Skills and Spills: Pathways from research to innovation
Matthew B. Ross¹, Akina Ikudo², Joseph Staudt³, Bruce A. Weinberg⁴,⁵, Julia I. Lane¹

Abstract: An existing body of research has identified a strong relationship between university research activity and local economic growth. That literature has primarily examined spillovers resulting from vendor/supplier relationships, technology transfer and commercialization licenses, coauthorship networks, or patent citations. Less attention has been paid to the idea that highly skilled workers serve as a conduit for transmitting ideas between the academy and private industry. Here, we explore this important channel for knowledge spillovers using data detailing employment on sponsored research projects for a subset of UMETRICS universities linked LEHD employment and earnings. We characterize individuals by their primary role in the laboratory, source of funding support, and the topical focus of their research. For those workers who transition out of academia, we trace out the short-run employment linkages in terms of detailed industry sector, geography, and firm characteristics. Using this rich mapping between academia and industry, we examine the implications of an exogenous increase in highly skilled workers resulting from aggregate shocks to scientific research funding. In particular, we quantify the direct effect to a particular industry and location in terms of firm births, revenues, employment, and earnings. In addition, we explore secondary effects to related industries and neighboring geographies as well as the crowding out/in of both high and low skilled workers. The findings from this study provide new and exciting insights about how the mobility of highly skilled workers facilitates the cross-pollination of knowledge, particularly from research universities to local economies. Our findings also provide important policy implications for considering the broader impact of national investments in basic scientific research.

¹ New York University, Wagner Graduate School of Public Service; New York, NY
² University of California Los Angeles, Economics Department, Los Angeles, CA
³ United States Census Bureau, Center for Economic Studies, Washington D.C.
⁴ University of California Los Angeles; Department of Economics; Los Angeles, CA
⁵ The Ohio State University; Department of Economics; Columbus, OH
⁶ National Bureau of Economic Research; Boston, MA

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. This research was supported by the National Center for Science and Engineering Statistics. NSF ScisIP Awards 1064220 and 1262447; NSF Education and Human Resources DGE Awards 1348691, 1547507, 1348701, 1535399, 1535370; NSF NCSES award 1423706; NIHPO1AG039347; and the Ewing Marion Kaufman and Alfred P. Sloan Foundations. Data were generously provided by the Committee on Institutional Cooperation and its member institutions. The research agenda draws on work with many coauthors, but particularly with Jason Owen Smith.
1. INTRODUCTION

There is an enormous literature on the link between investments in human capital and economic growth at the national, state and regional levels. As Romer and others have pointed out, those investments appear to be greater the greater the level of human capital because technological change “comes from things that people do” (1). The notion that there are such spillovers associated with high levels of human capital is examined in a similarly large literature studying the link between a particular type of human capital – that derived from university research - and economic growth. Universities do appear to be a key driver of regional economies (2–4). An important insight is that distance from a university is a key factor: geographical closeness affects the transmission of tacit knowledge(5) and spatial proximity is significant in regional knowledge production functions(6, 7). Geographic location affects both network connections and the strategic benefits of location in dense alliance networks (8).

However, much of the literature has been based on aggregate data with links inferred rather than directly measured (9–11). An ideal analysis would construct direct measures of the inputs that flow into economic activity: stock measures of human capital that were directly measured from the flows of high human capital into regional economies. While such data do not currently exist, new information is becoming available that does directly measure the flows of one type of high human capital – research trained individuals – into local economies.

It is these new data that form the basis of our work. We use these data to estimate the flows of human capital into regional economies – both at the aggregate and the detailed industry level. We find, using a shift-share approach and using counties as the unit of analysis, that

a. The closer counties are to a research-intensive university, the greater the flows of research trained workers to the county
b. greater levels of human capital affect economic activity in more R&D intensive industries
c. There are measurable spillovers to other industries in the county

There is a literature that suggests that firm and economic growth can be significantly affected by highly skilled workers (12, 13) Particularly useful in this context is economy-wide linked employer-employee data, such as the LEHD data (14). Such data have been used in the past to generate different measures of worker experience at different types of businesses (15). Barth et al., for example, show that there are returns to experience at R&D performing firms (16); Abowd et al. also use linked data to compute person specific measures of human capital (17).
An extensive and parallel literature links regional economic development clusters with the presence of active research universities (9, 18–20). Qualitative evidence suggests that universities are engines of regional economic development—the existence of Silicon Valley has been traced to its propinquity to Stanford, Boston’s growth has been attributed to its great research universities, and the Research Triangle to the research performed at Duke, University of North Carolina and North Carolina State University (21). Part of this effect could be due to the effects of educational investments. Although it has been difficult to identify these relationships, research in Sweden has shown that universities have important effects on the productivity of workers (22), and research in the United States suggests that a 10% increase in higher education spending increases local non-education sector labor income by about 0.5% (10).

The spillover effect is thought to occur through two mechanisms. One is human contact: when people are physically close to each other, skills are easier to acquire and knowledge is easier to exchange (5, 23). The second is that the flows of students to jobs in regional businesses directly increase the human capital in the workforce, resulting in more productive businesses (24, 25). As Hausman points out, many of the mechanisms, particularly hiring, patenting and spinoffs, have a local bias (9).

The empirical analysis has used both firms and regional geographies as units of analysis. An excellent example of the analysis at the firm level is a thorough analysis of some 475 German IPOs, the spillover mechanisms for firms are captured by the level of scientific activity (the number of articles by scientists of a university, the citations, third party funding, the number of university patents and the number of students) at the closest university (26). An example of work at the county level is an analysis of a sample of counties selected through propensity score matching which suggests that counties that host new universities grew around 20 per cent more than those that did not (27).

The input and output measures are also constructed differently. Human capital measures are variously defined. One measure is “degree production intensity—calculated as the number of degrees produced in a metropolitan area relative to the number of working-age people” (28); another is to proxy human capital by local academic R&D expenditures (28, 29). Output measures range from patents to regional GDP per capita (30) as well as invention and licensing records (31).

In general, the findings are consistent with the notion that an important source of knowledge transfer is the flows of research-experienced workers from universities to regional economies (32, 33), but do not directly measure those worker flows.
2. CONCEPTUAL FRAMEWORK

The approach that we use is to examine regional output as a function of human capital and labor. Consistent with Romer’s approach, we think of human capital (produced by universities) as an input separate from labor and physical capital. Our core model is thus

\[ Y_{cilt} = H_{cilt}^a L_{cilt}^\beta K_{cilt}^\gamma \varepsilon_{cilt} \]  

(1)

\( Y \) is produced in each county (c) and industry (i) and time period (t) by combining human capital – as proxied by university trained individuals (H) – labor (L) and physical capital (K).

We expect the level of human capital flowing from a university to an industry in each county to be affected by the distance of the county from that university, the research intensity of the county (to capture agglomeration effects) and the research intensity of the university

\[ H_{ucilt} = D_{uc}^\delta CRI_{cilt}^\theta URI_{ut}^\theta \omega_{ucilt} \]  

(2)

Here, the human capital flows from a university (u) to an industry (i) in a county (c) is driven by the distance (D) of each university from each county, the research intensity of the industry in that county (CRI) and the cumulative research intensity of the university (URI). The second model is essentially a test of the role of geographical proximity in the transmission of knowledge. Our expectation from estimating model 2 is that the closer counties are to a research-intensive university, the greater the flows of research trained workers to the county. Our expectation from estimating model 1 is both that greater levels of human capital affect economic activity in more R&D intensive industries and that there are measurable spillovers to other industries in the county.

3. DATA

The data used in this paper quantifies a specific type of human capital obtained from formal and informal scientific research training (34). While the data have not been used in the particular context of this paper, previous work suggests that research trained workers are more likely to work at firms with characteristics closely linked to productivity (35). This work was based on the UMERTICS 2017Q4a dataset for research(36). This dataset contains annual data from 26 IRIS member universities...
including coverage between 2001 and 2017 (this coverage varies by institution). The core files include university financial and personnel administrative data pertaining to sponsored project expenditures at IRIS member universities during a given year. IRIS core files are based on administrative data drawn directly from sponsored projects, procurement, and human resources data systems on each IRIS member university’s campus. Individual campus files are de-identified, cleaned and aggregated by IRIS to produce these core files. The 2018 release includes transactions from about 300,000 unique federal and non-federal awards including wage payments to 480,000 individuals. The Catalog of Federal Domestic Assistance (CFDA), which is included in each award identifier, allows us to classify projects by the funding agency.

The key file of interest in this project is the employee file. For each funded research project, both federal and nonfederal, the file contains all payroll charges for all pay periods (identified by period start date and period end date). This includes links to both the federal award id (unique award number) and the internal university identification number (recipient account number). In addition to first name, last name and date of birth, the data include the employee’s internal de-identified employee number, and the job title (which we mapped into broad occupational categories). We make use of data on 959,031 individuals employed on sponsored research projects for 21 universities.

We are able to directly link employment on sponsored research projects to W2 and LEHD employment records from the US Census Bureau(25). Placement and earnings are derived from a match of UMETRICS data to the data at the US Census Bureau. These data have been provided to the Census Bureau in order for a Protected Identification Key (PIK), Census’s internal individual identifier, to be assigned based on the employing university, the employee last name, first name, and (in some cases) date of birth. The Census Bureau’s Person Identification Validation System (PVS) is used to assign an anonymous, unique person identifier to university employees(37). UMETRICS employee name, address, and date of birth when available are parsed, standardized and geocoded during the input process for the PVS. Next, a probabilistic match is performed between the UMETRICS data and PVS reference files that are based on the Social Security Administration’s Numerical Identification File (Numident). When possible PVS assigns this person identifier, the protected identification key (PIK). Because PVS is a probabilistic match, it is possible for a UMETRICS employee to receive multiple PIK values. UMETRICS employee data is historic and spans multiple years. Thus, a custom PVS process with many years of associated reference files for each university is used. For detailed information about reference files in PVS or the matching algorithm, see (37)
Not all universities provide employee date of birth, resulting in higher rates of multiple PIKs than when date of birth is present. A filter is applied to all university employee PIKs in order to select the correct PIK from the multiple values when possible as well as to screen false one-to-one matches. W-2 data used for the filter is limited to records for the years that the university employee data spans, the EIN(s) associated with the university, and addresses within a 200 mile radius of the university campus address. A match to the W-2 data must occur for that employee to be retained in the sample. For multiple PIK values, only the PIK that appears in the W-2 data is retained for the employee. Filtered data are output to employee crosswalk data file.

Once each UMETRICS employee has been assigned a PIK, these data are matched to employment and earnings records in the LEHD. Using the UMETRICS data, we characterize individuals by their primary role in the laboratory, source of funding support, and the topical focus of their research. For those workers who transition out of academia (i.e. those who transition out of NAICS 614), we trace out the short-run employment linkages in terms of detailed industry sector, geography, and firm characteristics. Using this rich mapping between academia and industry, we are able to develop an analytical dataset that allows us to examine the impact of funding shocks to specific agencies and topical areas on the supply of highly trained workers. Specifically, we can explore whether an increase in the supply of highly trained workers to a particular location and industry sector crowds out/in existing workers. In addition, we examine whether these workers generate positive externalities benefiting local economies in terms of increased growth, sales, and business starts. The findings from our work provide new and exciting insights about how workers cross-pollinate ideas between the academy and private industry.

4. EMPIRICAL STRATEGY

We proceed by first characterizing individuals by their primary role in the laboratory, source of funding support, and the topical focus of their research. We then trace out linkages with private industry by matching these individuals to three sources of administrative records including the Decennial Census, IRS W2 wage records, and LEHD employment and earnings. Although we only have UMETRICS data for 21 universities of which we restrict our sample to 10, we are able to obtain national estimates on flows of research-trained workers from all universities by implementing a two-sample two-stage least squares procedure (TS2SLS) following a procedure proposed in Angrist and Krueger(39).
We operationalize the TS2SLS by first modeling employment outcomes for the UMETRICS sample as a function of distance, location, industry sector, time, university characteristics, fluctuations in university research funding by source, and industry-funding source interactions. Specifically, we model employment outcomes for a subset of 11 UMETRICS universities as

\[
\text{emp}_{t,c,n,a,u} = \beta_0 + \tau_t + \chi_c + \nu_n + \alpha_a + [\nu_n * \alpha_a] + \sum_{i=1}^{6} \beta_i \delta_{c,u} + \beta_7 \widehat{\text{fund}}_{t,a,u} \\
+ \beta_8 [\alpha_a * \widehat{\text{fund}}_{t,a,u}] + \beta_9 \text{enroll}_{t,u} + \beta_{10} \text{locale}_{t,u} + \beta_{11} \text{admit}_{t,u} \\
+ \beta_{12} \text{tuition}_{t,u} + \eta_{t,c,n,a,u}
\]  

(3)

where \(\text{emp}_{t,c,n,a,u}\) is the number of research-trained workers from university \(u\) with dominant funding source \(a\) who find employment in industry \(n\) and county \(c\) at time \(t\). We estimate research-trained workers as a function fixed-effects for time \(\tau_t\), county \(\chi_c\), industry \(\nu_n\), agency \(\alpha_a\), and the interaction \([\nu_n * \alpha_a]\) of industry with agency. Table 1 contains a ranking of the top industries for the major agencies in our data, most of which conform with prior expectations about where workers trained in these areas would eventually seek employment. Additional independent variables include a sixth-order polynomial of distance \(\sum_{i=1}^{6} \delta_{c,u}^i\), a shift-share measure of funding \(\widehat{\text{fund}}_{t,a,u}\), the interaction of the funding measure and agency fixed-effects \([\alpha_a * \widehat{\text{fund}}_{t,a,u}]\), university enrollment \(\text{enroll}_{t,u}\), university local population \(\text{locale}_{t,u}\), university admittance \(\text{admit}_{t,u}\), and tuition cost \(\text{tuition}_{t,u}\).

[Insert Table 1]

To capture exogenous changes in funding for scientific research, a shift share measure of funding is constructed as

\[
\widehat{\text{fund}}_{t,a,u} = \left[ \frac{\text{fund}_{2001,a,u}}{\text{fund}_{2001,a}} * \ln(\text{fund}_{t,a,u}) \right]
\]

---

4 See Equation 1 in the Technical Appendix.
where $\text{fund}_{2001,a,u}$ is total federal funding to university $u$ from source $a$ in a base period of 2001 and $\text{fund}_{2001,a}$ is the aggregate funding from source $a$ across the top 100 research universities, as ranked in terms of total funding using the 2015 NSF HERDS. This measure has a similar intuition to the so-called ethnic enclave instrument featured in prominent works by Angrist Krueger (1992, 1995) and Inoue and Solon (2010).

We apply TS2SLS where we use the coefficients from Equation 2 to predict research-trained employment outcomes for the remaining 100 non-UMETRICS universities as well as the coefficient from the first-stage in equation 2 to rescale that data. We then aggregate the predicted research-trained employment $\hat{emp}_{t,c,n}$ from the 11 university UMETRICS and Non-UMETRICS samples from equation 1 such that

$$rtw_{t,c,n} = \sum_u \sum_a \text{emp(UMETRICS)}_{t,c,n,a,u} + \sum_u \sum_a \text{emp(Non-UMETRICS)}_{t,c,n,a,u}$$

Using the coefficients obtained from modelling employment outcomes for the UMETRICS sample, we are able to estimate the number of research-trained workers by industry and location for the entire population of UMETRICS and non-UMETRICS universities listed in the top 100 ranking of research funding. These top 100 universities represent 81 percent of the total federal research funding transferred to universities in the United States. Note that trends in research funding for the UMETRICS universities and non-UMETRICS are quite similar and that the UMETRICS universities, see Figures 1. The UMETRICS universities we utilize are also fairly evenly distributed throughout the NSF’s ranking of research intensity where we note that trends among different levels of research intensity are relatively consistent, see Figure 2

[Insert Figures 1 and 2]

Having obtained national estimates of the number of research-trained workers by location and industry sector, we then examining the spillover effects resulting from plausibly exogenous supply

---

5 These population of universities in our study includes the top 100 universities listed on the 2015 edition of the NSF HERDS survey.
shocks of these workers.\textsuperscript{6} We are able to estimate a second-stage regression of local labor market productivity on a plausibly exogenous measure on changes to research-trained workers such that

\[
prod_{t,c,n} = \lambda_0 + \lambda_1 rtw_{t,c,n} + \tau_t + \chi_c + \nu_n + \varepsilon_{t,c,n}
\] (4)

where \(rtw_{t,c,n}\) is the predicted number of research-trained workers estimated from Equation 3. To capture the direct effect of an exogenous increase in the number of research-trained workers, we also include fixed-effects for time, county, and industry. The dependent variable \(prod_{t,c,n}\) captures various aspects of productivity within a county-industry cell including aggregate (1) employment, (2) payroll, (3) establishments, (4) employment per firm, (5) payroll per worker, (6) payroll per firm, and (7) revenue.

RESULTS

We have been delayed in getting out full quantitative results out of the RDC. Qualitative results approved for disclosure are contained below.

Qualitative Results Summary for 1\textsuperscript{st} Stage:

Our model of employment transitions finds a statistically significant positive relationship between changes to federal funding (both raw and as a shift-share) and research-trained employment.

Qualitative Results Summary for 2\textsuperscript{nd} Stage:

We find a statistically positive relationship between predicted research-trained employment and an IHST of number of establishments, employees, payroll, employees per establishment, payroll per establishment, and payroll per employee.

\begin{itemize}
  \item Robust to constructing the predicted number of research trained workers using raw vs. shift-share in federal research funding
  \item Robust to limiting UMETRICS data to all vs. the sub-sample of undergraduate, graduate, and post-docs.
  \item Robust to clustering on county and 4-digit NAICS industries
  \item Robust to 2 and 3-year lag structures on our measure of research-trained employment
\end{itemize}

\textsuperscript{6} We obtain these plausibly exogenous supply shocks using a shift-share style measure of funding. See the Technical Appendix for details.
CONCLUSIONS

[Our insightful conclusions will eventually go here!]
REFERENCES

2. A. Saxenian, Regional advantage (Harvard University Press, 1996).
doi:10.3386/w22512.


REFERENCES

Figure 1: Trends in Research Funding UMERICRS and Non-UMERICRS Samples
Figure 2: Trends in Research Funding by NSF HERDS Ranking
Table 1: Industry-Funding Source Linkages from Combined UMERTICS, LEHD, and W2 Data

<table>
<thead>
<tr>
<th>Rank</th>
<th>NAICS</th>
<th>Industry</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HHS/NIH</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6221</td>
<td>Medical and Surgical Hospitals</td>
</tr>
<tr>
<td>2</td>
<td>5417</td>
<td>Scientific Research and Development</td>
</tr>
<tr>
<td>3</td>
<td>5613</td>
<td>Employment Services</td>
</tr>
<tr>
<td>4</td>
<td>7221</td>
<td>Food Services</td>
</tr>
<tr>
<td>5</td>
<td>6211</td>
<td>Officers of Physicians</td>
</tr>
<tr>
<td><strong>DOD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5417</td>
<td>Scientific Research and Development</td>
</tr>
<tr>
<td>2</td>
<td>5415</td>
<td>Computer Systems Design</td>
</tr>
<tr>
<td>3</td>
<td>5413</td>
<td>Architectural and Engineering</td>
</tr>
<tr>
<td>4</td>
<td>3344</td>
<td>Semiconductor and Electronic Manufacturing</td>
</tr>
<tr>
<td>5</td>
<td>5613</td>
<td>Employment Services</td>
</tr>
<tr>
<td><strong>NSF</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5417</td>
<td>Scientific Research and Development</td>
</tr>
<tr>
<td>2</td>
<td>5415</td>
<td>Travel Arrangement Services</td>
</tr>
<tr>
<td>3</td>
<td>5413</td>
<td>Architectural and Engineering</td>
</tr>
<tr>
<td>4</td>
<td>5613</td>
<td>Employment Services</td>
</tr>
<tr>
<td>5</td>
<td>6221</td>
<td>Medical and Surgical Hospitals</td>
</tr>
<tr>
<td><strong>All Non-Fed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6221</td>
<td>Medical and Surgical Hospitals</td>
</tr>
<tr>
<td>2</td>
<td>5613</td>
<td>Employment Services</td>
</tr>
<tr>
<td>3</td>
<td>5417</td>
<td>Scientific Research and Development</td>
</tr>
<tr>
<td>4</td>
<td>5413</td>
<td>Architectural and Engineering</td>
</tr>
<tr>
<td>5</td>
<td>7221</td>
<td>Food Services</td>
</tr>
</tbody>
</table>