Economic History Student Working Papers No: 009

Measuring Human Capital in the United States using Copyright Title Pages, 1790-1870

Tancredi Rapone

Submitted as partial fulfilment of the MSc in Political Economy of Late Development 2020-21

January 2022

Economic History Department, London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, London, UK. T: +44 (0) 20 7955 7084.

Measuring Human Capital in the United States using Copyright Title Pages, 1790-1870

Tancredi Rapone

Abstract

This paper uses optical character recognition (OCR) to analyze the production of books in the US over 1790 to 1870 using copyright title pages taken from the online archives of the Library of Congress. We construct national time series of book production over this period which show an uptake in per-capita terms in 1830, around the starting point of the US' industrial revolution. We break down the production of books into topics using keywords for 8 topics: science, religion, novel, invention, diffusion, business, philosophy and textbook. On this basis we show that the composition of book production by topics is stable over time, except for textbooks and novels which show a persistent increase over the whole period both in relative and absolute terms. This pushes back the beginning of the growth in US human capital before the first reliable data on schooling and literacy starting in 1870. We thus offer mild support to an interpretation of US growth over the 19th century based on the expansion of knowledge and capabilities, while conceding that the link between the content of books and industrialization is tenuous.

Introduction

Paul David defined the period before 1840 the "statistical dark age of the United States".[1] Economic data on industrial production, literacy rates, demography and many other indicators are notoriously lacking for the early 19th century US. This is problematic since scholars broadly agree that the US' "take-off" in industrial and economic performance occurred somewhere between 1820 and 1850.[2] The transition from an agrarian Malthusian to an industrial economy is arguably the

[1] Paul A. David, "New light on a statistical dark age: US real product growth before 1840." *The American Economic Review* 57, no. 2 (1967): 294-306.

[2] Ibid., Joseph H. Davis, "An annual index of US industrial production, 1790–1915." *The Quarterly Journal of Economics* 119, no. 4 (2004): 1177-1215; Walter Rostow,ed. *The economics of take-off into sustained growth*. Springer, 1963, see in particular the contribution by North on American industrialization.; Robert E. Gallman, "Gross National Product in the United States, 1834–1909," in Dorothy S. Brady (ed.), *Output, Employment, and Productivity in the United States After 1800*, 1966; Robert E. Gallman, "Economic growth and structural change in the long nineteenth century" in Gallman, R. and Engerman, S. eds., *The Cambridge economic history of the United States* vol. 2.

single watershed moment in the past which has defined human history.[3] While scholars disagree on the causes behind this transition, not only in the American case, they agree on a set of general facts: Malthusian economies were characterized by low and stagnant life expectancy, education, and productivity levels; whereas industrial economies display the mirror image of these conditions.[4] For the US, thanks to the painstaking work of several scholars we now have estimates of capital and output series spanning the 19th century which allow us to understand the timing and dynamics of this transition.[5] However, data on education, literacy and other metrics proxying human capital, are missing for the greater part of this century.[6]

This is a crucial piece of the puzzle as human capital, broadly construed as the knowledge, skills and techniques which raise the productivity of workers,[7] is a key element of many theories of the industrial revolution,[8] making metrics proxying its evolution in high demand. This paper addresses this gap in the knowledge of the early economic development of the United States by studying the Copyright Title pages of works entered for copyright registration in local district courts over 1790-1870, a dataset which is argued to approximate the universe of the early intellectual production of the American economy.[9]

---

[3] Gregory Clark. *A farewell to alms*. Princeton University Press, 2008, introduction.
[4] Ibid.
[5] See Gallman, "Growth and structural change" n 2.
[6] See Claudia Goldin. "A brief history of education in the United States." *NBER WP* 119 (1999); Claudia Goldin. "The human-capital century and American leadership: Virtues of the past." *The Journal of Economic History* 61, no. 2 (2001): 263-292; Goldin, Claudia. "Human capital" *in C. Diebolt, M. Haupert*, *Handbook of Cliometrics* (2016): 55-86.
[7] Supra Goldin "Human Capital" n 6.
[8] Prominent proponents of the importance of human capital include Mokyr's work: Joel Mokyr. *The lever of riches: Technological creativity and economic progress*. Oxford University Press, 1992; Mokyr, Joel. *The gifts of Athena*. Princeton University Press, 2011; see also Oded Galor, and Omer Moav. "From physical to human capital accumulation: Inequality and the process of development." *The Review of Economic Studies* 71, no. 4 (2004): 1001-1026; and for a more general overview Gregory Clark. "Human capital, fertility, and the industrial revolution." *Journal of the European Economic Association* 3, no. 2-3 (2005): 505-515. Nathan Nunn. "History as evolution." In *The Handbook of Historical Economics*, pp. 41-91. Academic Press, 2021.
[9] See the collection's webpage at the Library of Congress' website at: https://www.loc.gov/collections/early-copyright-materials-of-the-united-states/about-this-collection/

Our dataset includes over 50 thousand individual works penned over the reference period and is, to the best of our knowledge, the first time the Library of Congress' (LoC henceforth) raw data on copyright title pages is used in the economic history literature. Using the resources of the American National Biography (ANB henceforth), we also trace the life stories of the 240 most prolific American authors of the nineteenth century gathering biographical information on them concerning their social background, educational status, political views and the relevance of their contributions to economic activity.

We have three main findings. Firstly, the correlation between book production, patent activity and economic growth is strong: a take-off in book production per capita can be observed slightly prior to the growth of inventive activity and in correspondence with the beginning of capital stock and industrial production growth. Secondly, we find that publications become slightly more scientific only in the second half of the 19th century, supporting the theory that productivity was not a central component of the US' early growth record.[10] Thirdly, we argue that, based on biographical data from the ANB on American authors and the change in composition of topics over time, the upper-tail of the human capital distribution became "fatter" (more inclusive) over the 19th century; a finding which is consistent with prominent theories of the Great Enrichment and long-run economic development.[11]

The rest of this paper is structured as follows: in the next section I review the literature on the role of human capital in economic growth. In Section III, I present the dataset and its estimation methodology in detail, highlighting potential limitations which should be addressed by further contributions and refinements. In Section IV, I discuss the findings gleaned by analyzing the dataset assembled

[10] See Moses Abramovitz and Paul A. David. *American macroeconomic growth in the era of knowledge-based progress: The long-run perspective*. In S. Engerman and RE Gallman eds, *The Cambridge economic history of the United States* vol 3, 1999; also Peter Temin. *Causal Factors in American Economic Growth in the Nineteenth Century*. Macmillan International Higher Education, 1975.

[11] Clark, supra n 3. Deirdre N. McCloskey, *The bourgeois virtues: Ethics for an age of commerce*. University of Chicago Press, 2010. Joel Mokyr. *The enlightened economy: an economic history of Britain, 1700-1850*. New Haven: Yale University Press, 2009.

in this paper. Section V concludes with a summary and suggestions for future research.

Section II

Easterlin[12] famously argued that the reason why the whole world isn't developed is because of differences in human capital which persist across countries. Adopting new technologies which raise productivity and spread the industrial revolution requires skills and knowledge which are simply lacking in the developing world, preventing income convergence even in the presence of liberalized capital and trade markets.[13]

However, empirical research on the role of human capital in the industrial revolution has not been kind to the simple formulation of this hypothesis, i.e. that improved schooling and literacy were behind the rise in productivity levels which characterized the transition to modern economic growth.[14] Similarly, the unified growth theory of Galor and his co-authors,[15] which suggests that a change in the premium to skilled labour induced a quantity-quality trade-off in fertility and consequently a build-up of human capital and sustained economic progress, is not supported by either the historical wage premiums to skilled labour or the qualitative literature on technological change during industrialization which generally led to less skill-intensive production methods.[16] In response to the seemingly missing correlation between human capital and economic growth,

[12] Richard A. Easterlin, "Why isn't the whole world developed?." *The Journal of Economic History* 41, no. 1 (1981): 1-17.
[13] Ibid.
[14] David Mitch, "The role of skill and human capital in the "British" industrial revolution." In J. Mokyr ed, *The British Industrial Revolution: An Economic Perspective, Boulder, Colorado* (1999): 241-279.
[15] Galor and Moav, supra n 8; see also Oded Galor. *Unified growth theory*. Princeton University Press, 2011.
[16] Clark, supra n 3, at 181; and Gregory Clark. "The industrial revolution." In S. Durlauf and P. Aghion eds, *Handbook of economic growth*, vol. 2, pp. 217-262. Elsevier, 2014; Acemoglu, Daron. "Technical change, inequality, and the labor market." *Journal of economic literature* 40, no. 1 (2002): 7-72;

scholars such as Mokyr,[17] Clark,[18] McCloskey[19] and others have taken a more nuanced approach.

Mokyr argues that the upper-tail of the human capital distribution is what matters in long-run economic performance as opposed to the literacy rate or the knowledge of the median worker, while features such as health and technical dexterity in the labour force are necessary but insufficient conditions for an economic take-off.[20] This resolves the paradox whereby increasing productivity levels were accompanied by a de-skilling of the labour force during early industrialization in Britain.[21] In his later work, he focuses endogenous technological change turning to culture as a determinant of economic growth insofar as it affects the incentives facing potential innovators, and the willingness of a society to adopt foreign ideas.[22] While monetary incentives often eluded the inventors who contributed to the industrial revolution,[23] widespread cultural attitudes toward new knowledge and discoveries arguably reduced implicit costs of innovation such as the stigma of breaking with tradition.[24]

Authors such as Landes, O' Brien and McCloskey have elsewhere made similar points. McCloskey argues that "liberalism" understood as the set of values and beliefs which came to prominence in Britain and the Low Countries prior to the industrial revolution was instrumental in creating an environment where innovation and breaking with tradition were not only tolerated, but actively

[17] Mokyr, supra n 8; also Joel Mokyr. *A culture of growth*. Princeton University Press, 2016.
[18] Clark, supra n 3.
[19] McCloskey, supra n 11; and sequels: Deirdre N. McCloskey, *Bourgeois dignity: Why economics can't explain the modern world*. University of Chicago Press, 2010; Deirdre N McCloskey. *Bourgeois equality*. University of Chicago Press, 2021.
[20] Morgan Kelly, Joel Mokyr, and Cormac Ó. Gráda. "Precocious Albion: a new interpretation of the British industrial revolution." *Annu. Rev. Econ.* 6, no. 1 (2014): 363-389.
[21] Acemoglu, supra n 16; Kelly, Mokyr and O' Grada, supra n 20; productivity of the labor force was already much higher in Britain in a comparative perspective before the industrial revolution. Hence, Kelly, Mokyr and O' Grada argue that Britain was able to capitalize on the technologies of the industrial revolution when they appeared.
[22] Mokyr, "A culture" n 17, chapter 10 in particular.
[23] Clark, supra n 3 at 235.
[24] Mokyr, "A culture" n 17, chapter 10 in particular. Also McCloskey supra n 11, chapters 3, 11 and 14 in particular.

encouraged.[25] This idea reaches back to Landes'[26] and O' Brien's[27] comparison of Europe and China which emphasized the reluctance of the Chinese to embrace foreign knowledge and the intellectual curiosity of Europeans as reasons why the world diverged into rich and poor nations around the end of the 18th century. Hence, according to these scholars, it was the intangible knowledge and capabilities which Europeans accumulated in the form of institutions, culture and human capital which made them successful. The challenge is explaining why these changes in people which lead to the accumulation of knowledge occurred at a particular place and point in time.[28]

This is made even more demanding by the fact that such a change in people's attitudes toward knowledge is hard to document empirically in the first place.[29] Prominent candidates which could point in such a change of preferences are the consumption of new goods such as books or direct evidence on behavioural changes such as fertility patterns.[30] For instance, Clark shows that when one uses a "modern" basket of goods (including books, among other commodities) to weight productivity gains by sector, efficiency growth in Britain took off long before the industrial revolution. [31] This finding is consistent with his long-run analysis of interest rates, violence, literacy and fertility patterns of the rich in the preindustrial period which suggest that the break from Malthusian constraints which obtained in England at the end of the 18th century was anticipated by, and causally related to, a centuries long process of cultural evolution which affected

[25] McCloskey, supra n 11 and 19.
[26] David S Landes. "Why Europe and the west? Why not China?." *Journal of Economic Perspectives* 20, no. 2 (2006): 3-22.
[27] O' Brien, Patrick. "The Needham question updated: a historiographical survey and elaboration" *History of technology* 29 (2009).
[28] See Mokyr, "A culture" n 17, chapters 9-10-16 in particular.
[29] See for instance Greg Clark's critique of Mokyr 2009: Gregory Clark. "The Enlightened Economy: An Economic History of Britain 1700-1850: Review Essay." *Journal of Economic Literature* 50, no. 1 (2012): 85-95.
[30] On new goods see the theory of the industrious revolution of Jan De Vries, which hypothesizes that the availability of "modern" goods lead consumers to increase their labor input during the commercial revolution; Jan De Vries, "The industrial revolution and the industrious revolution." *The Journal of Economic History* 54, no. 2 (1994): 249-270.
[31] See Clark, Gregory. "The industrial revolution: A cliometric perspective." In In C. Diebolt, M. Haupert eds, *Handbook of Cliometrics* (2016): 197-235.

the preferences and behaviour of key economic actors.[32] This hypothetical mechanism is to be distinguished from those formulated by authors such as Allen,[33] who believe that relative factor endowments which became salient in a quasi-random way at the end of the 18th century with the development of coal using technologies were responsible for Britain's economic success. According to Allen's theory, cultural changes should be endogenous to the growth process, which results from investments driven by the relative prices of factors of production.

The timing of a change in preferences, evidenced in book production time series such as those developed in this paper, is therefore crucial to establishing whether growth follows or precedes cultural change. Furthermore, using keywords to assign a category to publications (e.g. religious, scientific, fiction etc.), we are able to test whether the proportion of books in various categories changes over time, for instance whether religious topics become less popular with respect to secular ones during economic development, potentially indicating a change in social preferences.[34]

Not only does the production of books indicate the "sophistication" in a society's human capital base and its preference for "modern" consumer goods,[35] it also conveys information on the content of its intellectual production, where data on titles and authors can be obtained. Previous literature has looked at specific publications, such as encyclopaedias in the case of Squicciarini and Voigtlander[36] or institutions of educational learning and advancement such as universities,[37] to

[32] Clark, supra n 3 chapter 9. See also for a similar argument focusing on the generational transmission of time preference rates, Matthias Doepke and Fabrizio Zilibotti. "Occupational choice and the spirit of capitalism." *The Quarterly Journal of Economics* 123, no. 2 (2008): 747-793.
[33] Robert C. Allen. *The British industrial revolution in global perspective*. Cambridge University Press, 2009.
[34] McCloskey, "Bourgeois virtues" n 11, chapter 12 in particular.
[35] Baten and Van Zanden, "Book production" infra n 44 at 218.
[36] Mara P. Squicciarini and Nico Voigtländer. "Human capital and industrialization: Evidence from the age of enlightenment." *The Quarterly Journal of Economics* 130, no. 4 (2015): 1825-1883.
[37] Jeremiah Dittmar and Rolf Meisenzahl. "The research university, invention, and industry: Evidence from german history." *Working Paper LSE*, 2021; Davide Cantoni and Noam Yuchtman. "Medieval universities, legal institutions, and the commercial revolution." *The Quarterly Journal of Economics* 129, no. 2 (2014): 823-887.

isolate the type of knowledge which is argued to stimulate economic activity. For instance, Dittmar and Meisenzahl estimate that the research university contributed to industrialization in Germany by expanding the supply of qualified engineers and technicians able to adapt technologies to industrial purposes.[38] The contribution of research universities according to these authors is not limited to the direct link between scientific knowledge and industry, but, as also emphasized by Mokyr and McCloskey,[39] encompasses the cultural openness and critical spirit of a society which embraces modernization.[40] The university as an institution is also studied by Cantoni and Yuchtman in their paper on the effects of the papal schism on the supply of universities in Germany.[41] These authors find that a change in distance to universities was negatively associated with city growth in the late medieval period. Their hypothesized channel, eschewing the scientific focus in Dittmar and Meisenzahl, looks at the supply of lawyers and judges which arguably served to improve institutions of contract enforcement and reduce transaction costs.

The link between human capital and economic progress therefore is not necessarily parasitic on the role of scientific advancement in the development process, however several authors have supported this position as well. In 19th century France, Squicciarini and Voigtlander, find that regions which had higher rates of subscribers to the encyclopaedia, which they take as a proxy for density of high-skilled individuals able to implement and adapt foreign technology, also saw the most pronounced increase in real wages and industrial production during France's late industrial revolution.[42] The role of scientific knowledge in economic development is also emphasized by Chaney's work on Islamic civilizations, which shows that, using library catalogues, a declining number of scientific publications coincided with the economic decline of the Islamic world viz-à-viz the West.[43]

[38] Dittmar and Meisenzahl, supra n 37.
[39] McCloskey, "Bourgeois virtues" n 11; Mokyr, "A culture" n 17.
[40] Dittmar and Meisenzhahl, supra n 37 at 10-12.
[41] Cantoni and Yuchtman, supra n 37.
[42] Squicciarini and Voigtlander, "Human capital" n 36.
[43] Eric Chaney, "Religion and the rise and fall of Islamic science." *Work. Pap., Dep. Econ., Harvard Univ., Cambridge, MA* (2016).

In a paper which takes a similar research strategy to that used here, Baten and Van Zanden,[44] use library catalogues to estimate historical time series for various nations in western Europe, showing a positive association between book production in the early modern period and economic growth over the 19th century. They interpret this as evidence that human capital supported the spread of the industrial revolution, asserting that book production comprises information on both the demand and supply of knowledge in a society which can have positive effects on economic growth irrespective of whether publications are scientific.[45] To understand whether the relationship between book production and economic growth is driven by the increased supply of scientific knowledge or another channel, such as a cultural change in consumer preferences, the content of publications must be investigated, which is what we attempt in this paper.[46]

For the US specifically, previous work has focused on schooling and patent statistics as measures of human capital over the nineteenth century. Sokoloff and his collaborators,[47] have made the case that intellectual property was an essential ingredient to American development from the early nineteenth century. They find that patents were positively related to the growth of markets since the beginning of industrialization,[48] although conceding the critique that the patent system was not exigent enough, in terms of originality and usefulness of inventions, until its reform in 1836 to be relied on as a source of information on scientific capabilities. Nonetheless, it is clear from their data that the class of inventors who made repeated innovations grew over the nineteenth century, particularly in those areas experiencing more rapid industrialization. While they attribute this to the patent

[44] Joerg Baten and Jan Luiten Van Zanden. "Book production and the onset of modern economic growth." *Journal of Economic Growth* 13, no. 3 (2008): 217-235.
[45] Ibid. at 218.
[46] If science is the main conduit which links books to economic activity, we would expect to see a large number of publications dealing specifically with applied sciences and industrial technology.
[47] Kenneth L. Sokoloff "Inventive activity in early industrial America: evidence from patent records, 1790–1846." *The Journal of Economic History* 48, no. 4 (1988): 813-850; Naomi Lamoreaux and Kenneth L. Sokoloff. "Market trade in patents and the rise of a class of specialized inventors in the 19th century United States." *American Economic Review* 91, no. 2 (2001): 39-44; Naomi Lamoreaux Kenneth L. Sokoloff, and Dhanoos Sutthiphisal. "Patent alchemy: the market for technology in US history." *Business History Review* 87, no. 1 (2013): 3-38.
[48] In particular, Sokoloff, "Inventive activity" n 47.

system itself, it is conceivable that the effectiveness of intellectual property rights interacts with the presence of human capital.[49] As mentioned previously, we find a strong correlation between patent statistics and book production (corr = .81 over 1790-1870), which could suggest that productivity led growth preceded what Goldin refers to as the "human capital -20th- century".[50] This is somewhat at odds with the widely shared belief in the literature that American economic growth over the 19th century was concentrated at the extensive margin, in the growth of land, labor and capital used in production due to the physical expansion of the country; whereas the era of "knowledge-based-progress" only began in the 20th century.[51]

Goldin's study of schooling in the nineteenth century supports the latter view: primary schools only became free in the 1870s and were unevenly spread, whereas more skill intensive secondary schooling became prevalent only in the beginning of the 20th century.[52] However, given that schooling was largely a grassroots movement in the early 19th century born out of the interest of parents in small communities, it is not clear to what extent primary schooling was present in the US' early economic development or whether it grew substantially before the transition to federal funding in 1870.[53]

Data on book production, including textbooks and scientific publications, can help fill this gap in the literature and shed light on the growth and importance of human capital in the early development of the US.

---

[49] See for a summary of the contemporary literature on the effectiveness of intellectual property protection as an economic policy tool: Daniele Archibugi and Andrea Filippetti. "The globalisation of intellectual property rights: four learned lessons and four theses." *Global Policy* 1, no. 2 (2010): 137-149.
[50] Goldin, "Human capital century" n 6.
[51] Abramovitz and David, supra n 10; Temin, supra n 10; Richard Bensel, *The political economy of American industrialization, 1877-1900*. Cambridge University Press, 2000; for a contrarian perspective, see the work of Sokoloff and his collaborators, supra n 47; and Douglass C. North, "Industrialization in the United States (1815–60)." In *The Economics of Take-off into Sustained Growth*, pp. 44-62. Palgrave Macmillan, London, 1963.
[52] Goldin, supra n 6.
[53] Ibid.

Our paper also contributes to the debate between Acemoglu, Johnson and Robinson[54] and Glaeser et al.[55] on whether institutions or human capital were the main mechanism through which colonizers from the old world influenced economic outcomes. By finding an exponential increase in book production in the US over the 19th century, when immigrants from the old world poured into the country, we can moderately support Glaeser et al.'s[56] hypothesis that settlers from the old world did not bring *only* their institutions with them but also an increase in knowledge and human capital. Furthermore, by avoiding the use of either key publications[57] or library catalogues,[58] the coverage of our data is arguably superior to that of previous studies which use book production as their main metric of interest.[59]

Copyright registration data is arguably more comprehensive as it includes non-published work and, given the relatively low cost of obtaining registration, is made reliable by the substantial incentive that authors had to register their work.[60]

Section III

Our source of raw data on copyright registration is the LoC's collection of "copyright title pages".[61] In order for a copyright application to be processed by a district court, applicants had to submit a registration form, a title page and pay

[54] Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." *American economic review* 91, no. 5 (2001): 1369-1401.
[55] Glaeser, Edward L., Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. "Do institutions cause growth?." *Journal of economic Growth* 9, no. 3 (2004): 271-303.
[56] Ibid.
[57] Squicciarini and Voigtlander, supra n 36.
[58] Chaney, supra n 43; and Baten and Van Zanden, supra n 44;
[59] However, for some countries such as England, high quality data describing the number of books written each year exists (see the ESTC (English Short Title Catalog) and the EEBO (Early English Books Online) collections) and is commonly used in the literature. These collections also offer samples of fully transcribed books, which has led to interesting topic analysis work, see Dittmar, "The welfare impact" infra n 78.
[60] Price of copyright registration was 60 cents, approximately equal to half a bushel of wheat. Sources: Zorina Khan, *The Democratization of Invention: Patents and Copyrights in American Economic Development, 1790-1920* (New York: Cambridge University Press, 2005) for the copyright registration fee; and US Dept. of Agriculture, *Farm Prices in two Centuries*, (1892) for historical commodity price series.
[61] See an example in the appendix figure A.1.

the required fee.[62] The title page generally contains the name of the author, the publishing company and the title of the book. In some cases, in addition to the title page, authors also submitted following pages (usually these are blank or contain information on the copyright registration),[63] so we have approximately 50 thousand unique works in over 97 thousand images downloaded from the LoC.[64] We then use OCR to read all images and store the text contained in each on a single line of a csv file. Once blank lines[65] and registration pages[66] are deleted, this leaves us with approximately 50 thousand images, one per copyrighted work.

We extract the year of publication from the text contained in the images by matching the numerical characters in the text on possible dates. Since we know that the date will start with either "17" or "18" we retrieve this string where present followed by the two subsequent characters and store it as a separate variable. When roman numerals are used, extracting the date is more challenging as spaces between letters can get in the way of matching the start of a date and dates aren't always four characters in length. Nonetheless, we successfully extracted dates for over 43 thousand copyrighted works.[67] We then proceed to identifying the location of publication by matching strings in title pages on possible locations, from a list of the 20 most populous cities in the country which

[62] See supra n 60. The fee was not substantial; hence we have no reason to believe that it discouraged authors from registering. Furthermore, the protection granted by copyright was effective and piracy risks were considerable. See Kahn, "The Democratization" n 60.
[63] Such as which district court processed the registration.
[64] Code used to download the images is available in the online appendix.
[65] Denoting blank pages.
[66] Registration pages are those which were annexed to a title page, but do not represent an additional work. We identify these as containing the string "Act of Congress" as these pages generally start with "Entered According to Act of Congress" and then provide the name of the publisher or the person who filed the application and state his rights pursuant to copyright legislation.
[67] Equivalently, over 85% of the expected number of individual works. The procedure for extracting dates when roman numerals are used is similar to that used for Arabic numbers, being based on the extraction of MDCC or MDCCC (with variations allowing for spaces between letters) and the subsequent X, V or I characters and then converting to Arabic numerals the resulting date.

is updated every decade.[68] This drastically reduces the chances of false matches compared to a baseline list of all US cities as possible locations.[69]

To extract the name of the author from title pages we first try to extract all human names from the images using natural language processing,[70] however this is not always effective as natural language analysis algorithms are trained on traditional texts to extract names based on their syntactical position in a sentence.[71] Since the syntax of title pages differs substantially from that of a standard page of text, extracting names is challenging for these software packages. Where we can successfully extract names, we take the first full name to be the author (subsequent names are generally the publishers). Where this procedure doesn't work, we simply extract the two strings following the string "by" in the page. This gives us an approximately 60% success rate in extracting author names.[72]

To classify publications into 8 topics,[73] we use lists of keywords which are likely to identify a given topic: if a title page contains at least one of these words it is assigned to a given topic. The choice of topics is guided by the existing theoretical literature. Firstly, we are interested in the macro topic of scientific publications, which is further broken down into 3 smaller subsets: invention (specifically targeting publications on applied science and engineering), textbook and

[68] E.g. location for works copyrighted between 1791 and 1800 is obtained by matching on the list of 20 most populous cities in 1790; for 1801-1810 the list is updated to the 20 most populous cities in 1800 and so forth.
[69] Using a list of over 28000 US cities as the list of possible locations shows an implausible geographical distribution of copyrighted works due to the high risk of false matches.
[70] We use Python NLTK and Stanza, finding that Stanza gives us better (although still not entirely satisfactory) results.
[71] See the Stanza documentation, Stanza, "A python NLP package for many human languages" *Stanford University*. Available at: https://stanfordnlp.github.io/stanza/#about
[72] Based on a random check for 50 pairs of author names and raw text in the title page. In some cases, two strings following by is not enough, for instance when authors have a middle name; however, we still mark these as correct attributions since this is easily corrected at later stages (e.g. if "john f" keeps coming up, we check the raw data to see who "john f" is and complete the name).
[73] The categories are (1) science, (2) religion, (3) novel/entertainment, (4) invention, (5) diffusion, (6) business, (7) philosophy and (8) textbooks.

diffusion.[74] We also classify non-scientific publications as religious, philosophical and novel/entertainment. Our last category, "business", is designed to capture non-fiction essays and books concerning business related topics such as accounting, corporate finance and managerial techniques, following Dittmar and Seabold who find these kinds of publications to have been significantly correlated with city growth in early modern Europe.[75]

Using "relatedwords.org", a search engine dedicated to providing lists of words related to a common root, we assemble lists of keywords for 6 of the 9 roots in the first row of Table III.1, the remaining 3 (diffusion, invention and textbook) being more specific subsets of the "science" category. The list of keywords includes the root word. The *relatedwords* search engine uses dictionary entries to assemble a list of words which are most likely to have the root word of a given topic in their definition, hence providing a baseline list of keywords.[76] This list provided by this procedure is then filtered manually to exclude words potentially leading to high rates of false matches.[77]

We allow lists of keywords to overlap in all analysis, except when expressing the composition among these topics of the total publications in a time frame to avoid double counting.[78] This approach differs from that taken by Chaney[79] by including a larger number of keywords to maximize the match rate between publications and topics. While this could potentially lead to false matches, by using restricted subsets of the lists of keywords we double check our results finding them to be

[74] There is some overlap between these last two, however, textbook (the more restricted) should only capture books specifically written for classroom teaching.
[75] Their methodology for identifying business related publications is however admittedly less error-prone than ours as they have a baseline list of all business publications which they match to city printing data. See Jeremiah Dittmar and Skipper Seabold. "New media and competition: printing and Europe's transformation after Gutenberg." *LSE Working paper* (2019).
[76] See for instance for the words science, https://relatedwords.org/relatedto/science.
[77] For instance, the string "bio" can designate both a bio-science or a bio-graphy, hence it is excluded.
[78] This is consistent with the approach taken by Dittmar where the topic of a publication is treated as a hidden distribution of topics, i.e. a work can be scientific but also philosophical or religious just like a paper can be part economics and part history. Jeremiah Dittmar. "The welfare impact of a new good: The printed book." *Department of Economics, American University* (2011).
[79] See Chaney, supra n 42.

consistent.[80] The list of complete keywords is provided online,[81] whereas the most popular words for each category are reported in table III.1.

Table III.1: Keywords for ten categories of topics

science (1)	**religion (2)**	**novel (3)**	**invention (4)**	**diffusion (5)**
principles	rev.	magazine	practical	school
natural	god	acts	patent	journal
system	church	broadway	steam	course
dr.	sabbath	music	iron	dictionary
knowledge	christ	songs	manufactur	manual

Table III.1: continued

business (6)	**philosophy (7)**	**textbook (9)**
money	think	textbook
rational	truth	reader
efficient	thought	instruction
cost	moral	principal
industrial	treatise	grammar school

We choose this procedure since alternative approaches to topic classification used in the literature, such as Bayesian modelling, require large amounts of text to work and are hence more suitable to classifying fully transcribed works instead of title pages.[82]

Table III.2 reports descriptive statistics on the distribution of topics for the whole timeframe, whereas time varying patterns will be explored in Section IV.

[80] Using only the word "science" as a keyword for matching shows similar trends in the data at lower baseline levels.
[81] See the data appendix.
[82] See Dittmar, "The welfare impact" supra n 78.

Table III.2: Descriptive statistics, showing only categories with more than 5000 works classified.

	non overlapping	**overlap**	**% total**
science	5389	23717	9,2%
diffusion	n.a.	5333	
textbook	n.a.	9865	
invention	n.a.	3280	
novel/entertainment	3182	10419	5,4%
religious	4486	14387	7,7%
business	4642	16305	7,9%
not classified	26000	n.a.	44,4%
ambiguous	n.a.	14487	24,7%
total	58498	n.a.	100%

Note: The first column shows the number of works falling unambiguously in a category (e.g. scientific works only classified as such), the second shows 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.

As we can see, works are approximately equally distributed between scientific, religious and business, with science being the larger category and novel/entertainment being the smallest. We are however troubled by the large portion of unclassified works and the potential for misclassification (as evinced by the substantial portion of works falling in more than one category).[83]

Section IV

While an official industrial sector did not appear before 1825,[84] the process of industrialization in the US began in New England before independence, when "peddlers", who were forbidden from selling their manufactured goods internationally by the British crown, found a market in the rich north-eastern

[83] This isn't in itself a cause for concern as works can pertain to more than one subject matter in principle. However, when we see works being classified in unlikely combinations (e.g. science/novel or business/religion) we are more hesitant to attribute a classification. Hence, we perform robustness checks of the findings in Section IV (table IV.I) excluding works which fall into these suspect multi-topics, finding the results to be consistent.

[84] Davis, "An index" supra n 2.

agricultural region.[85] Given this start, industrialization in the US developed initially as a small-scale, labour and skill intensive process with low fixed-capital requirements.[86]

While international trade was constrained, it was more difficult for the British crown, and other European countries, to prevent the migration of skilled workers to the US, which picked up substantially in the decades after 1830.[87] As noted by Russell Smith (1924) "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".[88] Slater's case was likely not an isolated one. Sequeira, Nunn and Qian find a robust association between migration and industrialization over 1850-1920, arguing a similar channel of causality. This is supported by a large literature on the economic effects of migration,[89] as well as Glaeser et al.'s[90] hypothesis that the human capital brought by settlers was the conduit through which economic growth spread from the old to the new worlds.

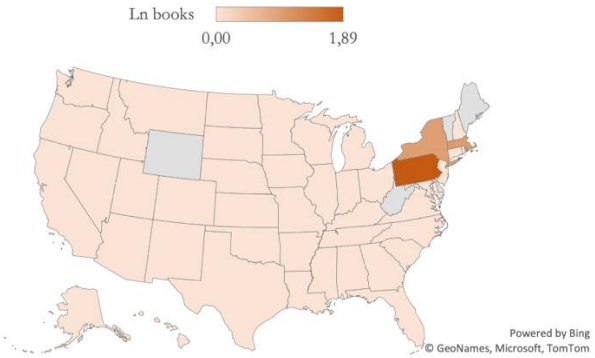
Due to the New England and Middle colonies' favourable geography, their unique puritan culture which praised education of the young and the active selection of migrants by established settlers,[91] we expect that book production should be concentrated in the North-East over the time dimension of our dataset, while gradually expanding westward. This is confirmed in Figure IV.1.

---

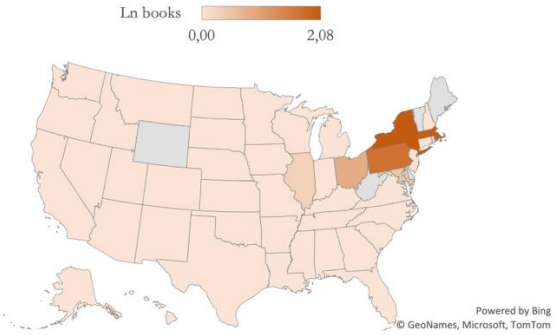
[85] See David R. Meyer, *The roots of American industrialization*. JHU Press, 2003; and J Russell Smith, *North America*, London: G. Bell and Sons, 1924 at 70-75.
[86] Ibid.
[87] Although most immigrants worked as unskilled laborers, a smaller number provided invaluable contributions in terms of knowledge and skill to early American industrialization. See Joseph P. Ferrie, *Yankeys now: Immigrants in the antebellum US 1840-1860*. Oxford UP, 1999; on the importance of skilled immigrant labor see Ran Abramitzky, Leah Platt Boustan, and Katherine Eriksson. "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration." *Journal of Political Economy* 122, no. 3 (2014): 467-506; Sandra Sequeira, Nathan Nunn, and Nancy Qian. "Immigrants and the Making of America." *The Review of Economic Studies* 87, no. 1 (2020): 382-419.
[88] Smith, supra n 85 at 72.
[89] See supra n 87.
[90] See Glaeser et al., supra n 55.
[91] See supra n 87; and Smith, supra n 85. See also in particular, David H. Fisher, *Albion's seed Vol 1. America: A cultural history*, OUP 1989, 219-234 on the "Puritan ethic of learning" and 64-74 on the "Social origins of the puritan migration" which discusses the selection of migrants.

Figure IV.1: Geographical distribution of published books over 1790-1870[92]

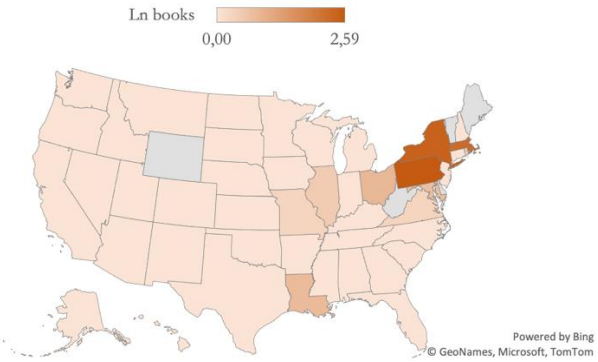
1790-1800



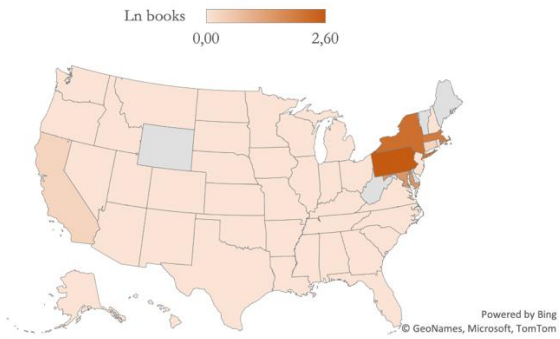
1800-1810



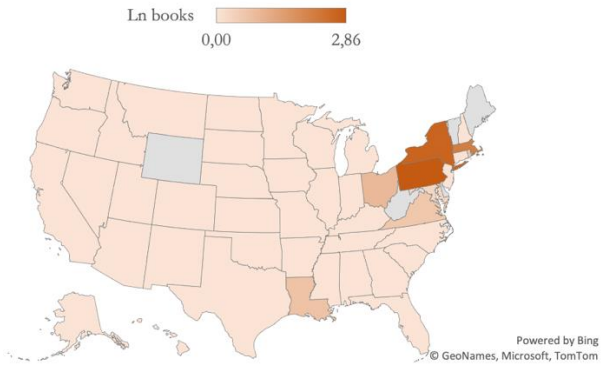
1810-1820



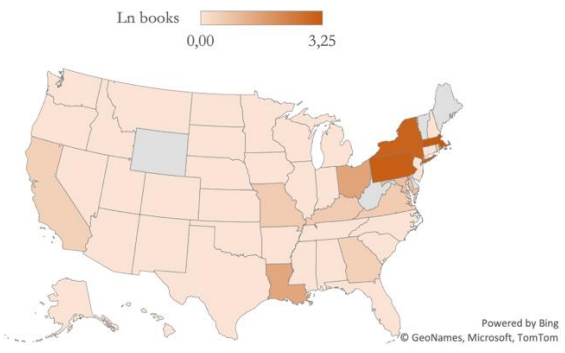
1820-1830



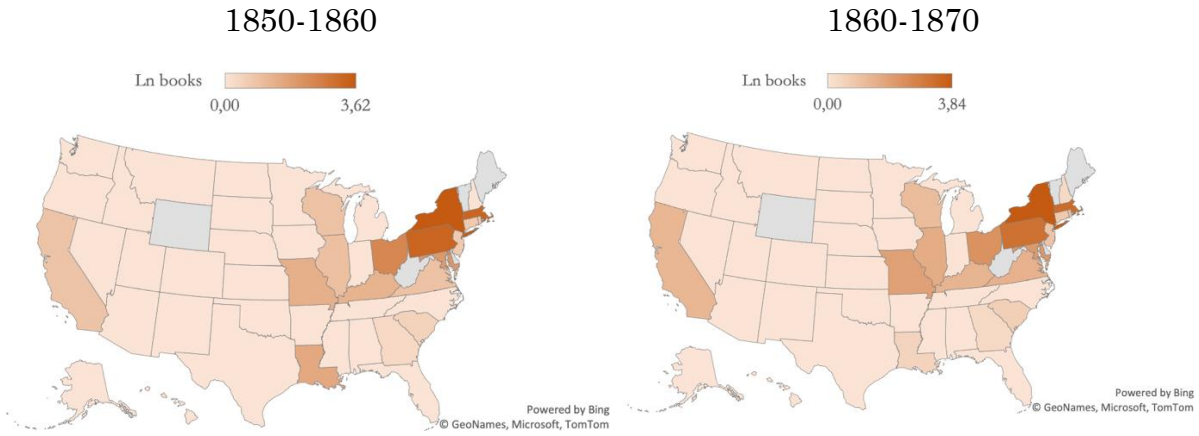
1830-1840



1840-1850



[92] We are generally not able to infer location for non-published books.



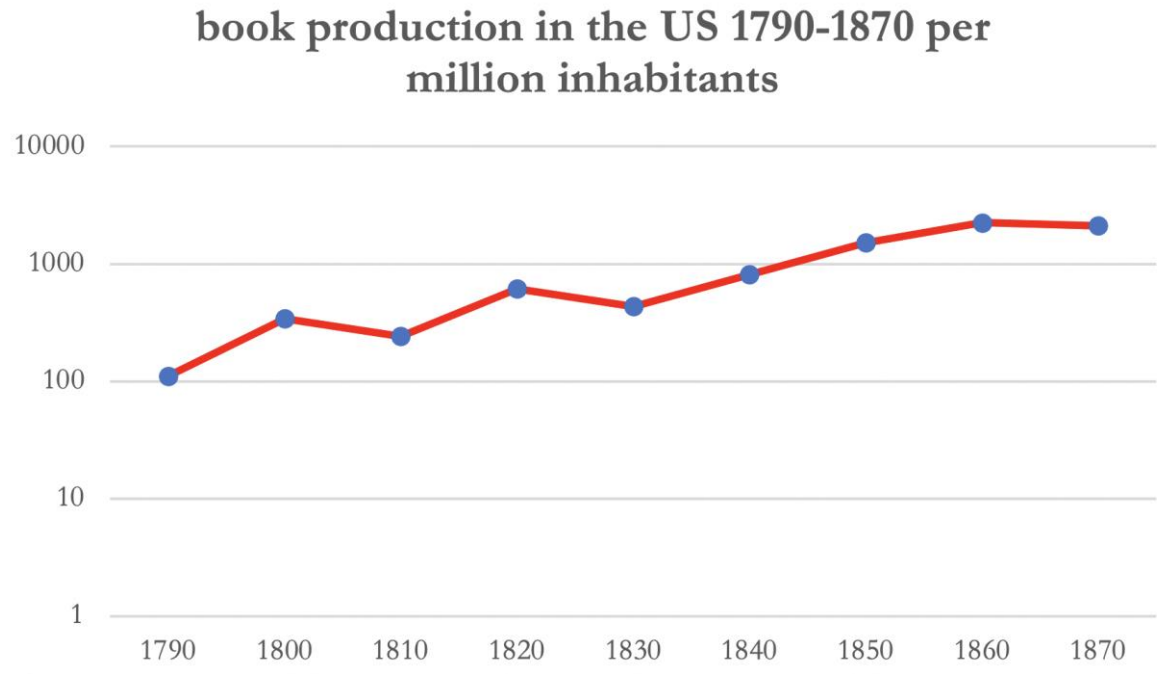
The log-scale facilitates comparability between states as over 90% of all registrations in each decade for which we were able to infer location occurred in one of New York, Massachusetts or Pennsylvania. This is not surprising as these states also accounted for most large cities in the US over this timeframe in addition to being where industrialization took hold.

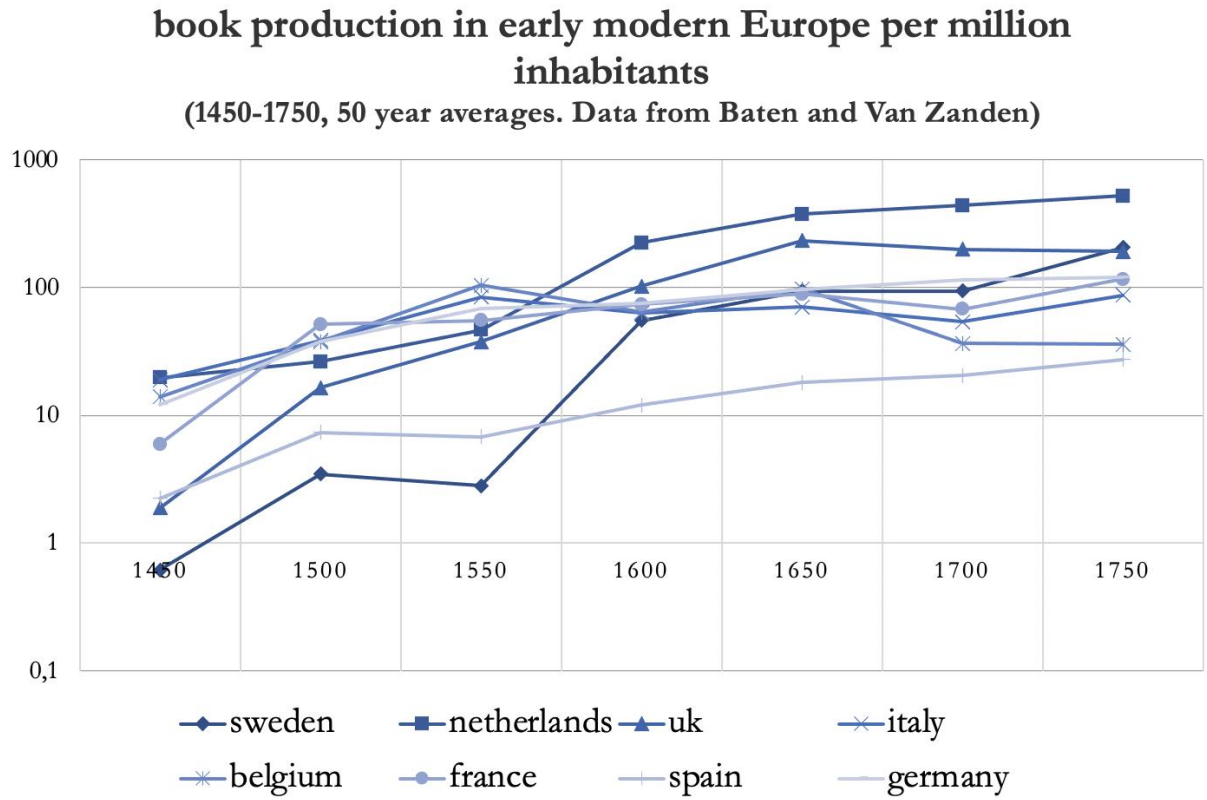
While this cross-sectional data is suggestive, its coverage and accuracy aren't sufficient to test a causal role of human capital in industrialization. Causal inference would require assembling spatial data on book production at the county level and predicting growth over the nineteenth century using initial levels of book production, controlling for potential confounders such as abundance of natural resources and energy sources. This would be a similar research design to Baten and Van Zanden,[93] Squicciarini and Voigtlander[94] and Dittmar,[95] who assemble datasets with wide cross-sectional components. Unfortunately, due to the nature of our raw data we cannot work with such a wide spatial dimension.[96] However, we can place the US in an international comparison, which we do in Figure IV.2

[93] Baten and Van Zanden, supra n 44.
[94] Squicciarini and Voigtlander, supra n 36.
[95] See Dittmar's work on the printing press: , Jeremiah E. Dittmar. "Information technology and economic change: the impact of the printing press." *The Quarterly Journal of Economics* 126, no. 3 (2011): 1133-1172; also Dittmar, Jeremiah, and Skipper Seabold. "New media and competition: printing and Europe's transformation after Gutenberg." *LSE* WP, 2019.
[96] This could be achieved if one were able to infer location based on the author's name using population censuses. However, this is not a small feat considering the difficulties we encounter below in identifying the most popular authors in the dataset using the ANB, WorldCat and other online sources of biographical data. Using census data would likely lead to many matches for popular names; meaning that one would need to work with a probabilistic model of location assignment. Thanks to an anonymous supervisor for raising this to my attention.

using the data of Baten and Van Zanden for early modern European book production over 1450-1750. The figure shows that the US in 1790 had a level of book production comparable to that of the richest countries in Europe at the beginning of the 17th century but surpassed them in the span of a few decades, accomplishing in a generation the increase in book production which took them centuries.

Figure IV.2: The US in international comparison



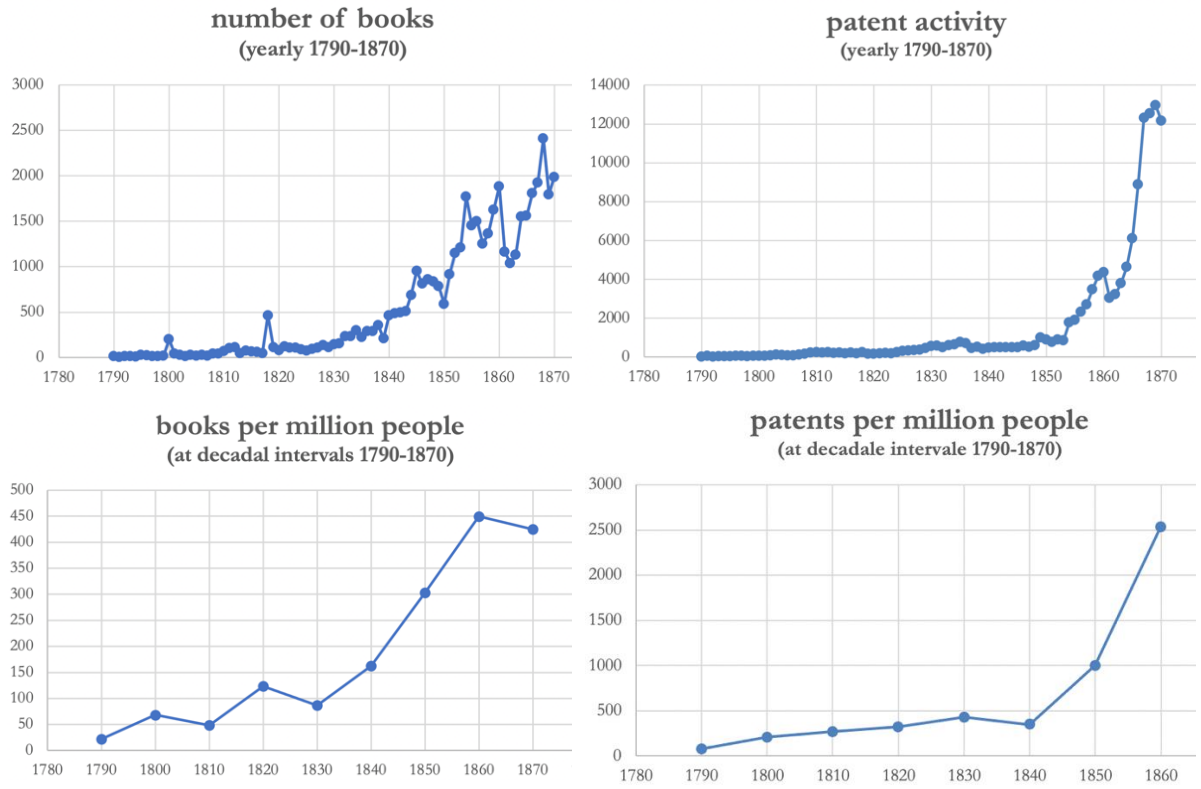


Note: Log-scale is used on the left axis in both panels. The second panel shows data from early modern Europe taken from Baten and Van Zanden (2009), displaying the number of new editions over each half-century per million inhabitants. To make the figures comparable, the first panel shows book production in the US by decade per million inhabitants linearized to a 50-year interval (e.g. the datapoint for 1790 is new editions over 1781-1790 times 5 divided by population in millions in 1790).

As we can see in Figure IV.2, the rise in book production relative to population becomes sustained only after 1830 (see also below Figure IV.3). The timing of this increase is not sufficiently distant from the start of the US' industrial revolution to suggest whether human capital is endogenous to economic development. However, assuming human capital plays a causal role, we would expect the mechanism linking it to output growth to pass through inventive activity leading to a correlation with patent statistics. This is confirmed in Figure IV.3. While the timing of the increase in book production per capita appears to coincide with the beginning of industrialization instead of preceding it, it does precede the surge in patenting by about a decade (corr = 0.81). We also perform granger causality tests on the relationship between patenting and book production to understand the direction of this relationship, which, however, are inconclusive.[97]

[97] See the Appendix, Table A.1.

Figure IV.3 number of books/patents total and per million people



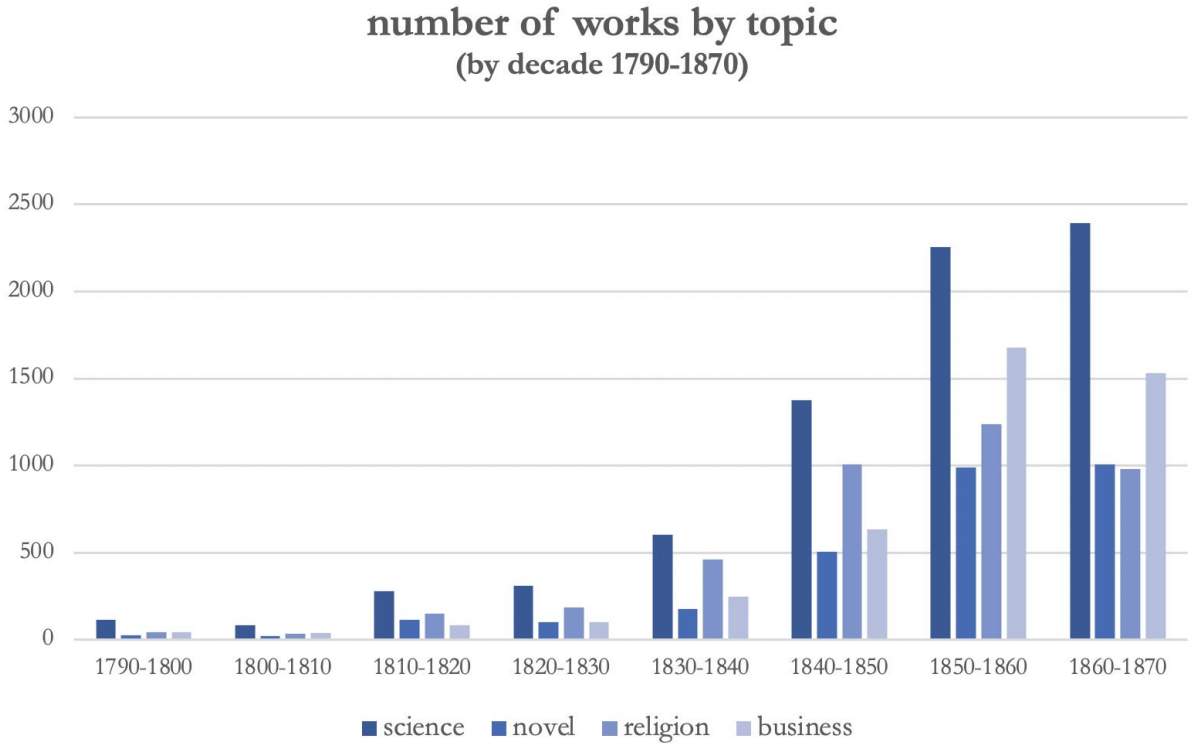
Note: Patent data is from the *US Patent Statistics Report*. Data on population is from the Michael Haines' *Historical, Demographic, Economic and Social Data: The United States, 1790-2002*. The lull in book production in the 1860s coincides with the Civil War years.

In his seminal study of inventive activity in the US over 1790-1846, Sokoloff argued that the rise in patent activity was due to the closure of foreign markets amidst international embargoes and US trade policy which raised demand for locally produced manufactured goods. While these factors affected the entire US, he argues manufacturing activity was concentrated in the northeast due to proximity to waterways and favourable geography.[98] Hence, we are not surprised that the rise in book production occurring approximately a decade before the take-off in inventive activity should also be concentrated in the northeast. However, this can indicate either that the growth of knowledge and human capital proxied by book production was causally related to inventive activity (i.e. more knowledge leading to more inventions), or that a change in consumption patterns took place and consequently affected both book production and inventive activity. To distinguish between these two hypotheses, we need to look at the content of books

[98] Sokoloff and others, supra n 47.

being produced and establish whether it bears any relation to inventions with an industrial application. We turn to this issue in Figure IV.4, focusing on topics which had at least 5000 books classified in them over the sample period.

Figure IV.4: number of works classified in each topic by decade



While Figure IV.4 shows us a strong increase in the absolute number of works dedicated to the broad categories of science, business and novel, two caveats apply: firstly, this increase is not sufficient to outstrip the aggregate growth in book production, hence the relative shares of scientific works in the total stays approximately constant. This is confirmed by testing the changing composition of books statistically in the following logistical regression model:

$$\Pr(\boldsymbol{t}_i = 1) = F(\boldsymbol{\beta}_{0,i} + \boldsymbol{\beta}_{t,i}'\boldsymbol{decade} + \boldsymbol{\varepsilon}_i) \quad (1)$$

Where $\boldsymbol{t}_i$ is an $i \times 1$ vector of dummy variables equal to 1 if a work in our dataset is classified in the i^{th} topic, $\boldsymbol{\beta}_{t,i}$ is an $i \times t$ matrix (one coefficient per topic per decade), $\boldsymbol{decade}$ is a $t \times 1$ vector of dummy variables equal to 1 if a work is

classified in the t^{th} decade, $\boldsymbol{\varepsilon_i}$ is a classic error term and $F()$ is the cumulative logistic distribution. The results in Table IV.1 confirm that, for most categories, the change in composition of topics is not stable over time or statistically distinguishable from zero. However, we do observe a statistically significant and persistent increase in the proportion of textbooks and novels over time.

Table IV.1: logistical regression of topic on decades (1791-1800 is the reference category)

	science	religion	business	novel	textbook
1800	-.016	.081	.098	.224	.404**
	(.107)	(.126)	(.123)	(.148)	(.157)
1810	-.050	.030	-.165	.211*	.418***
	(.083)	(.099)	(.099)	(.118)	(.127)
1820	.016	.174*	-.086	.147	.606***
	(.082)	(.097)	(.097)	(.117)	(.125)
1830	-.062	.258***	-.176*	.058	.651***
	(.077)	(.091)	(.091)	(.112)	(.120)
1840	-.014	.197**	-.119	.202*	.633***
	(.075)	(.089)	(.088)	(.107)	(.117)
1850	-.001	.127	.198**	.296***	.260**
	(.074)	(.088)	(.086)	(.106)	(.117)
1860	.128*	.021	.360***	.492***	.306***
	(.074)	(.089)	(.087)	(.106)	(.117)
Observations	50,298	50,298	50,298	50,298	50,298
Pseudo R2	0.0008	0.0010	0.0069	0.0031	0.0046

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Note: Only topics which had at least over 5000 works classified in them (allowing for overlap) were considered. While the topic "textbook" is contained in "science", we express it as a separate category to show the significant increase in textbooks as a proportion of copyrighted works over time. The textbook category also has more than 5000 individual works.

The second caveat is that, most scientific publications relate to the medical sciences, as is clear from the biographical data on the most popular authors presented below and the meagreness of the "invention" category.[99] Most of those who wrote on scientific topics were physicians. While Mokyr maintains the importance of medically trained professionals in the development of useful

[99] This category is discarded in Table IV.1 due to its small size and insignificant results.

knowledge (since they were usually also trained chemists), their life stories show few direct links between their work and any industrial application.[100]

With regards to the first caveat, a possible limitation of this approach is that there is likely selection bias in the reading public at earlier times. While we have no reliable literacy statistics prior to the mid-19th century, scholars broadly agree that being able to read in 1870 was much more common than in 1790.[101] Hence, if we are interested in how the importance of works addressed to the upper tail of the human capital distribution involving science, philosophy or religion change over time in relative terms, we must take into account the changing composition of the reading public. Observing a declining share of scientific works in the total can therefore be misleading, since this is precisely what we would expect as the economy becomes more literate, assuming there is selection bias of upper-tail human capital individuals in the reading public at earlier times when literacy is less common.

To investigate this hypothesis, we use data on urbanization[102] as a proxy for literacy to track the growth of the reading public and predict the composition of books between "upper-tail" (i.e. science, religion, business, philosophy) and laymen (novels, poetry, popular culture[103]) over time and then compare our predictions to the actual values.

We start by assuming a simple model of book production: individuals are split between upper-tail and laymen. Upper-tail individuals can read with probability

[100] A notable exception being Joseph Priestley, the scientist, minister and physician who discovered oxygen. See the full dataset online, following the link in the data appendix. We compile detailed information on all authors, including a 200-300 word summary of their biography from the ANB, where possible. See for the claim above, Mokyr, "A culture" supra n 17 at 90-91 and 134. Mokyr argues that the medical sciences provided fertile grounds for the development of professionals and knowledge which played a key role in developing the technologies of the industrial revolution and were emblematic of the new "Baconian ethic" which overtook Britain and the Low Countries, inspiring a new approach to science and nature in general.

[101] See Goldin, "A brief history" n 6 at 4.

[102] Data on urbanization is from Michael Haines' *Historical, Demographic, Economic and Social Data: The United States, 1790-2002.*

[103] We treat the residual of books we are not able to classify as laymen publications. This means that, to be conservative, the importance of upper-tail publications is biased downward.

1 and are assumed to be x% of the total population. Laymen read with probability 0.8 if they live in the city and with probability 0.2 if they live in the countryside.[104] Books are produced with the following production function:[105]

$$y_i = \sqrt{l_i}$$

$$l_{upper} = 1(x)l_{total}$$

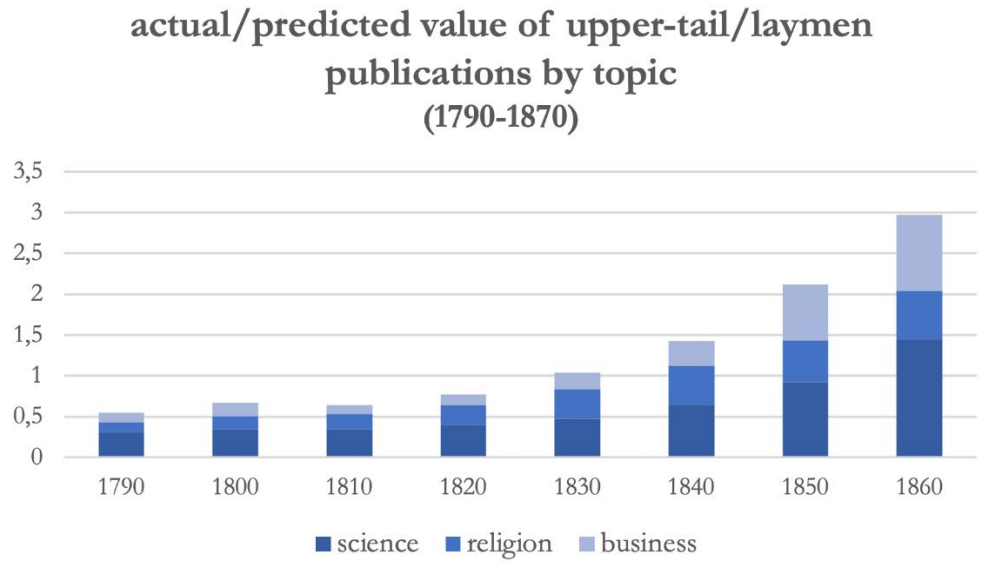
$$l_{laymen} = 0.2\left(p_{rural,t}(1-x)l_{total}\right) + 0.8\left(p_{urban,t}(1-x)l_{total}\right)$$

Where i denotes type, l_i is the reading public per type, l_{total} is the total population of the US and p_t is the time changing probability that a layman lives in a rural or urban area. As the economy urbanizes the number of laymen who can read changes, and hence so does the expected value of upper-tail to laymen publications. The changing selectivity of the reading public thus affects our interpretation of the changing composition of copyrighted works. Comparing the model's predictions with the actual data thus tells us how upper-tail publications increased or decreased in importance starting from our baseline expectation that they should proportionally decline with the growth of the reading public. The results of this exercise are shown in Figure IV.5 where we divide the actual value of upper-tail to laymen publications by the model's predictions with $x = 0.01$.

[104] This is a reasonable assumption given the urban concentration of schools. See Claudia Goldin "A brief history" n 6 at 2.

[105] One could also consider a demand and supply model using data on income per capita to calibrate the demand side, however this would require additional assumptions on the income distribution and does not alter the general result that as the economy gets richer and more urbanized more people learn how to read hence diluting the concentration of upper-tail human capital individuals in the reading public.

Figure IV.5: actual/predicted values of upper-tail to laymen publications

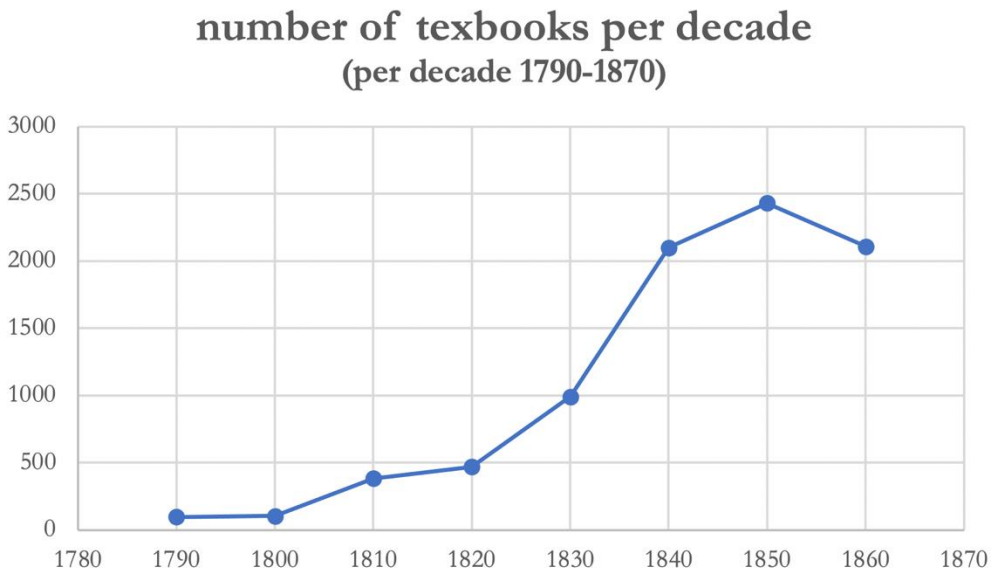


These results suggest that, while as confirmed in Table IV.1, there is no observable trend in the share of works being classified as scientific or any other of the upper-tail topic categories, starting from a baseline expectation of a declining share of these works in the total our data suggests that upper-tail publications increased in importance over time. One way to reconcile the growing discrepancy between the predicted share of upper-tail works to laymen works and the data, is by expanding the class of upper-tail individuals over time (changing x). This would imply a 2% yearly growth rate in the share of individuals belonging to the upper tail of the human capital distribution,[106] which is comparable to the estimated growth rates of output and physical capital over this period.[107]

This result pushes back the beginning of the spread of knowledge to the broader population before the first reliable data on federal schooling and literacy.[108] This interpretation is supported by the impressive growth in the number of textbooks being copyrighted, observable in Figure IV.6 and Table IV.1.

[106] This can be computed by setting the growth rate of x at a level which makes the actual/predicted values of upper-tail to laymen publications approximately equal over time.
[107] See Gallman, supra n 2.
[108] See Goldin, supra n 6.

Figure IV.6: number of textbooks per decade



To add more rigor to this conclusion, we directly analyze the biographies of the 30 most prolific authors for each decade over 1790-1870, paying particular attention to their socio-economic background. Not all authors have entries in the ANB, hence we use multiple sources to trace their life stories from their names.[109] When we are not able to find any information concerning an author, as a last resort we infer from their written work what field they were interested in and whether they held a particular position (professor, headmaster, reverend…).

This allows us to identify the most intellectually active American authors over time. Traditionally the study of intellectual history in economic development has focused on pivotal figures, such as enlightenment philosophers, inventors and scientists.[110] This poses several challenges when one seeks to ascertain to what extent cultural changes or scientific advances in knowledge were widespread, given the selectivity of the individuals which make it into that sample. By constructing a list of the 30 most *prolific* authors by decade we eschew this problem for two reasons: firstly, we have a decently large sample of over 240 individuals evenly distributed over time and, secondly, the sampling technique is

[109] See the full table in the online appendix. This comprises 240 authors in total.
[110] See Mokyr, supra n 8 and 17; and McCloskey, supra n 11 and 19.

based on how many works an author penned, not his or her saliency as a figure of significance in economic and intellectual history. The results of this exercise can be summarized below. Figure IV.7 shows the distribution of authors by area of interest over time. The variation in total number of authors per decade is due to the fact that some authors fall in more than one category (e.g. they are both teachers and scientists).

Figure IV.7: distribution of authors by decade over 5 categories: science, legal, religious, teacher and rest (including novelists, poets, illustrators etc.)

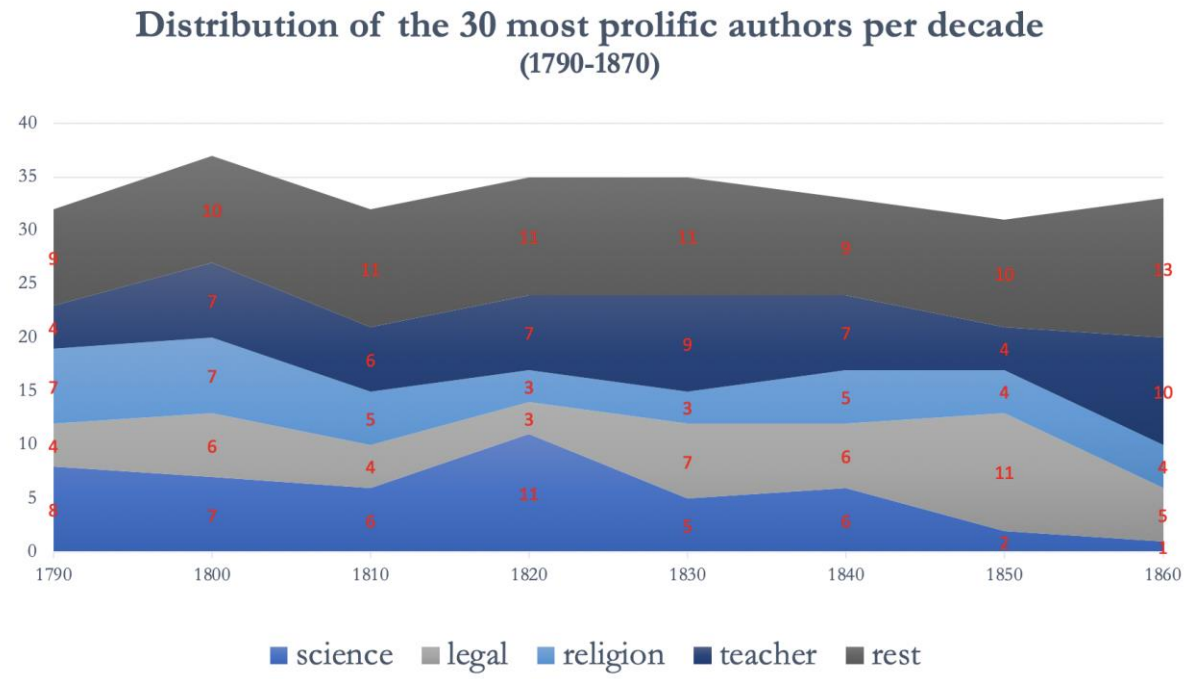
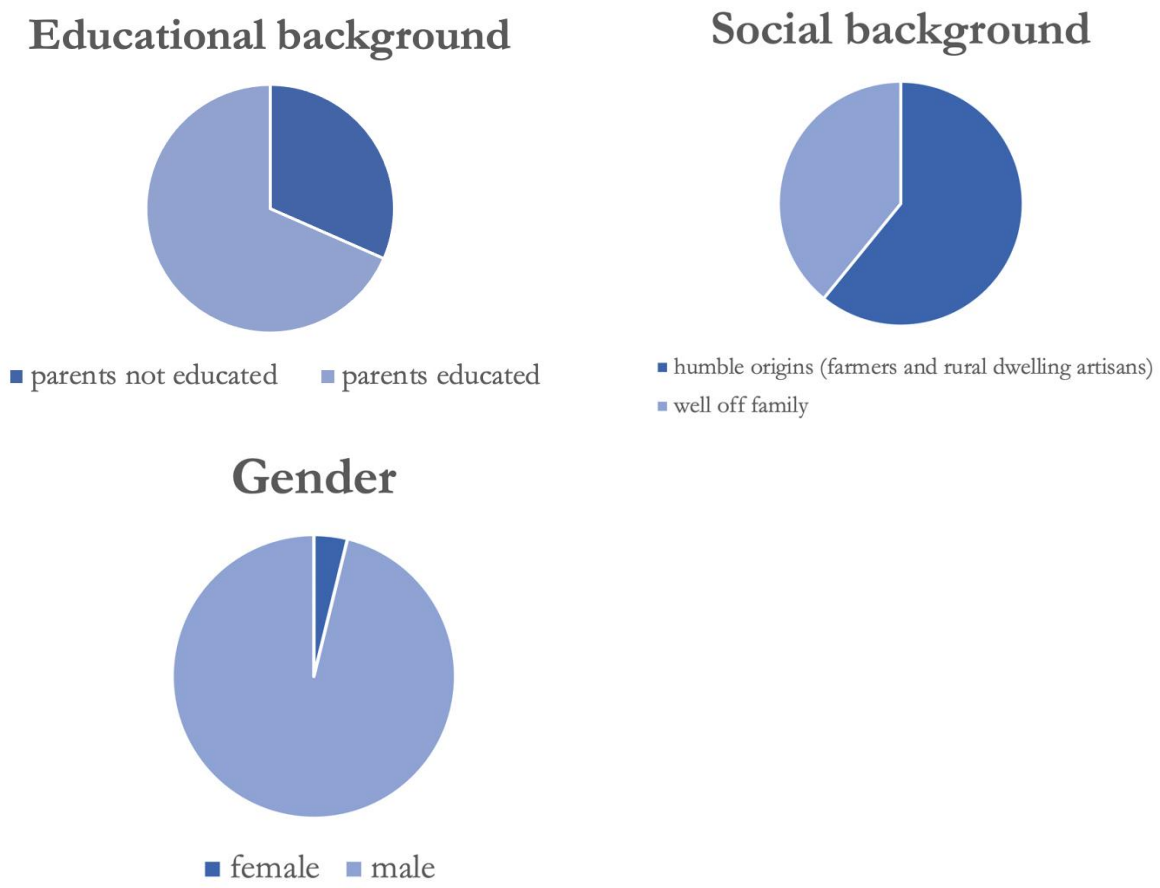


Figure IV.8: gender and socio-educational background of American authors over 1790-1870



Note: Gender is assigned for all 240 authors. Social and educational background statistics are based on a total of respectively 46 and 38 authors for whom matching with the ANB database provided viable information.

Figure IV.7 shows a growing number of teachers among the most prolific authors of the 19th century at the expense of scientists and religious figures. These results support the findings in Table IV.1 and Figure IV.5/6. The growing number of legal scholars, which account for the largest group among the most prolific authors of 1850 and are a sizeable group throughout this timeframe suggests that legal texts, court reports and opinions were in high demand. This gives some anecdotal support to those who emphasize the importance of institutions governing disputes and establishing property rights as key elements of economic development.[111] Due to space limitations, we refrain from presenting the life stories of these 240

[111] See North, supra n 51 and Douglass C North, *Institutions, institutional change and economic performance*. CUP, 1990.

American authors.[112] However, one important finding from this research which emerges in Figure IV.8 is the remarkable inclusivity of this class of American intellectuals with respect to social and educational background.[113] Although meagre overall, the proportion of female authors increases slightly over time.[114] These results are consistent not only with our findings in Figure IV.5/6, but also with widely shared perceptions of the early US as being a locus of remarkable dynamism, opportunity and intellectual depth which are commonly found in the literature.[115]

Conclusion

In this paper we presented the first time series of book production for the United States over 1790-1870. Our findings show a sustained increase in book production starting at around the beginning of the US' industrial revolution. We analyze the content of books produced using lists of keywords which we match on words extracted from the title pages of books written over the reference period of our dataset. This shows a growing number of books dedicated to science, business related topics, novels and a somewhat slower growth of religious topics. The only category of books to grow in a sustained way over the 19th century as a proportion of total books are textbooks, which show a pronounced increase starting in 1800. We also use urbanization data as a proxy for the spread of literacy to predict the shares of topics over time. Intuitively, our model predicts that the share of scientific, religious and business-related books should decline as the reading public becomes less selective and expands beyond the upper-tail of the human

[112] See the Online Appendix. We are able to get full "life-stories" from the ANB for only 25% of the 240 baseline list.
[113] This finding is stable over time.
[114] See also Figure A.2 in the appendix. Using word frequency distributions, we extract the 500 most common words per decade, comparing the relative position of "mr." and "mrs." (which in a title page are likely to correlate with author sex) in this ranking shows a remarkable convergence over time.
[115] See Clayne Pope, "Inequality in the nineteenth century" in R Gallman and S Engerman eds., *Cambridge economic history of the United States: Vol2 The long nineteenth century*, CUP 2008; Fisher, "Albion's seed" supra n 91; on social mobility in the nineteenth century see the work of Joe Ferrie e.g. Joseph P. Ferrie, "History lessons: The end of American exceptionalism? Mobility in the United States since 1850." *Journal of Economic Perspectives* 19, no. 3 (2005): 199-215.

capital distribution. Comparing the actual data on "upper-tail" publications to "laymen" publications suggests that the relative size of the upper-tail of the human capital distribution (the proportion of individuals interested in scientific, religious or business-related books) grew substantially over the 19th century. We also construct a dataset of the 240 most prolific authors of the 19th century (decade by decade 1790-1870) finding in particular those dedicated to scientific research to decline, benefitting the categories of teachers and legal scholars. This is consistent with an expansion of the upper-tail class of individuals dedicated to science: while the number of scientific works is rising, it is spread over an increasing number of people leading to less authors figuring in the top 30 in later decades.

This paper has only scratched the surface of what can be gleaned from the analysis of American intellectual production over the 19th century. Further research should use author names to assign each publication to a US county based on the provenance of the author. This would likely need to be done manually as names alone are not informative enough for census matching. Doing so would enable the construction of a high-density spatial dataset which would allow performing causal inference of hypotheses regarding the role of book production in economic growth.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration." *American Economic Review* 102, no. 5 (2012): 1832-56.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration." *Journal of Political Economy* 122, no. 3 (2014): 467-506.

Abramovitz, M., and P. A. David. "Growth in the Era of Knowledge-Based Progress." In S. Engerman and RE Gallman eds, *The Cambridge Economic History of the United States Vol3: The twentieth century*, CUP 2000.

Acemoglu, Daron. "Technical change, inequality, and the labor market." *Journal of economic literature* 40, no. 1 (2002): 7-72.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." *American economic review* 91, no. 5 (2001): 1369-1401.

Allen, Robert C. *The British industrial revolution in global perspective*. Cambridge University Press, 2009.

Archibugi, Daniele, and Andrea Filippetti. "The globalisation of intellectual property rights: four learned lessons and four theses." *Global Policy* 1, no. 2 (2010): 137-149.

Baten, Joerg, and Jan Luiten Van Zanden. "Book production and the onset of modern economic growth." *Journal of Economic Growth* 13, no. 3 (2008): 217-235.

Cantoni, Davide, and Noam Yuchtman. "Medieval universities, legal institutions, and the commercial revolution." *The Quarterly Journal of Economics* 129, no. 2 (2014): 823-887.

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

Clark, Gregory. "Human capital, fertility, and the industrial revolution." *Journal of the European Economic Association* 3, no. 2-3 (2005): 505-515.

Clark, Gregory. *A farewell to alms*. Princeton University Press, 2008.

Clark, Gregory. "The Enlightened Economy: An Economic History of Britain 1700-1850: Review Essay." *Journal of Economic Literature* 50, no. 1 (2012): 85-95.

Clark, Gregory. "The industrial revolution: A cliometric perspective." In In C. Diebolt, M. Haupert eds, *Handbook of Cliometrics* (2016): 197-235.

Clark, Gregory. "The industrial revolution." In S. Durlauf and P. Aghion eds, *Handbook of economic growth*, vol. 2, pp. 217-262. Elsevier, 2014.

David, Paul A. "New light on a statistical dark age: US real product growth before 1840." *The American Economic Review* 57, no. 2 (1967): 294-306.

Davis, Joseph H. "An annual index of US industrial production, 1790–1915." *The Quarterly Journal of Economics* 119, no. 4 (2004): 1177-1215.

De Vries, Jan. "The industrial revolution and the industrious revolution." *The Journal of Economic History* 54, no. 2 (1994): 249-270.

Dittmar, Jeremiah E. "Information technology and economic change: the impact of the printing press." *The Quarterly Journal of Economics* 126, no. 3 (2011): 1133-1172.

Dittmar, Jeremiah. "The impact of a new good: The printed book." *Department of Economics, American University*(2011).

Dittmar, Jeremiah, and Skipper Seabold. "New media and competition: printing and Europe's transformation after Gutenberg." *Working Paper LSE,* 2019.

Dittmar, Jeremiah, and Rolf Meisenzahl. The research university, invention, and industry: Evidence from german history. *Working Paper LSE*, 2021.

Doepke, Matthias, and Fabrizio Zilibotti. "Occupational choice and the spirit of capitalism." *The Quarterly Journal of Economics* 123, no. 2 (2008): 747-793.

Easterlin, Richard A. "Why isn't the whole world developed?." *The Journal of Economic History* 41, no. 1 (1981): 1-17.

Ferrie, Joseph P. *Yankeys now: Immigrants in the antebellum US 1840-1860*. Oxford UP, 1999

Ferrie, Joseph P. "History lessons: The end of American exceptionalism? Mobility in the United States since 1850." *Journal of Economic Perspectives* 19, no. 3 (2005): 199-215.

Fischer, David Hackett. *Albion's seed: Four British folkways in America. Vol. 1. America: A Cultural History*, OUP 1989.

Gallman, Robert E. "Gross National Product in the United States, 1834–1909," in Dorothy S. Brady (ed.), *Output, Employment, and Productivity in the United States After 1800*, 1966.

Gallman, Robert E. "Economic growth and structural change in the long nineteenth century" in Gallman, R. and Engerman, S. eds., *The Cambridge economic history of the United States* vol. 2, 2008.

Galor, Oded, and Omer Moav. "From physical to human capital accumulation: Inequality and the process of development." *The Review of Economic Studies* 71, no. 4 (2004): 1001-1026.

Galor, Oded. *Unified growth theory*. Princeton University Press, 2011.

Goldin, Claudia. "A brief history of education in the United States." *NBER WP* 119 (1999).

Goldin, Claudia. "Human capital" In C. Diebolt, M. Haupert eds, *Handbook of Cliometrics* (2016): 55-86.

Goldin, Claudia. "The human-capital century and American leadership: Virtues of the past." *The Journal of Economic History* 61, no. 2 (2001): 263-292.

Glaeser, Edward L., Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. "Do institutions cause growth?." *Journal of economic Growth* 9, no. 3 (2004): 271-303.

Kelly, Morgan, Joel Mokyr, and Cormac Ó. Gráda. "Precocious Albion: a new interpretation of the British industrial revolution." *Annu. Rev. Econ*. 6, no. 1 (2014): 363-389.

Khan, B. Zorina. *The Democratization of Invention: patents and copyrights in American economic development, 1790-1920*. Cambridge University Press, 2005.

Lamoreaux, Naomi R., and Kenneth L. Sokoloff. "Market trade in patents and the rise of a class of specialized inventors in the 19th-century United States." *American Economic Review* 91, no. 2 (2001): 39-44.
Lamoreaux, Naomi R., Kenneth L. Sokoloff, and Dhanoos Sutthiphisal. "Patent alchemy: the market for technology in US history." *Business History Review* 87, no. 1 (2013): 3-38.
Landes, David S. "Why Europe and the west? Why not China?." *Journal of Economic Perspectives* 20, no. 2 (2006): 3-22.
McCloskey, Deirdre N. *The bourgeois virtues: Ethics for an age of commerce*. University of Chicago Press, 2010.
McCloskey, Deirdre N. *Bourgeois dignity: Why economics can't explain the modern world*. University of Chicago Press, 2010.
McCloskey, Deirdre N. *Bourgeois equality*. University of Chicago Press, 2021.
Meyer, David R. *The roots of American industrialization*. JHU Press, 2003.
Mitch, David. "The role of skill and human capital in the "British" industrial revolution." in *Mokyr J., The British Industrial Revolution: An Economic Perspective, Boulder, Colorado* (1999): 241-279.
Mokyr, Joel. *A culture of growth*. Princeton University Press, 2016.
Mokyr, Joel. *The enlightened economy: an economic history of Britain, 1700-1850*. New Haven: Yale University Press, 2009.
Mokyr, Joel. *The gifts of Athena*. Princeton University Press, 2011.
Mokyr, Joel. *The lever of riches: Technological creativity and economic progress*. Oxford University Press, 1992.
North, Douglass C. "Industrialization in the United States (1815–60)." In *The Economics of Take-off into Sustained Growth*, pp. 44-62. Palgrave Macmillan, London, 1963.
North, Douglass C. *Institutions, institutional change and economic performance*. CUP, 1990.
Nunn, Nathan. "History as evolution." In *The Handbook of Historical Economics*, pp. 41-91. Academic Press, 2021.
O' Brien, Patrick. "The Needham question updated: a historiographical survey and elaboration" *History of technology 29* (2009).
Pope, Clayne. "Inequality in the nineteenth century" in R Gallman and S Engerman eds., *Cambridge economic history of the United States: Vol2 The long nineteenth century*, CUP 2008.
Rostow, Walter ed. *The economics of take-off into sustained growth*. Springer, 1963.
Schultz, Theodore W. "Investment in human capital." *The American economic review* 51, no. 1 (1961): 1-17.
Sequeira, Sandra, Nathan Nunn, and Nancy Qian. "Immigrants and the Making of America." *The Review of Economic Studies* 87, no. 1 (2020): 382-419.
Slauter, Will. "Toward a history of copyright for periodical writings: Examples from nineteenth-century America." HAL Archives Ouvertes (2014).
Smith, J Russell. *North America*, London: G. Bell and Sons, 1924.
Sokoloff, Kenneth L. "Inventive activity in early industrial America: evidence from patent records, 1790–1846." *The Journal of Economic History* 48, no. 4 (1988): 813-850.

Squicciarini, Mara P., and Nico Voigtländer. "Human capital and industrialization: Evidence from the age of enlightenment." *The Quarterly Journal of Economics* 130, no. 4 (2015): 1825-1883.
Temin, Peter. *Causal Factors in American Economic Growth in the Nineteenth Century*. Macmillan International Higher Education, 1975.

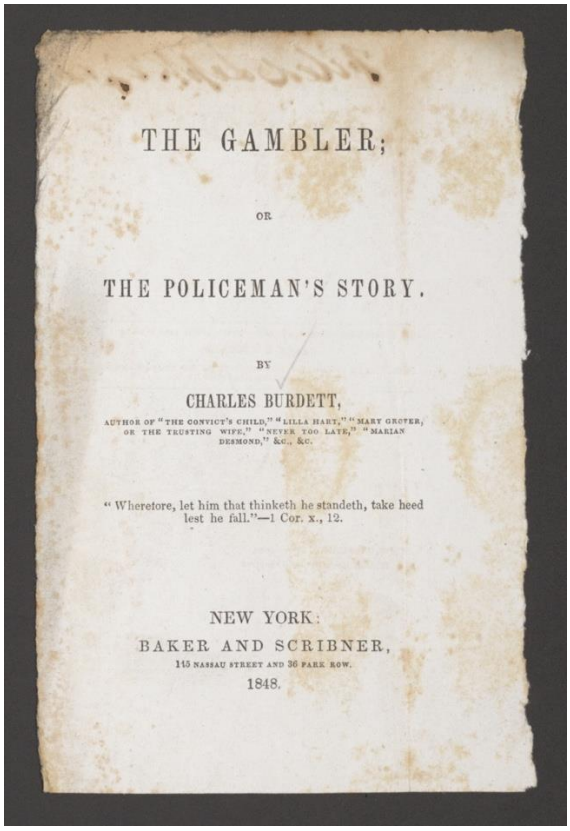
Appendix

All code[116] (used to download, pre-process and read the title page images) and data files (raw and polished) is available online at the following link. Slight changes were made to the code files to preserve anonymity.

https://drive.google.com/drive/folders/1F2h__9d9micMttZHJtWXbBW8xMyCJLzm?usp=sharing

This includes the list of the 240 most prolific authors over the reference period along with a summary of their life based on the ANB entries (where available). Below, we show an example of a title page image in Figure A.1. As can be seen, the quality of the images is generally high, which allows us to extract text relatively easily.

Figure A.1; example of a title page image



The collection of downloaded images is also available for analysis. It is not available online as it is too large to be stored on a google drive. If you would like to see all the raw images, please contact the author. Finally, the results of the granger causality test of book production and patenting statistics is shown below in Table A.1.

[116] Minor modifications were made to preserve anonymity

Table A.1

equation	excluded	Chi-squared	df	p>chi2
books	patents	19.66	2	0.000
books	all	19.66	2	0.000
patents	books	4.24	2	0.12
patents	all	4.24	2	0.12

As we can see the test suggests that the causal relationship runs from patents to books (not the other way around). However, books also appear to have an effect on patenting which falls just short of statistical significance. This test is shown for completeness, but it's not clear to what extent a granger causality test can shed light on this relationship since we would expect that production have an effect on economic activity which appears and persists after many decades. Hence, the appropriate test would be to use cross-sectional (or panel) data examining the relationship between two variables at two distant intervals, similarly to what is done by other authors in the literature mentioned in Section II. Supplemental analysis is provided in Figures A.2 and A.3 which look at the frequency of words in each topic/category appearing by decade. As we can see, a notable finding is that the rank of the string "mrs." relative to the rank of the string "mr." declines substantially over time, indicating that their relative frequencies are converging. Although this could indicate a growing proportion of female authors, it may also indicate a growing number of works addressed to a female public (such as fashion periodicals which become more common over time).

Figure A.2

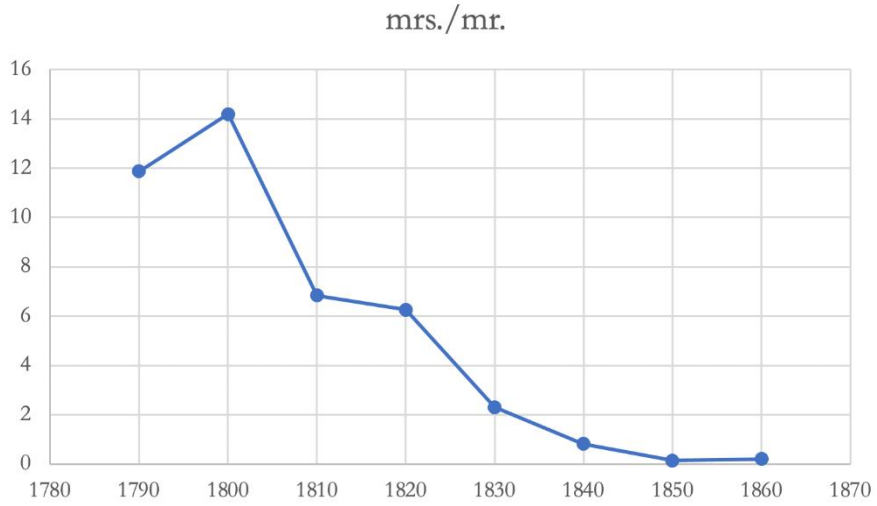
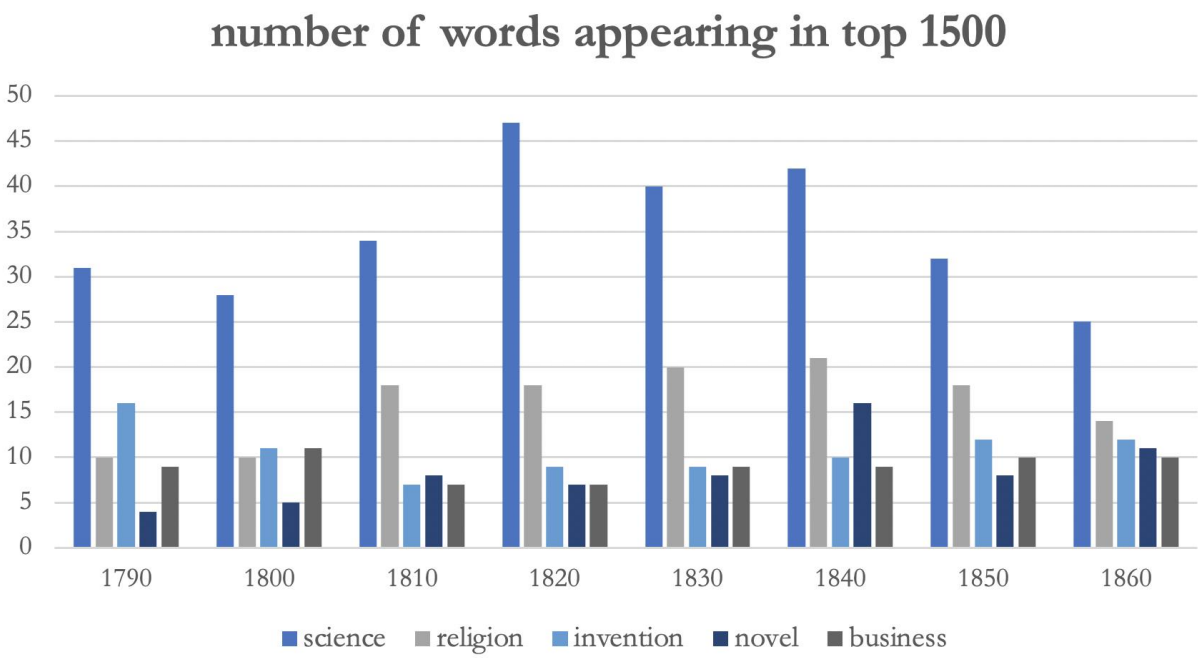


Figure A.2 shows the rank of the string "mrs." in the list of 1500 most common words per decade relative to the rank of the string "mr.". The downward trend indicates a growing proportion of female authors.

Figure A.3 frequency of words by topic appearing in the top 900 words per decade



The list of most popular words starts from a baseline of 1500 from which stop-words (prepositions, articles, pronouns etc) are removed, resulting in a list of approximately 900 words for each decade. This figure shows that scientific words consistently appear among the top words in each decade to a greater extent than other categories. Also notable is the large representation of the "invention" category, which is not as well represented in terms of the number of works published by decade, meaning that the "invention" keywords are repeated many times in a small number of works.