Using Published Dissertations to Identify Graduates' Countries of Origin[1]

By Megan MacGarvie
Boston University and NBER

Prepared for the NBER Conference: "Career Patterns of Foreign-Born Scientist and Engineers, Trained and/or Working in the U.S."

November 2, 2007

Recent years have seen a dramatic increase in the foreign-born share of the science and engineering (S&E) doctorate recipient pool, a majority of whom remain in the U.S. immediately after doctorate receipt. The foreign proportion of the S&E doctoral recipients increased by 74% between 1985 and 2005. (See NSF Science and Engineering Doctorate Award 2005, Bound, Turner and Walsh 2006, National Academies 2005.) Several countries, in particular China and India, saw even bigger increases in the number of students they sent to obtain doctoral degrees in the U.S.

We do not yet have a complete picture of the effects of this increase on science and innovation (and consequently, on economic growth) in the U.S. and abroad. This may be related to some degree to the limitations of existing datasets. This document describes a new source of data on foreign-born recipients of Ph.D.s at American universities. It discusses the advantages and disadvantages of this data source relative to other existing data sources, and presents the results of some preliminary investigations as examples of how the data may be used.

Existing data on foreign-born students at U.S. universities is available through the NSF's SESTAT program. While this data is extremely detailed, comprehensive, and reliable, it has two limitations. One is that, in the interest of protecting the anonymity of survey respondents, there are restrictions on matching external datasets to NSF data by student

[1] I am grateful to the NSF for providing access to the *Survey of Earned Doctorates* micro-data. The use of NSF data does not imply NSF endorsement of the research, research methods, or conclusions reached in this paper.

name. This means that detailed data on career histories, publications, and patents cannot easily be combined with the SESTAT data. While the SESTAT data does include publication counts, researchers wanting to investigate issues relating to knowledge diffusion, networks, peer effects, or other questions requiring access to an individuals' detailed publication or patenting record, must look for alternative sources. Another limitation of the NSF data is that the longitudinal *Survey of Doctoral Recipients* does not follow subjects who leave the United States. This means that it is difficult to perform comparisons of those students who remained in the U.S. with those who moved to their home countries or another location.

One alternative data source that is not subject to these limitations (but which has its own drawbacks, as detailed below) is a dataset I am developing based on published Ph.D. dissertations at a sub-set of American universities. Several universities require students to list biographical information in the front matter of the dissertation. Table 1 lists these universities, which were identified by checking dissertations filed at the universities that are major producers of engineers in the United States. There may be additional universities that require biographical information in the dissertation.

At some universitiees, the information includes a full biographical sketch (e.g., Ohio State, NC State), but in most cases, the information is limited to a list of previous degrees. ProQuest's *Dissertations & Theses* database, available online through most university libraries, is a database of almost all dissertations filed at over 700 U.S. universities. Starting in the late 1990's, ProQuest began publishing online the full text of the first 24 pages of the dissertation. Appendix A presents examples of this information drawn from dissertations filed at UC Berkeley, the University of Illinois, and the Ohio State University.

The biographical information contained in these dissertations can be used to identify the country of origin of the student.

Using this information as a proxy for the nationality of the student will of course introduce some error, since not all students receiving undergraduate degrees do so in their country of origin. However, evidence from the NSF's *Survey of Earned Doctorates* suggests that the country of undergraduate degree is a very good proxy for the country of origin. For students completing doctorates in 2003 and 2004, the *SED* lists the country of undergraduate degree. For 84.9% of students, the country of undergraduate degree is the same as the country of citizenship. However, there is considerable heterogeneity across countries in the

extent to which students pursue undergraduate studies outside their countries of origin. Table 2 presents, for a selected list of countries, the share of students responding to the *SED*'s questions who remained in their home country for undergraduate study. Students from Germany and Japan have the lowest rates of staying at home among the major producers of U.S. graduate students (73% and 74%, respectively). However, the countries that send the most students (China, India, Taiwan, Korea, and Canada) have high stay-at-home rates for undergraduate study (98%, 93%, 89%, 76%, and 82%, respectively). Furthermore, counts of the number of doctoral recipients by country of origin, university and year computed from the ProQuest sample described here have a correlation of 0.948 with analogous counts obtained from the *SED*.

Relative to existing datasets, the ProQuest data have advantages and disadvantages. The most significant advantage is that these data allow for name matching with other datasets. In principle, one could also construct a dataset containing foreign students who left the U.S. upon completion of studies. Another advantage is that and the data are freely available over the internet. Finally, in addition to the information described above, there is information easily downloadable in digital format on the student's institution, year of degree, advisors, and fields of study. Additional information may be obtained from the full text of the dissertation, though this would require considerably more effort.

The main disadvantage of this dataset is that it contains very little information relative to, for example, the SESTAT data. Thus, it is essentially useless unless matched to other datasets. Another major issue is that the data on country of origin must be hand collected. In addition, the data on country of origin is only available beginning in the late 1990's when universities began submitting digital copies of dissertations to be posted on the web by ProQuest. Table 3 shows that in 1996, a large number of dissertations published by ProQuest for OH State and UIUC contained no data on country of origin (because digital copies are not available for those dissertations). However, by 1997 almost all dissertations are available in digital format.

Preliminary data containing the patents and citations of 1,720 students who completed PhD programs in engineering at the University of California, the University of Illinois and the Ohio State University between 1996 and 1999. We obtained information on the institution and year of the previous degrees and the thesis title.

These data were then matched to information on inventors and patents from the USPTO and the NBER patent dataset. We matched students to inventors who had the same last name, first name and middle initial (where a middle name or initial was listed). We drop all foreign-invented patents in order to remove bias associated with the fact that inventors with foreign names living in foreign countries may not be exact matches to the PhD graduates from the U.S. university. This procedure resulted in a match to 271 inventors.

Table 2 contains the results of a regression of the number of forward citations (by country) to the patents held by students graduating from U.S. universities from the year of graduation to 2004. We employ a Negative Binomial regression model in which the unit of observation is a student-country pair, and we include controls for the log of the number of patents held by individual *i* and the log of the number of patents invented in country *j*.

The main variable of interest is a dummy equal to 1 if the citing country is the same as the student's country of origin. Thus we compare the number of citations by patents in the student's home country to the number of citations associated with other countries. We cut the data in several ways: the first column contains results from the pooled sample of all countries. The second and third columns focus on students from Organization for Economic Co-operation and Development (OECD) countries, either including or excluding American students and citations. The results in the final column are from a restricted sample of students from China, India, Singapore and Taiwan.

These results show that, for students from OECD countries, there is a clear "home country bias" in forward citations. That is, students are approximately twice as likely to be cited by patents from their home countries, after controlling for the overall tendency of those students and countries to patent. However, PhD recipients from the Asian countries that are the largest "exporters" of students to the U.S. are not more likely to be cited at home.

Table 1

Universities requiring information on prior degrees in dissertations

University of California
Syracuse
U Illinois
U Colorado
Fordham
NC State
Ohio State
U Virginia
Boston U
U Massachusetts
Cornell

Table 2: Share of Ph.D. students at U.S. universities who received undergraduate degrees in their countries of citizenship

AUSTRALIA	85.00%
BRAZIL	96.02%
CANADA	82.51%
CHINA	98.35%
EGYPT	96.38%
FRANCE	82.05%
GERMANY	73.05%
GREECE	80.51%
INDIA	92.71%
IRAN	88.33%
ISRAEL	88.46%
JAPAN	73.51%
MEXICO	89.19%
NIGERIA	60.61%
PHILIPPINES	87.23%
SOUTH KOREA	76.33%
TAIWAN	89.19%
THAILAND	87.28%
TURKEY	95.57%
U.K.	63.64%
Weighted average across these countries	89.50%
Weighted average across all countries	84.79%

Source: *Survey of Earned Doctorates* micro-data and author's calculations

Table 3: Engineering Ph.D.s by University, Year of Degree, and Country of Undergraduate Degree

	Ohio State University					UC Berkeley					UIUC				
Country	1996	1997	1998	1999	Total	1996	1997	1998	1999	Total	1996	1997	1998	1999	Total
ARGENTINA						1	0	0	0	1					0
BANGLADESH	0	0	1	1	2					0	0	1	0	0	1
ARMENIA					0	1	0	0	0	1	0	0	1	0	1
AUSTRALIA					0					0	0	0	0	1	1
AUSTRIA					0	0	0	0	1	1					0
BRAZIL	0	0	0	1	1	1	1	1	2	5	1	0	1	1	3
BULGARIA					0					0	0	1	0	0	1
CANADA	1	0	0	0	1	0	3	6	2	11	1	2	0	1	4
CHILE					0					0	0	1	0	0	1
CHINA	3	8	13	12	36	13	11	16	11	51	12	12	18	14	56
COLOMBIA	0	1	0	0	1					0	0	1	2	1	4
COSTA RICA					0	0	0	0	1	1	0	0	0	1	1
DOMINICAN REPUBLIC					0					0	0	1	1	0	2
EGYPT	0	3	3	4	10	0	0	1	1	2	2	2	1	1	6
ETHIOPIA	1	0	0	0	1					0					0
FRANCE					0	0	0	0	2	2	1	0	0	0	1
GERMANY	0	1	1	0	2	1	0	1	0	2	0	1	0	0	1
GREECE	1	0	0	0	1	1	1	0	1	3	2	2	4	1	9
HUNGARY					0					0	0	0	1	0	1
HONG KONG					0	0	0	0	1	1					0
ICELAND					0	1	0	0	0	1					0
INDIA	3	16	6	11	36	13	10	11	8	42	15	13	12	9	49
INDONESIA					0	1	0	0	0	1	0	1	0	0	1
IRAN	1	0	0	0	1	1	0	0	0	1	1	1	1	1	4
ISRAEL					0					0	1	0	0	0	1
IRELAND					0	0	1	1	0	2					0
ITALY					0	1	0	0	0	1					0
JAPAN	0	1	0	0	1	1	0	4	2	7	2	0	1	1	4
JORDAN	1	1	0	0	2					0	1	3	1	1	6
KENYA	0	1	0	0	1					0					0

	Ohio State University					UC Berkeley					UIUC				
Country	1996	1997	1998	1999	Total	1996	1997	1998	1999	Total	1996	1997	1998	1999	Total
KOREA	2	4	4	5	15	7	6	2	7	22	3	6	8	2	19
KUWAIT					0					0	0	0	1	0	1
LEBANON	0	1	0	0	1					0					0
MEXICO	1	0	1	2	4	2	1	0	0	3					0
NEW ZEALAND					0	1	0	0	0	1					0
PAKISTAN					0					0	1	1	0	1	3
PALESTINE	0	0	1	0	1					0	0	1	0	0	1
PORTUGAL					0					0	0	1	0	0	1
ROMANIA	0	1	0	0	1	0	0	1	0	1	0	1	0	1	2
RUSSIA	0	1	0	0	1	0	1	0	0	1	1	0	0	1	2
SLOVAKIA					0					0	0	1	0	0	1
SAUDI ARABIA	0	0	2	1	3	0	1	0	0	1					0
SINGAPORE					0	0	0	0	1	1					0
SOUTH AFRICA	0	0	0	2	2	1	0	1	1	3	1	0	0	0	1
SPAIN					0	1	1	0	0	2	0	1	0	1	2
SUDAN					0	0	0	1	0	1					0
SWEDEN					0	0	1	0	0	1	0	0	0	1	1
SWITZERLAND					0	0	0	0	1	1					0
TAIWAN	9	10	6	10	35	11	6	9	12	38	4	4	3	3	14
THAILAND	0	0	0	1	1	1	0	0	0	1					0
TRINIDAD					0					0	0	2	0	0	2
TURKEY	0	3	5	5	13	0	1	1	0	2	1	3	2	0	6
UK	0	1	0	0	1	1	0	2	1	4					0
UNITED ARAB EMIRATES	0	1	0	0	1					0					0
USA	8	29	26	17	80	92	85	121	85	383	69	92	77	70	308
VENEZUELA					0					0	0	2	0	1	3
NO DATA	53	3	4	2	62	3	4	0	0	7	26	1	1	0	28
Total	84	86	73	74	317	156	134	179	140	609	145	158	136	114	553

Source: *ProQuest Dissertations & Theses* and author's calculations

Table 4: Individual-level analysis of forward citations to patents invented by Ph.D. recipients from U.S. universities, by country of origin

	(1)	(2)	(3)	(4)
	Full Sample	OECD Countries	OECD countries, excl USA	China, India, Korea, Singapore & Taiwan
Dummy=1 if country is student's home country	0.646	1.173	0.966	0.252
	(0.083)***	(0.148)***	(0.479)**	(0.552)
Student's patent count	1.595	1.559	1.586	1.587
	(0.048)***	(0.062)***	(0.178)***	(0.072)***
Country's patent count	1.026	0.954	1.319	0.981
	(0.035)***	(0.043)***	(0.130)***	(0.067)***
Constant	-18.462	-17.778	-22.454	-17.564
	(0.496)***	(0.576)***	(2.077)***	(0.950)***
Observations	67044	37560	1326	19929

Robust standard errors in parentheses, clustered by inventor.

* significant at 10%; ** significant at 5%; *** significant at 1%

Appendix A:
Examples from Proquest

A Comparison of Freight Distribution Costs for Combination and Dedicated Carriers in the Air Express Industry

by

Max Karl Kiesling

B.S. (Texas Tech University) 1989
B.A. (Texas Tech University) 1989
M.S. (University of Texas at Austin) 1991

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Civil Engineering

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Mark M. Hansen
Professor Carlos F. Daganzo
Professor David W. Gillen
Professor Pablo T. Spiller

1995

ALGORITHMS AND ARCHITECTURES FOR SOFT-DECODING REED-SOLOMON CODES

BY

ARSHAD AHMED

B.E., Regional Engineering College, Trichy, 1998
M.E., Indian Institute of Science, Bangalore, 2000

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2006

Urbana, Illinois

VITA

January 31, 1973	Born – Da-An, Jilin Province, China
September 1989 – July 1993	Bachelor of Science in Electrical Engineering, Nanjing University of Science and Technology, Nanjing, China
September 1993 – April 1996	Master of Science in Electrical Engineering, Nanjing University of Science and Technology, Nanjing, China
Septmeber 2002 – present	Ph.D student, Analog VLSI Laboratory, Department of Electrical and Computer Engineering, the Ohio State University, Columbus, Ohio
Since June 2006	RFIC design engineer, Freescale Semiconductor Inc., Boca Raton, Florida

PUBLICATIONS

Research Publications

P. Zhang, and M. Ismail "A New RF Front-End and Frequency Synthesizer Architecture for 3.1–10.6 GHz MB-OFDM UWB Receivers", *Proc.* 48th *Midwest Symposium on Circuit and System*, vol.2, pp.1119–1122, August 2005.

C. Garuda, X. Cui, P. Lin, S. Doo, P. Zhang, and M. Ismail "A 3–5 GHz Fully Differential CMOS LNA with Dual-gain Mode for Wireless UWB Applications", *Proc.* 48th *Midwest Symposium on Circuit and System*, vol.1, pp.790–793, August 2005.

Y. Yu, L. Bu, S. Shen, B. Jalali-Farahani, G. Ghiaasi, P. Zhang, and M. Ismail "A 1.8V Fully Integrated Dual-band VCO for Zero-IF WiMAX/WLNA Receiver in 0.18μm CMOS", *Proc.* 48th *Midwest Symposium on Circuit and System*, vol.1, pp.187–190, August 2005.