferences for human capital and technology can generate a long-lasting divergence in skill accumulation between regions which are ex-ante identical except for the relative scarcity of skill-using individuals. More work is needed to consider how this model can be brought to the data to perform counterfactual exercises which would illustrate how settlement patterns from centuries ago may still affect the dispersion of skills and other economic outcomes as a result of the intergenerational transmission of technology and preferences for human capital.

# References


Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, *102*(5), 1832–1856.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, *122*(3), 467–506.

Abramovitz, M., & David, P. (1996). American macroeconomic growth in the era of knowledge based progress: The long run perspective. In S. Engerman & R. Gallman (Eds.), *The cambridge economic history of the united states* (pp. 1–93). Cambridge University Press.

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of economic literature*, *40*(1), 7–72.

Acemoglu, D. (2003). Factor prices and technical change: From induced innovations to recent debates. *Knowledge, information and expectations in modern macroeconomics*, 464–491.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American economic review*, *91*(5), 1369–1401.

Ager, P., & Brückner, M. (2013). Cultural diversity and economic growth: Evidence from the us during the age of mass migration. *European Economic Review*, *64*, 76–97.

Alesina, A., & La Ferrara, E. (2005). Ethnic diversity and economic performance. *Journal of economic literature*, *43*(3), 762–800.

Allen, R. C. (2009). *The british industrial revolution in global perspective*. Cambridge University Press.

Arellano, M., & Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of econometrics*, *68*(1), 29–51.

Atack, J., Bateman, F., & Parker, W. N. (2008). Northern agriculture and the westward movement. In S. Engerman & R. Gallman (Eds.), *The cambridge economic history of the united states* (pp. 285–329). Cambridge University Press.

Bartel, A. P. (1989). Where do the new us immigrants live? *Journal of labor economics*, *7*(4), 371–391.

Baten, J., & Van Zanden, J. L. (2008). Book production and the onset of modern economic growth. *Journal of Economic Growth*, 217–235.

Bazzi, S., Ferrara, A., Fiszbein, M., Pearson, T., & A. Testa, P. (2022). Sundown towns and racial exclusion: The southern white diaspora and the "great retreat". *AEA Papers and Proceedings*, *112*, 234–238.

Bazzi, S., Ferrara, A., Fiszbein, M., Pearson, T. P., & Testa, P. A. (2021). The other great migration: Southern whites and the new right.

Bazzi, S., Fiszbein, M., & Gebresilasse, M. (2020). Frontier culture: The roots and persistence of "rugged individualism" in the united states. *Econometrica*, *88*(6), 2329–2368.

Bazzi, S., Fiszbein, M., & Gebresilasse, M. (2021). "rugged individualism" and collective (in) action during the covid-19 pandemic. *Journal of Public Economics*, *195*, 104357.

Becker, S. O., & Woessmann, L. (2009). Was weber wrong? a human capital theory of protestant economic history. *The quarterly journal of economics*, *124*(2), 531–596.

Bisin, A., & Verdier, T. (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic theory*, *97*(2), 298–319.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, *87*(1), 115–143.

Cantoni, D. (2015). The economic effects of the protestant reformation: Testing the weber hypothesis in the german lands. *Journal of the European Economic Association*, *13*(4), 561–598.

Cantoni, D., Dittmar, J., & Yuchtman, N. (2018). Religious competition and reallocation: The political economy of secularization in the protestant reformation. *The Quarterly Journal of Economics*, *133*(4), 2037–2096.

Chaney, E. (2016). Religion and the rise and fall of islamic science. *Work. Pap., Dep. Econ., Harvard Univ., Cambridge, MA*.

Clark, G. (2008). A farewell to alms. In *A farewell to alms*. Princeton University Press.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of econometrics*, *92*(1), 1–45.

Cozzens, P. (2016). *The earth is weeping: The epic story of the indian wars for the american west*. Vintage.

Davis, J. H. (2004). An annual index of us industrial production, 1790–1915. *The Quarterly Journal of Economics*, *119*(4), 1177–1215.

De Long, J. B., & Shleifer, A. (1993). Princes and merchants: European city growth before the industrial revolution. *The Journal of Law and Economics*, *36*(2), 671–702.

Dittmar, J. E. (2011). Information technology and economic change: The impact of the printing press. *The Quarterly Journal of Economics*, *126*(3), 1133–1172.

Doepke, M., & Zilibotti, F. (2008). Occupational choice and the spirit of capitalism. *The Quarterly Journal of Economics*, *123*(2), 747–793.

Fernandez, R., & Rogerson, R. (1996). Income distribution, communities, and the quality of public education. *The Quarterly Journal of Economics*, *111*(1), 135–164.

Fernandez, R., & Rogerson, R. (1998). Public education and income distribution: A dynamic quantitative evaluation of education-finance reform. *American Economic Review*, 813–833.

Ferrie, J. P. (1999). *Yankeys now: Immigrants in the antebellum us 1840-1860*. Oxford University Press, USA.

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

Gallman, R. E. (2000). Economic growth and structural change in the long nineteenth century. *The Cambridge economic history of the United States*, *2*, 1–55.

Galor, O. (2011). *Unified growth theory*. Princeton University Press.

Galor, O., & Moav, O. (2002). Natural selection and the origin of economic growth. *The Quarterly Journal of Economics*, *117*(4), 1133–1191.

Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth? *Journal of economic Growth*, *9*, 271–303.

Goldin, C. D. (2016). Human capital.

Haines, M. R., for Political, I.-u. C., & Research, S. (2010). Historical, demographic, economic, and social data: The united states, 1790-2002. https://doi.org/10.3886/ICPSR02896.v3

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169–15211.

Jones, C. I. (2001). Was an industrial revolution inevitable? economic growth over the very long run. *The BE Journal of Macroeconomics*, *1*(2).

Jones, C. I. (2022). The past and future of economic growth: A semi-endogenous perspective. *Annual Review of Economics*, *14*, 125–152.

Kelly, M. (2020). Understanding persistence.

Kelly, M., Mokyr, J., & Gráda, C. Ó. (2014). Precocious albion: A new interpretation of the british industrial revolution. *Annu. Rev. Econ.*, *6*(1), 363–389.

Khan, B. Z. (2005). *The democratization of invention: Patents and copyrights in american economic development, 1790-1920*. Cambridge University Press.

Khan, B. Z., & Sokoloff, K. L. (1993). "schemes of practical utility": Entrepreneurship and innovation among "great inventors" in the united states, 1790–1865. *The Journal of Economic History*, *53*(2), 289–307.

Kremer, M. (1993). Population growth and technological change: One million bc to 1990. *The quarterly journal of economics*, *108*(3), 681–716.

Library of Congress. (1868). Early collection of copyright title pages: 1790-1870 [https://www.loc.gov/item/2019713455/, Last accessed on 2022-10-25].

Mao, Y., & Wang, J. J. (2022). Access to finance and technological innovation: Evidence from pre-civil war america. *Journal of Financial and Quantitative Analysis*, 1–51.

Mitch, D. (1993). The role of education and skill in the british industrial revolution. In J. Mokyr (Ed.), *The british industrial revolution: An economic perspective*. Westview Press.

Mokyr, J. (2005). The intellectual origins of modern economic growth. *The Journal of Economic History*, *65*(2), 285–351.

Mokyr, J. (2009). *The enlightened economy: An economic history of britain, 1700-1850*. Yale University Press New Haven, CT.

Mokyr, J. (2016). A culture of growth. In *A culture of growth*. Princeton University Press.

Montalvo, J. G., & Reynal-Querol, M. (2003). Religious polarization and economic development. *Economics Letters*, *80*(2), 201–210.

Montalvo, J. G., & Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. *American economic review*, *95*(3), 796–816.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the us labor market. *The Quarterly Journal of Economics*, *118*(2), 549–599.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204.

Porter, R. P., Gannett, H., & Hunt, W. (1895). Progress of the nation, 1790 to 1890. *Report on Population of the United States at the Eleventh Census: 1890*.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, *98*(5, Part 2), S71–S102.

Sequeira, S., Nunn, N., & Qian, N. (2020). Immigrants and the making of america. *The Review of Economic Studies*, *87*(1), 382–419.

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). Layoutparser: A unified toolkit for deep learning based document image analysis. *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, 131–146.

Smith, J. R. (1925). *North america: Its people and the resources, development, and prospects of the continent as an agricultural, industrial, and commercial area*. Harcourt, Brace.

Sokoloff, K. L. (1988). Inventive activity in early industrial america: Evidence from patent records, 1790–1846. *The Journal of Economic History*, *48*(4), 813–850.

Squicciarini, M. P., & Voigtländer, N. (2015). Human capital and industrialization: Evidence from the age of enlightenment. *The Quarterly Journal of Economics*, *130*(4), 1825–1883.

USPTO. (2022). *U.s. patent statistics report*. Retrieved September 30, 2010, from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/h_counts.htm

Vandenbroucke, G. (2008a). The american frontier: Technology versus immigration. *Review of economic dynamics*, *11*(2), 283–301.

Vandenbroucke, G. (2008b). The us westward expansion. *International Economic Review*, *49*(1), 81–110.

Voigtländer, N., & Voth, H.-J. (2006). Why england? demographic factors, structural change and physical capital accumulation during the industrial revolution. *Journal of economic growth*, 319–361.

# 9 Appendix

## 9.1 Human Capital and Agricultural Diversity

Fiszbein (2022) suggests that agricultural diversity played a determinant role in the US' structural change by increasing skill variety and the development of human capital. In Figure 12 I replicate Figure 7 from Fiszbein (2022) adding pre-industrial human capital, measured by the log of total copyright registrations over 1840-1860, to the IV regression. In this Figure, we plot the coefficients of 1860 agricultural diversity and pre-industrial human capital from regressions where the outcome variable is changing over time. The model we are testing is therefore the following:

$$y_{i,t} = \alpha_s + \beta_t \ln(Copyright_{1840-1860}) + \gamma Agridiv_{1860} + \Gamma_t' X_{i,1860} + \varepsilon_{i,t} \tag{9}$$

Where I instrument for both $\ln(Copyright_{1840-1860})$ and $Agridiv_{1860}$. The instrument for agricultural diversity is the same one constructed by Fiszbein (2022) so that, when I exclude $\ln(Copyright_{1840-1860})$ from the regression, Figure 12 is an exact replication of Figure 7 in Fiszbein (2022). As we can see the relationship between agricultural diversity and manufacturing disappears once we control for pre-industrial human capital and is substantially weakened in the case of population density. In the case of the labor share, the coefficients show exactly the opposite pattern that one would expect, being insignificant or close to zero prior to 1970 and positive thereafter. Repeating this exercise for pre-industrial human capital as a sanity check shows that controlling for agricultural diversity has no effect: estimates are similar to those shown in Figure 8. This shows that the relationship between economic development and pre-industrial human capital is particularly strong and, most likely, does not merely reflect a selection effect whereby knowledge producers settled in more geographically amenable locations. If this were the case, we would have expected the coefficients in panels (c) and (d) of Figure 12 to drop substantially when including Fiszbein (2022) variable capturing agricultural diversity.

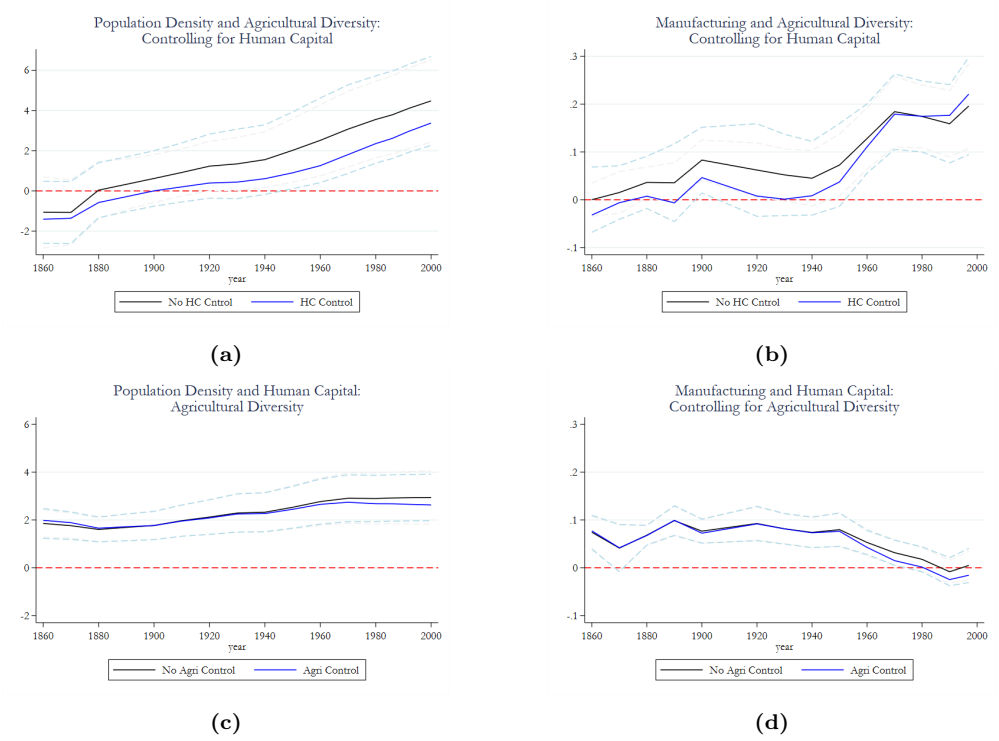



**Figure 12:** Replication of Fig. 7 from Fiszbein (2022). Regression results of the IV estimates of the impact of agricultural diversity, in panels (a) and (b), and pre-industrial human capital, in panels (c) and (d), on population density and manufacturing labor shares over time. Panels (a) and (c) show results for population density and panels (b) and (d) for the labor share of manufacturing.

## 9.2 Monte Carlo Exercise: Difference between IV and OLS Estimates

Large differences between IV and OLS estimates can be cause for concern. This is because, if the instrument is endogenous, bias in IV coefficients can be substantial even for small violations of the exclusion restriction. This especially the case for weak instruments as the formula for the IV estimator is:

$$\hat{\beta_{iv}} = \frac{cov(z,y)}{cov(z,x)} = \frac{cov(z, a + \beta x + \varepsilon)}{cov(z,x)} = \beta + \frac{cov(z,\varepsilon)}{cov(z,x)}$$

Hence, in the case of a weak instrument (which implies a low $cov(z,x)$) the bias of the IV estimator can be large even for small sample covariances of the instrument with the error term ($[cov(z,\varepsilon)] \neq 0$). It is therefore important to justify large differences between OLS and IV estimates, commenting in particular on the direction of the bias (of the OLS estimate). In our case, the OLS estimate is biased downward, by a factor of around 5-7. While this is large, I argued in Section 4 that measurement error in the variable of interest could generate this. Consider the following Monte-Carlo experiment. Suppose the d.g.p. for an outcome variable $y$ is as follows:

$$y = a + bx + \varepsilon$$

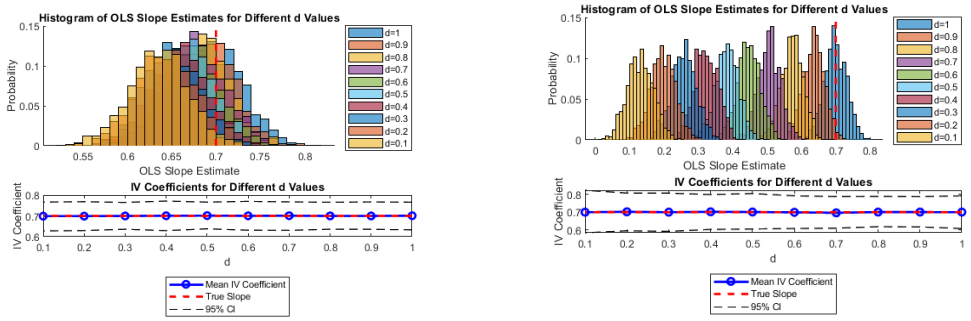
Where $x$ is the total amount of knowledge produced in a given geographical unit. However, part of this knowledge is unobservable (to the econometrician) which leads to measurement error. The true extent of knowledge production is divided among observable (to the econometrician) and unobservable:

$$x = d\tilde{x} + (1-d)f$$

Where $d \in (0,1)$ and $\tilde{x}$ is the observable knowledge and $f$ is the unobservable component. Suppose that $\tilde{x}$ and $f$ come from a multivariate normal distribution with covariance equal to $\rho$. Moreover, suppose the econometrician disposes of an instrumental variable which is equal to the sum of observable and unobservable knowledge plus an error term:

$$z = \tilde{x} + f + \epsilon$$

Where $\epsilon$ is distributed normal with mean zero and variance $\sigma_e = 1$ (increasing $\sigma_e$ increases the confidence intervals around $\hat{\beta_{iv}}$ but does not introduce bias). The higher the covariance $\rho$ and the fraction of observable knowledge $d$, the lower will be the downward bias of OLS estimates. Supposing that the true $b$ is 0.7 and with a correlation of 0.9, Figure 13a shows the distribution of the OLS estimates and the mean of the IV estimates for different levels of $d$.




**(a)** **(b)**

**Figure 13:** Monte-Carlo exercise showing the impact of measurement error on OLS and IV coefficients with $\rho = 0.9$, panel (a), and $\rho = 0.1$, panel (b), with $\sigma_e = 1$. The true parameter value is denoted by the red dotted line.

As we can see, for high correlations of observable and unobservable knowledge, the OLS estimator is informative of the true parameter value even when the fraction $d$ of observable knoweldge is small. If this correlation is lower however, perhaps due to imperfect author matching, the extent of downward bias rises rapidly. Figure 13b shows the same exercise for $\rho = 0.1$. As the fraction of observable knowledge production $d$ is likely to be small, a large downward bias in the OLS coefficients is not unexpected. As we can see, as long as the instrumental variable is sufficiently precise (low $\sigma_e$) concerning the sum of observable and unobservable knowledge, we can recover the true parameter estimate using the IV.

## 9.3 Supplementary Results

Up until 1850 women were excluded from the census, which was only directed at (male) household heads. An important source of bias could ensue if a large fraction of authors prior to 1850 were women. Figure 14 shows that this was not the case, over the whole period approximately 1.3% of works were penned by women. While some female authors, such as Harriette Baker, have certainly left their mark on American literary history, when looking at the general picture female authorship appears to be markedly limited.

Turning to the match quality, Table 17 shows that over 70% of all matches are "exact matches", which in this context means that for a given name, surname and state combination I observe only one person satisfying the occupational criteria. The list of 48 (out of 235) occupations to which I restrict the sample of potential matches is shown in Table 18 which also shows the frequency of each occupational category. It is important to note however that an "exact match" in this sense does not mean a "correct match" as I could still have made an error in extracting the author name from the title page. It is therefore important to consider how the occupational restriction of potential could drive the results. I do not show this here, but adding the log of

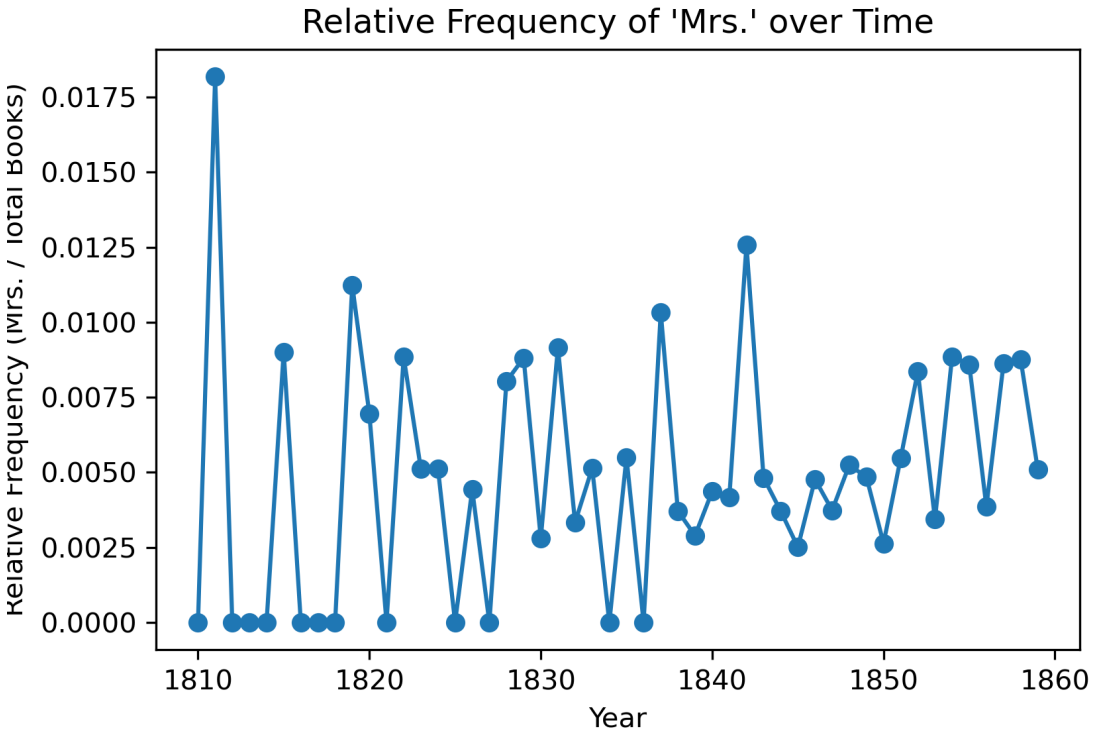



**Figure 14:** Relative Frequency of works containing the string "Mrs." or "Miss" (or non-capitalized variations thereof). Only relative frequencies from years with more than 100 copyright registrations are shown.

total "potential matches", meaning people who fulfill the occupational criteria, to the regressions in Section 6 does not qualitatively change results.

**Table 17:** Math Quality

| Precision | Freq. | Percent |
|---|---|---|
| 1 | 6927 | 71.69 |
| 2 | 990 | 10.25 |
| 3 | 585 | 6.055 |
| 4 | 364 | 3.767 |
| 5 | 255 | 2.639 |
| 6 | 180 | 1.863 |
| 7 | 133 | 1.377 |
| 8 | 88 | 0.911 |
| 9 | 90 | 0.931 |
| 10 | 50 | 0.517 |
| Total | 9662 | 100 |

Table showing the precision of author matches. The first column shows the number of potential matches and the following columns show the frequency and percentage of authors being matched with that level of potential matches.

**Table 18:** Potential Matches: Occupations

| | Freq. | Percent |
|---|---|---|
| Accountants and auditors | 1215 | 0.802 |
| Architects | 572 | 0.377 |
| Artists and art teachers | 2238 | 1.477 |
| Authors | 96 | 0.0634 |
| Chemists | 436 | 0.288 |
| Clergymen | 27698 | 18.28 |
| College presidents and deans | 79 | 0.0521 |
| Agricultural sciences | 1 | 0.000660 |
| Chemistry | 21 | 0.0139 |
| Engineering | 2 | 0.00132 |
| Geology and geophysics | 2 | 0.00132 |
| Mathematics | 32 | 0.0211 |
| Medical sciences | 11 | 0.00726 |
| Physics | 3 | 0.00198 |
| Natural science (n.e.c.) | 8 | 0.00528 |
| Social sciences (n.e.c.) | 12 | 0.00792 |
| Nonscientific subjects | 160 | 0.106 |
| Subject not specified | 539 | 0.356 |
| Dentists | 3029 | 1.999 |
| Designers | 113 | 0.0746 |
| Draftsmen | 104 | 0.0686 |
| Editors and reporters | 1526 | 1.007 |
| Engineers, civil | 812 | 0.536 |
| Engineers, electrical | 1 | 0.000660 |
| Engineers, industrial | 3 | 0.00198 |
| Engineers, mechanical | 49 | 0.0323 |
| Engineers, mining | 5 | 0.00330 |
| Engineers (n.e.c.) | 37 | 0.0244 |
| Lawyers and judges | 25129 | 16.58 |
| Musicians and music teachers | 3808 | 2.513 |
| Nurses, professional | 6 | 0.00396 |
| Agricultural scientists | 7 | 0.00462 |
| Biological scientists | 55 | 0.0363 |
| Geologists and geophysicists | 27 | 0.0178 |
| Physicists | 13 | 0.00858 |
| Miscellaneous natural scientists | 13 | 0.00858 |
| Pharmacists | 6209 | 4.098 |
| Physicians and surgeons | 44721 | 29.51 |

| | | |
|---|---|---|
| Religious workers | 406 | 0.268 |
| Statisticians and actuaries | 4 | 0.00264 |
| Teachers (n.e.c.) | 31217 | 20.60 |
| Technicians, medical and dental | 3 | 0.00198 |
| Technicians, testing | 23 | 0.0152 |
| Technicians (n.e.c.) | 3 | 0.00198 |
| Therapists and healers (n.e.c.) | 370 | 0.244 |
| Veterinarians | 142 | 0.0937 |
| Professional, technical and kindred workers (n.e.c.) | 371 | 0.245 |
| Opticians and lens grinders and polishers | 200 | 0.132 |
| Total | 151531 | 100 |

Table showing the occupations which are considered potential matches and their frequency. Source: US Census of Population 1850.

Relating to the relationship between religious fractionalization and polarization and pre-industrial human capital, Figure 2 showed a, respectively, positive and negative relationship with copyright registrations. The indices of religious fractionalization and polarization are constructed following the methodology proposed by Montalvo and Reynal-Querol (2003) using data on churches at the county level. These indices are formally calculated as follows:

$$FRAC_i = 1 - \sum_{j=1}^{J} (\frac{n_{i,j}}{N_i})^2$$

$$POLi = 1 - \sum_{j=1}^{J} (\frac{0.5 - \pi_{i,j}}{0.5})^2 \pi_{i,j}$$

Where $i$ refers to counties and $j$ to a particular religious denomination, $J$ being the total number of denominations. $N_i$ is the total number of churches in county $i$ and $n_{i,j}$ is the number of churches in county $i$ of denomination $j$. Intuitively, polarization is maximized when two denominations each hold 50% of the churches in a county and fractionalization increases when the number of groups increases. In Table 19 I run a series of regressions where the outcome variable is the explanatory variable used throughout the paper: the log of the total copyright registrations between 1840-1860 in each county. This exercise shows that religious polarization and fractionalization are robustly related to copyrighting activity even when controlling for the entire set of controls used throughout the paper.

Another issue which should deserve considerable attention is the demographics of authors. While the paper focuses mostly on aggregate measures of authorship at the county level, it is also interesting to ask the question: Who were the individuals who produced copyrighted works in the 19th century. Answering this question is not only important to complement the analysis carried out in the rest of the paper, but can

**Table 19:** OLS Regressions: Log Copyright

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| FRAC | 0.582*** | 0.558*** | 0.523*** | 0.523*** | 0.268*** |
| | (7.30) | (8.05) | (7.92) | (7.92) | (5.59) |
| POL | -0.505*** | -0.476*** | -0.445*** | -0.445*** | -0.215*** |
| | (-6.38) | (-6.73) | (-6.23) | (-6.23) | (-5.14) |
| N | 1511 | 1506 | 1505 | 1505 | 1504 |
| State FE | yes | yes | yes | yes | yes |
| Lat/Lon Polynomial | yes | yes | yes | yes | yes |
| Geography Controls | no | yes | yes | yes | yes |
| HC Controls | no | no | yes | yes | yes |
| Social Controls | no | no | no | no | yes |

$t$ statistics in parentheses.

Dependent variable is log of total copyright registrations over 1840-1860. Conley standard errors used with a 100km cutoff. Definitions of variables and sources are contained in the Data Appendix.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

also assuage concerns regarding the accuracy of the matching process. If the group of authors is a random subset of the population (due to purely spurious matching) the we should not observe systematic differences between authors and non-authors. Table 20 shows that this is not the case, indicating that the author names used in the matching process are adding a valuable source of information. Authors were richer (in terms of property value), older, more foreign and less white than other people in the potential match group (defined as individuals belonging to one of the occupations in Table 18).

**Table 20:** Author Descriptive Statistics

| | Author | | Non-Author | | Diff |
|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | t |
| White | 0.990 | 0.098 | 0.995 | 0.073 | (3.105) |
| Real estate value | 2395.759 | 11704.767 | 1679.085 | 10034.188 | (-4.343) |
| Foreign | 0.158 | 0.364 | 0.118 | 0.323 | (-7.658) |
| Age | 38.810 | 12.924 | 35.834 | 12.731 | (-16.271) |
| Observations | 5162 | | 146369 | | 151531 |

Descriptive statistics for authors and non-authors. The non-authors reference category is composed of individuals belonging to one of the occupations in Table 18 who do not figure in the copyright registrations. The variables "Foreign" and "White" take on value 1 if an individual is foreign or white and 0 otherwise. Real estate value is defined in 1850 dollars. Source: US Census of Population 1850.

I do not show the distribution of occupations for authors as it is very similar to that presented in Table 18. This effectively means that the differences between authors and non-authors reported in Table 20 is not due to occupational selection.

**Table 21:** Controls

| Variable | Geography | Social | Human Capital |
|---|---|---|---|
| Potential Yields | ✓ | | |
| Average Rainfall | ✓ | | |
| Temperature | ✓ | | |
| Distance to Oceans or Great Lakes | ✓ | | |
| Terrain Ruggedness | ✓ | | |
| Log of Population | | ✓ | |
| Urbanization | | ✓ | |
| Fraction of Foreigners | | ✓ | |
| Internal Migrants | | ✓ | |
| Population Density | | ✓ | |
| Literacy Rate | | | ✓ |
| School Enrollment Rate | | | ✓ |

## 9.4 Data Appendix

In this paper I use three main data sources. First, the data on copyright registrations is obtained from the website of the Library of Congress (see Section 4 for more information). Second, data relating to all controls and outcome variables (excepting urbanization literacy and school enrollment rates) comes from Fiszbein (2022). Lastly, data on religious denominations, city growth, literacy and school enrollment rates comes from the ICPSR (Haines et al., 2010). I also check whether the data in Fiszbein (2022) is consistent with the data from the ICPSR where the two provide the same indicators and find this to be the case. As discussed in Section 6, there are three types of controls relating to: (i) geography, (ii) human capital and other socially determined variables. It is important to distinguish between the first and remaining groups as the latter will be potentially endogenous to human activity and hence may inadvertently capture a channel through which the variable of interest affects the dependent variable. To limit this concern, these controls are always measured at baseline (1850). Table 21 shows the variables falling in each group of controls.

As far as the classification of academic/non-academic works, table 22 contains the list of keywords used to classify a work as academic.

The classification procedure is done as described in Section 4.

**Table 22:** Keywords

| Keywords |
|---|
| textbook |
| reader |
| instruction |
| principal |
| grammar |
| school |
| Sundayschool |
| Sabbath school |
| manual |
| teacher |
| university |
| professor |