A Tale of Two Markets? Knowledge, Skill, and STEM Career Outcomes

Xuan Jiang, Joseph Staudt, Bruce A. Weinberg *

February 23, 2022

Abstract

Is the United States producing too many or too few Ph.D. recipients? The answer to this question hinges crucially on whether demand for Ph.D. recipients is a function merely of their high levels of general skill or also depends on their field-specific knowledge gained during graduate training. Indeed, there is growing concern that, outside of increasingly long and poorly-compensated post-doc positions, many Ph.D. recipients are placing at jobs that do not utilize their specialized knowledge accumulated during graduate training and are instead placing in jobs that mostly value their high levels of general skill. This paper takes a first step toward untangling the determinants of labor demand by using novel data to analyze the labor market for recently minted Ph.D. recipients. We combine administrative data on people employed on research projects as graduate students from universities (UMETRICS), confidential, individual-level, career information from the U.S. Census Bureau (including the universe of W2 tax records), and a newly-developed measure of the relevance of a Ph.D. recipient's dissertation-specific knowledge for different types of industries. We construct this dissertation-industry relevance measure by comparing the text of a Ph.D. recipient's dissertation to the text of the universe of USPTO patents, for which we observe the assignees. From the similarity of dissertations and the patents assigned to an industry, we generate an individual-specific measure of the relevance of their knowledge (as represented in their dissertation) to innovation (i.e. patenting) in every industry. We first validate our new dissertation-industry relevance measure by relating it to real labor market outcomes for Ph.D. recipients, such as placement and earnings. We then use this dissertation-industry relevance score to analyze the labor market for Ph.D. recipients from different fields and placed in different industries. Finally, We bring O*NET's skill indices to shed light on how much the employment is valued by skills but not knowledge.

JEL Classification: F16, I20, J20, O15
Keywords: STEM Ph.D.,knowledge, Industry Placement, ProQuest, USPTO

*Xuan Jiang: Department of Economics, The Ohio State University, jiang.445@osu.edu; Joseph Staudt, U.S. Census Bureau Bruce A. Weinberg, Ohio State University, IZA, and NBER

"I know we all want the U.S. to continue to be the world's center for innovation. But our position is at risk. There are many reasons for this but two stand out. First, U.S. companies face a severe shortfall of scientists and engineers with expertise to develop the next generation of breakthroughs.... I believe the country is at a crossroads. For decades innovation has been our engine or prosperity. Now economic progress depends more than ever on innovation. Without leadership from Congress and the President to implement policies like those I've outlined today and the commitment of the private sector to do its part, the center of progress can shift to other nations that are more committed to the pursuit of innovation."—Bill Gates, Testimony to Congress, 2008.[1]

"I think in all good conscience, PhD programmes now know that they're taking on more people than have a high probability of getting research positions...we just can't keep on growing these PhD programmes without good outlets... I think if you train a lot of people who don't end up in research positions, the PhD, as I've understood it and as we've always discussed it, is to train people in research, so that's how I think it should be assessed."—Paula Stephan, *Nature*, 2019.[2]

1 Introduction

Is the United States producing too many or too few research doctorates? On the one hand are policymakers, institutions and practitioners who see the STEM labor market as experiencing the worst of times. They (rightly) note that there has been a sustained increase in the share of Ph.D. recipients placing in increasingly long and poorly-compensated post-doc positions, while an increasing number appear to be pushed by excess supply into a wide range of jobs that do not utilize the specialized knowledge accumulated during graduate training (Cyranoski et al., 2011; Stephan, 2012; Alberts et al., 2015; Gould, 2015). This issue is especially acute in fields such as biology, where the fraction of Ph.D. recipients taking post-doc positions has increased substantially (Powell, 2015).

[1] https://news.microsoft.com/2008/03/12/bill-gates-testimony-before-the-committee-on-science-and-technology-u-s-house-of-representatives/

[2] https://www.nature.com/articles/d41586-019-03439-x

On the other side are those who see the STEM market as the experiencing the best of times. They emphasize that research and innovation are crucial inputs for long-term economic growth (Romer, 1990) and (rightly) note that Ph.D. recipients have low rates of unemployment after receiving their degree. Far from viewing the wide range of jobs in which STEM PhDs place as evidence of over supply, they view it as evidence of demand pull for STEM knowledge across a myriad of sectors of the economy, with STEM being a new, multi-purpose training. This camp views these labor market facts as illustrating a need to, if anything, train more researchers. They view the increasing incidence and length of post-doc positions as a response to the increasing complexity of science (Jones, 2009).

Thus, advocates of these polar positions point to a similar set of facts, but provide starkly different interpretations for them. This paper seeks to apply economic logic and unique, new data to help resolve this long-standing controversy.

At a conceptual level, we argue that the widespread placement of Ph.D. recipients outside of academia is indeed consistent with both the demand-pull, "best of times" view and with the supply-push, "worst of times" view of the market. To distinguish these hypotheses one therefore requires additional information, with wages and other measures of job quality being obvious candidates. Intuitively, if supply is pushing researchers out of academia and into industry, one should see weak labor market outcomes and low wages. In contrast, if demand is pulling researchers into industry then one ought to see strong labor market outcomes and high wages.

While low initial wages may indicate low demand, they can also reflect heavy early investments in human capital. Here too, economic theory points to clear testable hypotheses – if wages are low because human capital investment is important early in the career, then wages should grow rapidly and be higher later on, but if supply exceeds demand then wages should continue to be low even as researchers'

careers progress.

Obtaining a Ph.D. requires a substantial investment in human capital, so another way to assess the quality of placements is to analyze whether research doctorates are making use of the specialized knowledge and skills that they possess in their jobs. We see this analysis as informative for two reasons. First, If research doctorates are taking jobs that do not require their specialized skill or knowledge, then it seems that (aside from signaling value), they might have obtained similar positions with smaller and/or different investments. Second, there is evidence that research trained individuals view the ability to conduct research as a valuable job attribute.

One of our contributions is to construct person-specific measures of the extent to which researchers are flowing into jobs that are likely to use the specialized scientific knowledge that they obtained in graduate school. Specifically, we combine administrative data from universities, confidential individual-level career information from the U.S. Census Bureau (including the universe of W2 tax records), and a newly developed measure of the relevance of a Ph.D. recipient's dissertation-specific knowledge for different types of industries. We construct this new dissertation-industry relevance measure by comparing the text of a Ph.D. recipient's dissertation to the text of the universe of USPTO patents. Then, using the subset of patents assigned to U.S. firms, we are able to measure, for each individual Ph.D. recipient, the relevance of their field-specific knowledge (as encoded in their dissertation) to particular firms or to the innovative (i.e. patenting) portion of every industry.

Our first set of results will examine the relationship between our dissertation-industry relevance measure and real labor market outcomes for Ph.D. recipients. Specifically, we will estimate whether a Ph.D. recipient is more likely to become employed in an industry for which their knowledge is highly relevant. We then examine whether highly relevant employment matches are associated with higher

wages.

Our next set of results will use our dissertation-industry relevance score to analyze the labor market for Ph.D. recipients in different fields of study (e.g., physics, biology, etc.). First, we will compute the fraction of Ph.D. recipients in each field that obtain jobs as faculty, post-docs, and in industry. As suggested, large shares of placements as faculty in academia are seen as an indication of strong academic demand. Large shares of placements in postdocs may indicate strong demand but that graduates require more training or weak demand, which we assess using wage growth. We seek to distinguish high demand for scientific knowledge from high demand for skills from low demand by looking at wages and research match. If a high fraction of Ph.D. recipients within a field go into a highly relevant industry, then we interpret this as evidence that the individuals in this field are valued for their specific knowledge. Conversely, if Ph.D. recipients within a field tend to go into industries with low relevance while still getting relatively high pay, we interpret this as evidence that, while firms may value the high level of skills these people bring to the table, they do not value their field-specific knowledge per se. If Ph.D. recipients go into industries with low relevance and receive low pay, we interpret this as firms valuing neither their general skill nor their field-specific knowledge.

2 Data and Measures

This paper develops and deploys unique, emerging data to construct new metrics that represent a first step toward quantifying the knowledge of STEM PhDs. We employ (1) new administrative data from the UMETRICS project, which provides unprecedented detail on all people employed on research projects at participating universities; (2) the full set of doctoral dissertations in ProQuest's Dissertation

and Theses Database (2005–2015) for participating universities, including titles and abstracts; and (3) W-2 records on covered workers that are housed at the US Census Bureau and that provide earnings information and link workers to their employers. These data, in turn, have been linked by the Census Bureau to the Decennial Census and to the wealth of data it houses on business, including the US patents assigned to each business. (4) The text similarity measure constructed for ProQuest dissertation and patents in USPTO Database (2000–2017). (5) Firm information from Business Register (BR).

2.1 Summary Statistics on the Ph.D. and University Sample

The UMETRICS data are administrative data on all transactions at 30 major research universities, including public and private universities, which collectively account for more than one-third of federally funded academic R&D.[3] While these institutions are a convenience sample, they provide a valuable window into the academic research enterprise, especially at research-intensive institutions. Data on the full teams of researchers (not just faculty or authors on articles) supported by individual research grants are captured in the UMETRICS data (Lane et al., 2013, 2015). This coverage is reliable because the data are drawn directly from payroll records. The data also include data on job titles (i.e. Faculty, Graduate Student, Research Assistant, etc.) allowing us to identify graduate students and postdocs (and even undergraduates) (Ikudo et al., 2019). In this study, we focus on graduate students who obtained Ph.D. degrees from theses UMETRICS universities. The UMETRICS data have been used successfully on numerous research projects

[3]Universities included in UMETRICS are University of Michigan, The Ohio State University, ...

(Chang et al., 2019; Ikudo et al., 2019; Buffington et al., 2016; Lane et al., 2013, 2015; Weinberg et al., 2014; Zolas et al., 2015). These data have already been accessed by over 110 researchers from 30 institutions. We focus on the 12 STEM fields in our analysis because 1) they are the fields that received the most federal funding so that they are the largest fields; 2) our knowledge to industry relevance measures is based on innovation intensity encoded in US patents. Table 1 summarizes the individual characteristics, dissertation characteristics, and employment outcome for the analysis sample.

Table 1: PIK/Dissertation Summary Statistics

	Industry		Assistant Prof.		Post-Doc	
	Mean	SD	Mean	SD	Mean	SD
Individual Characteristics						
Female	0.5322	0.4992	.4945	.5009	0.5426	0.4984
White						
Asian						
Hispanic						
Birth Year						
Dissertation Characteristics						
Degree year						
Age at degree year						
Fraction with 0 matched patents						
Number of patents matched						
Mean (of mean) cosine similarity						
Mean (of max) cosine similarity						
Employment Outcomes						
Mean starting wage						
Mean arcsine starting wage						
Mean 3-year wage growth						
PIKs (rounded)	1300	0	300	0	1000	0

Table 2 shows the by-field share of placement (in faculty, postdocs, and industry), starting wage, and year 3 wage since degree.

Table 2: Placement Shares and Wages

	Faculty	Postdocs	Industry	Observations	Faculty	Postdocs	Industry
Placement					**Mean Cosine**		
Agriculture	xx	xx	xx	xxxx	xx	xx	xx
Biology	xx	xx	xx	xxxx	xx	xx	xx
Chemistry	xx	xx	xx	xxxx	xx	xx	xx
Information sciences	xx	xx	xx	xxxx	xx	xx	xx
Ecological sciences?	xx	xx	xx	xxxx	xx	xx	xx
Engineering	xx	xx	xx	xxxx	xx	xx	xx
Environmental Sciences	xx	xx	xx	xxxx	xx	xx	xx
Geographic sciences	xx	xx	xx	xxxx	xx	xx	xx
Medical Sciences	xx	xx	xx	xxxx	xx	xx	xx
Math	xx	xx	xx	xxxx	xx	xx	xx
Physics	xx	xx	xx	xxxx	xx	xx	xx
Social sciences	xx	xx	xx	xxxx	xx	xx	xx
Starting Wage					**Year 3 Wage**		
Agriculture	xx	xx	xx	xxxx	xx	xx	xx
Biology	xx	xx	xx	xxxx	xx	xx	xx
Chemistry	xx	xx	xx	xxxx	xx	xx	xx
Information sciences	xx	xx	xx	xxxx	xx	xx	xx
Ecological sciences?	xx	xx	xx	xxxx	xx	xx	xx
Engineering	xx	xx	xx	xxxx	xx	xx	xx
Environmental Sciences	xx	xx	xx	xxxx	xx	xx	xx
Geographic sciences	xx	xx	xx	xxxx	xx	xx	xx
Medical Sciences	xx	xx	xx	xxxx	xx	xx	xx
Math	xx	xx	xx	xxxx	xx	xx	xx
Physics	xx	xx	xx	xxxx	xx	xx	xx
Social sciences	xx	xx	xx	xxxx	xx	xx	xx

Notes:

2.2 Measure of Relevance between Ph.D.'s Specific Knowledge and Industry

We develop a measure of relevance between each Ph.D. recipient's field-specific knowledge and the possible industries and companies for which individual researcher's knowledge is valuable. Because of the importance of individually differentiated, advanced research knowledge possessed by Ph.D. recipients, we utilize the content of Ph.D.'s dissertation to represent their novel and field-specific knowledge. We utilize the content of US patents possessed by US companies within industries to represent the frontier of knowledge valued and applied by each industry.

Specifically, we build a similarity index between the text of individual PhD recipient's dissertation abstracts to the abstract of patents issued by the USPTO, which have in turn been matched to their assignees. This index is a linguistic distance metric that allows to evaluate how closely related one dissertation and one patent are, based on their subject matter. This similarity index, cosine similarity, is a basic measure (ranges [0, 1]) of linguistic proximity between documents that commonly used in computer sciences; where 0 means the two abstracts have no similarity while 1 means they are identical. This measure has been getting increasing attention in social science studies, especially economics of innovation (Koffi, 2021; Biasi and Ma, 2020). For each ProQuest dissertation abstract, we compute its cosine similarities to all patents and keep the top 1,000 cosine similarities associated with the most closely related U.S. patent abstracts (i.e. the 1000 patents with the highest non-zero similarity scores).[4]

[4]There are some dissertations match to less than 1000 patents due to zero cosine similarity scores.

2.2.1 Cosine Similarity

Cosine similarity is the basic measure of linguistic proximity between documents. It is computed as

$$cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are components of vector representations of documents A (dissertation) and B (patent). Theoretically, cosine similarity ranges from -1 (meaning exactly opposite) to 1 (meaning exactly the same). In our case, there is no direction of meaning, thus the cosine similarity in our data ranges from 0 to 1. The values in between show the degree of similarity or dissimilarity. Table B1 provides examples of text of dissertation, text of patent, and the corresponding cosine similarity scores.

We use U.S. patent data from 1976 to 2017, which is accessible through USPTO PatentsView project (www.patentsview.org), to construct the measure. And we have use all ProQuest dissertation from 2000 to 2017, including the dissertations of PhDs from non-UMETRICS University (for comparison if necessary) (WE MAY WANT TO GENERATE SOME SUM STATS TO COMPARE UMETRICS UNIV. VS. NON-UMETRICS UNIV).

For each ProQuest dissertation abstract, similarity cosine was produced for the 1,000 most closely related U.S. patent abstracts. Table below shows three examples of the ProQuest dissertation abstract and US patent abstract pairs and the corresponding cosine similarity score. We particularly choose a high, a middle, and a low cosine similarity score for comparison purpose. Looking at the content of the abstracts, we can tell that the similarity measure we generate is meaningful and convincing.

Table 2 includes the distribution of cosine similarity index by placement sector

and how it varies across research fields. Overall, cosine similarity varies from xxx to xxx. The mean is XXX and standard deviation is xxx. This suggests that, overall, PhD dissertations are closely related to patents. Among the 12 research fields, XXX field has the highest mean cosine similarity, indicating the largest relevance of field-specific knowledge to industries. In other words, Ph.D. recipients from XXX field have largest opportunity to find industry jobs that value their field-specific knowledge.

Among three categories of placement, faculty positions, postdoctoral researchers, and industry positions, we find that xxx has the largest mean cosine similarity, suggesting that ...

2.3 Match Quality between Ph.D. Recipients and Industry

In this section, we elaborate how we bridge Ph.D.'s knowledge to each industry and generate the match quality of the Ph.D.'s actual job placement. Figure 1 shows the data linking process. Basically, we identify graduate students who had been funded through federal fundings and obtained Ph.D. degrees from the 30 UMETRICS universities, then link them to their Ph.D. dissertation in the ProQuest database. Each Ph.D. dissertation is matched to (up to) 1000 US patents based on their cosine similarities, as noted above; and each US patent is associated with an assignee (i.e. a firm), which is identified by a firm ID. Further, the Business Register (BR) within Census Bureau provide detailed information on each firm, including the employment, payroll, and location of its each establishment.[5] Each establishment is classified into one industry, associated with a North American Industry Classification System (NAICS) code. To fix mind, at this point, we can consider that one Ph.D. recipient is matched to up to 1000 US patents, each possessed by

[5]A firm might have multiple establishments. In our sample, the number of establishment of each firm varies from 1 to xxx.

one firm that is dealing products or services in a series of industries. For example, hypothetically, a computer science Ph.D., whose dissertation is on face recognition, is matched to 1000 patents, most of which are related to face recognition technology. The best matched patent is possessed by Apple Inc., which associates to two NAICS code: 334220 (NAICS Title: Radio and Television Broadcasting and Wireless Communications Equipment Manufacturing) and 334111 (Electronic Computer Manufacturing). Similarly, the rest of 999 patents are each possessed by a company associates with multiple NAICS codes. Due to the relevance in products and services of these companies, they might have many overlaps in NAICS codes. Eventually, we will observe that this Ph.D. recipient is matched to multiple NAICS codes. Figure 1 illustrates the data structure and linkages.

Table 3 summarizes the relationship between cosine similarity and placement outcome by NAICS code. Specifically, columns 2 and 3 show the unconditional probability of a PIK/Dissertation getting a job in the n-digit NAICS code that corresponds to/suggested by the maximum cosine score. It is equivalent to first computing the fraction of PIKs that place in a particular industry and then taking the mean of those fractions over industries. Columns 4-7 show the means and standard deviations of the mean and max cosine similarity, over patents, between a dissertation and NAICS code of firm to which the patent was assigned. Columns 8-11 show the means and standard deviations of the mean and max cosine similarity score conditional on the PIK having actually placed in the NAICS code.

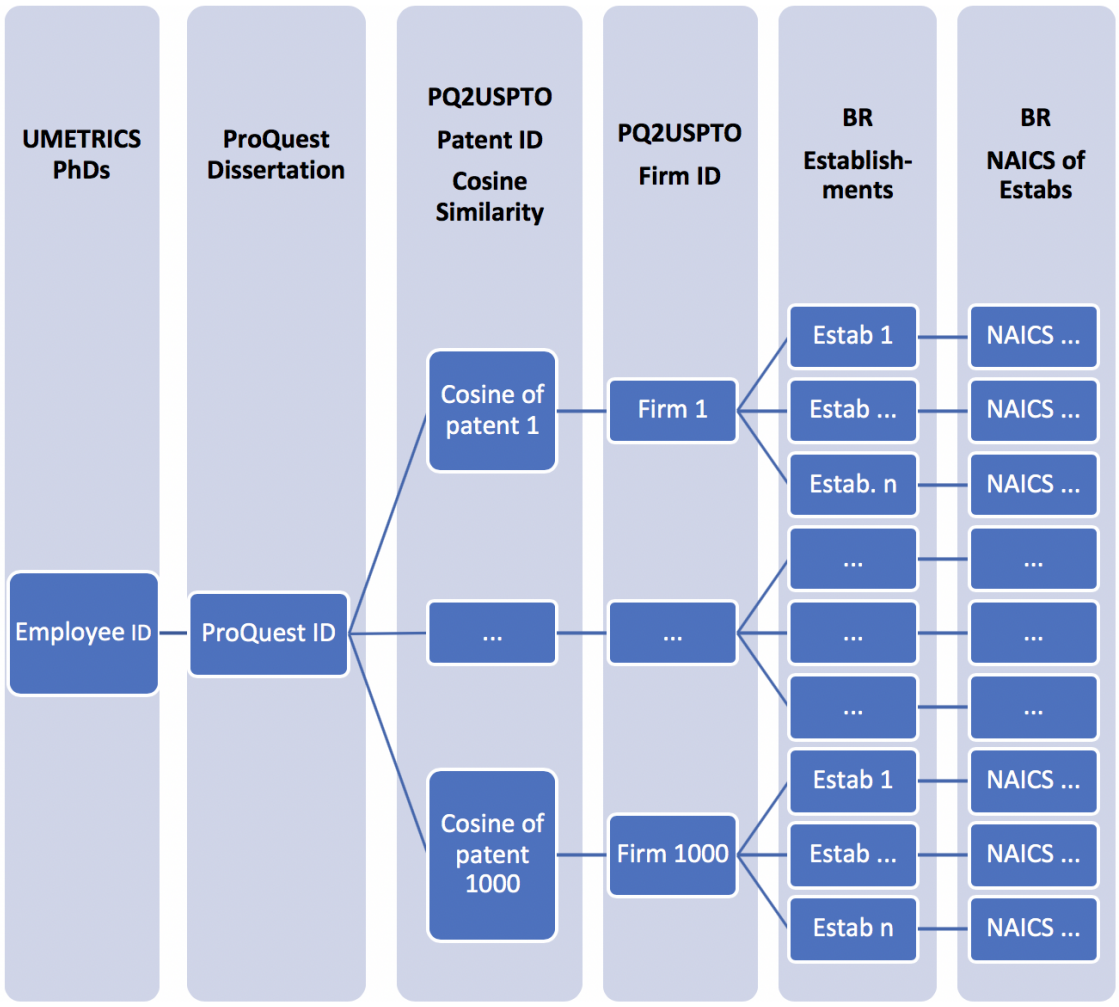


Figure 1: Illustration of Data Linkages

Table 3: Dyad Summary Statistics: PIK/Dissertation Job Placements

			Unconditional				Conditional on Placement			
	Placement for Max Cosine		Mean Cosine		Max Cosine		Mean Cosine		Max Cosine	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2-Digit NAICS	0.5322	0.4992	.4945	.5009						
3-Digit NAICS	0.5322	0.4992	.4945	.5009						
4-Digit NAICS	0.5322	0.4992	.4945	.5009						
5-Digit NAICS	0.5322	0.4992	.4945	.5009						
6-Digit NAICS	0.5322	0.4992	.4945	.5009						
PIKs (rounded)	1300	0	300	0						

Notes – Placement is the unconditional probability of a PIK/Dissertation getting a job in a given n-digit NAICS code. It is equivalent to first computing the fraction of PIKs that place in a particular industry and then taking the mean of those fractions over industries. Cosine Mean (Max) is the average (maximum) similarity, over patents, between a dissertation and NAICS code of firm to which the patent was assigned. Cosine Mean (Max) is this measure conditional on the PIK having actually placed in the NAICS code. Indicator for whether the PIK/Dissertation placed in their most similar NAICS code.

Next, we generate the individual-industry relevance by aggregating the cosine similarity scores from firm level to industry level. Generally speaking, each firm has one or multiple establishment(s), each corresponding to an industry (identified by a unique NAICS code). based on different establishment weights. We explore three ways of weighting establishment: equal weight, employment weight, and payroll weight.

We construct a measure of individual-industry relevance in the following way:

$$TotalWeightedCosine_{in} = \sum_{j,e} cosine_{ij} * weight_{je}$$

i: individual; j: patent/patented firm; e: establishment in a firm; n: NAICS code associated with each establishment. We explore 3 types of $weight_{je}$:

1) Weighed each establishment equally: $weight^1_{je} = \frac{1}{\#e}$

2) Weighed by employment: $weight^2_{je} = \frac{emp_{je}}{\sum_e emp_{je}}$

3) Weighed by annual payment: $weight^3_{je} = \frac{ap_{je}}{\sum_e ap_{je}}$

Based on the industry (NAICS code) of actual placement and the ordering of the most relevant industries, we create 3 metrics of "match quality": 1) Match Dummy: is the placement in a relevant industry (i.e. NAICS)? 2) The total weighted cosine of the placement NAICS 3) The percentile of the placement industry in the total weighted cosine distribution.

3 Results

3.1 Relevance Score and Labor Market Outcomes

In this section, we examine the relationships between our newly developed dissertation-industry relevance score and the early-career labor market outcomes of Ph.D. recip-

ients. We first examine whether the relevance between a dissertation-industry pair predicts whether a Ph.D. recipient actually places in that industry. To do this, we construct a data set of all pairwise combinations of of Ph.D. recipients/dissertations and industries. Then for each combination, we compute the relevance score and construct an indicator for whether the Ph.D. recipient placed in that industry.[6] We then regress the placement indicator on the relevance score.

Table 4 shows these regression results at the 2-digit to 6-digit NAICS (industry levels). Reassuringly, a higher relevance score between a Ph.D. recipient/dissertation and an industry is strongly related to whether the Ph.D. recipient will place in that industry. Specifically, a one standard deviation increase in the relevance score increases the probability of placement by XX percentage points, which is XX% of the mean. Unsurprisingly, the relationship is stronger for 6/4 digit NAICS...

Next, we restrict our attention to the Ph.D. recipient/dissertation and industry combinations in which the Ph.D. recipient actually placed. That is, the values for which the placement indicator equals 1. We then examine the relationship between the relevance score for a given Ph.D. recipient's actual job placement and their wages, both levels and growth. Table 5 shows these estimates. Each column is a regression specification while each row is a different specification cosine similarity score based on a NAICS level. Each cell is a separate regression. Column1 include degree year fixed effects, NAICS fixed effects, and field fixed effects. Column 2 adds in a control variable indicating the number of patents associated with each PIK/dissertation. Column 3 include an interaction term of degree year, NAICS,

[6]How to handle multiple jobs. Should everyone belong to one and only one industry? Can some people belong to no industries? Can some people belong to more than one industry?

Table 4: Dyad Regressions: Placement Outcome

2-Digit NAICS					
Average Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Maximum Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
3-Digit NAICS					
Average Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Maximum Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4-Digit NAICS					
Average Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Maximum Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
5-Digit NAICS					
Average Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Maximum Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
6-Digit NAICS					
Average Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Maximum Cosine	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PIK Count					
Degree Year FEs	x	x			
NAICS FEs	x	x			
Field FEs	x	x			
Degree Year x NAICS x Field FEs			x	x	
Patent Controls		x		x	
Degree Year x NAICS x Field x Patent Controls					x

Notes – Unit of observation is a PIK-NAICS. Outcome=1 if the PIK placed in the industry (i.e. received positive earnings from a firm in that industry) and 0 otherwise. Placement must take place within three years of receiving degree. Each cell is a separate regression. May make into a graph and put the table in the appendix. LAST COLUMN DEPENDS ON RESULTS AND NEED TO CUT STATS.

and field fixed effects. Column 4 includes a control for the number of patents, in addition to column 3. Finally, column 5 presents a most restricted specification, including one interaction term of degree year, NAICS, and field fixed effects, and patent control.

3.2 Industry Placement Match Quality by Field

Table 6 shows the most commonly placed industry and the industry-level mean wage. We do so by separating PIKs into four categories depending on two dimensions: 1) match quality: well-matched (above average cosine similarity) and bad-matched (below average cosine similarity). 2) wage level: among each match quality category, separating PIKs by high paid (above average wage) and low paid (below average wage). This table provides direct evidence of industries that value knowledge, industries that value skills, and the rest.

Table 5: Dyad Regressions: Arcine Wage Outcome

2-Digit NAICS						
Average Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Maximum Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
3-Digit NAICS						
Average Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
Maximum Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
4-Digit NAICS						
Average Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
Maximum Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
5-Digit NAICS						
Average Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
Maximum Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
6-Digit NAICS						
Average Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
Maximum Cosine	0.48	0.48	0.48	0.48	0.48	0.48
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	
PIK Count						
Field FE	x	x	x			
Degree Year FEs	x	x	x			
NAICS FEs		x	x			
Patent Controls			x		x	
Degree Year x NAICS x Field FEs				x	x	

Notes – Unit of observation is a PIK-NAICS. However, the sample is restricted to dyads where the PIK actually placed in the industry. WE DISCUSSED TAKING THIS TO THE PIK LEVEL... Each cell is a separate regression.

Table 6: Common Industries Placed by Pay-Match Quality

Industry	Mean Cosine	Share	Mean wage	Industry	Mean Cosine	Share	Mean wage
Well-matched, high paid				**Well-matched, low paid**			
Investment Banks	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Portfolio Management	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Internet Publisher	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Portfolio Management	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Internet Publisher	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Bad-matched, high paid				**Bad-matched, low paid**			
Investment Banks	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Portfolio Management	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Internet Publisher	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Portfolio Management	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx
Internet Publisher	xxxx	xx%	xxxx	Investment Banks	xxxx	xx%	xxxx

Notes – This table shows the top 5 industries that placed the largest share of Ph.D.s conditional on the match quality and wage.

3.3 O*NET Evidence on Skills

Based on evidence from Table 6, Ph.D. recipients who are well paid while badly matched to their employer's industry in terms of knowledge relevance must possess skills that are valued by the employers. In order to measure skills valued by industries, we borrow two occupation-level measures from O*NET, "knowledge"[7] and "skill".[8] To begin, we use O*NET's "knowledge" measure to cross-validate our relevance measure. Once validated, we turn to the occupation-level skill measures by O*NET to analyze the by-field and by-industry of placement skill levels, and estimate of the skill premium for those who have lower industry relevance placements. We are particularly interested in how much skill can explain the placement of Ph.D.s who are poorly-matched in term of knowledge but highly paid.

O*NET knowledge measure is originally on knowledge-occupation level. The categories of knowledge provided by O*NET is very close to research disciplines. We crosswalk our 12 fields to the 11 O*NET knowledge categories categories of knowledge that are most relevant for our analysis[9]. We can view this knowledge categories as a rough one-on-one mapping to our research field and assume that phd recipients from a certain field should have the highest level of the corresponding knowledge. For example, biology phd should hold the highest level of biology knowledge than phd recipients from other fields. We then construct a knowledge-industry level measure by aggregating the knowledge-occupation measure using the occupation-industry shares in the ACS. After merging the knowledge measure via industry code to our main dataset, we show the consistency of our industry relevance measure (individual-industry level) with the O*NET knowledge (industry-level by each knowledge) measure by plotting the knowledge value by each field. Figure 2

[7] https://www.onetcenter.org/dictionary/26.2/excel/knowledge.html
[8] https://www.onetcenter.org/dictionary/26.2/excel/skills.html
[9] Administration and management, biology, chemistry,...

below shows the density of each knowledge by each field.

[Figure 2: heatplot for knowledge*field]

Notes: this figure shows the density of knowledge levels by field

We bring the O*NET skill measures into our dataset in the same fashion. There are 6 categories of skills in O*NET: Basic Skills, Complex Problem Solving Skills, Resource Management Skills, Social Skills, Systems Skills, and Technical Skills.[10] We use the four categories of skills that are most relevant to our analysis: Complex Problem Solving Skills, Resource Management Skills, Systems Skills, and Technical Skills. Figure 3 shows the heatplot of skill level by skill by field.

[Figure 3: heatplot for skill*field]

We then show how much each skill is valued by industry by regressing W2 wage on the skill measures of the PhD recipients whose dissertation are poorly matched. To further illustrate the industries that value the specific knowledge versus the general skills, we document the top 10 industries (by headcount or by pay) for the highly-matched phd recipients, the poorly-matched and well-paid ones, and the poorly-matched and poorly-paid ones, respectively.

To have a straightforward view of the occupational required skill possessed by PhD from each field, we present the mean skill level by field in Table 7.

[10]Basic Skills (10 elements) — Developed capacities that facilitate learning or the more rapid acquisition of knowledge. Complex Problem Solving Skills (1 element) — Developed capacities used to solve novel, ill-defined problems in complex, real-world settings. Resource Management Skills (4 elements) — Developed capacities used to allocate resources efficiently Social Skills (6 elements) — Developed capacities used to work with people to achieve goals. Systems Skills (3 elements) — Developed capacities used to understand, monitor, and improve socio-technical systems. Technical Skills (11 elements) — Developed capacities used to design, set-up, operate, and correct malfunctions involving application of machines or technological systems.

Table 7: Mean skill level by field

	Highly matched				Poorly matched			
Field	Complex	Management	System	Technical	Complex	Management	System	Technical
Agriculture	xxxx	xxxx	xxxx	xxxx				
Biology	xxxx	xxxx	xxxx	xxxx				
Chemistry	xxxx	xxxx	xxxx	xxxx				
Ecological sciences	xxxx	xxxx	xxxx	xxxx				
Engineering	xxxx	xxxx	xxxx	xxxx				
Geographic sciences	xxxx	xxxx	xxxx	xxxx				
Information sciences	xxxx	xxxx	xxxx	xxxx				
Math	xxxx	xxxx	xxxx	xxxx				
Physics	xxxx	xxxx	xxxx	xxxx				

Notes – This table shows the average of the four skill indices (complex skill, management skill, system skill, and technical skill) by match quality and by field.

For example, Field X has a lot of postdocs - Do they have high wage growth (relative to low postdoc fields, relative to industry, relative to faculty)? – move to point ii), Economics (Lots of faculty –we're done), Engineering (Lots of industry – move to point iii). In contrast, poorly-matched Ph.D.s from field Y, have placement in finance industry: high wage and high wage growth. The relatively lower match quality indicates that finance industry values math and field Y phd's general skills. And we do not see this type of high wage & low match quality placement in other fields.

3.4 Federal Funding Provided to Fields

We plan to do an input and output analysis to address that why some certain PhD fields receive as much as (or even higher) research funding but produce PhD recipients with less demand. STUDENTS MIGHT NOT BE FULLY INFORMED ABOUT THE JOB MARKET WHEN THEY ARE MAKING GRADUATE PROGRAM CHOICE.

4 Conclusion

References

Alberts, B., Kirschner, M. W., Tilghman, S., and Varmus, H. (2015). Opinion: Addressing systemic problems in the biomedical research enterprise. *Proceedings of the National Academy of Sciences*, 112(7):1912–1913.

Biasi, B. and Ma, S. (2020). The education-innovation gap. Technical report, Working paper.

Buffington, C., Cerf, B., Jones, C., and Weinberg, B. A. (2016). Stem training and early career outcomes of female and male graduate students: Evidence from umetrics data linked to the 2010 census. *American Economic Review*, 106(5):333–38.

Chang, W.-Y., Cheng, W., Lane, J., and Weinberg, B. (2019). Federal funding of doctoral recipients: What can be learned from linked data. *Research Policy*, 48(6):1487–1492.

Cyranoski, D., Gilbert, N., Ledford, H., Nayar, A., and Yahia, M. (2011). Education: the phd factory.

Gould, J. (2015). How to build a better phd. *Nature News*, 528(7580):22.

Ikudo, A., Lane, J. I., Staudt, J., and Weinberg, B. A. (2019). Occupational classifications: A machine learning approach. *Journal of Economic and Social Measurement*, 44(2-3):57–87.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Koffi, M. (2021). Innovative ideas and gender inequality. Technical report, Working Paper Series.

Lane, J., King, J., and Schwarz, L. (2013). The creation of new administrative data. *Available at SSRN 2213916*.

Lane, J. I., Owen-Smith, J., Rosen, R. F., and Weinberg, B. A. (2015). New linked data on research investments: Scientific workforce, productivity, and public value. *Research policy*, 44(9):1659–1671.

Powell, K. (2015). The future of the postdoc. *Nature*, 520(7546):144–148.

Romer, P. M. (1990). Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102.

Stephan, P. E. (2012). *How economics shapes science*, volume 1. Harvard University Press Cambridge, MA.

Weinberg, B. A., Owen-Smith, J., Rosen, R. F., Schwarz, L., Allen, B. M., Weiss, R. E., and Lane, J. (2014). Science funding and short-term economic activity. *Science*, 344(6179):41–43.

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F., Allen, B. M., Weinberg, B. A., and Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for ph. d. recipients. *Science*, 350(6266):1367–1371.

A Contribution

B Additional Tables and Figures

C Data Appendix

We use a novel dataset, which is constructed by linking a variety of high quality data sources, including UMETRICS, ProQuest's Dissertation and Theses Database, PubMed, the universe of confidential W2 tax records, demographic characteristics and the universe of unemployment insurance wage records from the Longitudinal Employer-Household Dynamics (LEHD) program, and the Decennial Census.

We start with all Ph.D. recipients in ProQuest who were employed on a sponsored research project between 2001 and 2016 at one of the 31 universities participating in the UMETRICS program.[11] ProQuest also provides information on each Ph.D. recipient's research fields, dissertation topics, and degree year; UMETRICS also allows us to create measures related to each Ph.D. recipient's individual ability and training environment, including the total length of time they were funded and the total number of grants from which they were paid. UMETRICS, combined with restricted-use demographic information from the LEHD program and the Decennial Census, enable us to construct individual-level demographic information on the Ph.D. recipient and lab-level demographic information on peers. The gender of PIs are imputed based on their names using *Ethnea*.

Published scientific articles indexed in PubMed have been matched to UMETRICS awards via research funding acknowledgements. This match is covers awards funded by the National Institutes of Health (NIH), and so we restrict our analysis to Ph.D. recipients funded by NIH grants. Using these article-award links, along with journal quality measures from Scimago Journal & Country Rank (SJR), we construct measures for both the quantity and quality of research output. Specifically, we compute publication counts (quantity) and average SJR per publication (quality) for each lab.

The restricted-use universe of W2 tax records and LEHD unemployment insurance wage records, housed at the U.S. Census Bureau, provide information on Ph.D. recipients' job placements, earnings, and employer characteristics. In this paper, we restrict our focus to Ph.D. recipients whose first job placement after leaving their UMETRICS university is at an academic institution, which we are able to identify using universities' Federal Employer Identification Numbers (EINs) available,

[11] UMETRICS data capture transactions from university grants to individuals providing labor (e.g. faculty, post-docs, students, etc.) and vendors providing goods and services (e.g. lab space, IT services, microscopes, lab mice, etc.). They are housed at the Institute for Research on Innovation and Science (IRIS) at the University of Michigan.

Table B1: Sample Dissertation and Patent (hypothetical; no disclosive information)

Proquest	US Patent	Similarity
ProQuest #AAI3319598. Therapeutic potential of radial glial RG3.6 cells in rat spinal cord injury. Spinal cord injury (SCI) triggers a cascade of pathophysiological changes that lead to secondary tissue damage after the mechanical insult. Early after SCI, cells are disrupted and excitotoxic amino acids (e.g. glutamate) are released. Inflammatory cytokines and chemokines are quickly induced...	U.S. Patent No. 9717804. Regenerating functional neurons for treatment of disease and injury in the nervous system. Methods for producing new neurons in the brain in vivo are provided according to aspects of the present invention which include introducing NeuroD1 into a glial cell, particularly into a reactive astrocyte or NG2 cell...	0.9xxx
ProQuest #AAI1432716. Multiuser TDMA channel estimation. Wireless communication systems require efficient utilization of the limited available spectrum. There are various methods in which division of spectrum between users have been done till date, including frequency division...	U.S. Patent No. 7110378. Channel aware optimal space-time signaling for wireless communication over wideband multipath channels. A method and system is described for more optimally managing the usage of a wideband space-time multipath channel. The wideband space-time multipath channel is decomposed into a plurality of orthogonal subchannels...	0.8xxx

This isn't really a table. It is a stand-in for having actual dissertation text compared to actual patent text. This will be in the appendix. —NEED TO FIGURE OUT DISCLOSURE

from the Integrated Postsecondary Education Data System (IPEDS). Note that IPEDS provides EIN information on nearly all U.S. universities, not just UMETRICS universities. This is important because, while a Ph.D. recipient may place at their UMETRICS university, they are much more likely to place at a different, non-UMETRICS university.

We measure the quality of a Ph.D. recipient's academic job placement (which we interpret as a proxy for the quality of training they received while at their UMETRICS university) using the amount of research funding received by the institution. The institution-level research funding information is obtained from the NSF's Higher Education Research and Development Survey (HERDS). The HERDS dataset provides the amount of annual federal research funding received by each institution from each funding agency, including Department of Defense (DOD), Department of Energy (DOE), Health and Human Service (HHS), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), United States Department of Agriculture (USDA), and others. We focus on two sources of funding as the main placement quality (i.e. training quality) measures – total federal research funding and federal research funding from the U.S. Department of Health and Human Services (HHS). We view total funding as representing the research capacity of the institution and we emphasize HHS funding because of our focus on Ph.D. recipients who were supported by an NIH (part of HHS) grant while at their UMETRICS university. Rankings on R&D are broadly similar to reputation ranks. We further categorize academic placement into two occupations, faculty and postdoctoral researcher. Note that we do not have occupation information of placement, as W2 only provides industry but not occupation. We define postdocs based on their wage and whether they stay at the same university after graduation. We define the academic position as postdoc if the doctoral recipient is staying in the same university (using match of UMETRICS university and IPEDS university), or W2 wage less than 40,000 (my memory is vague, maybe Joe should double check this number) one year after graduation.

Table XXX shows summary statistics for the key variables we use in this paper. Our final analysis sample contains about 1,300 Ph.D. recipients who were supported by an NIH grant at a UMETRICS university and subsequently placed at academic institution after leaving their UMETRICS university. About 300 of these, or 23%, were supported by a grant led by a female PI. About 53% of Ph.D. recipients placed at an academic institution as a post-doc (with the remaining 47% placed as faculty) and the typical Ph.D. recipient placed at an institution that receives about $29 million in annual total federal research funding and about $18 million in HHS research funding.[12] Foreshadowing regression results, Ph.D. recipients trained

[12]We categorize postdoc positions for those who 1) stay in the same institutions after receiving their Ph.D. degrees and W2 wage in the first year after graduation is no more than $40k. We expect that it is very unlikely for PhD recipients to get a faculty position in the 21 UMETRICS universities in our final sample. 2) place in another institution and W2 wage in the first year since graduation is less than $40k.

by female PIs placed at institutions with higher funding, which we interpret as suggestive of superior training. While at their UMETRICS university, the typical Ph.D. recipient worked in a lab that produced 18 articles with an average SJR of 1.6. Ph.D. recipients trained by female PIs worked in labs that produced fewer publications with a lower SJR, again foreshadowing our regression results, and suggesting that higher-quality research does not necessarily lead to higher-quality training. About half of the Ph.D. recipients are female (though this rises to about 62% when the PI is female) and have an average age of a little less than 30. The typical Ph.D. recipient is on a grant for about 660 days (1.8 years), is paid by 2.2 awards, collaborates with two 2.1 faculty members and 3.7 students, and receives their degree in 2011.

What is demand for knowledge and skills. For engineering and computer scientists it seems to be there. Kind of obvious here. Philosophy on the other end. Biomedical folks are in the middle. In contrast biomedical researchers don't make a tone. They take long post-docs and work in all kinds of jobs in industry. People there say the fact that they are getting a wide range of jobs is an indication of how valuable their skills are.

Are you prodding a skill set universally valuable or producing so many people they need to get off the beaten path jobs. Look at earnings.

To how many industries do a given field's dissertation match. Measure of how broadly applicable their skills. Do they get jobs in these areas. Earnings for people who get jobs in these areas.

How broadly applicable are my skills. Do I get a job that actually uses those skills. Does my earnings depend on.

Step 1: how broadly their dissertation matches Bruce: opportunity Step 2: Whether they go into fields they are well-matched. Step 3: Conditional on match relevance and field how much are you paid. Or earnings.

D Data Appendix (Joe)

Our ultimate goal is to create a full set of dissertation-firm dyads (i.e. all pairwise combinations) with a measure of how similar each dissertation is to the patenting portfolio of each firm as well as an indicator for whether the individual who wrote the dissertation works at the firm after they receive their degree (and, if so, how much they earn). This data will allow us to to examine how the probability of working at a firm varies with how closely an individual's field-specific knowledge (as encoded in their dissertation) matches the patenting portfolio of the firm. Moreover, subsets of the data (e.g. the subset of all individual/dissertation-firm pairs that actually have an employment relationship) will allow us to examine how earnings vary with match quality and how job placement patterns of different fields vary by match quality.

We create this dyadic data structure in a series of steps. First, we link individuals

to their dissertation and use confidential earnings data at the U.S. Census Bureau to determine their employment outcomes after they receive their Ph.D. Specifically, we determine the firms at which each individual/dissertation (for our purposes, an individual and the dissertation they wrote can be thought of interchangeably) work along with their earnings. Second, we identify, for each individual/dissertation, the 1,000 most closely related patents in terms of text similarity. Third, in a series of sub-steps, we link patents to the assignee firms and aggregate the similarity score for a given dissertation-patent pair to a similarity score for dissertation-firm pairs. This gives us our dyadic structure where each individual/dissertation is linked to every patenting firm and each pair has a similarity score as well as an indicator for whether the individual/dissertation works at the firm after receive their Ph.D. In addition, since each firm can be linked to an industry, we can also create dyadic data with individual/dissertation-industry pairs as an alternative to individual/dissertation-firm pairs. This will allow us to explore how the probability of working at an industry varies with how closely an individual's field-specific knowledge matches the patenting portfolio of that industry.

The rest of this section describes these linkages in greater details.

D.1 UMETRICS Employees and Employment Outcomes

We begin with all individuals paid on a grant at a UMETRICS university (i.e. UMETRICS employees).[13] A subset of these individuals are able to be linked by Census staff (using identifying information such as name and birth date) to a Protected Identification Key (PIK) which is an internal person-level identifier at Census.[14] Once an individual is assigned to a PIK, we can link them to a variety of confidential Census data – in our case, we link them to W2 tax records (2005-2018), which enable us to identify which firms (EINs) employed each PIK as well as earnings received by that EIN.

D.2 UMETRICS Employees, Dissertations, and PIKs

Though we link all UMETRICS employees that receive a PIK to their W2 tax records, this group does not comprise our main sample of interest. Rather, we focus on the subset of PIK'd UMETRICS employees that are also able to be linked to their dissertations in ProQuest. We focus on this group because we have, in the form of their dissertation, concrete information about the field-specific knowledge they accured during their graduate training. Thus, for this subset of individuals, we will be able to measure how similar their field-specific knowledge is to the patenting portfolio of firms or industries (see below for details). This will shed light on how well-matched each individual is to the firm or industry in which they are actually employed after receiving their Ph.D.

[13] We use the 2018 Q4 tranche of UMETRICS data delivered to Census.

[14] The PVS process is a probabalistic match...

ProQuest dissertations were linked to UMETRICS employees by IRIS staff. We link these dissertations to the subset of UMETRICS employees that receive a PIK. Unfortunately, some PIKs link to multiple dissertations and some dissertations are attributed to multiple PIKs. To simplify the analysis, we only keep the (vast majority) of PIKs that match, 1-to-1, to a ProQuest dissertation. By confining our analysis to these 1-to-1 links, PIKs and dissertations can be thought of interchangeably.

D.3 Dissertations, Patents, and Cosine Similarity

As noted, we want to measure how closely related a PIK's ProQuest dissertation is to the patenting portfolio of a firm or an industry. To do this, we identify the 1,000 USPTO patents that are most closely related to each dissertation on the basis of a cosine similarity score comparing dissertation and patent text. If the cosine similarity for a dissertation-patent link falls below a threshold (WHAT IS THIS THRESHOLD???), the link is discarded and the dissertation will be linked to fewer than 1,000 patents. In the extreme, if all of the most closely matched patents have a similarity score that falls below the threshold, the dissertation will not be linked to any patents.

Our final sample of individuals is a set of UMETRICS employees that 1) receive a PIK (and thus can be linked to employment outcomes at Census) 2) link 1-to-1 to a ProQuest dissertation (and thus their field-specific knowledge can be observed and compared to the patenting portfolio of firms or industries) and 3) whose dissertation links to at least 1 patent.

D.4 Aggregating Cosine Similarity to Dissertation-Firm Level

We are not directly interested interested in the links and similarity scores for dissertations and patents. Rather, we wish to use these links to construct a measure of how closely related each dissertation is to the firms assigned the patents (or industries in which these firms operate).

To do this, we link dissertation-patent pairs (and their accompanying cosine similarity score) to the patents' assignee EINs. This creates a data set with observations unique at the dissertation-patent-EIN level. We then collapse over the patents to obtain the average and maximum cosine similarity score for each dissertation-EIN pair. Thus, each dissertation-EIN pair inherits a similarity score that is comprised of a combination of several dissertation-patent similarity scores.

That is, for each EIN to which a dissertation can be linked through a matched patent...

Though conceptually straightforward, the data are structured in a way that requires us to undertake this aggregation after substantial pre-processing of the raw data.

Since firms in the W2 files are identified by EINs, we want to link each patent to the assignee EIN. However, the patent-firm (BDSPF) crosswalk at Census links patents to an alternative firm identifier called FIRMID.

D.5 Dissertations, Patents, and Cosine Similarity

A subset of UMETRICS employees are linked to their ProQuest dissertations. We link this group

Our next step is to identify which patents are most closely related (in terms of text) to each dissertation. Once a dissertation is linked to a set of patents, we will then be able to link it to firms using a patent-firm bridge created by Census. To link dissertations to similar patents, we write a Python script that takes the abstract of each dissertation and identifies the 1,000 patents that most closely match each dissertation on the basis of a cosine similarity score. If the cosine similarity for a dissertation-patent link falls below a threshold (WHAT IS THIS THRESHOLD???), the link is discarded and the dissertation will be linked to fewer than 1,000 patents. In the extreme, it is possible for a dissertation to not to link to any patents because all of the most closely matched patents have a similarity score that falls below the threshold. In the end, this gives us a set of ProQuest dissertations, each linked to 1,000 or fewer patents along with the cosine similarity score for each dissertation-patent pair. Note that we only retain the PIK/dissertations from step 1 if they match to at least one patent.

We then link the subset of UMETRICS employees with PIKs to W2 tax records for 2005-2018. This allows us to determine the EIN at which each UMETRICS PIK is employed and the earnings received by that EIN.

D.6 Aggregating Cosine Similarity to Dissertation-Firm Level

D.7 Individual/Disseration-Firm Dyads

First, we identify, for each dissertation in Proquest, the 1,000 most closely related patents in terms of text similarity. Second, we link individuals

link each ProQuest dissertation to a firm. We do this in a series of steps. First, we link person identifiers (PIKs) to dissertations. Second, we link dissertations to patents. Finally, we link patents to firms.

The ultimate goal is to link each Proquest dissertation to a federal tax identification number (EIN). In addition, we want a measure of the similarity of each dissertation-EIN pair – more specfically, a measure of how similar the text of a dissertation is to the text of patents assigned to the EIN. Finally, we want to identify the EINs at which each dissertation works. This data will allow us to measure whether the probability of an individual working at an EIN is related to how closely related that individuals dissertation is to the patenting of the EIN. More generally,

we can relate the cosine similarity between for a dissertation-EIN pair to employment outcomes for the Ph.D. recipient who wrote the dissertation.

Our focal sample of PIKs is a set of Ph.D. recipients for whom (1) we can match them 1-to-1 to a ProQuest dissertation and (2) we can link their dissertation to at least one patent based on a text cosine similarity score.[15]

First, we want to link ProQuest dissertations to Protected Idenfication Keys (PIKs) at the U.S. Census Bureau. A PIK is an internal person-level identifier at Census and allows us to link individuals to confidential data such as W2 or UI wage records. By linking a dissertation to a PIK, we are able to analyze various employment outcomes for the Ph.D. recipient who wrote the dissertation. Moreover, we can examine how these employment outcomes are related to various characteristics of the dissertation (e.g. how similar it is to the IP at different firms or industries).

Unfortunately, there is no direct link between ProQuest dissertations and PIKs. However, we are able to link PIKs to a subset of dissertations – namely those dissertations that were written by an individual in the UMETRICS database. UMETRICS is... We can link PIKs to dissertations in two steps. First, PIKs are linked to UMETRICS employee IDs (UMIDs) using a PIK-UMID crosswalk created at Census. Second, UMIDs are linked to ProQuest dissertations.[16]. These two steps allow us to assign a PIK to each dissertation written by a Ph.D. recipient in the UMETRICS database. In this data, some PIKs link to multiple dissertations and some dissertations are attributed to multiple PIKs. To simplify the analysis, we only keep the (vast majority) of PIKs that match, 1-to-1, to a ProQuest dissertation. Note that since PIKs and dissertations are linked 1-to-1, they can be thought of interchangeably.

Our next step is to identify which patents are most closely related (in terms of text) to each dissertation. Once a dissertation is linked to a set of patents, we will then be able to link it to firms using a patent-firm bridge created by Census. To link dissertations to similar patents, we write a Python script that takes the abstract of each dissertation and identifies the 1,000 patents that most closely match each dissertation on the basis of a cosine similarity score. If the cosine similarity for a dissertation-patent link falls below a threshold (WHAT IS THIS THRESHOLD???), the link is discarded and the dissertation will be linked to fewer than 1,000 patents. In the extreme, it is possible for a dissertation to not to link to any patents because all of the most closely matched patents have a similarity score that falls below the threshold. In the end, this gives us a set of ProQuest dissertations, each linked to 1,000 or fewer patents along with the cosine similarity score for each dissertation-patent pair. Note that we only retain the PIK/dissertations from step 1 if they match to at least one patent.

Our third step is to link each patent in the dissertation-patent crosswalk to its

[15] What is the similarity score cutoff that Evgeny used???

[16] These linkages were created by the staff at IRIS and the data were transferred to Census. DO WE HAVE AN OVERVIEW OF HOW THIS WAS DONE? WHO ACTUALLY DID THE LINKING AT IRIS?

assigned firm.

To identify this focal sample of PIKs, we first write a Python script that takes the abstract of every dissertation in the Proquest database and identifies the 1,000 patents that most closely match each dissertation on the basis of a cosine similarity score. If the cosine similarity for a dissertation-patent link falls below a threshold (WHAT IS THIS THRESHOLD???), the link is discarded and the dissertation will be linked to fewer than 1,000 patents. In the extreme, it is possible for a dissertation to not to link to any patents because all of the most closely matched patents have a similarity score that falls below the threshold. In the end, this gives us a set of ProQuest dissertations, each linked to 1,000 or fewer patents along with the cosine similarity score for each dissertation-patent pair.

Next, ProQuest dissertations are linked to UMETRICS employee IDs (UMIDs).[17] In this ProQuest-UMID crosswalk, each dissertation is uniquely linked to a UMID, but a UMID can be linked to multiple dissertations.[18] In order to link each dissertation to confidential Censsu data (e.g. W2 or UI wage records) we need to link each dissertation (through the UMID) to a Protected Identification Key (PIK) at Census.[19] To do this, we use a PIK-UMID crosswalk produced at Census. When this link is made, each dissertation is linked to a PIK. In this data, some PIKs link to multiple dissertations and some dissertations are attributed to multiple PIKs. To simplify the analysis, we only keep the (vast majority) of PIKs that match, 1-to-1, to a Proquest dissertation.

Once we have a set of 1-to-1 dissertation-PIK links (PIKs and dissertations can now be though of interchangeably), we subset the dissertation-patent links for which the dissertation matches 1-to-1 to a PIK. The reason we subset the Proquest data, which considerably reduces the sample size because the dissertation must belong to a UMETRICS individual and that individual must link to a PIK and the PIK must match 1-to-1 to a Proquest Dissertation. The reason we subset the data in this way is because we can only get labor market outcomes for dissertations linked to PIKs because we need the PIK to link to confidential Census data.

Thus, as mentioned, our focal sample of PIKs/dissertations will be those for which we have 1-to-1 links and those that match to at least one patent (above the threshold).

Ultimately, we are not interested in how similar each PIK's dissertation is to patents. Rather, we are interested in in how similar the dissertation is to the a firm (EIN) or industry. To do this we link patents to firms, which can be assigned to an industry (NACIS code). IF WE COULD DO THIS OVER AGAIN WOULD WE DIRECTLY CHECK HOW SIMILAR EACH DISSERTATION IS TO THE PATENT PORTFOLIO OF EACH FIRM? THAT IS, TAKE EACH DIS-

[17] These linkages were created by the staff at IRIS and the data were transferred to Census. DO WE HAVE AN OVERVIEW OF HOW THIS WAS DONE? WHO ACTUALLY DID THE LINKING AT IRIS?

[18] How often does someone actually have multiple disserations

[19] PIKs are...

SERTATION AND COMPUTE THE COSINE SIMILARITY FOR EACH FIRM'S PATENT PORTFOLIO IN THE FIRMID-PATENT XWALK AT CENSUS. THIS WOULD GIVE US A DIRECT SCORE OF EACH DISSERTATION TO EACH FIRM.

In addition to computing how similar a dissertation is to the patents in a FIRM/Industry, we also want to link each PIK/dissertatino to a set of employment outcomes using confidential data at the US Census Bureau.

D.7.1 Prepare W2/LEHD Files

D.7.2 Prepare CBPBR Files

The unit of observation in the CBPBR is an establishment, uniquely identified (within a year) by an `estabID`. We use CBPBR files for the years 2000-2018. When stacked, the unit of observation is an establishment-year (`estabID-year`).

For each `estabID-year`, we first create a firm-identifier, `firmID`, using a combination of information from the `alpha` and `EIN` fields.[20] Since firms in the the patent-firm (BDSPF) crosswalk (see above) are identified by the `firmID`, assigning this information to each `estabID-year` is a crucial step in linking patents to EINs.

D.7.3 Prepare FIRMID-Patent Crosswalk Files

D.7.4 Prepare Linked Dissertation-Patent Files

D.7.5 Prepare Linked Proquest-UMETRICS Files

D.7.6 Link UMETRICS Employees to W2/LEHD Records

D.7.7 Link UMETRICS-W2 EINs to CBPBR

[20] `alpha` is an XX-digit idnetifer....`EIN` is an XX-digit identifier for the federal tax idenfication number. The `firmID` is created from `alpha` by .... and is created from the `EIN` by....There are 2 fields with information on `alpha` and there are 3 fields with information on `EIN`. For each `estabID-year`, we create a list of up to 5 corresponding `FIRMIDs`.