How the Dawn of Public Higher Education (1900-1940) Shaped Access and Work

Collin Wardius

Department of Economics, UC San Diego
Approved by

October 30, 2025

Higher education in the US experienced its first major transformation in the early 1900s

- Many more students enrolled
- Public universities began to dominate in terms of enrollment

Questions

- **How did the founding of public colleges change access to college?**
- How did the founding of public colleges change the labor force of local economies?

Preview of identification approach

- **Identifying variation**: quasi-random founding date of a university
- Some people are lucky as they are born just late enough to access a new university
- Some people are unlucky as they are born too early to access a new university

Literature

- **History of US higher education (1900-1940)**: Goldin (1998), Goldin and Katz (1998), Goldin (2001)
  - → *My contribution:* Quantify the causal effect of university expansion on education access
- **Effects of university building in non-US countries**: Duflo (2001), Nimier-David (2023)
  - → *My contribution:* US university foundings and variation in public vs private control
- **How proximity to college affects attainment and earnings**: Card (1993), Acton et al. (2025)
  - → *My contribution:* Examine extensive margin of college access via new university foundings

BA Completion: 1900 vs 1936 Birth Cohorts

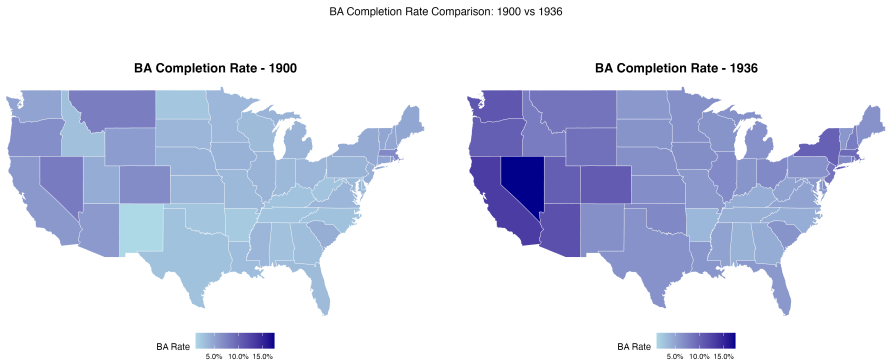

Figure: BA Completion: 1900 vs 1936 Birth Cohorts

College Founding Years by Region

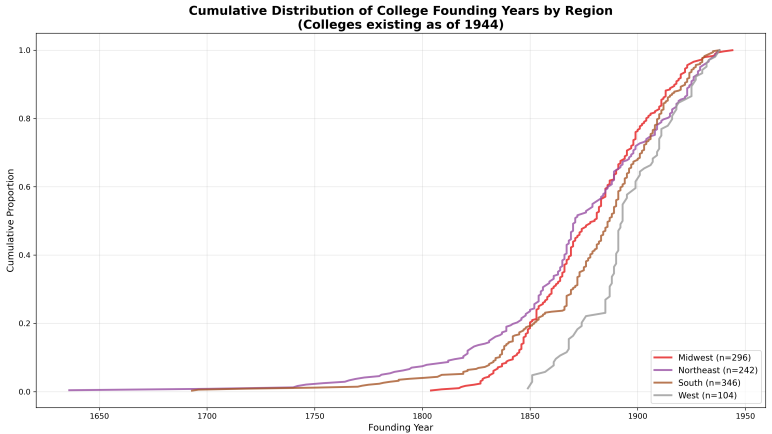

Figure: Regional Distribution of College Founding Years

College Founding Years by Control

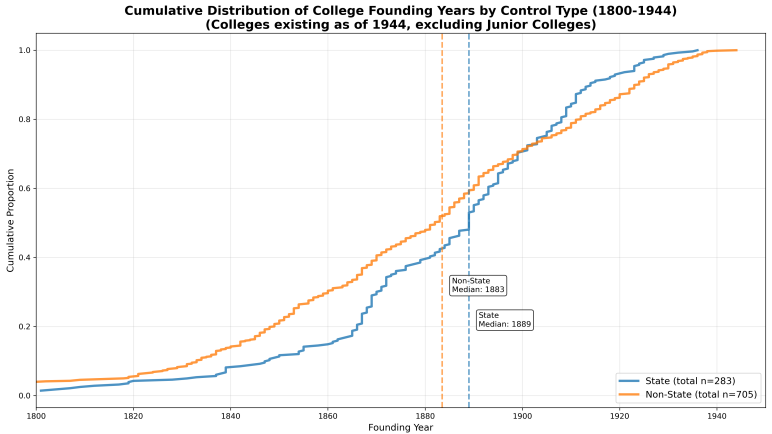

Figure: Regional Distribution of College Founding Years (1800+)

Estimating the effect of a university founding on college attainment

Cross sectional regression, identifying variation is at the cohort-by-county level.

$$y_{ick} = \alpha_c + \lambda_k + \beta \text{New college}_{ck} \times \lambda_k + \xi \boldsymbol{X}_{ick} + \epsilon_{ick} \tag{1}$$

- c: county, k: age cohort, i: person
- $\text{New college}_{ck} = \mathbb{1}\{\text{There is a college founded in } c \text{ that is available to } k\}$

The identification assumption

- Compare the gap in attainment between older cohorts and younger cohorts in counties that have a new college versus those that do not

- **Identifying assumption**: Conditional on controls, counties that gained a college and those that didn't would have experienced parallel trends in attainment across cohorts, absent the new college.

Isolating treated and control counties

- Restrict attention to “conventional” colleges: exclude junior colleges, normal schools, teachers colleges, and colleges with capacity ≤ 100.
- Treated counties gain exactly one college over this period
- Three natural control groups:
  - Counties that never get a college (never-treated)
  - Counties that get a college later in the period (not yet treated)
  - Counties that received a college before 1900 and do not receive a college during this period (already treated)

County spatial stability over 1900-1940 | County crosswalk construction

Quantifying the treated and control counties

Table: County Classification for College Analysis (1900-1940)

County Group	Count	Role in Analysis
Had college before 1900	320	—
Did not gain college 1900-1940	239	Potential Control
Gained college(s) 1900-1940	81	—
No college before 1900	2788	—
Gained exactly 1 college 1900-1940	72	**Treated**
Gained 2+ colleges 1900-1940	4	—
Never gained college by 1940	2712	Potential Control

Notes: Analysis excludes junior colleges, normal schools, teachers colleges, and colleges with capacity ≤ 100. Treated group consists of counties that had no college before 1900 and gained exactly one college 1900-1940. Potential control groups consist of (1) counties that had a college before 1900 but did not gain additional colleges 1900-1940, and (2) counties that never had a college by 1940.

Identifying treated individuals in the census

We only observe education in 1940, after individuals either received or did not receive a college education

1. Identify individuals who are at least 25 in the 1940 census
2. Link back to the latest census for which they are below the age of 18
3. Assign the individual that county of residence for the purposes of treatment assignment

Comparing linked versus unlinked individuals in the census

Table: Comparison of 1940 Characteristics: Linked vs Unlinked Individuals

	Linked Mean	Unlinked Mean	Difference
Female (%)	22.7	63.9	-41.2
Age	36.3	47.7	-11.4
College (%)	9.9	6.9	3.0
Married (%)	74.7	83.5	-8.7
White (%)	77.2	67.9	9.2
N	391,789	641,859	
% of Total	37.9%	62.1%	

Note: This table compares mean characteristics in 1940 for individuals age between 25 and 70 who were successfully linked to pre-age 18 observations versus those who were not linked.

Testing parallel trends: Event study specification

To test for pre-trends and trace out dynamic effects, estimate:

$$y_{ick} = \alpha_c + \lambda_k + \sum_{j \neq -1} \beta_j \mathbb{1}\{\text{Cohort } k \text{ born } j \text{ years relative to college founding in } c\} + \xi \boldsymbol{X}_{ick} + \epsilon_{ick} \tag{2}$$

- $j < 0$: Cohorts born *before* college founding (test for pre-trends)
- $j \geq 0$: Cohorts born *after* college founding (treatment effects)
- Omit $j = -1$ as reference category
- Null hypothesis: $\beta_j = 0$ for all $j < 0$ (no pre-trends)

County spatial stability over 1900-1940

Table: County Boundary Stability Between 1900 and 1940

	Reference Period	
Overlap Threshold	1940 Counties	1900 Counties
Total Counties	3108	2848
99% or more overlap	2852 (91.8%)	2538 (89.1%)
95% or more overlap	2941 (94.6%)	2616 (91.9%)
90% or more overlap	2976 (95.8%)	2647 (92.9%)
80% or more overlap	3005 (96.7%)	2681 (94.1%)

Notes: The 1940 Counties column shows the percentage of 1940 counties that overlap with a single 1900 county at the specified threshold. The 1900 Counties column shows the percentage of 1900 counties that overlap with a single 1940 county.

Back to Isolating treated and control counties

Creating a county crosswalk

We need consistent county boundaries to accurately assign college locations across census years.

Approach:

1. Use 1900 as the reference year (counties typically split into smaller units over time, rather than merging)
2. Spatially intersect 1910, 1920, 1930, and 1940 boundaries with 1900 boundaries
3. Match counties where the intersection exceeds 70% overlap
4. Retain only counties that appear consistently across all census years

Back to Isolating treated and control counties

My reference on how the spatial join is performed

Consider A and B from 1900 (the base year) and 1940 (the target year) respectively. Then get the area of $A \cap B$. We then compare this to the area of the target to calculate

$$\frac{A \cap B}{B} \tag{3}$$

and we map B to A if this is above some threshold. I am using 70% as of right now but this could be modified. In practice, almost 90% of the counties have close to 100% overlap.