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

Tancredi Rapone[1]

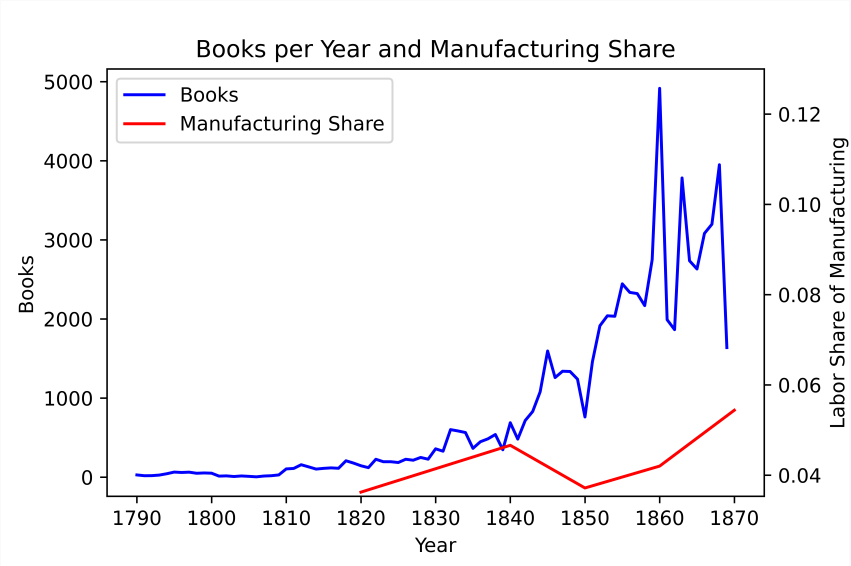
Macro Club, Winter 2025

[1]Barcelona School of Economics (BSE) and Universitat Autònoma de Barcelona (UAB)

Research Question

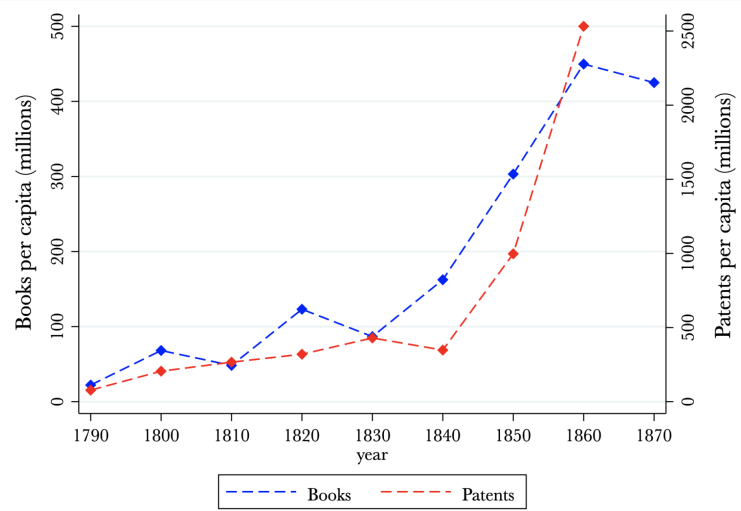
- **Q:** What role did knowledge creation play during the Second Industrial Revolution in the US?
  - What kind of knowledge correlates with economic development?
  - Who produces knowledge and why?
  - Cultural determinants of knowledge production?

Preview of the Results 1/2



Manufacturing Share

Preview of the Results 2/2



Literature Review

Existing Literature

Controversial debate over importance of human capital in the IR:

- Mitch (1993); Galor (2005); Clark (2007) and Acemoglu (2002):
  - Human capital was unimportant in the first phase of the Industrial Revolution.

  How could this be?
- Large literature finds robust association between HC and outcomes in modern data (e.g. Barro and Lee, 2013).

⇨ Squicciarini and Voigtlander (2015) were the first to test Mokyr's (2002) thesis that upper-tail human capital mattered for growth.

Roadmap

1. Data Collection
2. Descriptive results
3. Empirical Strategy
4. Results

Data Collection

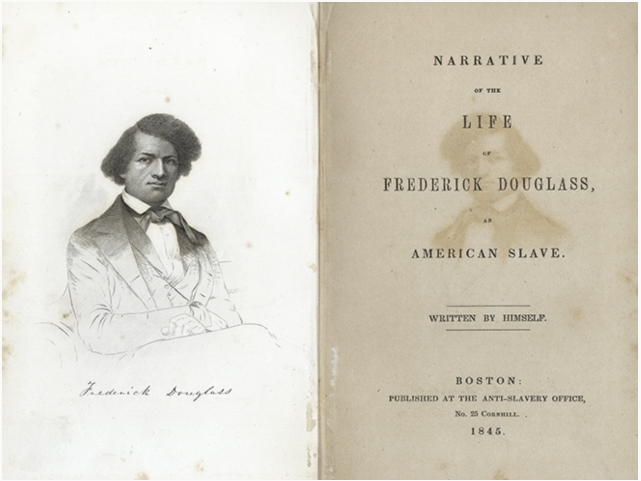
Primary Sources: LOC Early Copyright Title Pages

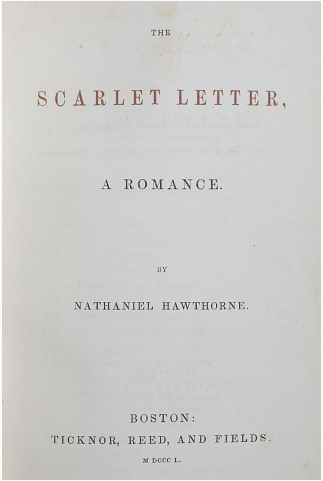
The data source for this project is the Library of Congress (LOC) "Collection of Early Copyright Title Pages"

- Recently digitized in 2020.
- Arguably comprises the *universe* of intellectual production, excepting patents, in the US over 1790-1870.
  - Books
  - Illustrations and designs
  - Articles
  - And much more
- Definitely more comprehensive than library catalogues used by comparable studies (e.g. Baten and Van Zanden (2008), Chaney (2016)).
- However, information is not as easily extracted $\rightarrow$ measurement error.

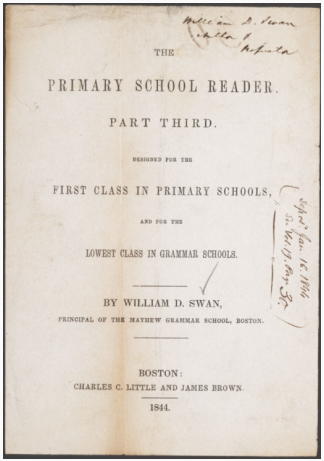
All other data is from Fiszbein (2022) and Haines (2010).

More on the Data: Example 1/2





More on the Data: Example 2/2



THE

PRIMARY SCHOOL READER.

PART THIRD.

DESIGNED FOR THE

FIRST CLASS IN PRIMARY SCHOOLS,

AND FOR THE

LOWEST CLASS IN GRAMMAR SCHOOLS.

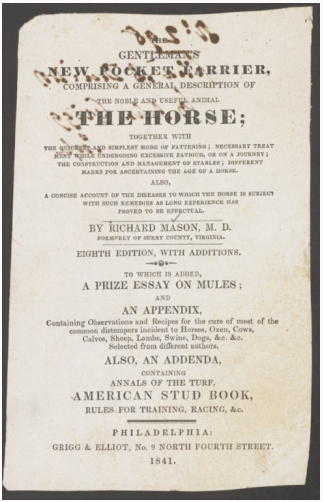
BY WILLIAM D. SWAN,

PRINCIPAL OF THE MAYHEW GRAMMAR SCHOOL, BOSTON.

BOSTON:

CHARLES C. LITTLE AND JAMES BROWN.

1844.



GENTLEMAN'S

NEW POCKET FARRIER,

COMPRISING A GENERAL DESCRIPTION OF

THE NOBLE AND USEFUL ANIMAL,

THE HORSE;

TOGETHER WITH

THE QUICKEST AND SIMPLEST MODE OF FATTENING; NECESSARY TREATMENT WHILE UNDERGOING EXCESSIVE FATIGUE, OR ON A JOURNEY; THE CONSTRUCTION AND MANAGEMENT OF STABLES; DIFFERENT MARKS FOR ASCERTAINING THE AGE OF A HORSE.

ALSO,

A CONCISE ACCOUNT OF THE DISEASES TO WHICH THE HORSE IS SUBJECT WITH SUCH REMEDIES AS LONG EXPERIENCE HAS PROVED TO BE EFFECTUAL.

BY RICHARD MASON, M. D.

FORMERLY OF SURRY COUNTY, VIRGINIA.

EIGHTH EDITION, WITH ADDITIONS.

TO WHICH IS ADDED,

A PRIZE ESSAY ON MULES;

AND

AN APPENDIX,

Containing Observations and Recipes for the cure of most of the common distempers incident to Horses, Oxen, Cows, Calves, Sheep, Lambs, Swine, Dogs, &c. &c. Selected from different authors.

ALSO, AN ADDENDA,

CONTAINING

ANNALS OF THE TURF,

AMERICAN STUD BOOK,

RULES FOR TRAINING, RACING, &c.

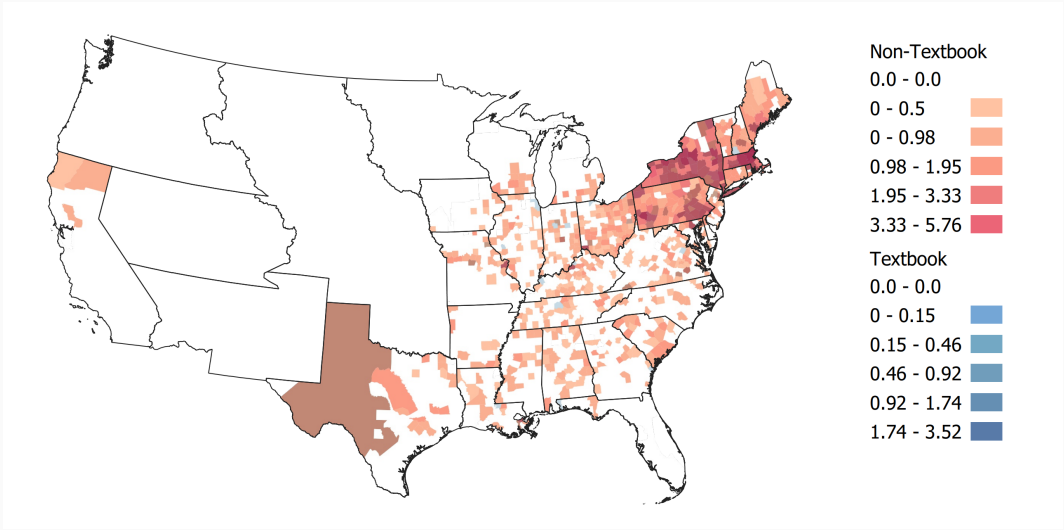
PHILADELPHIA:

GRIGG & ELLIOT, No. 9 NORTH FOURTH STREET.

1841.

Descriptives

Map: Copyrighting 1840-1860 (log. scale)



Knowledge

Topic Analysis

Methodology

Q: What do American authors produce over 1790/1870? Does this change over time, because of secularization, science, changing consumer preferences?

To answer this I use a mixed keyword/logistic classification method:

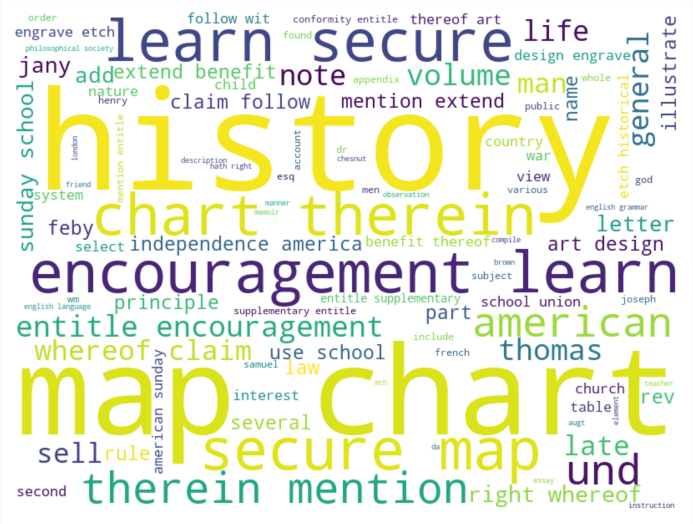
- Keyword strategy presents a tradeoff between short and long keyword lists. Long lists generate Type 1 errors, short lists generate Type 2 errors.
- Unsupervised topic models used for this kind of task (e.g. LDA) usually work best with large texts.

⇨ Apply a short keyword list and classify with a threshold (e.g. top 10% and bottom 10%) use this as a training dataset for a logistic classifier.

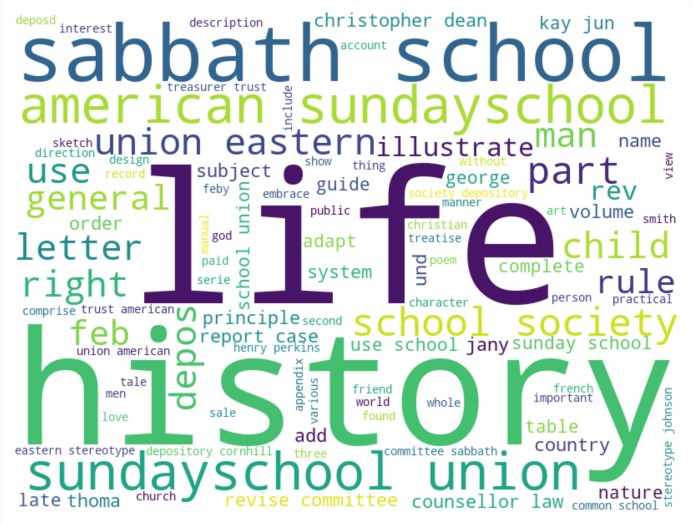
This is an original procedure which may result in some bias.

Wordcloud 1790-1810

Wordcloud 1810-1830

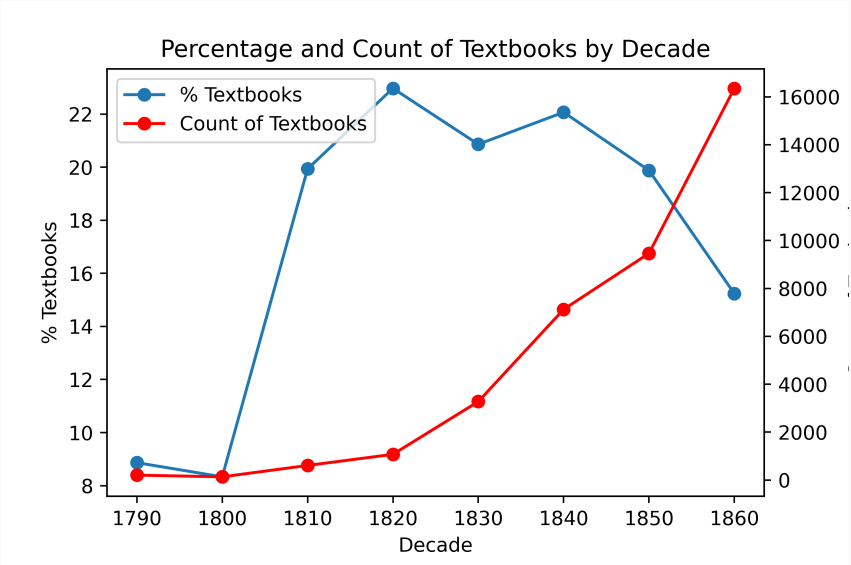


Wordcloud 1830-1850



Wordcloud 1850-1870

Textbooks and other Academic Works



Empirical Strategy

Empirical Models

The baseline empirical strategy follows Fiszbein (2022):

$$y_{i,t} = \alpha_s + \beta \ln(Copyright_{1840-1860}) + \Gamma' X_{1850} + \varepsilon_{i,t} \quad (1)$$

Where X is a vector of controls measured in 1850. Standard errors in 1 are calculated using the Conley (1999) correction with 100km cutoff.

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			✓

Instrument Construction

The instrument exploits vast internal migration patterns occurring in this time-period.

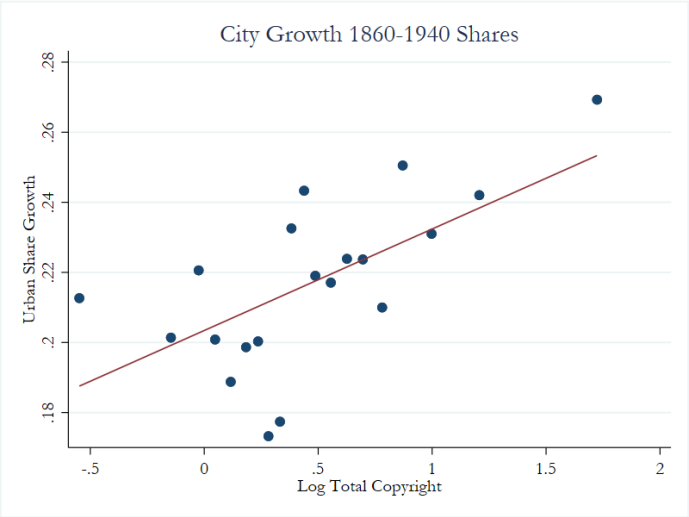
$$\overline{books}_{i,s} = \sum_{l \neq s} \left(\frac{authors_l}{people_l} - \frac{authors_{us}}{people_{us}} \right) \pi_{i,l} \tag{2}$$

Where i denotes county and s, l denote states. The instrument helps deal with selection bias and measurement error in knowledge production.

Migration Balance

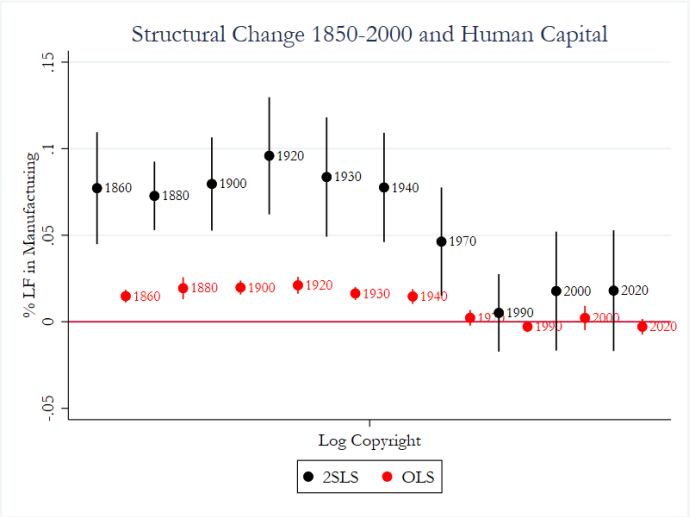
Results

Results: City Growth



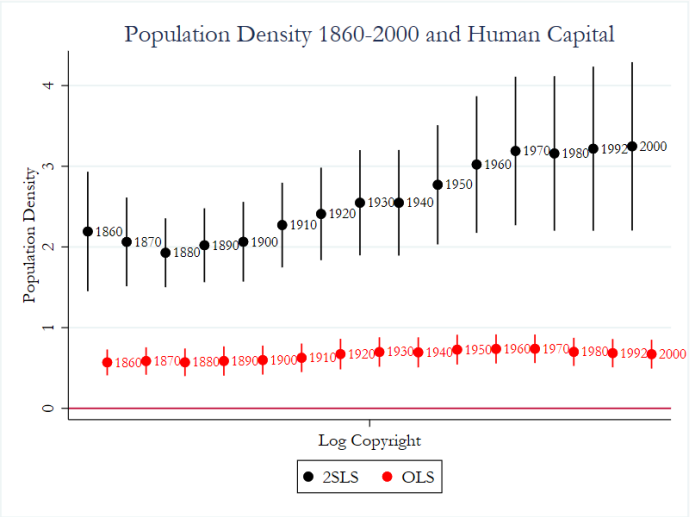
iv ols

Results: Structural Change



iv ols

Results: Population Density



iv ols

OLS: 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
Conley standard errors used with a 100km cutoff.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

IV: 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

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

Drivers of Copyrighting: Measuring Religious Diversity

Definition: Let π_i be the fraction of the population belonging to religious group i in 1870.

1. Reynal-Querol Index (RQ)

$$RQ = 1 - \sum_{i=1}^{N} \left(\frac{1/2 - \pi_i}{1/2} \right)^2 \pi_i \tag{3}$$

- This index measures the deviation of each group's share from an equal split ($\frac{1}{2}$).

2. Fractionalization Index (FRAC)

$$FRAC = 1 - \sum_{i=1}^{N} \pi_i^2 = \sum_{i=1}^{N} \pi_i (1 - \pi_i) \tag{4}$$

- The probability that two randomly chosen individuals belong to different religious groups. - Ranges from 0 (perfect homogeneity) to a maximum near 1 (high diversity).

Fractionalization and Polarization

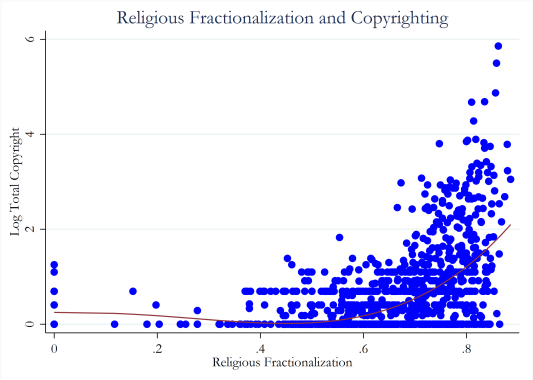


Figure 1: FRAC

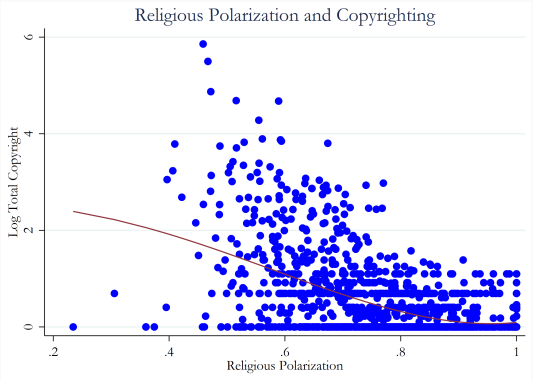


Figure 2: RQ

Conclusion

- The sustained increase in book production took place before industrialization in the US.
- The topic composition of copyrighted works suggests a strong interest learning useful "knowledge".
- Spatial distribution of copyrighting (in pre-industrial times) is strongly correlated with industrialization and population density over the Second Industrial Revolution.

Limitations/future research:

- Identification: instrument needs more work.
- Topic analysis: more fine grained classification would be useful to understand causal channels.

appendix

Tables and Figures: Supplementary Material

Appendix: Migration

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.

Back

Appendix: Balance

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

Back

OLS Structural Change Table: Incl. Soc. Controls

	(1) 1860	(2) 1880	(3) 1900	(4) 1920	(5) 1940
Log Copyright	0.00764***	0.00868***	0.00977***	0.0111***	0.00529*
	(5.62)	(3.60)	(4.62)	(3.50)	(2.32)
% In School 1850	-0.0122***	-0.00172	0.00729	0.0160**	0.0176**
	(-3.92)	(-0.53)	(1.50)	(2.69)	(3.05)
% Literate	-0.0139	0.0159	0.0315	0.0601	0.0816**
	(-0.74)	(0.92)	(1.13)	(1.58)	(2.92)
% Urbanized	-0.00537	0.0214	-0.00949	-0.0814***	-0.0779***
	(-0.40)	(0.96)	(-0.43)	(-3.51)	(-6.31)
Log Population 1850	-0.000170	0.00153	0.00259*	0.00127	0.00400*
	(-0.10)	(1.62)	(2.13)	(1.09)	(2.15)
N	1504	1504	1504	1504	1504
R2	0.292	0.282	0.210	0.204	0.198
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
Conley standard errors used with a 100km cutoff.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Back

IV Structural Change Table: Excl. Soc. Controls

Table 3: IV Regressions: Labor Share all Controls

	(1) 1860	(2) 1880	(3) 1900	(4) 1920	(5) 1940
Log Copyright	0.139**	0.109***	0.134***	0.185***	0.156***
	(3.04)	(4.24)	(4.00)	(3.88)	(3.31)
% In School 1850	-0.0356**	-0.0151	-0.00909	-0.00765	-0.000879
	(-2.99)	(-1.90)	(-1.13)	(-0.73)	(-0.09)
% Literate	-0.144*	-0.0778	-0.0869	-0.102	-0.0483
	(-2.28)	(-1.79)	(-1.62)	(-1.38)	(-0.71)
% Urbanized	-0.280*	-0.165*	-0.245**	-0.401***	-0.354**
	(-2.49)	(-2.57)	(-3.16)	(-3.42)	(-3.04)
N	1505	1505	1505	1505	1505
Kleibergen-Paap	13	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
Conley standard errors used with a 100km cutoff.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Back

OLS Pop. Density Table

	(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

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

Back

IV Pop. Density Table

Table 4: IV Regressions: Pop. Density 1860-1980

	(1) 1860	(2) 1890	(3) 1920	(4) 1950	(5) 1980
Log Copyright	3.266***	2.605***	3.305***	4.117***	5.235***
	(3.47)	(4.62)	(4.55)	(3.93)	(3.60)
% In School 1850	-0.0854	-0.257	-0.383	-0.356	-0.328
	(-0.36)	(-1.64)	(-1.92)	(-1.37)	(-1.00)
% Literate 1850	0.0205	-1.201	-3.056*	-3.711*	-4.411*
	(0.01)	(-1.05)	(-2.29)	(-2.23)	(-2.15)
% Urbanized 1850	-4.850*	-2.639*	-4.043*	-6.080*	-9.373**
	(-2.11)	(-2.03)	(-2.39)	(-2.48)	(-2.79)
N	1505	1505	1505	1505	1505
Kleibergen-Paap F	13	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
Conley standard errors used with a 100km cutoff.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Back

OLS City Growth Table

	(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

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

Back

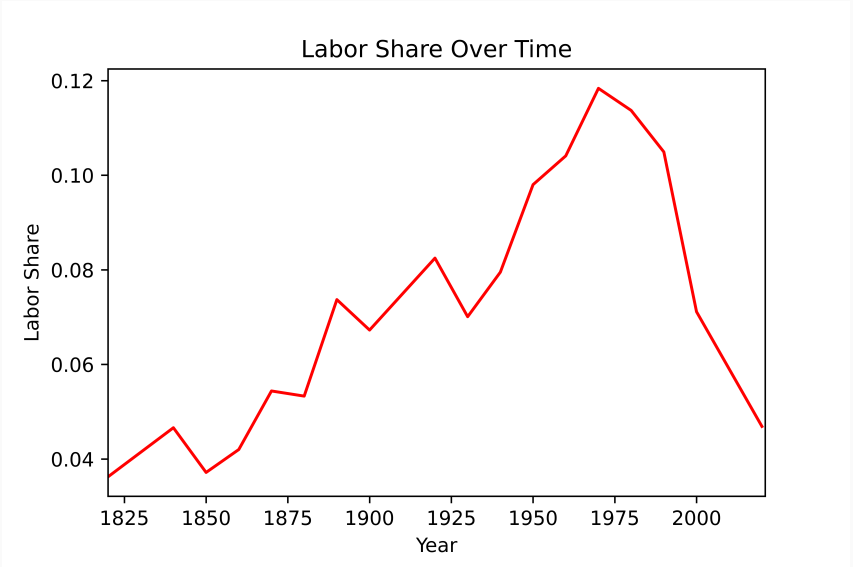
IV City Growth Table

	(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

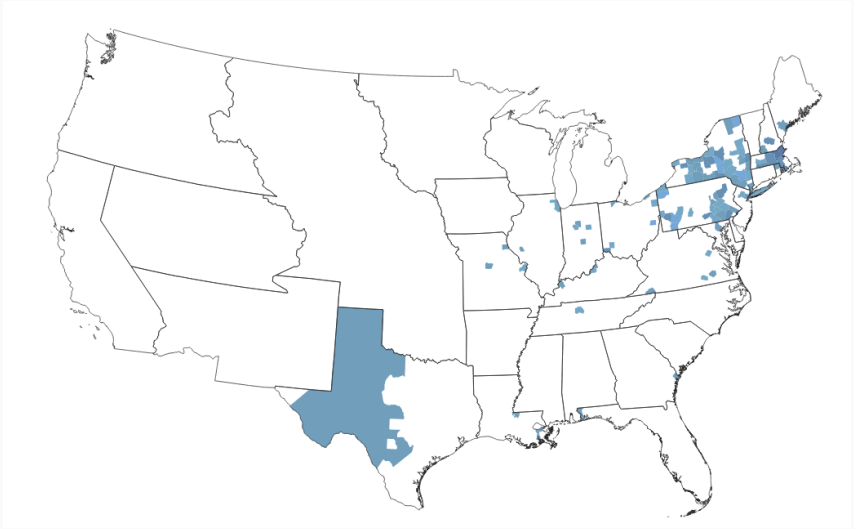
Back

Labor Share Manufacturing 1820-2020



Back

Textbooks Map



Back