# Innovation Nation: Evidence from Broadening Access to Ph.D. Training in the U.S.

Francisca Antman[1] Kirk Doran[2]
Xuechao (Jane) Qian[3] Bruce Weinberg[4]

[1]University of Colorado, Boulder

[2]University of Notre Dame

[3]Stanford University

[4]The Ohio State University

April 21, 2023

# US: An Innovation Nation

- The U.S. rose to become a global leader in scientific research and ideas in the early 20th century (Urquiola, 2020; MacLeod and Urquiola, 2021 ▸ Appendix )
  - This paper studies one aspect of how that happened: **The expansion of research doctoral education**
    - matching ProQuest and EPO's PATSTAT patent to full count censuses (1850-1940)
  - The question of how expanding access affected science and innovation participation is particularly timely:
    - for the U.S., as it seeks to remain competitive
    - for other countries, looking to expand their innovative workforces

# This paper

- Documents the expansion of research doctoral education in the early 20th century
- Estimates how it expanded access to and completion of research doctorates
  - pays particular attention to how expansions affected who obtains a doctorate and show that new populations were drawn into innovation
  - suggests the importance of tapping all sources of innovative potential
- Estimates how it further promoted patenting activities
- Contributes to the body of research on US innovation and higher education institution ▸ Literature

# Main Findings

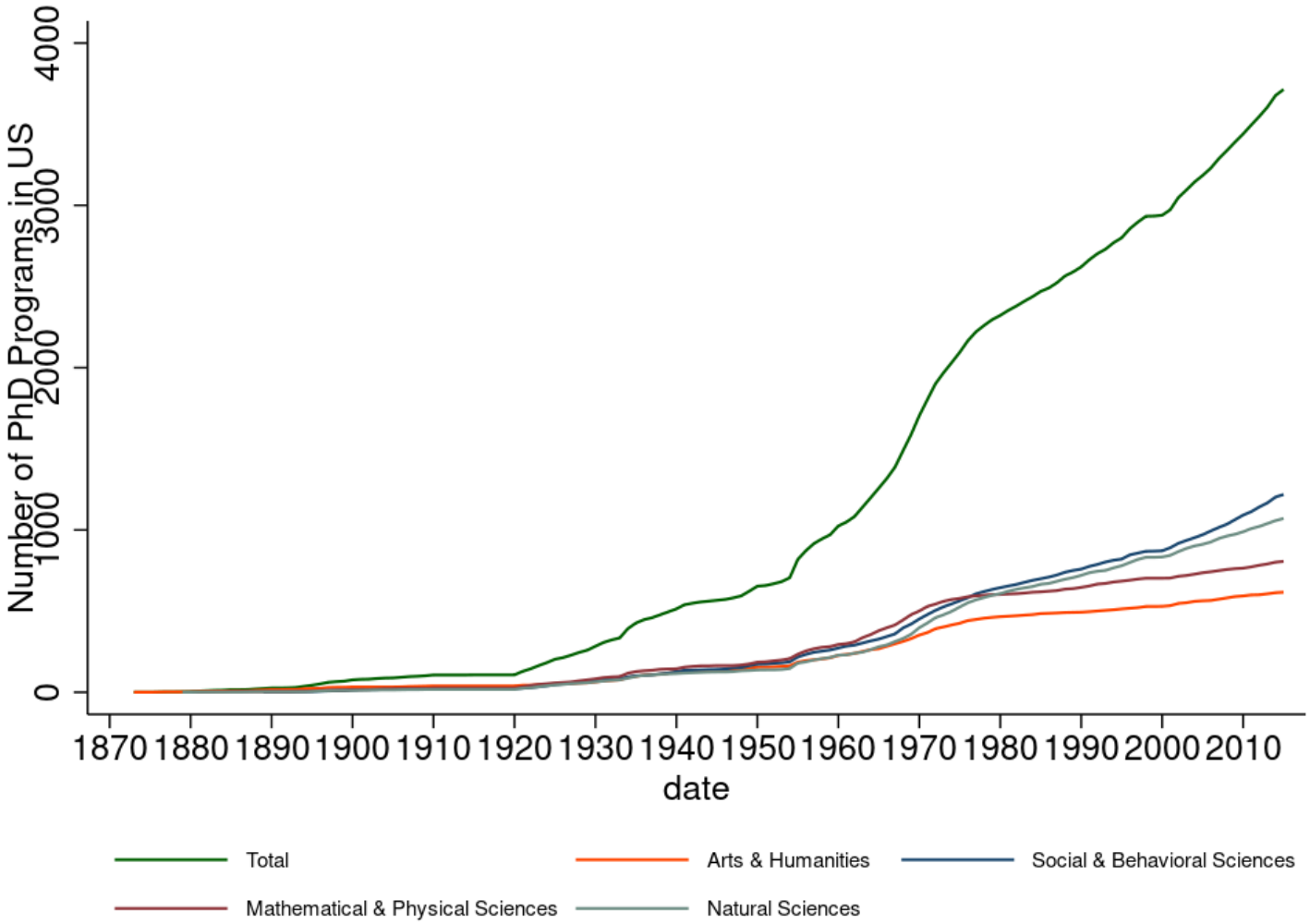
- The large expansion of PhD programs across the US since early 20th century:
  - supplied 36 more PhD recipients per 1 million people from birth states per year: 4.2% of 1940 base (855 PhDs per 1 million)
    - 12 new programs per year from 1861 to 1960
    - one new PhD program during the peak ages of graduate study generates nearly 3 more PhD recipients per 1 million people **born in that state** Extension
  - generated positive spillover effects on innovation during innovators' early life
  - improved access to doctoral training for underrepresented population groups

# ProQuest

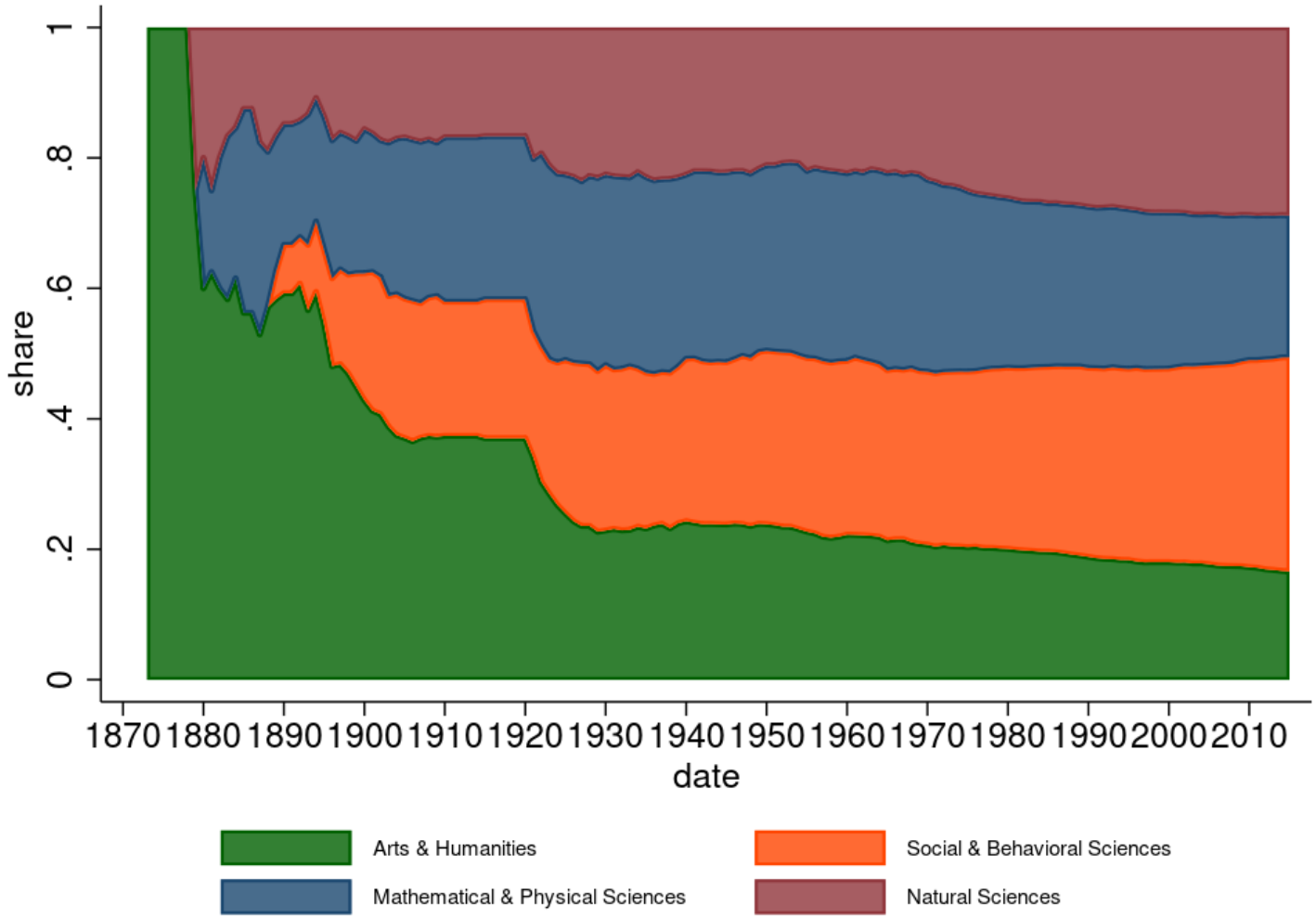
- The official offsite dissertation repository for the Library of Congress:
  - World's largest curated collection of dissertations and theses since 1861
  - We only focus on PhD recipients awarded in US institutions
- Main variables provided:
  - unique dissertation identifier
  - full author name
  - research subject
  - institution information (a school identifier and name)
  - degree information of the recipient (the degree and date).
- Identifying the opening of one PhD program:
  - using the year when first cohort of PhD recipients graduated (submitted the dissertations)
  - research field and institution level

# Number of PhD programs opened in US

▸ PhD V.S. College

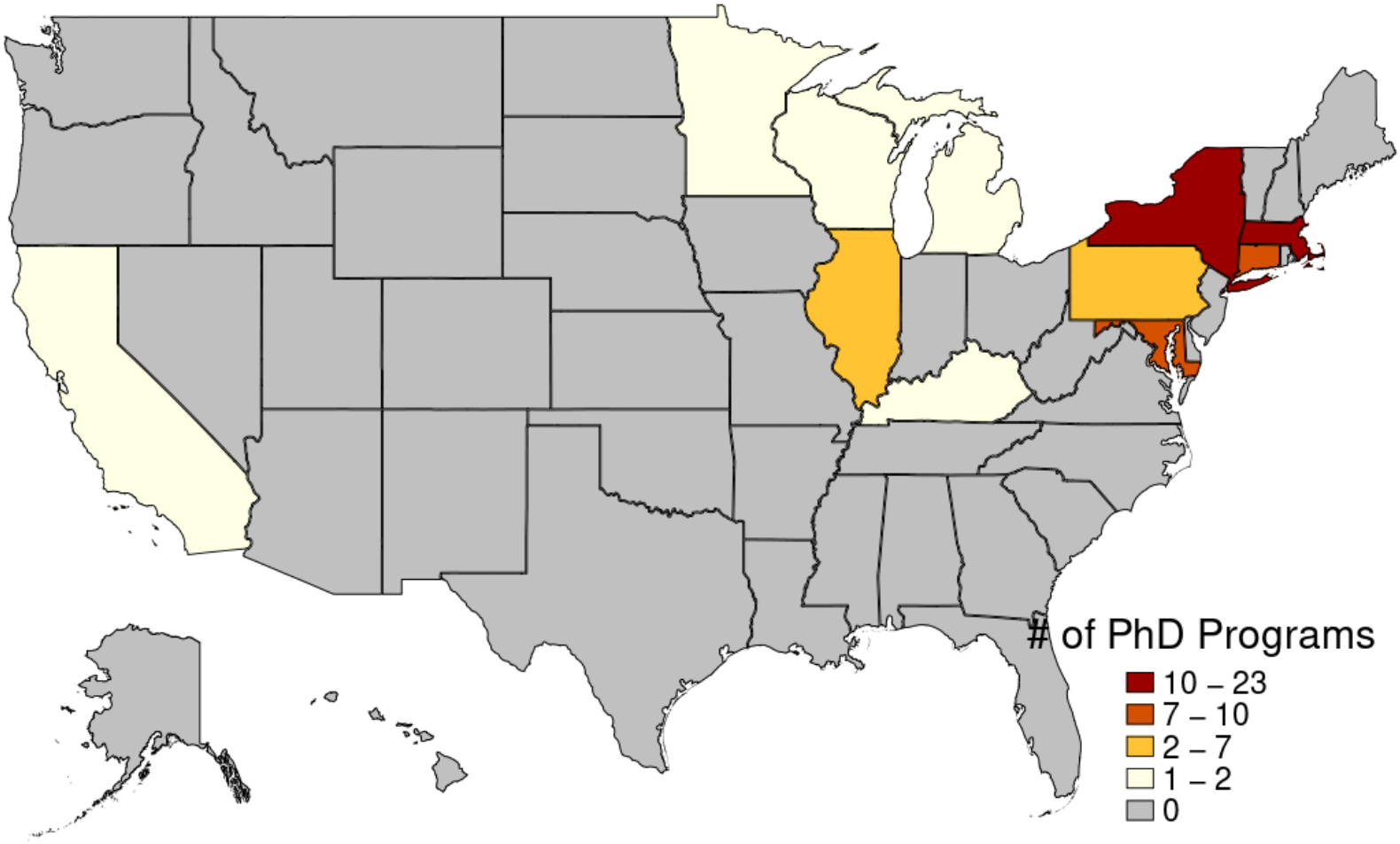


# Share of PhD programs opened in US by major field



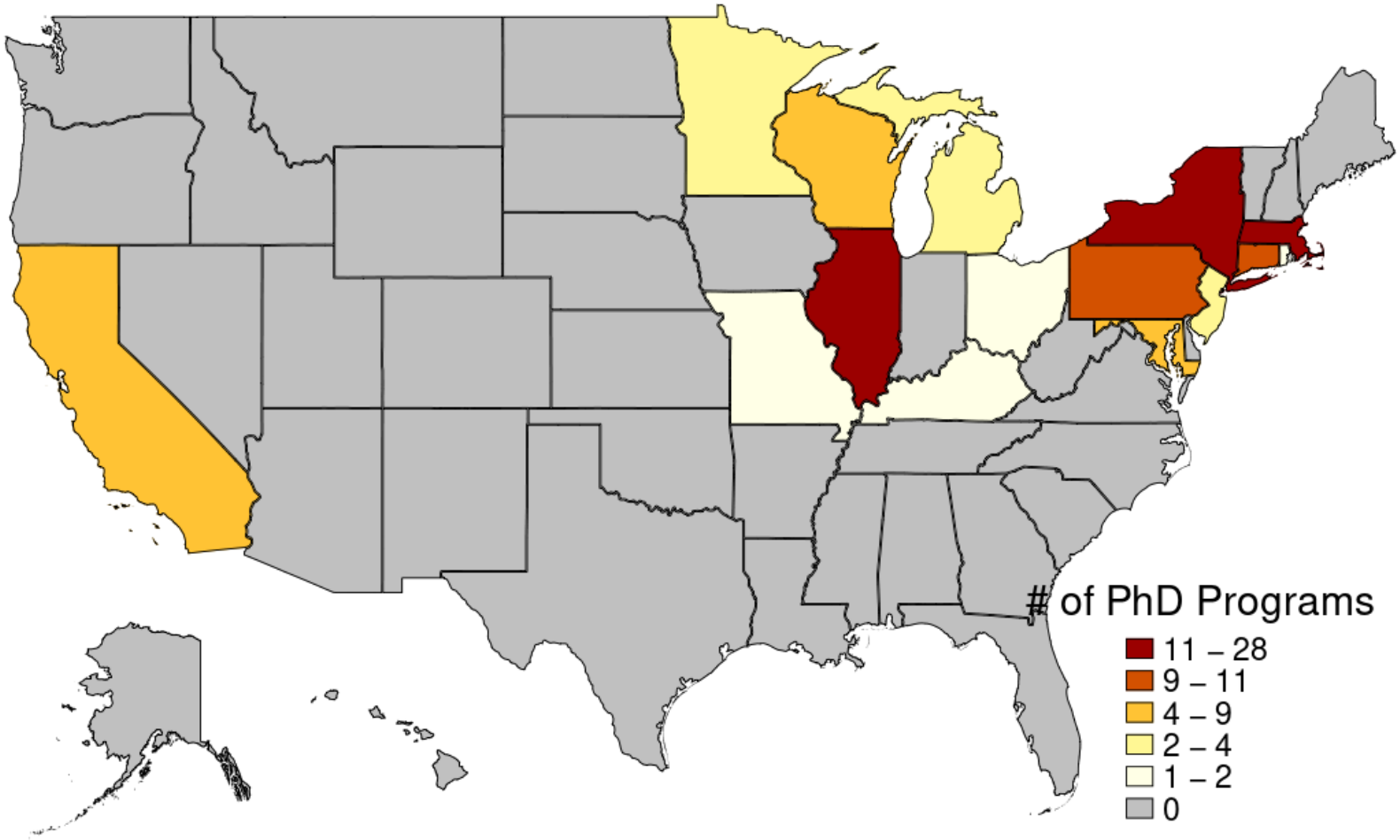
# Geographical distribution of PhD programs in US over time

By 1900



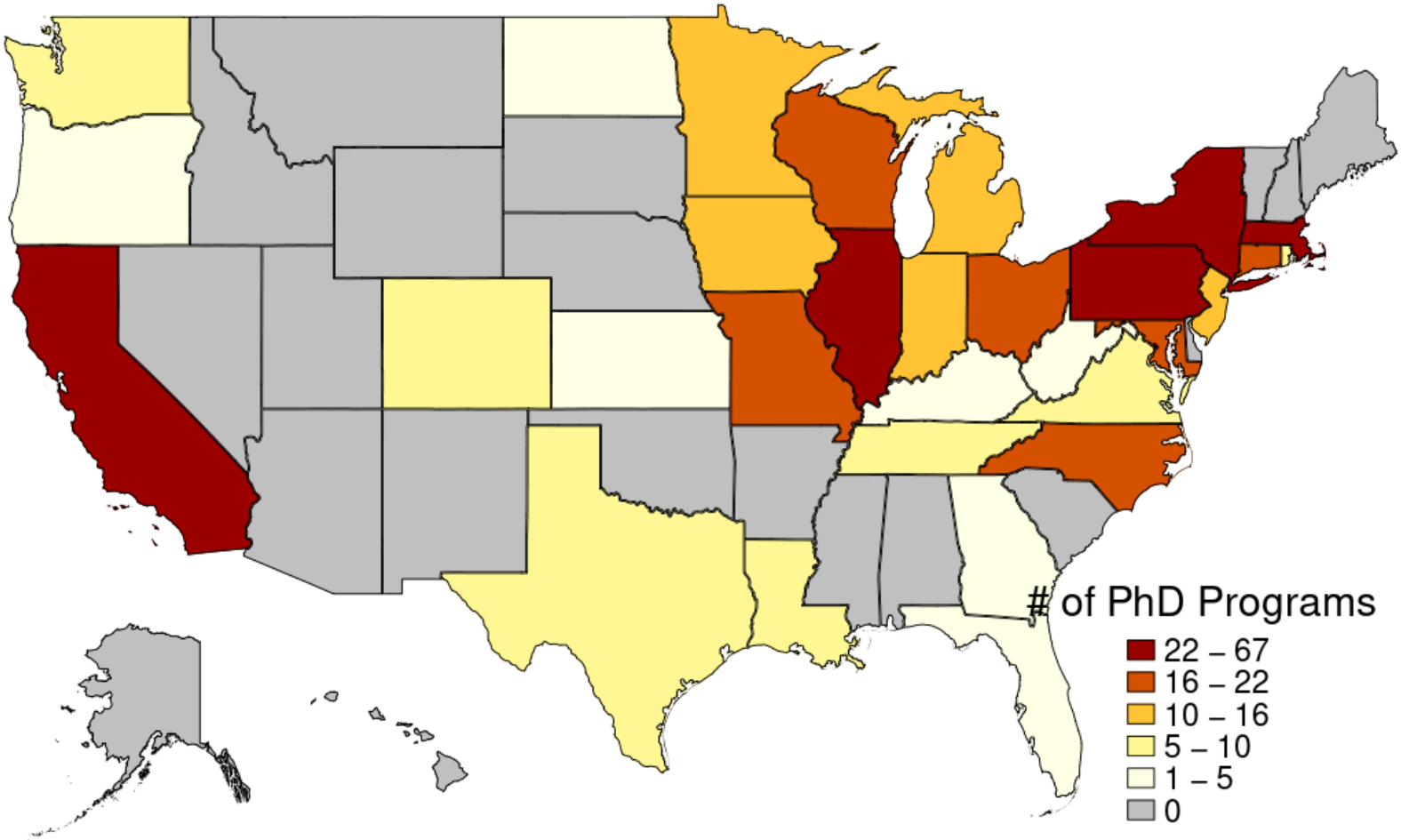
# Geographical distribution of PhD programs in US over time

By 1921



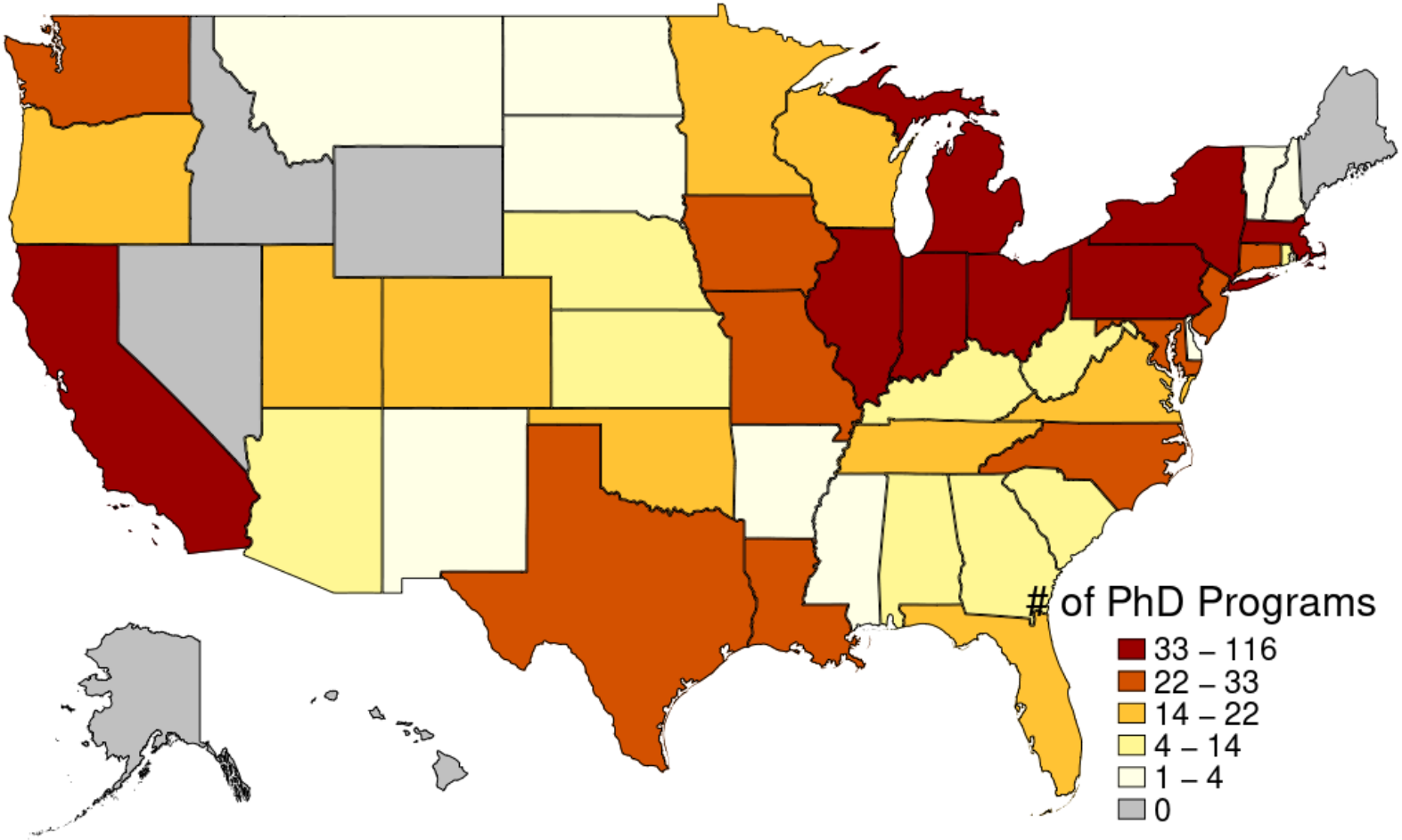
# Geographical distribution of PhD programs in US over time

By 1940



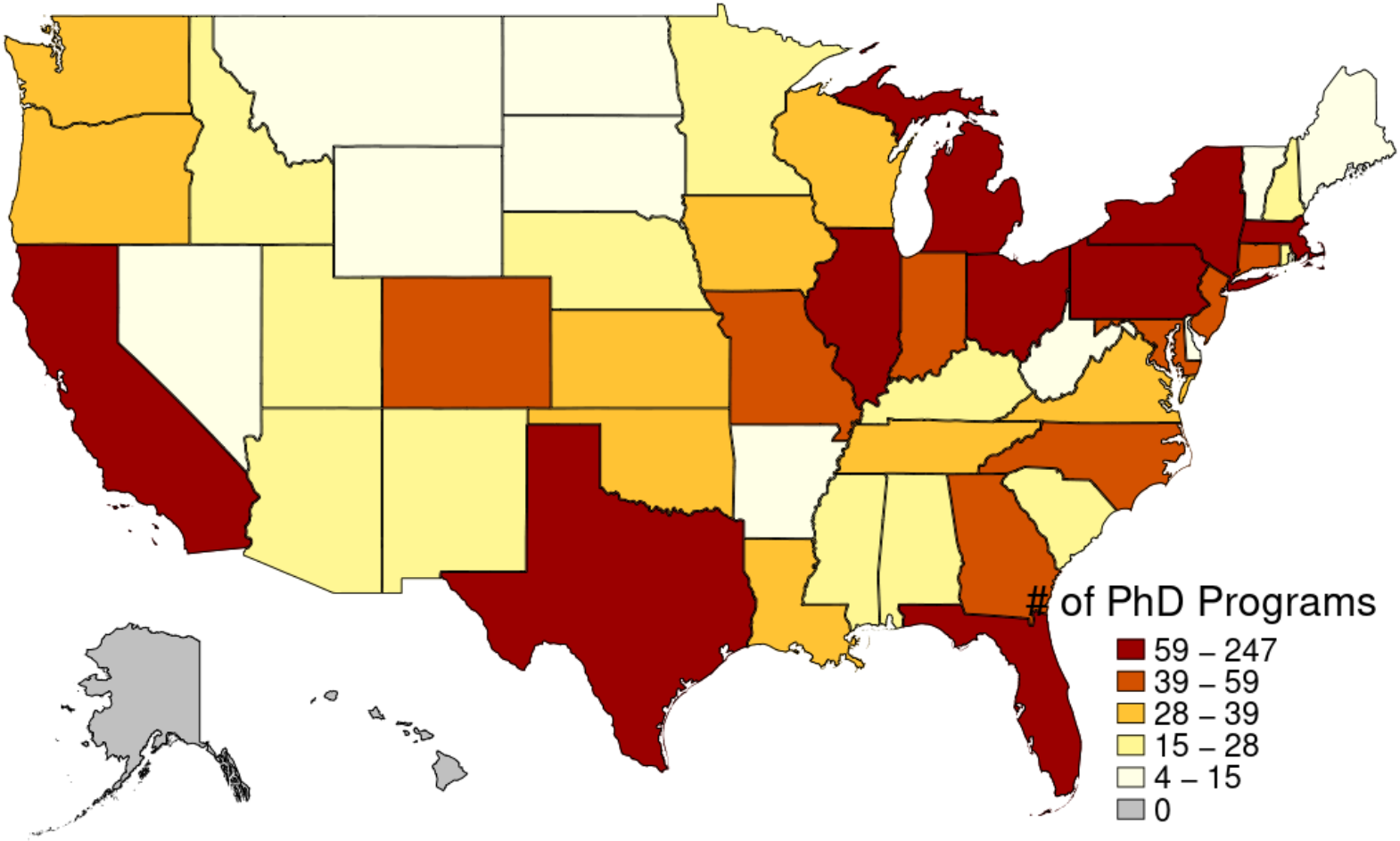
# Geographical distribution of PhD programs in US over time

By 1960



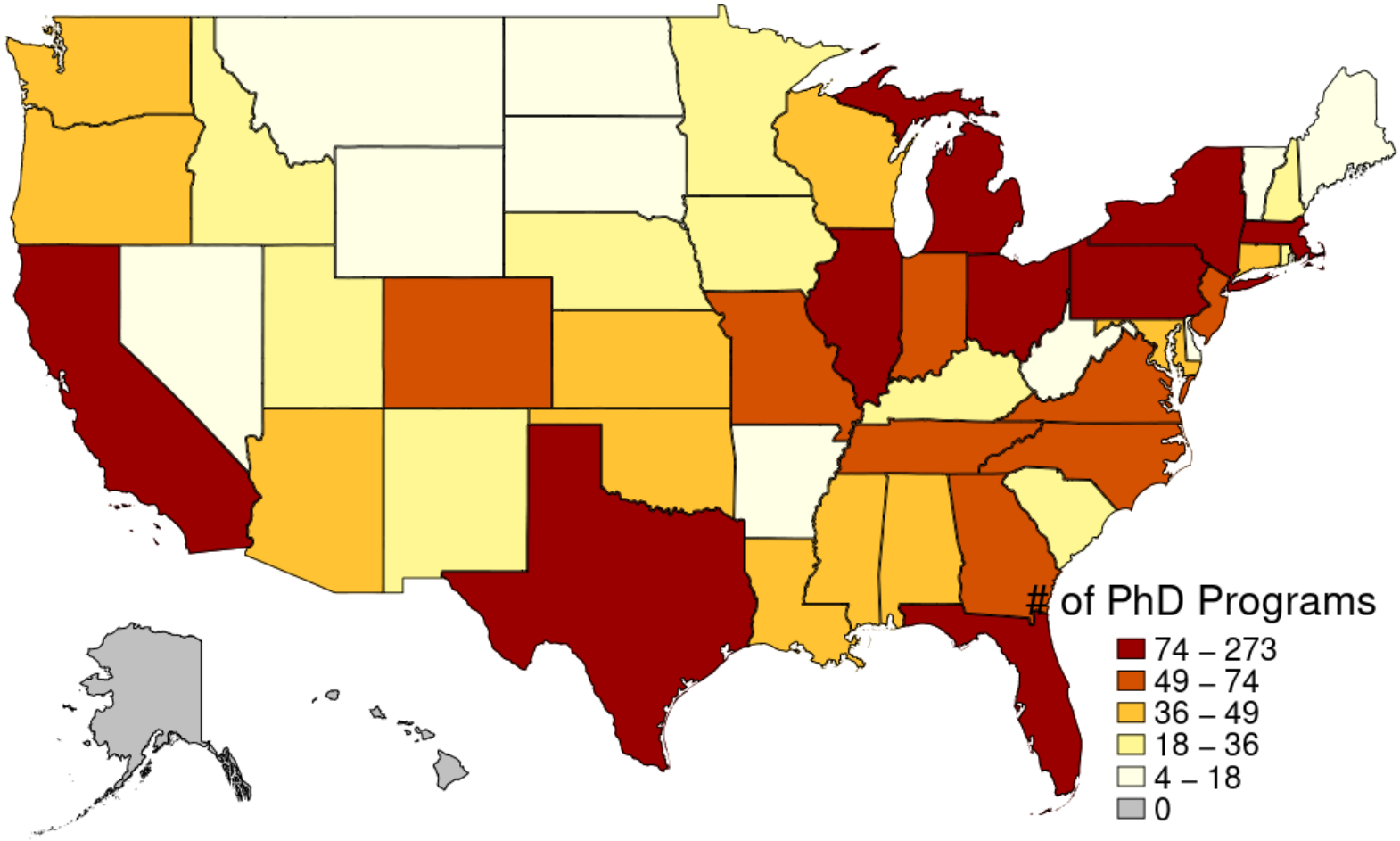
# Geographical distribution of PhD programs in US over time

By 1980



# Geographical distribution of PhD programs in US over time

By 2000



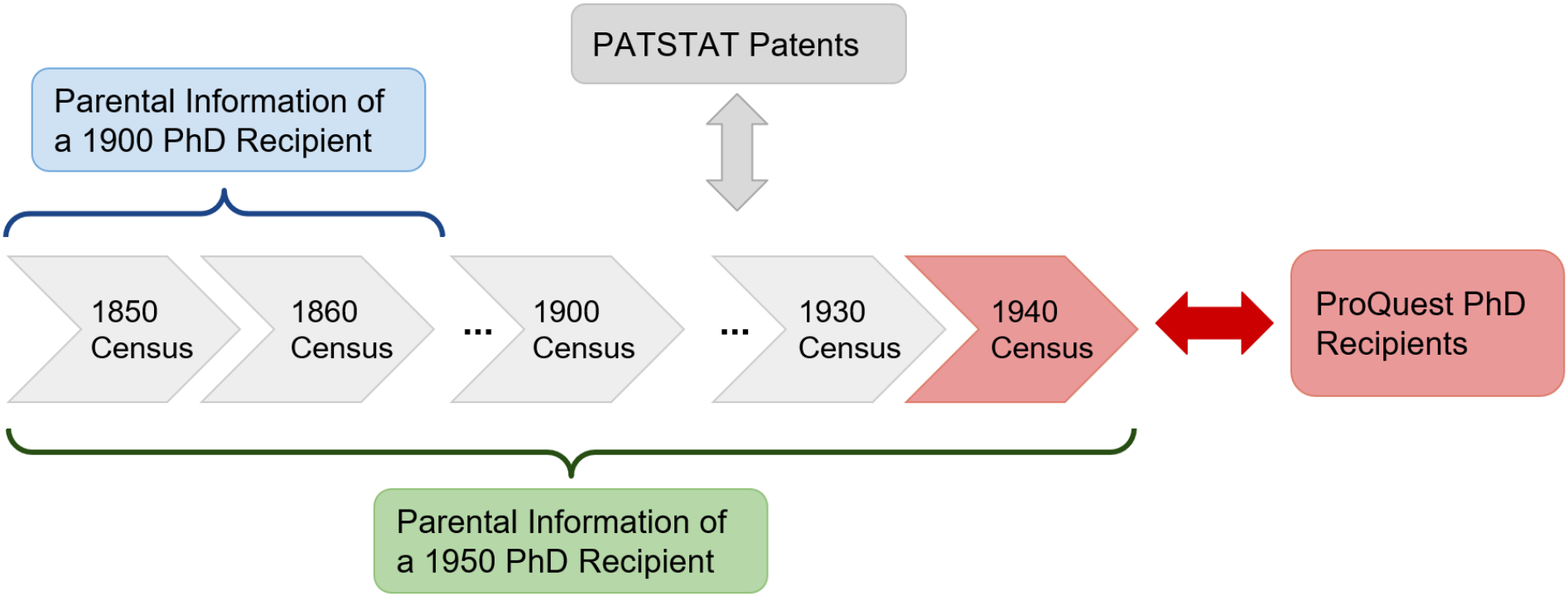
# PATSTAT

- The European Patent Office's PATSTAT database
  - patent application to the United States Patent Office from 1899 to the present
- Main variables provided:
  - inventor's full name
  - year of application
  - International Patent Classification (IPC)
  - number of citations

# Census (1850-1940) and Linked Outcomes

- Full count data of the entire U.S. population (citizens, non citizens, etc) from 1850 to 1940 (except 1890 destroyed in a fire).
- Linked across the whole period from 1850 to 1940 by the Census Linking Project (Abramitzky et al., 2022)
- Each census:
  - Full name, birth year, birth state, current location, demographics, occupation, co-residential family members
- 1940 census:
  - Highest education grade
  - Baseline to link ProQuest data
- Linked outcomes (▸ Algorithm):
  - ProQuest: ever completed a PhD degree
  - PATSTAT: ever invented, # of patents, # of citations

# Data Linkage Roadmap



# Summary Statistics of Main Demographics

| | Mean | S.D. | Mean | S.D. | Mean | S.D. |
|---|---|---|---|---|---|---|
| | 1940 population born in US (N=120,033,290) | | Matched PhD recipients (N=102,642) | | Matched inventors (N=446,043) | |
| PhD recipient | 0.00086 | 0.02923 | | | 0.00683 | 0.08237 |
| Male | 0.50 | 0.50 | 0.88 | 0.32 | 0.87 | 0.33 |
| White | 0.89 | 0.31 | 0.95 | 0.21 | 0.92 | 0.26 |
| Urban residence | 0.53 | 0.50 | 0.70 | 0.46 | 0.63 | 0.48 |
| Foreign born parent(s) | 0.09 | 0.29 | 0.14 | 0.35 | 0.18 | 0.39 |

- PhD education is extremely rare in early 20th Century US (855 per 1m)
- Overwhelmingly male and white.

# Identification Approach

- Time and geographic variations of PhD program opening
  - Time: the number of PhD programs over the course of life for each cohort (based on graduation age)
  - Geo: the number of PhD programs within distance rages from birth state
- Event study type of design
  - Main effect: the number of PhD programs in a person's **birth state** during the **peak ages** of graduate education (around 30 in deterministic and unique matches)
  - Check "parallel trends": whether the number of PhD programs and population rate of obtaining a PhD in a state jointly decided by other socioeconomic factors
  - Estimate effects of in and out state programs simultaneously

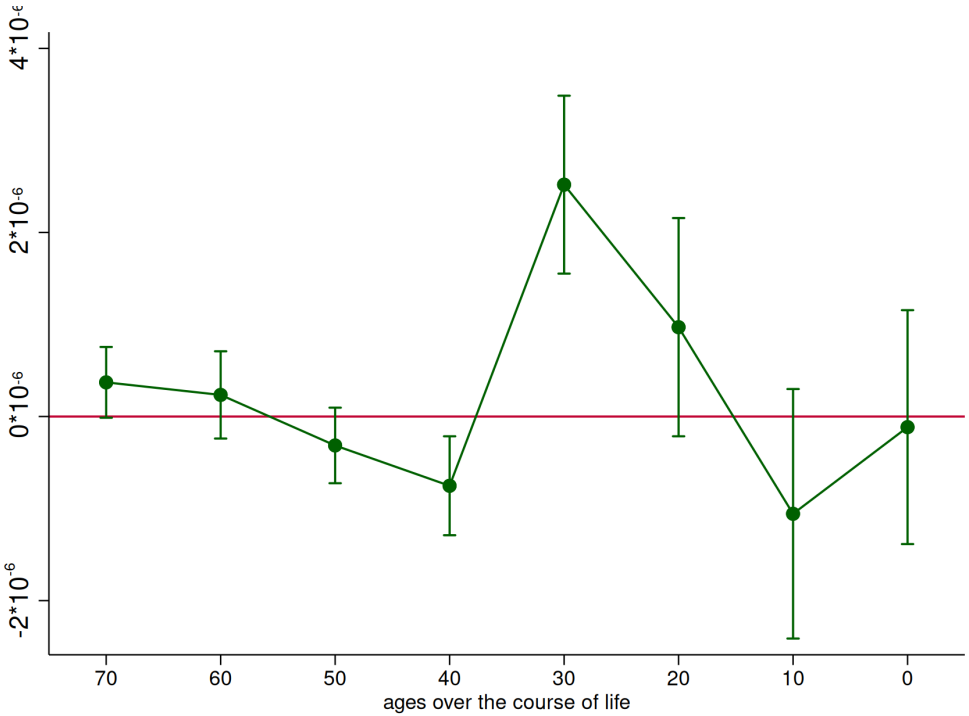
# Empirical Strategy

$$Outcome_i = \beta_0 + \sum_{k=0}^{7} \beta_{k,InState} N_{i,k,InState} + \sum_{k=0}^{7} \beta_{k,Nearby} N_{i,k,Nearby} + X_i \Upsilon + \mu_s + \lambda_c + \epsilon_i \quad (1)$$

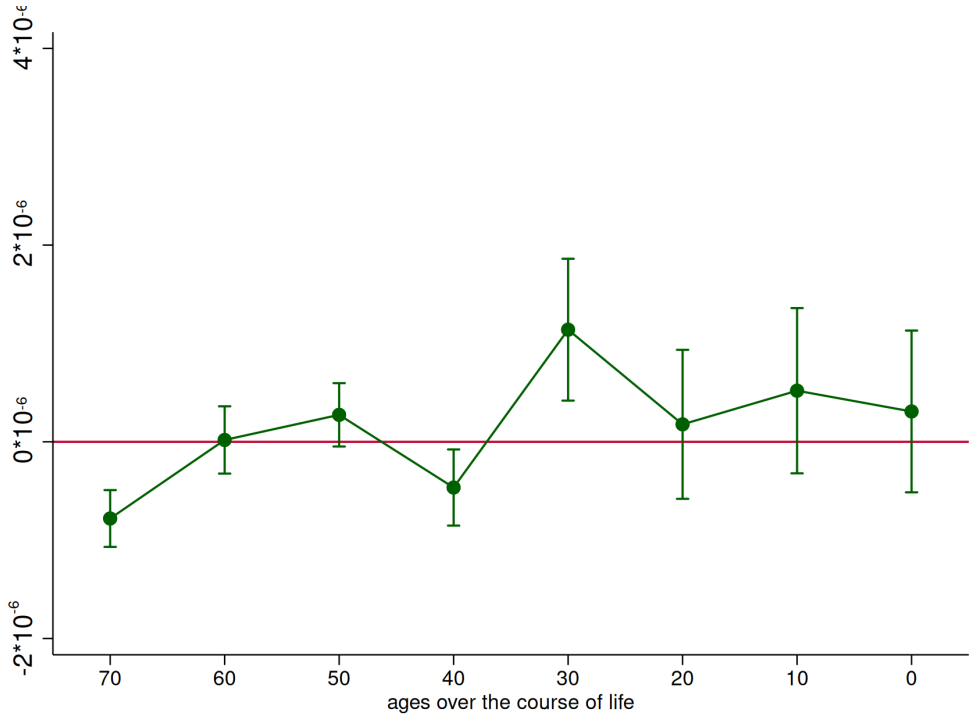
- $Outcome_i$ is the outcome of individual $i$, such as an indicator of **ever** receiving PhD degree in life $PhD_i$, an indicator of being an innovator $Inno_i$
- $N_{i,k,InState}, k = 0, 1, 2, ..., 7$ are the number of PhD programs in individual $i$'s birth state, from her birth to age 70 (at 10 year intervals).
  - $N_{i,k,Nearby}, k = 0, 1, 2, ..., 7$ are PhD program numbers in nearby-states within 200 miles
  - Number of faraway PhD programs outside the 200 mile distance is taken care by birth cohort fixed effect $\lambda_c$
  - We report the marginal effects
- $X_{it}$ is a set of individual characteristics: gender, race, urban residence, nativity indicators
- $\mu_s$ is the birth state fixed effect
- Standard errors clustered at birth state cohort level

# Probability of Earning a PhD degree (One regression)

$X$: # of PhD programs in birth state




$X$: # of PhD programs in nearby-states




- 1 more PhD program during the peak ages of graduate study induces nearly 3 more PhD recipients per 1 million people born in that state
- national level total: 4.2% of 1940 base (855 PhDs per 1 million)

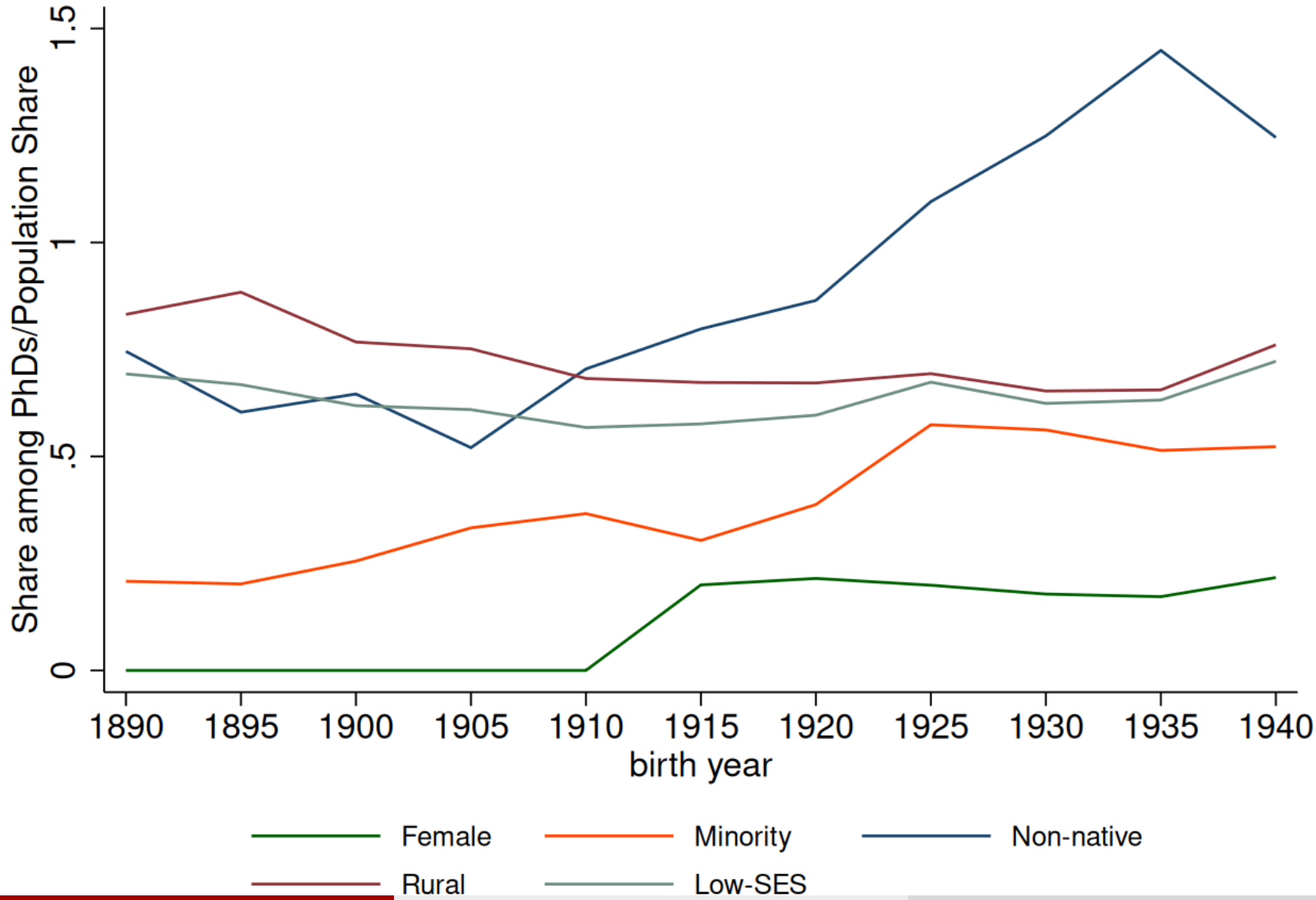
# Additional Estimates

- similar effects on outcomes measured in 1940 census ▸ Estimates
  - graduate degree
  - professional occupations
- positive spillover effects on patenting ▸ Estimates
- effects of PhD program expansion mainly comes from
  - private intuitions ▸ Estimates
  - Land Grant universities ▸ Estimates
- robust with additional controls, e.g. number of undergraduate colleges, birth region trends, birth division trends ▸ Estimates

# Empirical Strategy

- Heterogeneity effect: include interactions between the number of Ph.D. programs $N_{i,k,InState}, k = 0, 1, 2, ..., 7 / N_{i,k,Nearby}, k = 0, 1, 2, ..., 7$ and individual demographics $X_{it}$
  - Female
  - Non-white
  - Immigrant family
  - Rural residence in childhood: Census cross linking
  - Parental socioeconomic status
    - No income data before 1940
    - Census imputed socioeconomic measures, like:Duncan Socioeconomic Index, Occupational income score, Occupational prestige score
    - Find parents closest to child's age 10 using cross linked Censuses

# Share of underrepresented groups among PhDs



# Effects by demographic characteristics

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Prob(earning a PhD in life) | | | | | |
| # in-state programs around 30*Female | -0.0000078*** | | | | | -0.0000080*** |
| | (0.0000024) | | | | | (0.0000025) |
| # in-state programs around 30*Minority | | 0.0000071** | | | | 0.0000075** |
| | | (0.0000031) | | | | (0.0000033) |
| # in-state programs around 30*Non-native | | | 0.0000034 | | | 0.0000041* |
| | | | (0.0000022) | | | (0.0000023) |
| # in-state programs around 30*Rural | | | | 0.0000012 | | 0.0000013 |
| | | | | (0.0000012) | | (0.0000013) |
| # in-state programs around 30*Occ-income-score | | | | | -0.0000001* | -0.0000001 |
| | | | | | (0.0000001) | (0.0000001) |
| Observations | 56,155,926 | 56,155,926 | 56,155,926 | 56,155,926 | 56,155,926 | 56,155,926 |

- The effect of PhD training expansion is larger among minority, non-native, and low-SES population.

# Estimating PhD recipients' choices

- Choices made by PhD recipients (conditional on having obtained a PhD degree)
  - Where to pursue the doctoral training: in state, nearby, or faraway? Estimates
  - What subject to do research in? Estimates
  - What type of institution to attend: private, or public? Estimates
- Multinomial Logit Model (e.g.: location choice)
  - We report marginal effects

$$ln(\frac{Pr(Y_i = InState)}{Pr(Y_i = Faraway)}) = \sum_{k=0}^{7} \alpha_{k,InState}^{InState} N_{i,k,InState} + \sum_{k=0}^{7} \alpha_{k,Nearby}^{InState} N_{i,k,Nearby} + X_i \Upsilon^{InState} + \mu_s^{InState} + \lambda_c^{InState} + \epsilon_i^{InState} \quad (2)$$

$$ln(\frac{Pr(Y_i = Nearby)}{Pr(Y_i = Faraway)}) = \sum_{k=0}^{7} \alpha_{k,InState}^{Nearby} N_{i,k,InState} + \sum_{k=0}^{7} \alpha_{k,Nearby}^{Nearby} N_{i,k,Nearby} + X_i \Upsilon^{Nearby} + \mu_s^{Nearby} + \lambda_c^{Nearby} + \epsilon_i^{Nearby} \quad (3)$$

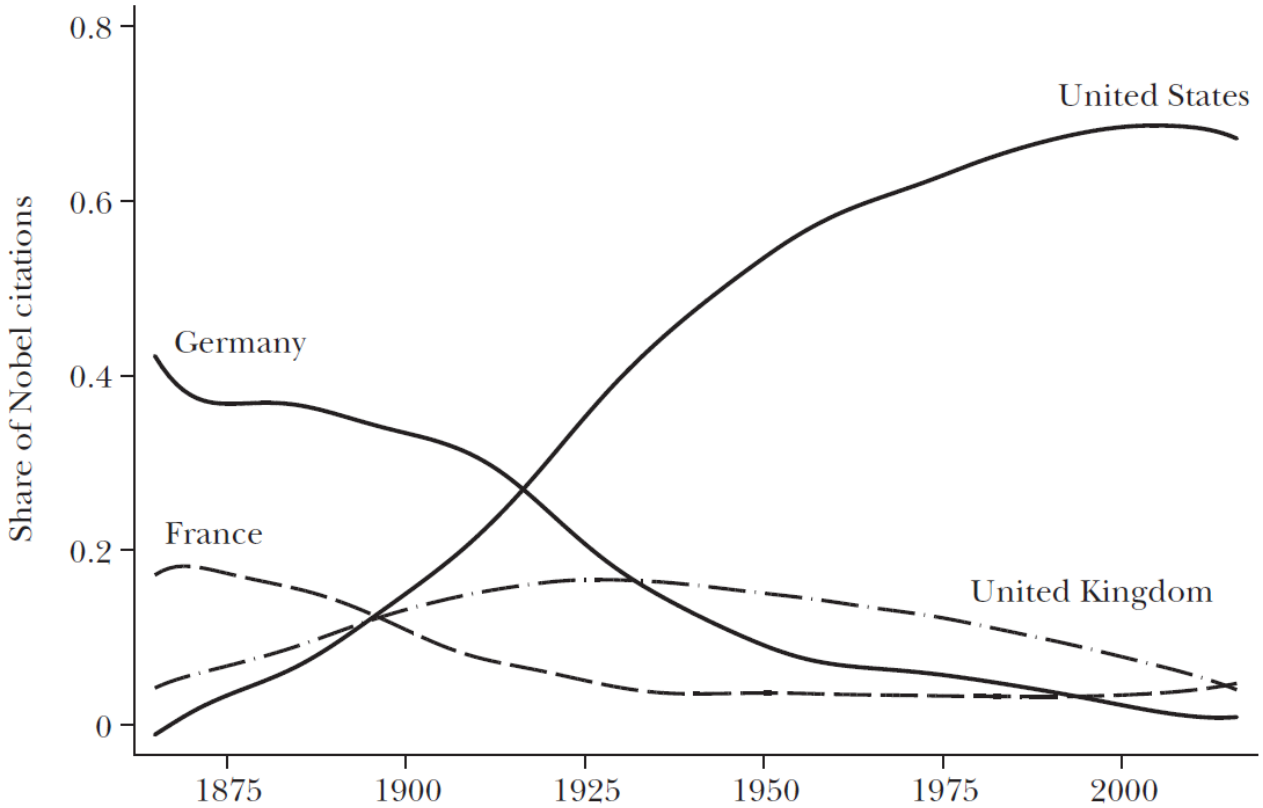
# Discussions

- The access to PhD training was broadened substantially around early 20th century
- The US population rate of pursuing a PhD degree increases after new doctoral granting programs opened
  - The location, research focus, and type of institution matter
- Early life exposure to PhD training intuitions increased patenting
- Democratization of opportunities for some underrepresented population groups
- Explore the impacts of other socioeconomic shocks on research training using the data in the future

# Share of universities mentioned in Nobel winners' biographies

Motivation

**University Nobel Prize Mentions, by Country**




*Source:* Urquiola (2020).

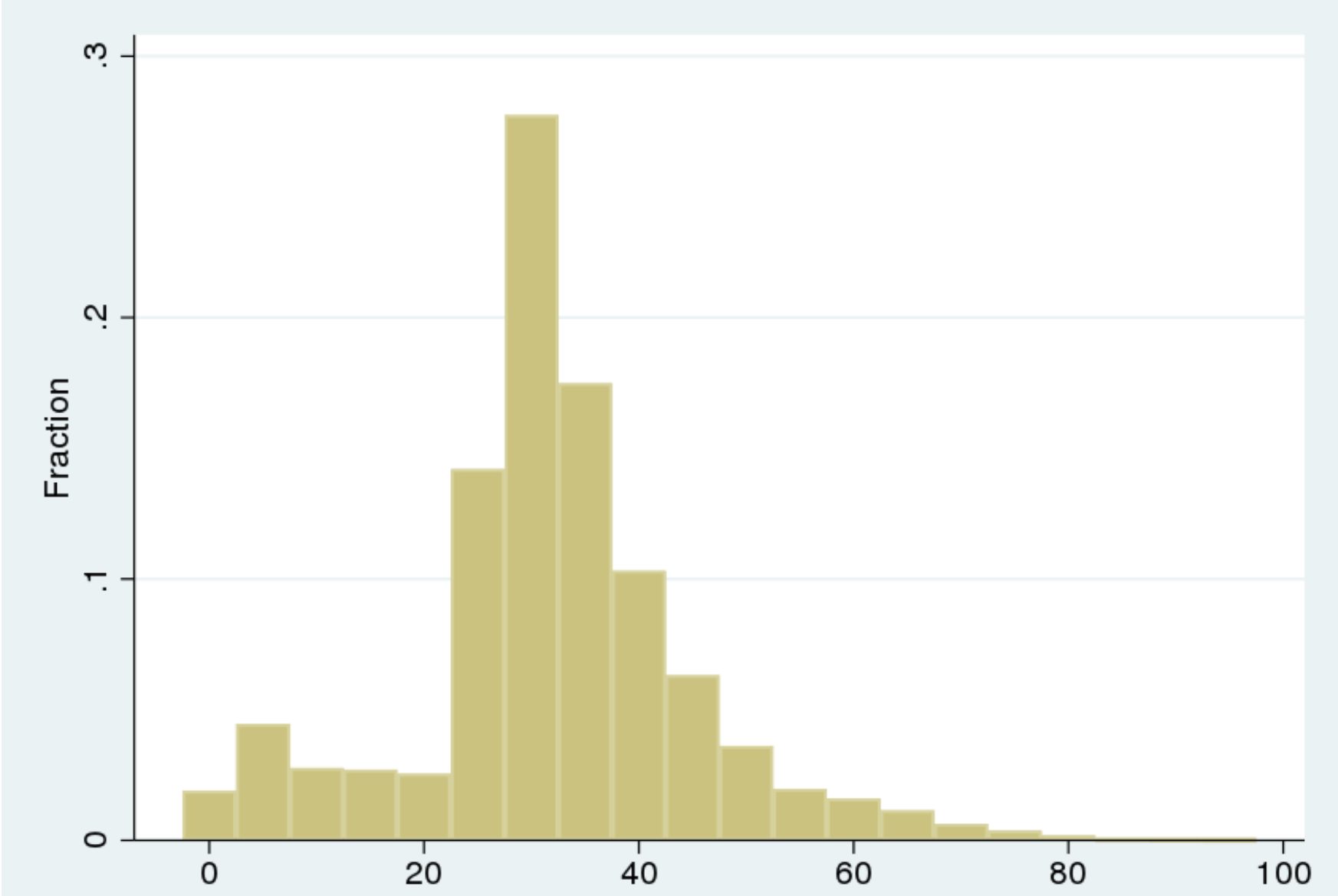
# Contributions

◂ Our paper

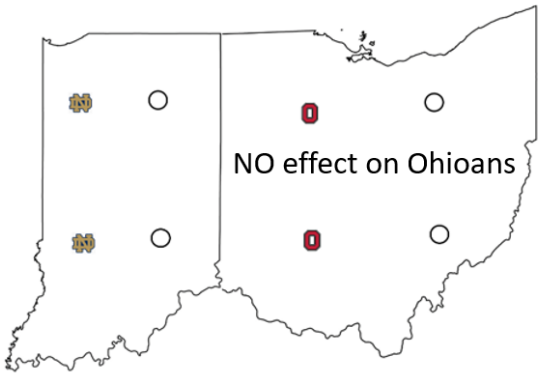
- **PhD training expands innovative workforce**:
  - global migration and U.S. institutional changes around World War II (Graham and Diamond 1997; Cole 2009; Gruber and Johnson 2019)
- **Among the first to examine the expansion of graduate (PhD) education**:
  - Before 1900, primarily educating the sons of the elite, mainly in the humanities at a relatively small number of traditional institutions (Geiger, 1986)
  - Primary education (Clay et al., 2012; Card and Giuliano, 2014); Secondary education (Goldin and Katz, 2009); College education (Goldin and Katz, 1999; Bound and Turner 2002; Chetty et al. 2017; Hoxby and Avery 2014)
- **Highlight the significance of supplying researchers through training**:
  - Other determinants: Demographic change (Borjas and Doran, 2012; Moser et al., 2014); Community (Bell et al., 2018; Berkes and Nencka, 2020)p; Family background (Aghion et al., 2017); Early career exposure (Azoulay et al., 2019); Market demand (Khan and Sokoloff, 1993)
- **Expansion of PhD training created more equality of opportunity?**:
  - Ethnic diversity and economic performance (Alesina and La Ferrara, 2005)
  - Firm diversity and innovation (Østergaard et al., 2011)
  - Underrepresented students innovate at higher rates (Hofstra et al., 2020)

# PhD age distribution of unique name matches

Preview

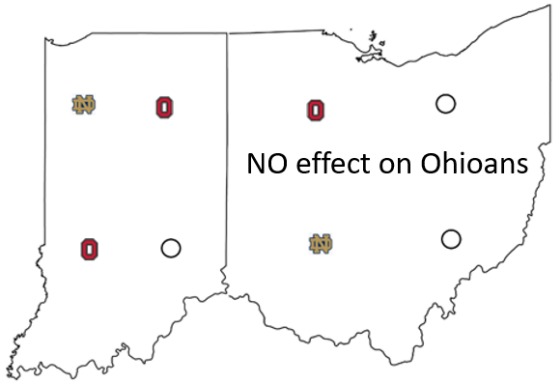


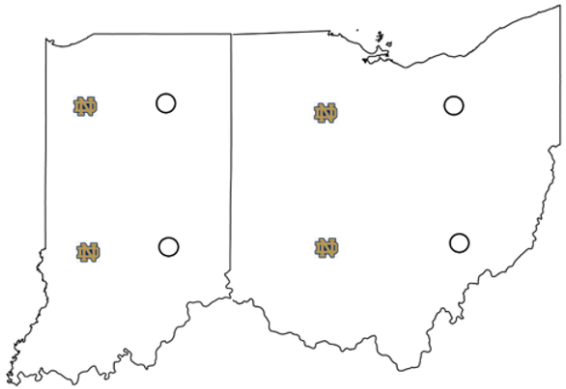
# Is it mechanical? ◂ Preview

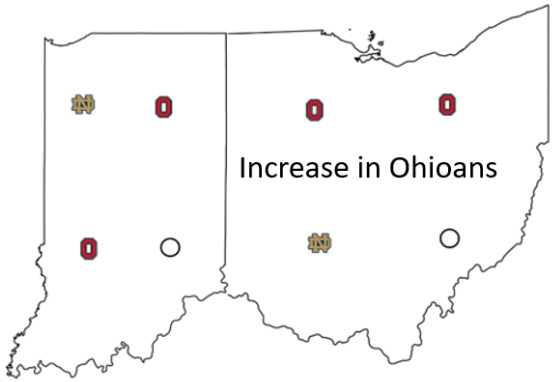


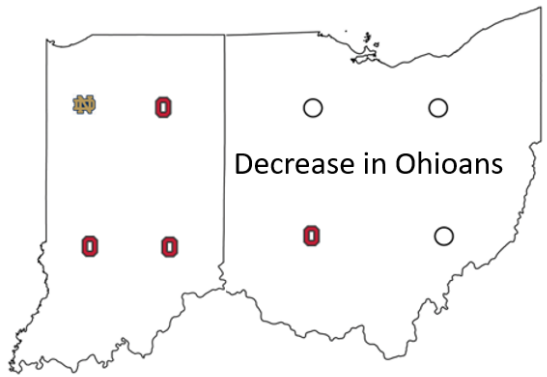

: 1842

: 1870



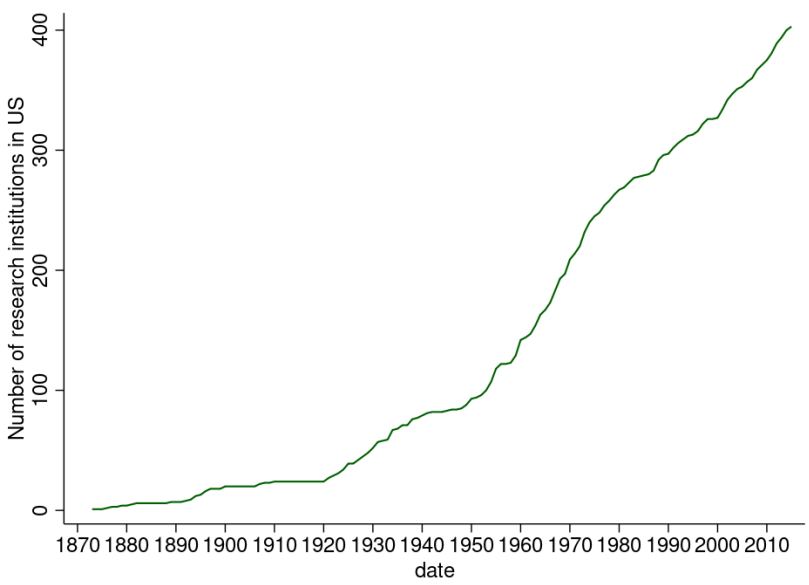


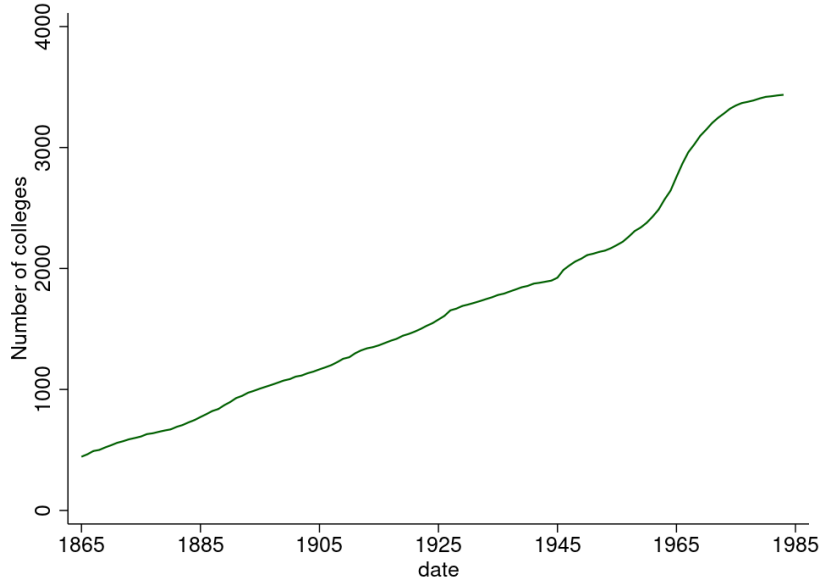
# Number of educational institutions opened in US

◂ Programs

PhD Granting Institutions (ProQuest)




All Colleges & Universities (IPEDS)



# Linking Algorithm

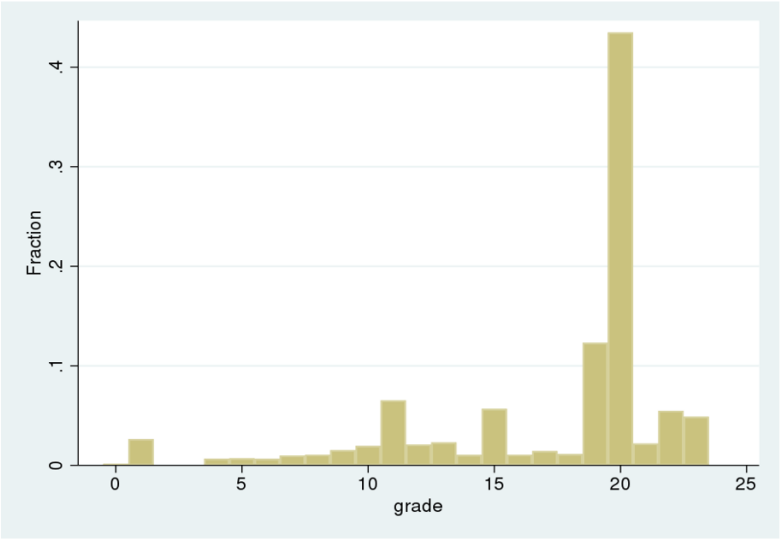
1. Deterministically match ProQuest and the 1940 Census using unique, exact full first name and last name matches
2. Generate distributions of each ▸ identifying variable from **unique** matches obtained in STEP 1 as "Gold Standards".
3. Assign weights for each identifying variable based on the distribution statistics in STEP 2.
4. Block first and last names, generate the matching score for each potential pair.
5. Choose the pair with the highest score in STEP 4 (provided that at least one pair is above the cutoff); Eliminate pairs with more than 10 close matches.
6. Link the full 1940 census with ProQuest variables to all previous full count censuses to find parental information using existing crosswalks.

# Identifying Variables

◂ Linking Algorithm

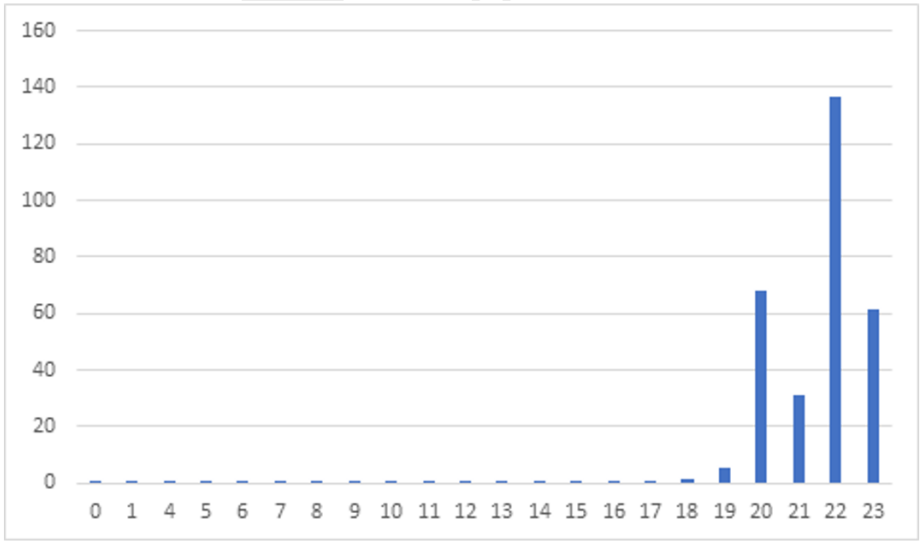
- Main identifying variables: Highest education grade; Distance between current residential county to university; Age awarded a PhD degree
- Three ProQuest cohorts: PhD awarded before 1940, 1940 to 1950, 1950 to 1960
  - Example: Highest education grade of PhDs awarded before 1940

**Share of highest education grade among unique matches**




| 1940 Census Code | |
|---|---|
| | College: |
| 16 | 1st year |
| 17 | 2nd year |
| 18 | 3rd year |
| 19 | 4th year |
| 20 | 5th year |
| 21 | 6th year |
| 22 | 7th year |
| 23 | 8th year or more |

**Ratio between the share of highest education grade among unique matches and 1940 population share**

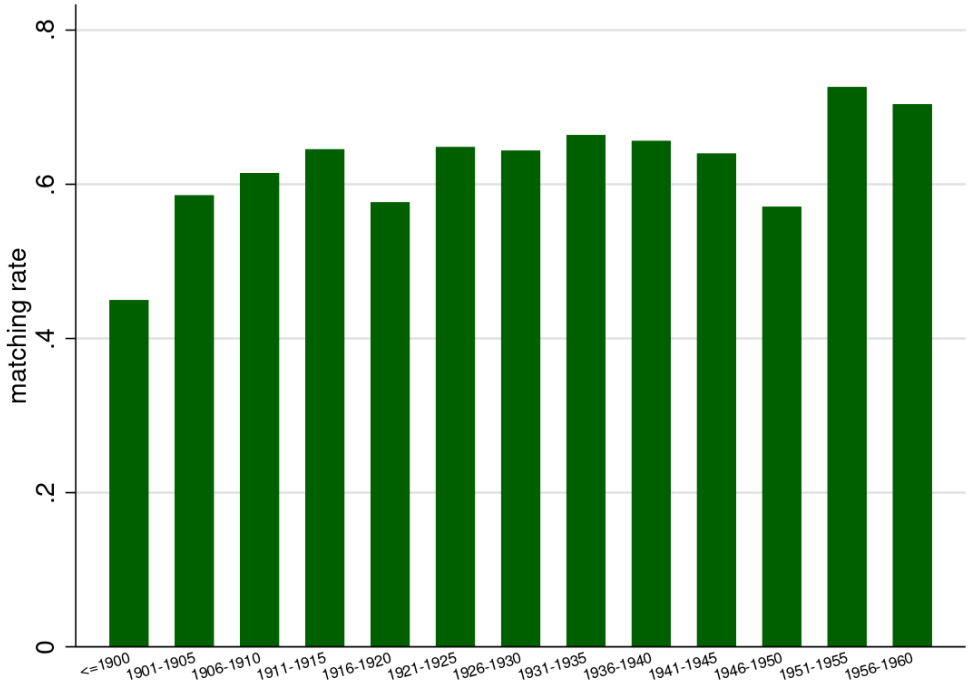


# Matching Rates

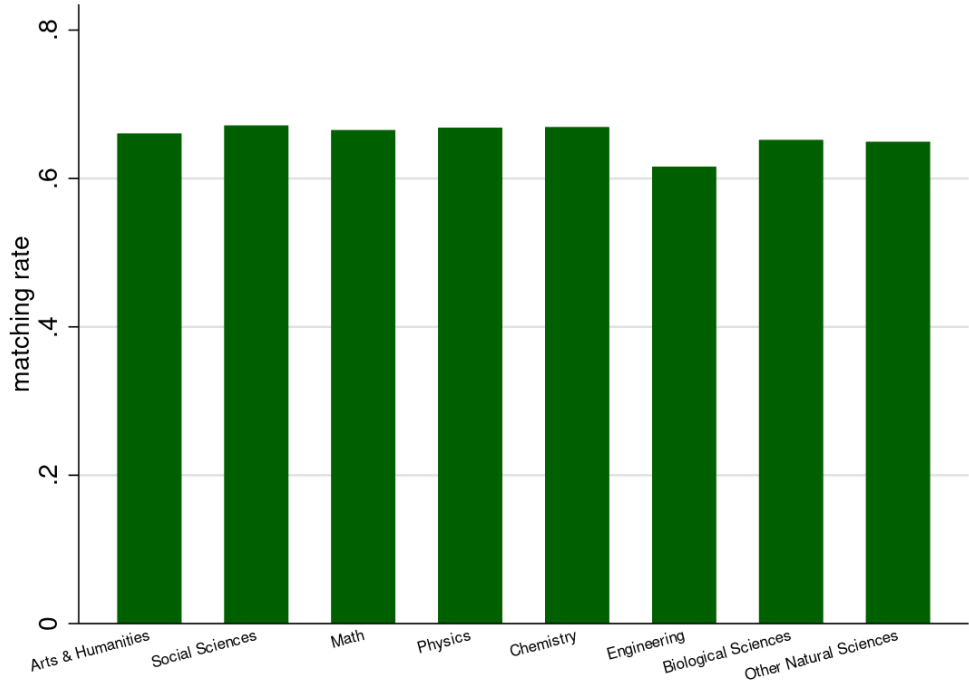
◂ Linked Census

- 102,642 matches between ProQuest PhD recipients and the 1940 population
  - matching scores above the cutoff
  - less than 10 matched per ProQuest recipient (88% unique matches)
- Overall matching rate out of ProQuest is 66%

By degree awarding date




By field

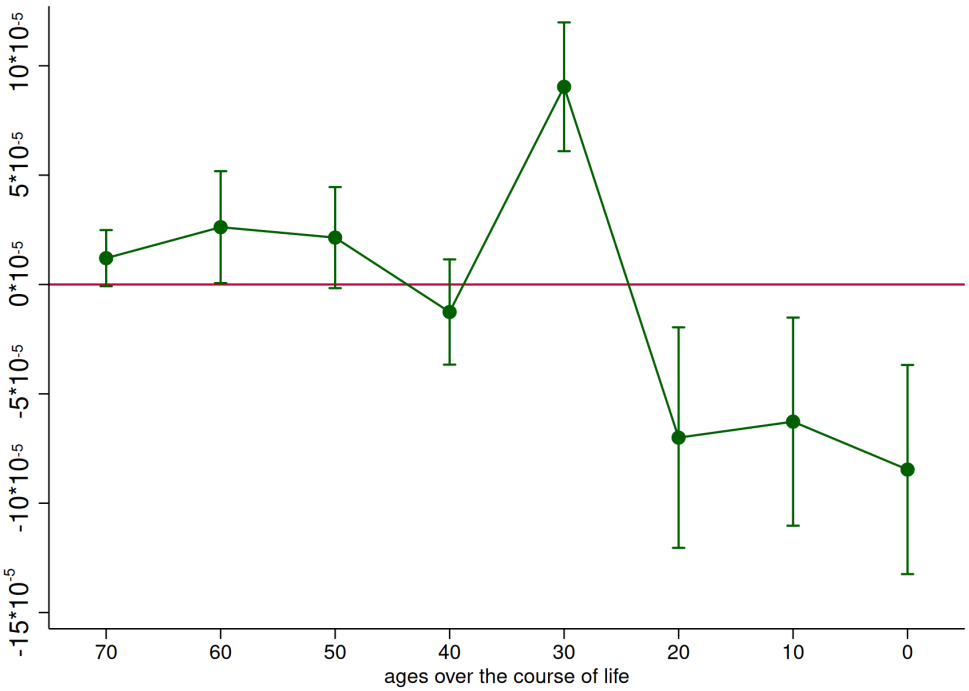


# Probability of Census Outcomes (above 40 years old)

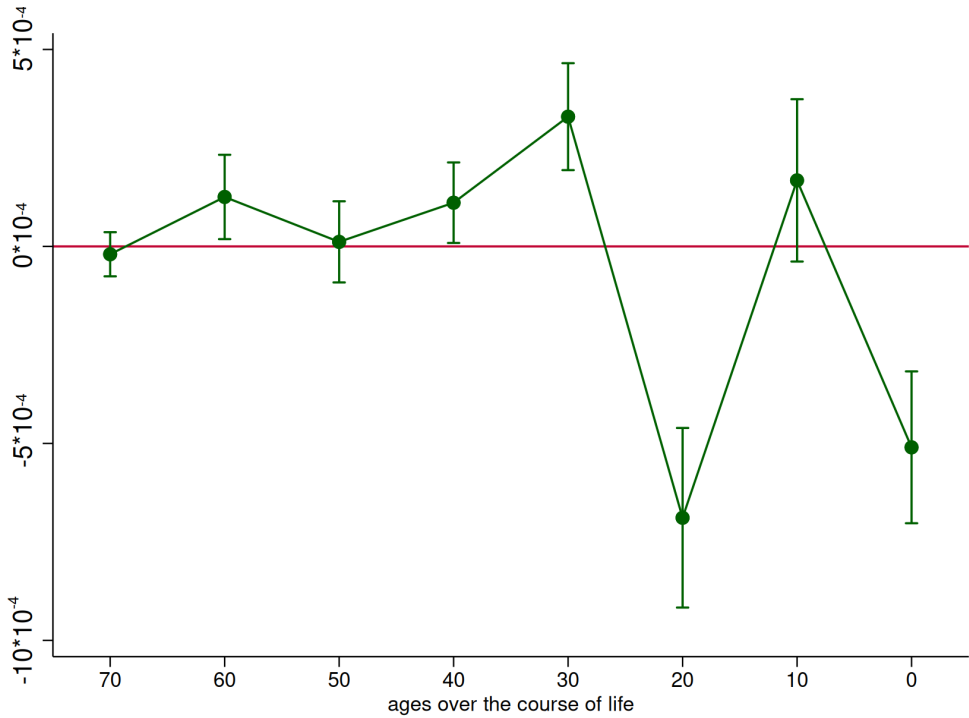
◂ Main estimates

$X$: # of PhD programs in birth state

Earning a graduate degree




Working in professional occupations




- relatively close to the pattern on ProQuest PhD outcome: main effect coming from in-state programs around age 30

# Probability of Census Outcomes (above 40 years old)

◂ Main estimates

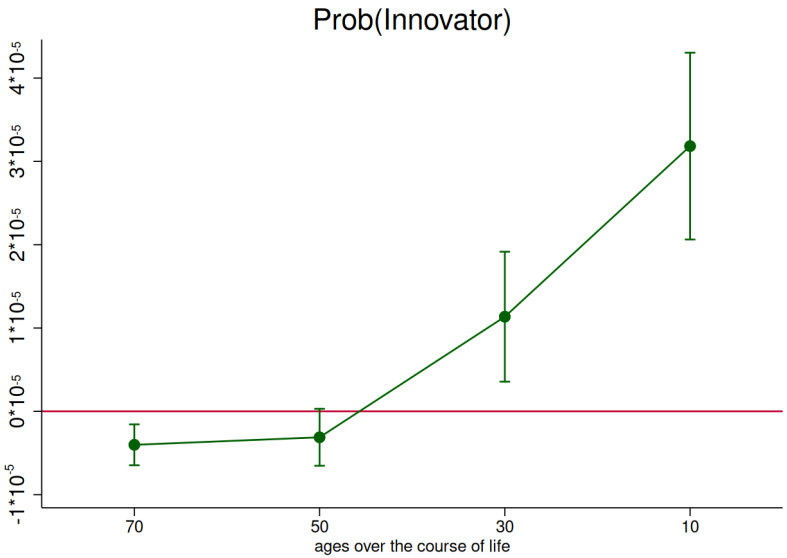
| | (1) Prob(earning a graduate degree) | (2) Prob(working in professional job) |
|---|---|---|
| # in-state programs around 30 | 0.0000904*** | 0.0003294*** |
| | (0.0000150) | (0.0000693) |
| Observations | 29,189,930 | 16,807,302 |

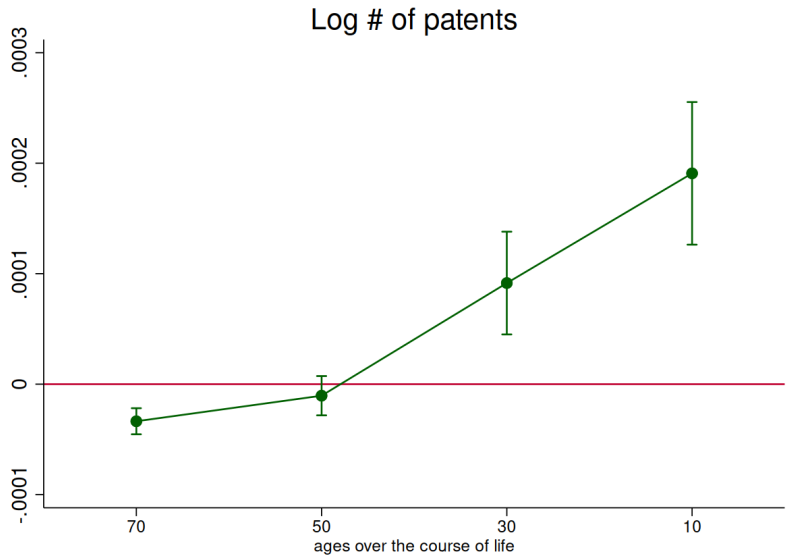
- relatively close to the pattern on ProQuest PhD outcome: main effect coming from in-state programs around age 30

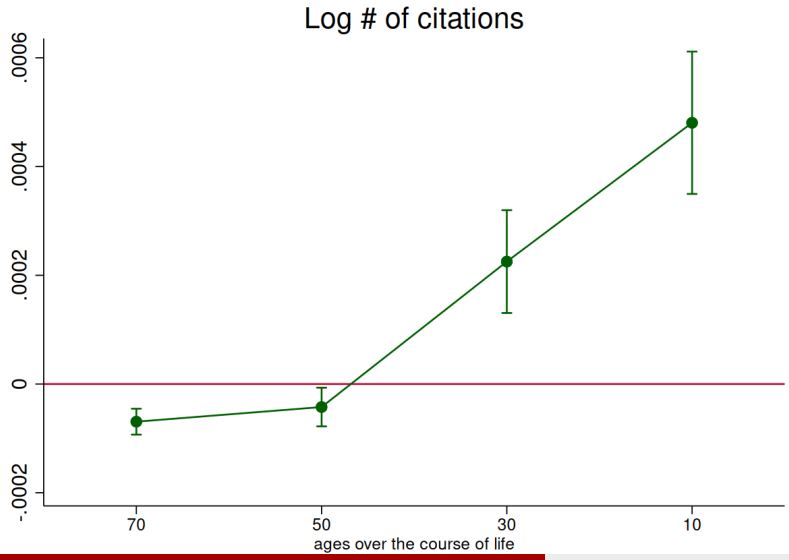
# Spillover Effects on Patenting

Main estimates

$X$: # of PhD programs in birth state



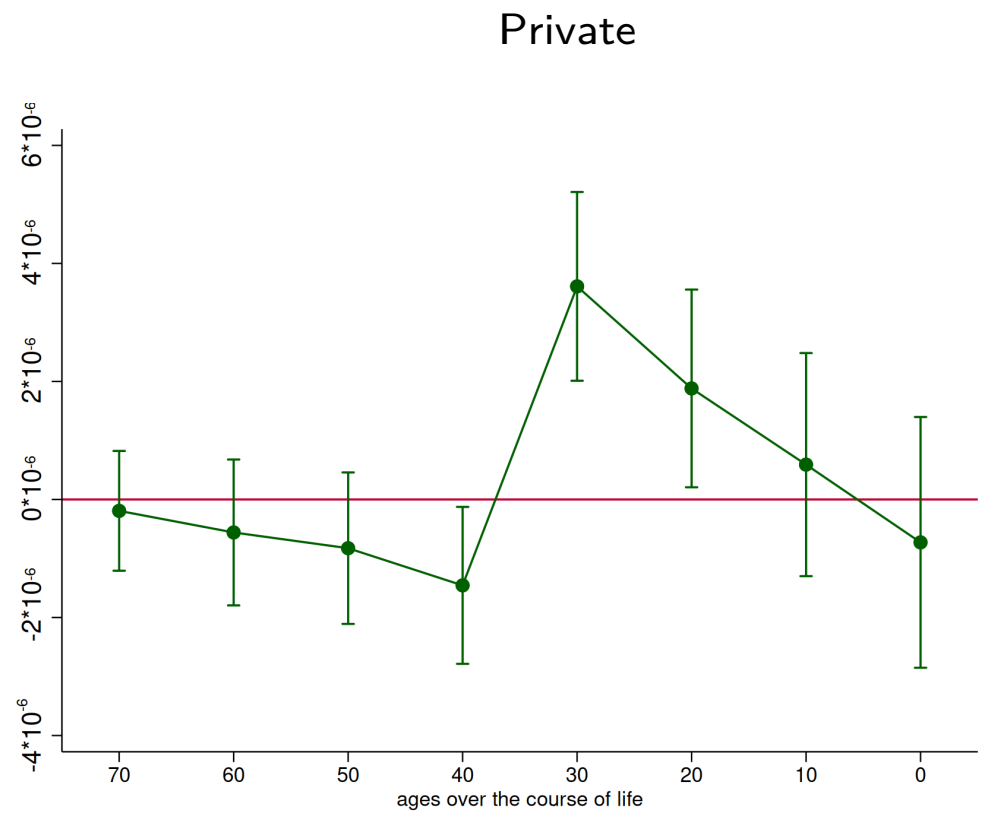



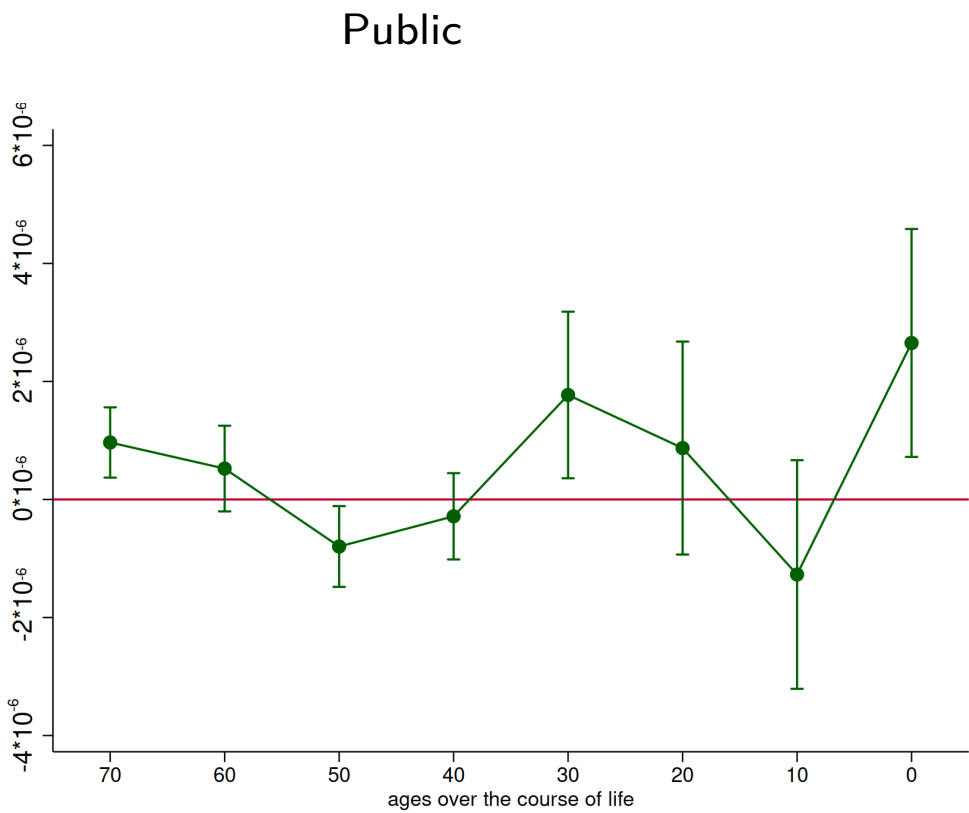

- Positive effect of early life exposure to research institutions on patenting
- Null effect from out-state doctoral training institution openings

# Effects of Different Institutions on Prob(PhD) (One regression)

Main estimates

$X$: # of PhD programs in _____ universities in birth state




- larger effects from private institutions
  - 5 times as many private universities as public ones opened (Goldin and

# Effects of Different Institutions on Prob(PhD) (One regression)

Main estimates

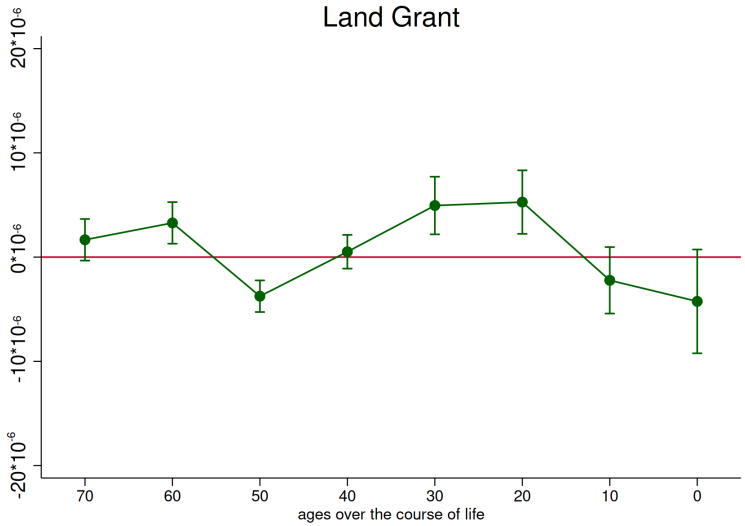
| | (1) Prob(earning a PhD degree) |
|---|---|
| # in-state private programs around 30 | 0.0000036*** |
| | (0.0000008) |
| # in-state public programs around 30 | 0.0000018** |
| | (0.0000007) |
| Observations | 115,004,798 |

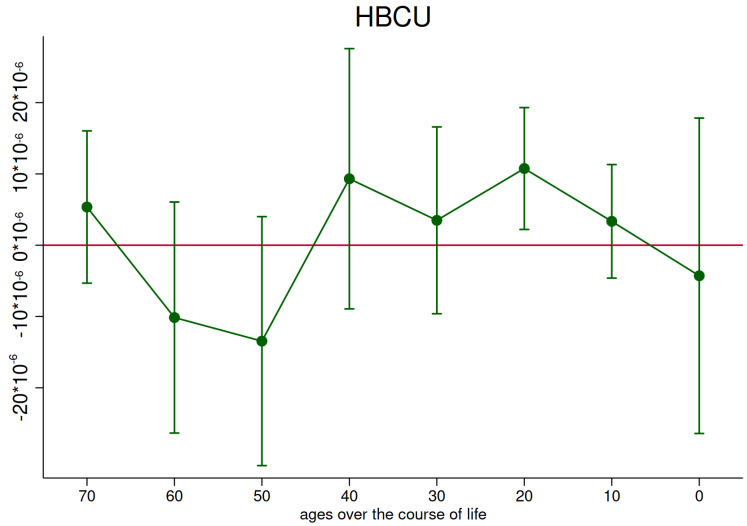
- larger effects from private institutions
  - 5 times as many private universities as public ones opened (Goldin and Katz, 1999)

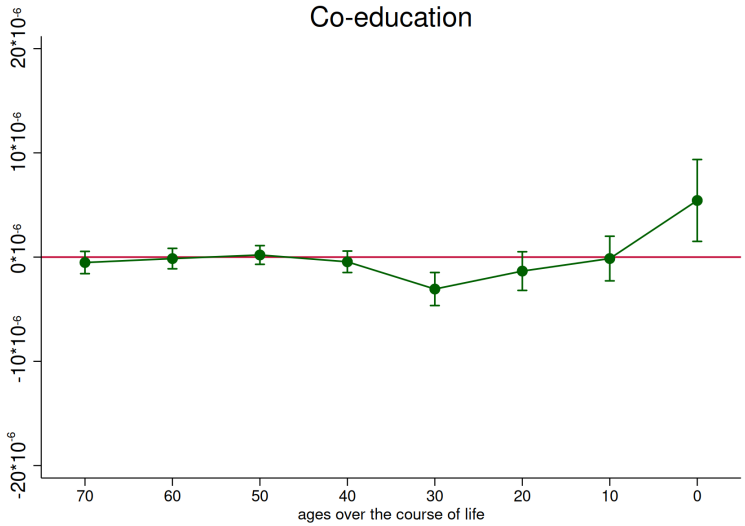
# Effects of Different Institutions on Prob(PhD) (One regression)

Main estimates

$X$: # of PhD programs in _____ universities in birth state




- mainly driven by Land Grant institutions, more research oriented
- associations between policy changes and
  number of graduate programs
- associations between policy changes and
  number of graduate students

# Effects of Different Institutions on Prob(PhD) (One regression)

Main estimates

| | (1) Prob(earning a PhD degree) |
|---|---|
| # in-state Land Grant programs around 30 | 0.0000049*** |
| | (0.0000014) |
| # in-state HBCU programs around 30 | 0.0000035 |
| | (0.0000067) |
| # in-state COED programs around 30 | -0.0000031*** |
| | (0.0000008) |
| Observations | 115,004,798 |

- mainly driven by Land Grant institutions, more research oriented
- associations between policy changes and number of graduate programs
- associations between policy changes and number of graduate students

# Robustness Check

◂ Main estimates

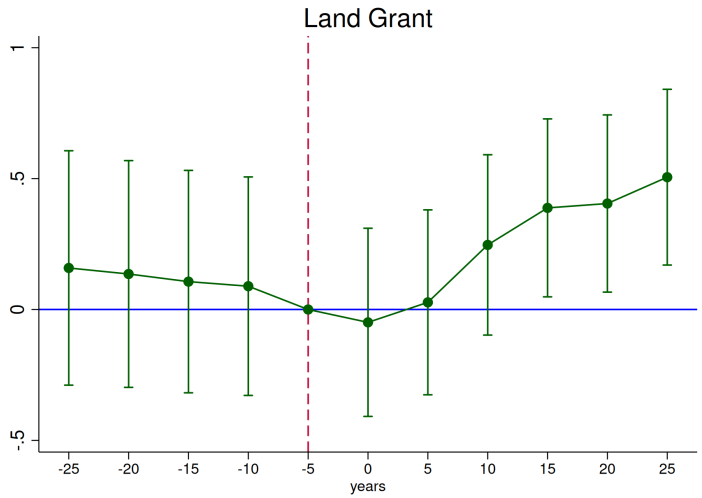
- Various Specifications

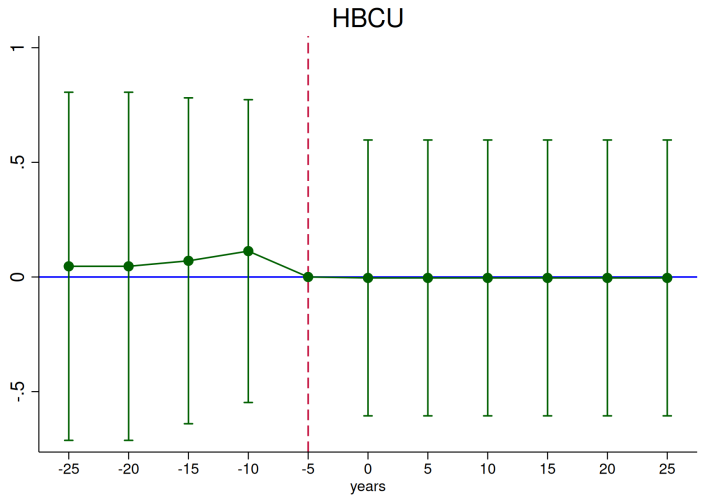
| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Prob(earning a PhD in life) | | | |
| # in-state programs around 30 | 0.0000025*** | 0.0000033*** | 0.0000026*** | 0.0000023*** |
| | (0.0000005) | (0.0000011) | (0.0000005) | (0.0000005) |
| # nearby-state programs around 30 | 0.0000011*** | 0.0000019** | 0.0000009** | 0.0000008** |
| | (0.0000004) | (0.0000007) | (0.0000004) | (0.0000004) |
| Number of colleges | | Y | | |
| Birth Region Trend | | | Y | |
| Birth Division Trend | | | | Y |
| Observations | 115,004,798 | 54,171,289 | 115,004,798 | 115,004,798 |

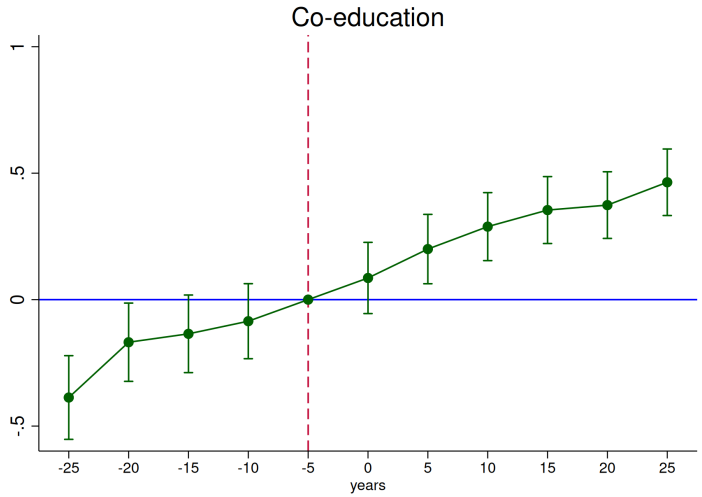
# Land Grant V.S. HBCU V.S. Coed

policy

Associations between each policy change and log number of graduate programs per institution (with institution and year F.E.)



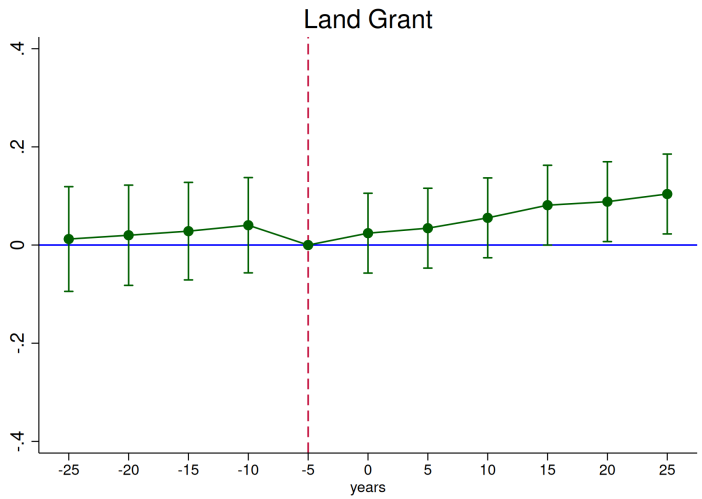



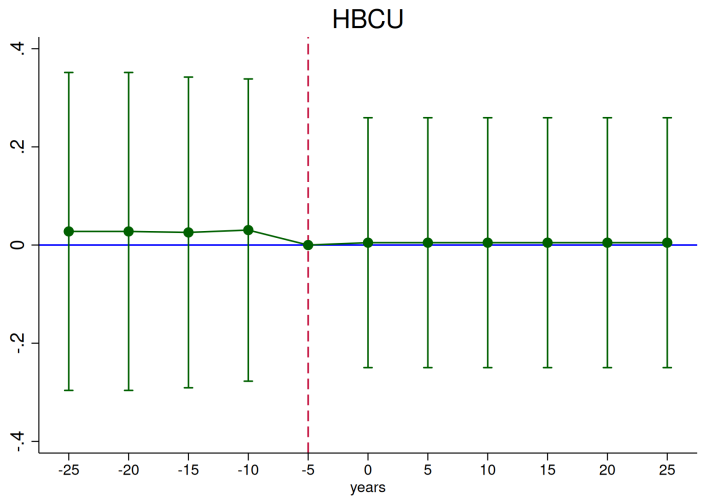

- mainly driven by Land Grant institutions, more research oriented

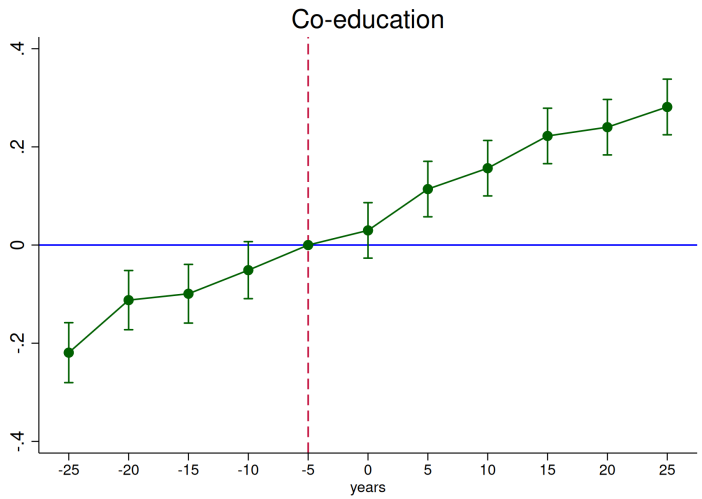
# Land Grant V.S. HBCU V.S. Coed

policy

Associations between each policy change and log number of graduate students per program (with institution and year F.E.)



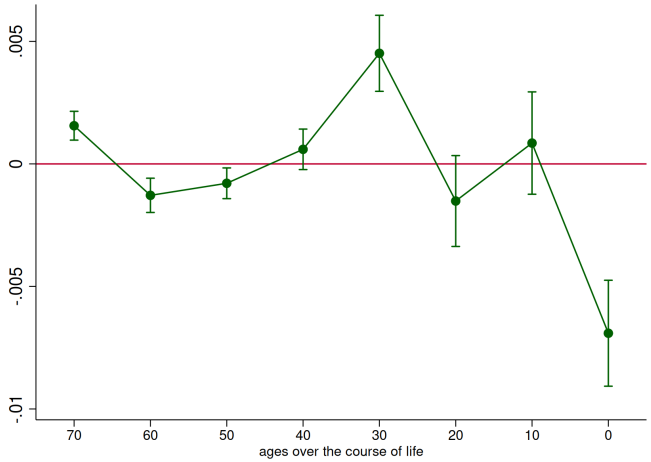




- mainly driven by Land Grant institutions, more research oriented

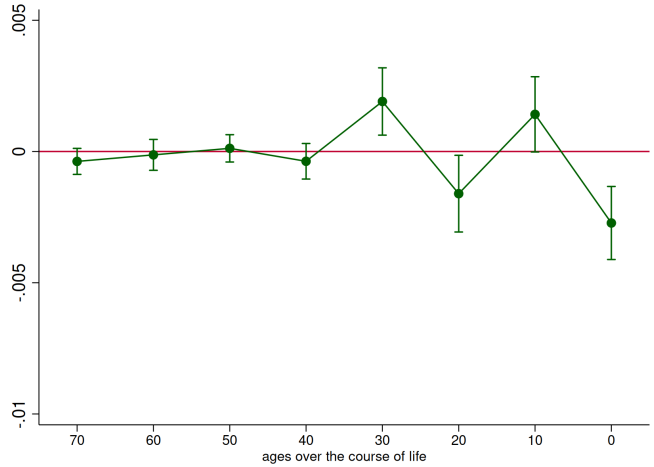
# Location Choices

Choice

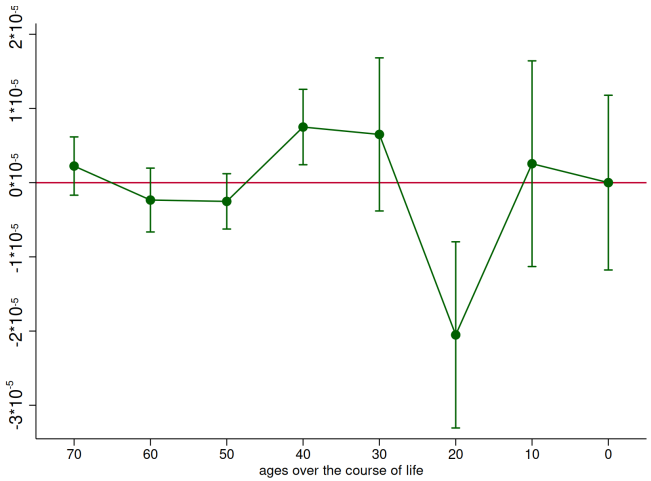
$X$: # of PhD programs in _____
**birth state**



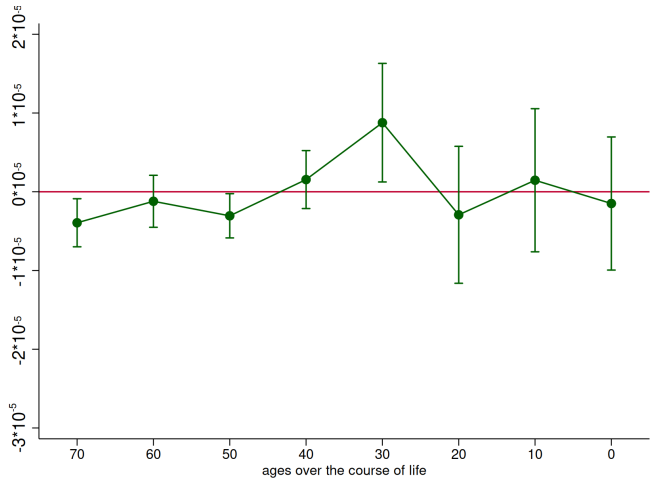

$Y$: earning a PhD in birth state
nearby-states




$X$: # of PhD programs in _____
birth state




$Y$: earning a PhD in nearby-states
**nearby-states**



# Location Choices

◂ Choice

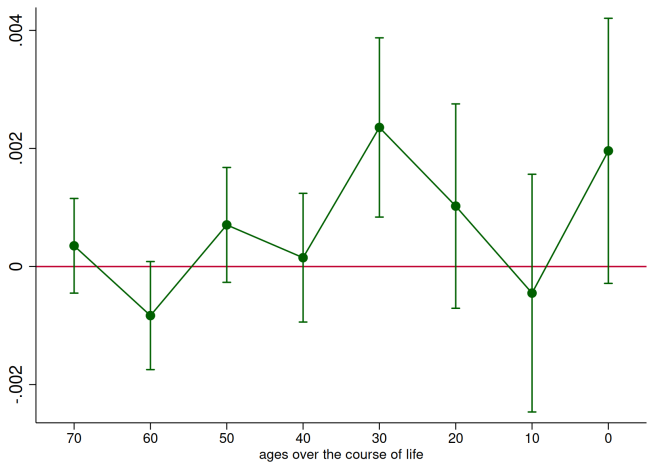
| | (1) Prob(earning an in-state PhD degree) | (2) Prob(earning a nearby-state PhD degree) |
|---|---|---|
| # in-state programs around 30 | 0.0045125*** | 0.0000065 |
| | (0.0007919) | (0.0000053) |
| # nearby-state programs around 30 | 0.0019060*** | 0.0000088** |
| | (0.0006542) | (0.0000038) |
| Observations | 101,082 | 101,082 |

# Field Choices

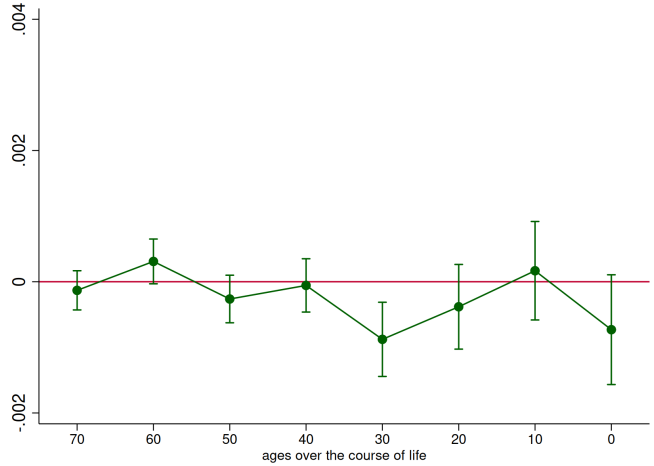
◂ Choice

$X$: # of PhD programs in _____ $Y$: earning a PhD in Arts and Humanities

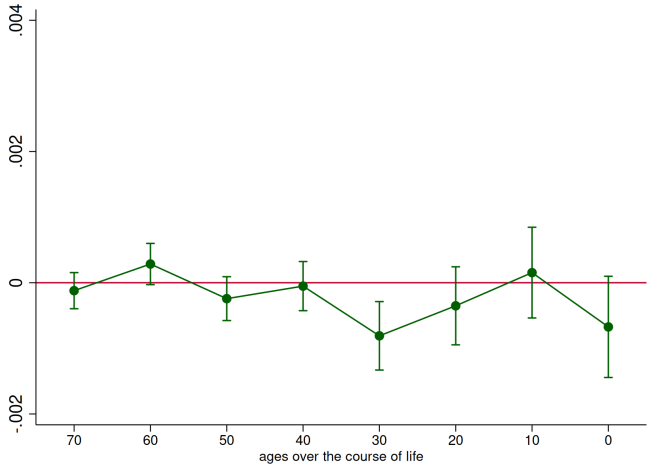
**Arts and Humanities**



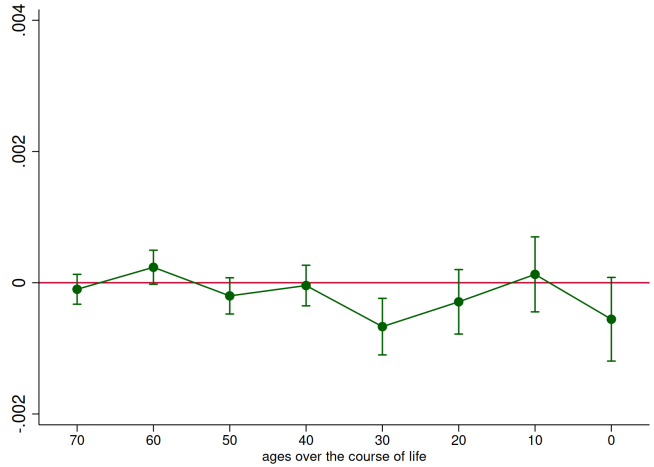

Social and Behavioral Sciences




Mathematical and Physical Sciences




Natural Sciences



# Field Choices

Choice

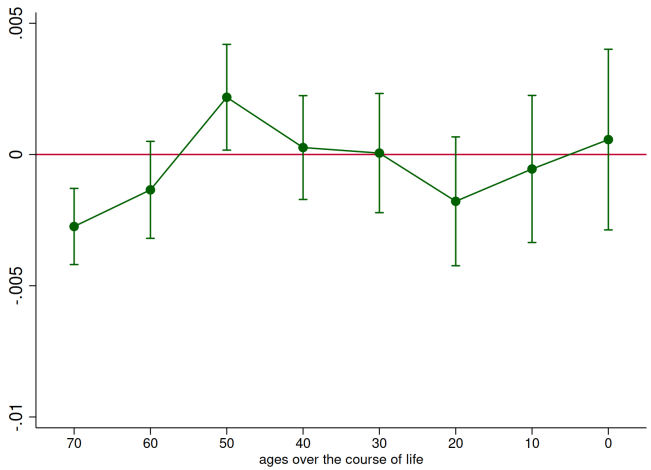
| | (1) |
|---|---|
| | Prob(earning a PhD degree in Arts and Humanities) |
| # in-state Arts and Humanities programs around 30 | 0.0023560*** |
| | (0.0007745) |
| # in-state Social and Behavioral Sciences programs around 30 | -0.0008781*** |
| | (0.0002885) |
| # in-state Mathematical and Physical Sciences programs around 30 | -0.0008090*** |
| | (0.0002662) |
| # in-state Natural Sciences programs around 30 | -0.0006689*** |
| | (0.0002199) |
| Observations | 101,116 |

# Choice of Attending Public Programs

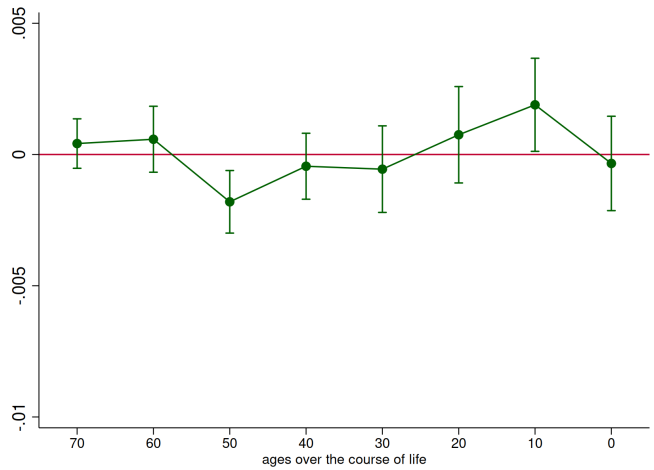
◂ Choice

$X$: # of PhD programs in _____ institutions

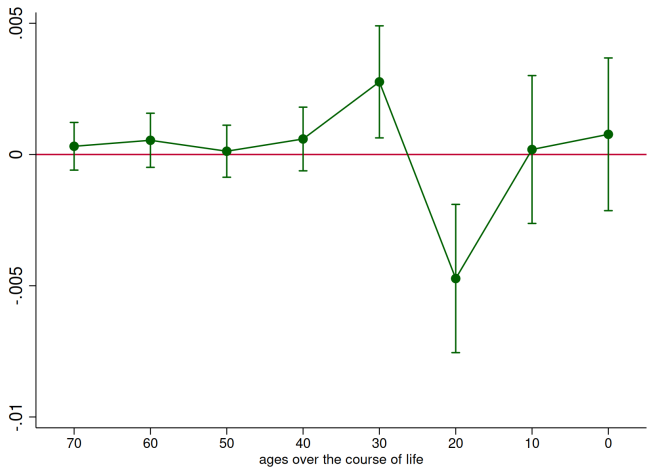
in-state private



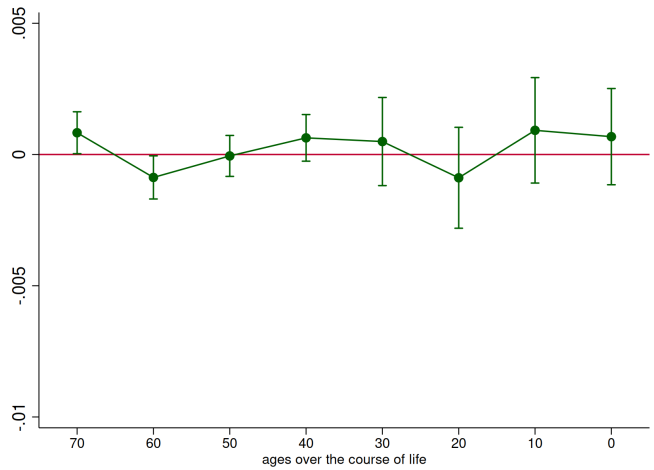

nearby private




**in-state public**




nearby public



# Choice of Attending Public Programs

◂ Choice

| | (1)<br>Prob(earning a PhD degree in public institutions) |
|---|---|
| # in-state private programs around 30 | 0.0000549 |
| | (0.0011588) |
| # in-state public programs around 30 | 0.0027680** |
| | (0.0010890) |
| # nearby-state private programs around 30 | -0.0005579 |
| | (0.0008415) |
| # nearby-state public programs around 30 | 0.0004938 |
| | (0.0008567) |
| Observations | 101,082 |