Competing for Scientific Talent: Industry vs. Academia

Justine Boudou
04/26/2024

Harvard Business School
Scientists drive innovation and economic growth
Scientists can create value in Academia
Scientists can create value in Industry
Markets don’t necessarily direct R&D to areas of highest value

How much friction is there in the allocation of scientists across sectors?
Universities’ AI Talent Poached by Tech Giants

Google, Baidu, Facebook lure away top computer science researchers responsible for training students
Universities’ AI Talent Poached by Tech Giants
Google, Baidu, Facebook lure away top computer science researchers responsible for training students

Why economists are flocking to Silicon Valley
And why big tech wants them
Controversies about the (mis)allocation of scientists

Universities’ AI Talent Poached by Tech Giants
Google, Baidu, Facebook lure away top computer science researchers responsible for training students

Why economists are flocking to Silicon Valley
And why big tech wants them

Why Silicon Valley needs humanities PhDs

INNOVATIONS
Controversies about the (mis)allocation of scientists

Universities’ AI Talent Poached by Tech Giants
Google, Baidu, Facebook lure away top computer science researchers responsible for training students

Why economists are flocking to Silicon Valley
And why big tech wants them

INNOVATIONS
Why Silicon Valley needs humanities PhDs

REPORT: Careers Outside of Academia are Richly Rewarding for PhD Physicists
Why do we care today?

- Knowledge-base economy: intellectual capital is a key driver of economic growth (Romer, 1990; Bell et al., 2017)
Why do we care today?

- Knowledge-base economy: intellectual capital is a key driver of economic growth (Romer, 1990; Bell et al., 2017)
- Competition for talent: firms need talent to maintain competitiveness; universities need talent for prestige and reputation
Why do we care today?

- Knowledge-base economy: intellectual capital is a key driver of economic growth (Romer, 1990; Bell et al., 2017)
- Competition for talent: firms need talent to maintain competitiveness; universities need talent for prestige and reputation
- Increasing earnings gap at graduation between Industry and Academia
Why do we care today?

Earnings gap at graduation in Industry and Academia, by cohort and field.
• How socially efficient is the allocation of scientists across sectors?
Research Question

- How socially efficient is the allocation of scientists across sectors?
- How do scientists perform in Industry vs Academia?
Research Question

- How socially efficient is the allocation of scientists across sectors?
- How do scientists perform in Industry vs Academia?

To answer this question we need:

1. Measures of private and social value
Research Question

• How socially efficient is the allocation of scientists across sectors?
• How do scientists perform in Industry vs Academia?

To answer this question we need:

1. Measures of private and social value
2. Identification strategy that accounts for selection and heterogeneity
This paper

• Measures of private and social value: focus on earnings and publications
  • Private value: earnings + non-pecuniary returns
  • Social value: Private value + Externalities (publications, ...)

• Administrative dataset (SED/SDR) that matches demographics and earnings information about doctorate holders who graduated between 1970 and 2013 to their publication output

• Identification strategy that accounts for selection and heterogeneity

• Model of selection

• Within department variation

• Instrumental variable (IV)

• Estimation of Marginal Treatment Effects (MTE)

• (Mis)Allocation - Preliminary results
This paper

- Measures of private and social value: focus on earnings and publications
  - Private value: earnings + non-pecuniary returns
  - Social value: Private value + Externalities (publications, ...)
- Administrative dataset (SED/SDR) that matches demographics and earnings information about doctorate holders who graduated between 1970 and 2013 to their publication output
This paper

• Measures of private and social value: focus on earnings and publications
  • Private value: earnings + non-pecuniary returns
  • Social value: Private value + Externalities (publications, ...)
  • Administrative dataset (SED/SDR) that matches demographics and earnings information about doctorate holders who graduated between 1970 and 2013 to their publication output

• Identification strategy that accounts for selection and heterogeneity
  • Model of selection
  • Within department variation
  • Instrumental variable (IV)
  • Estimation of Marginal Treatment Effects (MTE)
This paper

- Measures of private and social value: focus on earnings and publications
  - Private value: earnings + non-pecuniary returns
  - Social value: Private value + Externalities (publications, ...)
  - Administrative dataset (SED/SDR) that matches demographics and earnings information about doctorate holders who graduated between 1970 and 2013 to their publication output

- Identification strategy that accounts for selection and heterogeneity
  - Model of selection
  - Within department variation
  - Instrumental variable (IV)
  - Estimation of Marginal Treatment Effects (MTE)

- (Mis)Allocation - Preliminary results
Preliminary Findings

• Joining Industry (vs Academia) leads to:
  • A large increase in lifetime earnings, driven by the early-stage of the career
  • A large decrease in the overall stock of publications
Preliminary Findings

- Joining Industry (vs Academia) leads to:
  - A large increase in lifetime earnings, driven by the early-stage of the career
  - A large decrease in the overall stock of publications

- I find substantial heterogeneity and selection (ATT > ATE)

<table>
<thead>
<tr>
<th>Sector joined</th>
<th>$\beta^W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>85%</td>
</tr>
<tr>
<td>Academia</td>
<td>2%</td>
</tr>
</tbody>
</table>
Preliminary Findings

- Joining Industry (vs Academia) leads to:
  - A large **increase** in **lifetime earnings**, driven by the early-stage of the career
  - A large **decrease** in the overall **stock of publications**

- I find substantial **heterogeneity** and **selection** (ATT > ATE)

<table>
<thead>
<tr>
<th>Sector joined</th>
<th>$\beta^W$</th>
<th>$\beta^{Pub}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>85%</td>
<td>-60%</td>
</tr>
<tr>
<td>Academia</td>
<td>2%</td>
<td>-81%</td>
</tr>
</tbody>
</table>

The earnings premium in Industry appears relatively stable over time. The publications gap increased in the 90s and started to decrease again in the mid 2000s.
Preliminary Findings

- Joining Industry (vs Academia) leads to:
  - A large **increase** in **lifetime earnings**, driven by the early-stage of the career
  - A large **decrease** in the overall **stock of publications**

- I find substantial **heterogeneity** and **selection** (ATT > ATE)

<table>
<thead>
<tr>
<th>Sector joined</th>
<th>$\beta^W$</th>
<th>$\beta^{Pub}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>85%</td>
<td>-60%</td>
</tr>
<tr>
<td>Academia</td>
<td>2%</td>
<td>-81%</td>
</tr>
</tbody>
</table>

- The earnings premium in Industry appears relatively stable over time
- The publications gap increased in the 90s and started to decrease again in the mid 2000s
Contributions

• Literature on the role of Industry and Academia in the production of knowledge and innovation
  • Difference in institutional norms: (Dasgupta and David, 1994; Aghion et al., 2008; Stephan, 2012; Furman and Stern, 2011; Kahn and Ginther, 2017; Stephan, 1996; Stephan et al., 2004; Agarwal and Ohyama, 2013)
  • Differences in individuals’ preferences: (Stern, 2004; Roach and Sauermann, 2010, 2023, 2024; Sauermann and Roach, 2012; Sauermann and Stephan, 2013)
• Literature on the relationship between talent, innovation and economic growth
  • Frictions related to the extensive margin: Bell et al. (2017), Akcigit et al. (2020), Celik (2023)
  • Frictions related to the intensive margin: Akcigit and Goldschlag (2023)
• Literature on resource allocation within organizations and its consequences for performance
  • Literature on general and firm-specific human capital: Coff and Kryscynski (2011)
  • Selection: Shu (2016)
Data
Data Sources

- **Survey of Doctorate Recipients (SDR 2015):** *(during the career)*
  - Across cohorts, longitudinal
  - **Earnings (flow), Patents (flow, sub-sample)*

- **Survey of Doctorate Recipients Bibliometric Research Data:**
  - SDR 2015 matched with Web of Science
  - **Publication output (stock) from PhD graduation to 2017***
Data Sources

- **Survey of Doctorate Recipients (SDR 2015):** *(during the career)*
  - Across cohorts, longitudinal
  - **Earnings (flow), Patents (flow, sub-sample)*

- **Survey of Doctorate Recipients Bibliometric Research Data:**
  - SDR 2015 matched with Web of Science
  - **Publication output (stock) from PhD graduation to 2017**

- **Survey of Earned Doctorates (SED):** *(at graduation)*
  - Within cohort
  - Sector joined at graduation (+ earnings at start)

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="#" alt="SED 1970" /></td>
<td><img src="#" alt="SED 2007" /></td>
<td><img src="#" alt="SED 2013" /></td>
<td><strong>SDR 2015</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample and main variables

- **21,406 individuals**
  - Ind. with PhD cohort $\geq 1970$, with definite postgraduate commitment at the time of survey and start their career in the U.S.

- **Treatment:**
  - *Starting* in Industry vs Academia (incl. Government/Postdoc)

- **Outcomes:**
  - Earnings: total earned income before deductions in the previous year (2015 US dollars)
  - Publications: from PhD graduation to the time of survey (2017)
Supply

57 PhD majors, 5 fields (CS/Maths, Life Sc., Phys. Sc., Social Sc., Eng.)
Sorting at graduation, by PhD field

![Graph showing the share of individuals going to Industry (%) by PhD field from 1985 to 2015 for Computer Sciences/Mathematics, Life Sciences, Physical Sciences, Social Sciences, and Engineering.](image-url)
Raw differences - Earnings and Publications

(a) Cumulative number of publications

(b) Earnings (in 2015 USD th.)
Empirical Framework and Results
Overview

1. ‘Control’ for ability: within department variation

2. Selection at the individual level (LATE)

3. Selection at the individual level + Heterogeneity (MTE)

4. (Mis)allocation
1. ‘Control’ for ability: within department variation

2. Selection at the individual level (LATE)

3. Selection at the individual level + Heterogeneity (MTE)

4. (Mis)allocation
Within department variation

\[ Y_i = \alpha_0 + \alpha_1 \text{Industry}_i + \delta_{\text{mu}} + \alpha_2 X_i + \varepsilon_i \]  

---

- \( \delta_{\text{mu}} \): PhD major \( \times \) university FE
Within department variation

\[ Y_i = \alpha_0 + \alpha_1 \text{Industry}_i + \delta_{mu} + \alpha_2 \mathbf{X}_i + \varepsilon_i \] (1)

- \( \delta_{mu} \): PhD major × university FE

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry</td>
<td>-0.911***</td>
<td>-0.943***</td>
<td>-0.936***</td>
<td>-0.942***</td>
<td>-0.900***</td>
</tr>
<tr>
<td></td>
<td>(0.0200)</td>
<td>(0.0255)</td>
<td>(0.0255)</td>
<td>(0.0278)</td>
<td>(0.0296)</td>
</tr>
<tr>
<td>Observations</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
</tr>
<tr>
<td>Demographics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experience</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PhD major FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Doct. Inst. FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PhD major × Doct. Inst. FE</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taste</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standard errors in parentheses clustered at the doctoral institution level
Within department variation

\[ Y_i = \alpha_0 + \alpha_1 \text{Industry}_i + \delta_{mu} + \alpha_2 X_i + \varepsilon_i \]  

(2)

- \( \delta_{mu} \): PhD major × university FE

<table>
<thead>
<tr>
<th></th>
<th>Log(1+Publications (stock))</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>Industry</td>
<td>-0.911***</td>
<td>-0.943***</td>
<td>-0.936***</td>
<td>-0.942***</td>
<td>-0.900***</td>
</tr>
<tr>
<td></td>
<td>(0.0200)</td>
<td>(0.0255)</td>
<td>(0.0255)</td>
<td>(0.0278)</td>
<td>(0.0296)</td>
</tr>
<tr>
<td>Observations</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
</tr>
</tbody>
</table>

|                | Log(1+Earnings)             |          |          |          |          |
|                | (1)                         | (2)      | (3)      | (4)      | (5)      |
| Industry       | 0.367***                    | 0.296*** | 0.295*** | 0.302*** | 0.302*** |
|                | (0.0121)                    | (0.0134) | (0.0135) | (0.0149) | (0.0169) |
| Observations   | 21,406                      | 21,406   | 21,406   | 21,406   | 21,406   |

- Demographics: Yes
- Experience: Yes
- PhD major FE: Yes
- Doct. Inst. FE: Yes
- PhD major × Doct. Inst. FE: Yes
- Taste: Yes

Standard errors in parentheses clustered at the doctoral institution level
Overview

1. ‘Control’ for ability: within department variation

2. Selection at the individual level (LATE)

3. Selection at the individual level + Heterogeneity (MTE)

4. (Mis)allocation
**Instrumental Variable**

- **Ideal Variation**: PhDs graduating face an exogenous demand from firms’ and an exogenous demand from universities’

- **Empirical Variation**: exploits inter-temporal variation in firms’ vs universities’ demand for PhDs in the same major (within major across cohorts)
**Instrumental Variable**

- **Instrument**: exploits inter-temporal variation in firms’ vs universities’ demand for PhDs in the same major (within major across cohorts)

- **Empirically**: PhD major × PhD cohort FE, controlling for PhD major FE
  - I create a leave-one out continuous measure $Z_i = \text{share of individuals in major/cohorts going to Ind.}$
Instrumental Variable

• **Instrument**: exploits inter-temporal variation in firms’ vs universities’ demand for PhDs in the same major (within major across cohorts)

• **Empirically**: PhD major × PhD cohort FE, controlling for PhD major FE
  - I create a leave-one out continuous measure $Z_i = \text{share of individuals in major/cohort going to Ind.}$

• **Identifying assumption**: individuals that are more/less productive in Ind. (vs Acad.) do not time when they go on the market based on macro. conditions

• **Controls**:  
  - Doctoral institution FE  
  - Indicator for having published prior to graduation
**Instrumental Variable**

- **Instrument**: exploits inter-temporal variation in firms’ vs universities’ demand for PhDs in the same major (within major across cohorts)

- **Empirically**: PhD major $\times$ PhD cohort FE, controlling for PhD major FE
  - I create a leave-one out continuous measure $Z_i = \text{share of individuals in major/cohoot going to Ind.}$

- **Identifying assumption**: individuals that are more/less productive in Ind. (vs Acad.) do not time when they go on the market based on macro. conditions

- **Controls**:
  - Doctoral institution FE
  - Indicator for having published prior to graduation

- **Test**: conditions at the time of PhD entry *do not* predict sorting
### Instrumental Variable - First-stage

<table>
<thead>
<tr>
<th></th>
<th>Industry (1)</th>
<th>Industry (2)</th>
<th>Industry (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_i$</td>
<td>0.220***</td>
<td>0.214***</td>
<td>0.214***</td>
</tr>
<tr>
<td></td>
<td>(0.0385)</td>
<td>(0.0390)</td>
<td>(0.0389)</td>
</tr>
<tr>
<td>Demographics</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experience</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PhD major FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Doct. inst. FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Publ. bef. grad.</td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>F-stat</td>
<td>32.5</td>
<td>30.1</td>
<td>30.1</td>
</tr>
<tr>
<td>Observations</td>
<td>21,406</td>
<td>21,406</td>
<td>21,406</td>
</tr>
<tr>
<td>R-sq</td>
<td>0.163</td>
<td>0.185</td>
<td>0.185</td>
</tr>
</tbody>
</table>

**Notes:** Standard errors (in parentheses) are two-way clustered at the PhD major and PhD cohort level.

- $Z_i$ = share of individuals in major/cohoot going to Ind.
- If everyone in her major/cohoot goes to Industry, the focal individual has a 21% higher probability of going to Ind. (vs Acad.)
Instrumental Variable - 2SLS results

<table>
<thead>
<tr>
<th></th>
<th>Log(1+Publications)</th>
<th>Log(1+Earnings)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>2SLS</td>
</tr>
<tr>
<td>Industry</td>
<td>-0.900***</td>
<td>-1.227***</td>
</tr>
<tr>
<td></td>
<td>(0.0296)</td>
<td>(0.476)</td>
</tr>
<tr>
<td>F-stat</td>
<td>30.1</td>
<td>30.1</td>
</tr>
<tr>
<td>Observations</td>
<td>21,406</td>
<td>21,406</td>
</tr>
<tr>
<td>Demographics</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Experience</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PhD major FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Doct. Inst. FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Publ. bef. grad.</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Notes: Standard errors in parentheses clustered at the PhD major and PhD cohort level. * p < 0.10, ** p < 0.05, *** p < 0.01.

- OLS ≠ 2SLS because (Ishimaru, 2024):
  - Covariate weight difference
  - Treatment-level weight difference
  - Selection
Overview

1. ‘Control’ for ability: within department variation

2. Selection at the individual level (LATE)

3. Selection at the individual level + Heterogeneity (MTE)

4. (Mis)allocation
• I estimate **Marginal Treatment Effects** (MTE)
  
  • The ATE for groups of individuals who have the same **latent propensity** to be treated based on observables and unobservables

  • Is the micro-component of the ATE, ATT, LATE

• MTE for individuals with observables $X = x$ and unobservable $U_D = u$ is:

\[
\text{MTE}(X = x, U_D = u) = \mathbb{E}[Y^I - Y^A | X = x, U_D = u]
\]  
(3)
Heterogeneity on observables and unobservables - MTE

- I estimate **Marginal Treatment Effects (MTE)**
  - The ATE for groups of individuals who have the same **latent propensity** to be treated based on observables and unobservables
  - Is the micro-component of the ATE, ATT, LATE

- MTE for individuals with observables $X = x$ and unobservable $U_D = u$ is:

\[
MTE(X = x, U_D = u) = E[Y^I - Y^A | X = x, U_D = u]
\] (3)

- I can predict expected treatment effects for every individual based on their observables, treatment status and unobservable
Individual-specific expected treatment effects

Figure 2: Log(Earnings)

- Individuals who start their career in Industry have higher earnings gains of joining Ind. (vs Acad.)
Individual-specific expected treatment effects

**Figure 2:** Log(Earnings)
- Individuals who start their career in Industry have higher earnings gains of joining Ind. (vs Acad.)

**Figure 3:** Log(Publications)
- Individuals who start their career in Industry have a lower publication decrease of joining Ind. (vs Acad.)
Individuals who start their career in Industry have higher earnings in Industry and lower earnings in Academia than those who start their career in Academia.
Individuals who start their career in Industry have a higher publication potential in Industry than those who start their career in Academia.

In the private sector, individuals who start their career in Industry join bigger firms than those who start their career in Academia.
Individual-specific expected potential outcomes

- Individuals who start their career in Industry have the same publication potential in Academia than those who start their career in Academia.
- Individuals who start their career in Industry lack non-research related skills that are valued in Academia.
1. ‘Control’ for ability: within department variation → ATE

2. Selection at the individual level → LATE

3. Selection at the individual level + Heterogeneity → MTE

4. (Mis)allocation
Goal: think about the **wedge** between **private** and **social** returns

- Individuals’ sorting is influenced by their **private** returns in Ind. (vs Acad.)
- Individuals’ ‘best use’ is influenced by their **social** returns in Ind. (vs Acad.)
- The wedge between **private** and **social** returns comes from **externalities**
(Mis)Allocation

- Goal: think about the **wedge** between **private** and **social** returns
  - Individuals’ sorting is influenced by their **private** returns in Ind. (vs Acad.)
  - Individuals’ ‘best use’ is influenced by their **social** returns in Ind. (vs Acad.)
  - The wedge between **private** and **social** returns comes from **externalities**

- Goal: highlight **trends** and **patterns**
  - I cannot capture the entirety of private returns
  - I cannot capture the entirety of externalities
  - I can only discuss the **rate** of externalities, not the **direction**
Goal: think about the wedge between private and social returns
- Individuals’ sorting is influenced by their private returns in Ind. (vs Acad.)
- Individuals’ ‘best use’ is influenced by their social returns in Ind. (vs Acad.)
- The wedge between private and social returns comes from externalities

Goal: highlight trends and patterns
- I cannot capture the entirety of private returns
- I cannot capture the entirety of externalities
- I can only discuss the rate of externalities, not the direction

I am currently trying to build a simple model of social returns where I use the individual-level MTE for earnings and publications to shed light on misallocation
- I can compare the wedge in earnings with the wedge in publications
- I can characterize individuals based on where they are in the Earnings and Publications distribution in Industry and in Academia

All the extensions are the fun part!
In this paper, I study the **performance** of scientists in Industry and Academia after graduation:

- **Data**: I use administrative surveys that give me individual-level information about two measures of performance: **earnings** and **publications**

- **Identification**:
  - I account for individual-level **selection** with an instrument that leverage variation in firms’ vs universities’ demand over time for PhDs in the same major
  - I incorporate **heterogeneity** with the estimation of MTE

- **Next steps**: use the reduced-form individual-level estimates to highlight trends and patterns in the misallocation of scientists across sectors
Thank you!
\[ \text{ATE}(x) = \int_{0}^{1} MTE(x, u) du \]  
\[ \text{ATT}(x) = \int_{0}^{1} h_{\text{ATT}}(x, u) MTE(x, u) du \]  
\[ \text{ATUT}(x) = \int_{0}^{1} h_{\text{ATUT}}(x, u) MTE(x, u) du \]  
\[ \text{IV}(x) = \int_{0}^{1} h_{\text{IV}}(x, u) MTE(x, u) du \]
Figure 4: Propensity Score
MTE Curves

**Figure 5**: Log(1+Earnings)

**Figure 6**: Log(1+Publications)
References


