ISSN 2279-9362

Collegio Carlo Alberto

Italian Doctorate Holders and Academic Career. Progression in the Period 1986-2015

Massimiliano Coda Zabetta
Aldo Geuna

No. 629
December 2020

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Italian Doctorate Holders and Academic Career Progression in the Period 1986-2015

Massimiliano Coda Zabetta[1,2] *and Aldo Geuna*[3,4,5]

[1] *Spatial Dynamics Lab, University College Dublin*
[2] *GREThA – UMR CNRS 5113, University of Bordeaux*
[3] *Department of Cultures, Politics and Society, University of Turin*
[4] *Collegio Carlo Alberto*
[5] *Fellow CIFAR, Programme Innovation, Equity & The Future of Prosperity*

Corresponding author: Massimiliano Coda Zabetta
(massimiliano.codazabetta@ucd.ie)

4 December 2020

Abstract

This paper describes the Italian Doctoral Holder and Academic Career (IDH-AC) database, which includes unique information on the population of doctoral graduates from Italian universities, in all disciplines, in the period from the first cycle of doctorates (1983-86) to 2006. Doctoral graduates who pursued an academic career in Italy were identified by matching with the list of academics active in Italian universities in the period 1990-2015. These original data allows us to shed light on several issues related to the Italian academic labour market, such as gender, inbreeding, mobility, hiring and promotion patterns. The paper i) describes the record linkage between two datasets and ii) presents an exploratory statistical analysis of academic employment outcomes for the population of researchers who were awarded a doctoral degree from an Italian university over a 20 year period.

JEL classification: I23, J45, M51.

Keywords: PhD holders, Academic careers, Database, Record linkage.

1. Introduction

Doctoral graduates are key research and innovation actors. They have been trained to conduct research and are among the most important vectors for the diffusion of scientific knowledge. In addition, numbers of doctoral students provide a measure of a country's potential research capability. It is estimated that, in 2012, 717,000 doctoral graduates were working in the EU-28 countries, 492,000 in the US and 75,000 in Japan (European Commission, 2014).

Since the end of the 1990s, the number of PhD student places in Italy has increased from 21,128 in 2000-01 to 33,508 in 2013-14 and the number of PhD degrees awarded in Italy has risen from 4,087 to 10,745 (source: MIUR data).[1] However, Italy is at the bottom of the European ranking for number of doctoral graduates in the population. Italy has 28 doctoral graduates per 10,000 inhabitants compared to the European average of 65 per 10,000 inhabitants (Figure 1 - dashed line). One reason for Italy's poor performance is that formal doctoral training has been provided only since 1983, much later than in other European countries. There were opportunities for research training in Italy prior to 1983, but this was not formalized and was not associated to an official education degree.

Systematic information on doctoral students, doctoral awards and career outcomes is scarce in the case of Italy. For several years, Cineca was the only database which contained information on doctoral courses and students in Italian universities[2]. In 1998, the statistical office of the Italian Ministry of University and Research (MIUR), began to gather data on student enrolment and doctoral graduation for Italian universities.[3] There have also been some sporadic attempts to conduct surveys on the career outcomes of doctoral graduates from specific universities[4].

We constructed an Italian Doctoral Holder (IDH) database, which is the first attempt to collate data on the population of doctoral graduates awarded a doctoral degree from an Italian university in all disciplines including a scientific field. It covers the period 1983 to 2006. We obtained information from the repository of doctoral dissertations at the Italian national library in Florence (Biblioteca

---

[1] http://dati.ustat.miur.it

[2] https://cercauniversita.cineca.it/php5/dottorati/cerca.php

[3] http://dati.ustat.miur.it/dataset/formazione-post-laurea

[4] In 2006 CNVSU and MIUR conducted an analysis for the doctorate holders from universities of Pavia, Pisa, Siena and Salerno; in 2006 university of Trento conducted an analysis of the occupational outcomes of doctorate holders from universities of Milano, Milano-Bicocca and Trento; in 2009 Scuola Superiore Sant'Anna conducted an analysis of doctorates from seven universities (Bergamo, Brescia, Milano, Milano-Bicocca, Palermo, Pisa and Scuola Superiore Sant'Anna); in 2009 and 2014 ISTAT conducted a survey on the career of Italian doctorates; in 2018 University of Turin surveyed career outcome of a sample of doctorates graduated after 2007.

Nazionale Centrale di Firenze - BNCF), which was established by MIUR in 1980. We extracted the raw information from the BNCF archive, which we codified and cleaned. This allowed us to generate the following standardized data on: gender, university affiliation, year of graduation, thesis title, scientific field of the doctoral program and, in some cases, PhD supervisors. The resulting IDH database includes 76,093 doctoral dissertations representing the population of PhD graduates who graduated from an Italian university between 1986 and 2006.

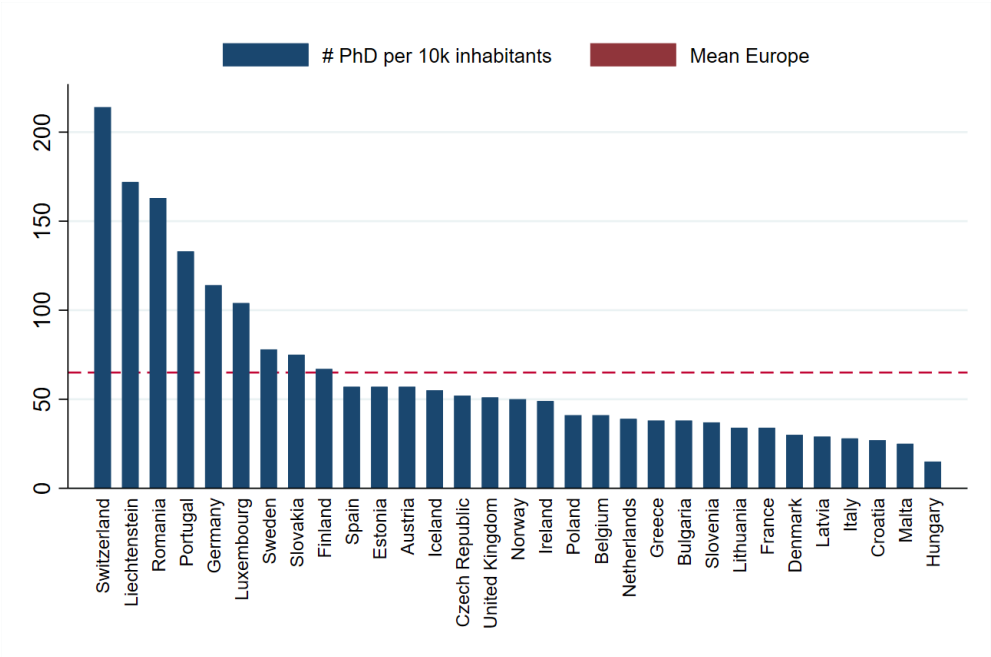

Figure 1: Number of PhD holders every 10000 inhabitants (Source: Auriol et al., 2013)

We identified Italian PhD graduates who pursued an academic career in Italy by linking the data to data on academics working in Italian universities in the period 1990-2015, available from MIUR. The final Italian Doctoral Holder and Academic Career (IDH-AC) database includes information on 25,412 doctoral graduates awarded their degree from an Italian university between 1986 and 2006 and who worked for at least one year in academia in Italy, as an assistant, associate or full professor, between 1990 and 2015. We use this information to conduct a descriptive analysis of the employment outcomes of the population of Italian PhD holders who chose an academic career in Italy, during the period 1986-2015.

The remainder of the paper is organized as follows. Section 2 presents the data sources and data retrieval procedure, and presents the related summary statistics. Section 3 describes the record linking exercise the IDH and MIUR data, used to create the IDH-AC database. Section 4 presents an explorative statistical analysis and discusses some stylized facts about the academic careers of Italian PhD holders. Section 5 concludes the paper.

2. Data Sources

In this section, we describe the two sources of information used The IDH dataset was constructed based on information available from the BNCF and official information on academics working in Italian universities provided by MIUR. Both sources provide a near-complete picture of the corresponding populations: Italian doctoral graduates are obliged to deposit a copy of their PhD thesis in the BNCF repository. The data are collected by public institutions following standardized procedures, which results in consistent records and, in turn, makes them suitable for research purposes. First, to create the IDH database, we extracted information on doctoral degree holders from the BNCF repository, and merged it with MIUR records. In what follows, we describe the data retrieval and creation processes and provide some descriptive statistics for these data sources.

2.1 Italian Doctorate Holders (IDH)

Doctoral training was provided in Italy much later than in other European countries: decree 382/1980, later modified by the Law 476/1984, was published in Italy in 1980. There were not substantial modifications to the doctorate legislation until 1998/1999 when Law 210/1998 was passed, making a doctoral degree mandatory for appointment as assistant professor at an Italian university. Law 224/1999 allowed Italian universities to offer doctoral training and to increase the number of PhD students by offering entry without a government scholarship (PhD grants are allocated by MIUR to the universities, which can accept offers from students with scholarships or who are able to fund themselves.) These reforms resulted in an increased number of doctoral courses (see Figure 2): between 1998 and 2006, the number of PhD programmes increased from 1,393 to over 2,177. Decree 124/2007 merged PhD courses into doctoral schools and introduced some specific requirements for new programmes. Then, Law 240/2010 required that doctoral programmes must be approved by the Italian National Agency for the Evaluation of University and Research Systems (ANVUR). These two actions led to a the decline of the number of courses and enrolments starting from 2008.

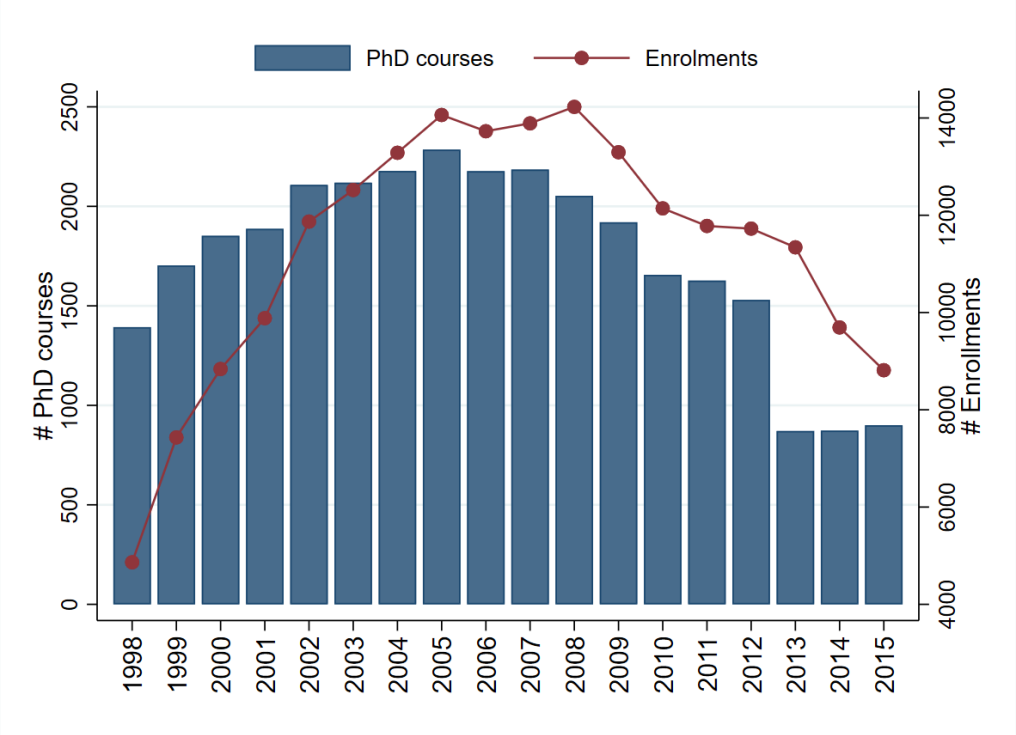

Figure 2: Number of doctoral courses, positions, scholarships and enrolments

There is no public central data recording Italian PhD students and doctoral graduates. Ministerial Decree 45/2013[5] makes MIUR responsible for organizing information on doctorates and doctoral theses, to enable investigation of career and occupational outcomes. The registry of doctorate holders, which is not publicly accessible and is used mainly administrative purposes, includes information starting from the academic year 2003/2004 on PhD enrolment and graduation. However, there is no obligation for university departments to provide or to keep information updated. Also, outcomes are not recorded for all students who are recorded as enrolled in a PhD programme, which makes it impossible to distinguish students who abandoned the PhD from those where the information is missing (i.e., students who graduated, but this outcome was not reported by the university).

However, since 1986, universities are legally obliged to deposit copies of students' doctoral theses at the BNCF.[6] These are catalogued and made accessible. Thus, this repository is more reliable, since data collection follows standardized procedures and the information start from the first cycle of PhD training. Also, it

[5] See: http://attiministeriali.miur.it/anno-2013/febbraio/dm-08022013-(1).aspx.

[6] In Section 2.1.1, we explain that DPR 252/2006 allows universities to choose between depositing doctoral theses with the BNCF repository or setting up their own open access repository which could be automatically harvested by BNCF, in which case, the theses are stored in a separate and not accessible repository. For more information, see: http://depositolegale.it.

is related only to doctorates. We therefore decided to use this source to retrieve data on Italian doctorate holders. The IDH database is the first to use this information.

2.1.1. Data Retrieval from BNCF

Using specific key query parameters to identify the doctoral theses, we harvested all the records from the BNCF Online Public Access Catalogue (OPAC), deposited since 1986, the first year that a PhD was awarded by an Italian university, to 2015. The BNCF OPAC allows downloads in XML format, but for one query at a time. Therefore, we constructed an automatic harvesting program to allow the collection of all of the theses in the database (more than 100,000). To reduce noise, we downloaded specific subfields of the records' XML structures, to collect information on author, year of registration in the BNCF repository,[7] university of affiliation, PhD cycle, thesis title, PhD supervisor/coordinator, scientific sector and additional notes. See Appendix Figure A1 for an example of the codification on a spreadsheet of XML tags for a thesis in the BNCF OPAC.

We further standardized and cleaned the information retrieved. For example, some XML tags were empty, but the missing information was recoverable from other tags (e.g., information on the grant awarding institution was attached to the thesis title, and, in some cases, information on PhD advisor or coordinator was recorded in the additional notes). We standardized university names. We associated a gender to the author names using the list of first names available in the MIUR database on Italian academics (see below for a description), which includes information on gender. In the case that we could not identify a match (i.e., the case of a foreign or uncommon Italian name), we conducted a manual web search. We grouped scientific field macro-areas based on the MIUR and Dewey classifications. If information on scientific field was missing (10% of cases), we classified the thesis on the basis of the dissertation title and the name of the doctoral programme. Finally, we cleaned the data, which identified several duplicate or incorrect records; these were deleted. See Appendix A for details of the cleaning process.

To check the robustness of our data collection process, we compared data extracted to number of doctorate holders by year, reported in the MIUR statistics. There are no official data related to the PhD cycles up to 1998. For these cases, we took the data from several studies (Brandi & Avveduto, 2004; Cipollone & Avveduto, 2004), that report MIUR statistics (starting in 1987) which are consistent with the IDH

[7] The year that the thesis was deposited with the BNCF might differ (by a few or several months) from the year that the thesis was successfully examined.

figures; for cycles in the period 1999-2014 we retrieved data on number of PhD graduates from the MIUR website.[8]

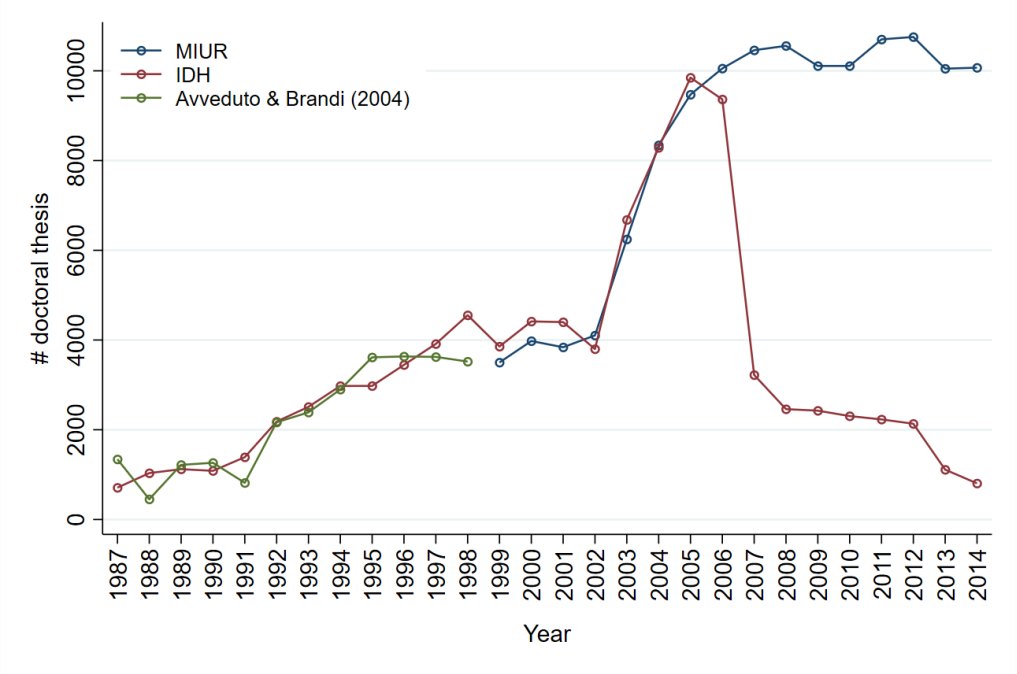

Figure 3: PhD holders (MIUR; Avveduto & Brandi, 2004) and PhD thesis (IDH) by year

Figure 3 shows that there was a substantial decline in the number of theses deposited in the national library of Florence after 2006. Following DPR 252/2006 and MIUR Communication 1746/2007, the BNCF launched a project aimed at automatically harvesting doctoral theses deposited in the open access repositories of participant universities. However, this was not a perfect substitute for the legal requirement to deposit a copy of the thesis with BNCF. Also, the BNCF OPAC only included theses sent by the universities and not those that had been harvested digitally; these latter were stored in a different, not publicly accessible repository. Therefore, to compare the IDH and the official statistics, we focus only on years 1987-2006 (there are no data available for 1986).

When we consider this period, we find small differences between the MIUR and IDH data. According to MIUR, the total number of PhD students who defended their theses in the period 1987-2006 was 76,473 whereas IDH records this number as 75,477, a 1.3 percentage points difference.

[8] We exclude so-called *istituti a ordinamento speciale* (institutes with a special statute): Normale of Pisa, Sant'Anna of Pisa, SISSA of Trieste, IMT of Lucca, IUSS of Pavia and SUM of Florence. Their PhD programmes are comparable to the doctoral courses offered by public and private Italian universities, but are not subject to requirement to deposit doctoral theses at the BNCF.

This small difference might be due to problems related to ministry data, which, for some of the years reported, did not receive information from some universities, or to the fact that the date on the thesis might vary with it cataloguing by BNCF (see fn. 7).

Differences between the IDH and MIUR data are reported by year (Table 1), university (Table 2) and field (Figure 4).

Table 1: Number of PhD holders (MIUR), PhD thesis (IDH) and differences by year

	IDH	MIUR	%Δ
1987	710	1342	47.1
1988	1034	452	-128.8
1989	1122	1218	7.9
1990	1085	1264	14.2
1991	1389	818	-69.8
1992	2182	2167	-0.7
1993	2511	2388	-5.2
1994	2976	2898	-2.7
1995	2978	3615	17.6
1996	3447	3634	5.1
1997	3914	3624	-8
1998	4553	3520	-29.3
1999	3856	3500	-10.2
2000	4415	3978	-11
2001	4400	3839	-14.6
2002	3799	4103	7.4
2003	6680	6246	-6.9
2004	8287	8343	0.7
2005	9344	9471	1.3
2006	6795	10053	32.4
TOT	75477	76473	1.3

The differences reported in Table 1 are indicative of the delay between PhD examination and registration in the BNCF OPAC. In 1988, BNCF registered 47.1% fewer theses than recorded by MIUR, while in 1991, BNCF includes 69.8% more theses than MIUR. The overall difference, across the whole period, is 1.3%. Table 2 shows the difference in the numbers of PhD theses in the IDH and MIUR data, by PhD granting institution. It is possible that some universities are more diligent about providing this information to BNCF. Also, some of the difference might be due to the way MIUR assigns the theses to universities if they operate as a consortium to offer PhD programmes (this applies to a few universities). The thesis title in the BNCF records shows the university that awarded the degree.

The trends related to scientific area of the theses, are similar in both the MIUR and the IDH data (see Figure 4). The small differences observed might be due to the fact that MIUR data take account of scientific disciplinary aggregations or Scientific Sector Disciplines (SSD) related to the doctoral courses, while IDH data include (i) MIUR SSDs, (ii) the Dewey Decimal Classification, which is a general classification system used by libraries, and (iii) the PhD program or the dissertation title. If the SSD is not reported in the BNCF record, we derived the scientific area from ii) or iii).

Table 2: PhD holders (MIUR), PhD thesis (IDH) and differences by institution – 1998-2006

University	IDH	MIUR	%Δ	University (cont'd)	IDH	MIUR	%Δ
Ancona	566	618	8.4	Napoli - "Parthenope"	154	167	7.8
Arcavacata di Rende	619	646	4.2	Napoli - "SOB"	21	16	-31.3
Bari - Politecnico	274	264	-3.8	Padova	2554	2195	-16.4
Bari	1931	1843	-4.8	Palermo	1756	1710	-2.7
Benevento	84	72	-16.7	Parma	819	792	-3.4
Bergamo	85	60	-41.7	Pavia	953	1060	10.1
Bologna	3311	3582	7.6	Perugia	970	946	-2.5
Brescia	277	398	30.4	Pisa	2055	2044	-0.5
Cagliari	478	567	15.7	Potenza	281	282	0.4
Camerino	195	193	-1	Reggio Calabria	364	371	1.9
Campobasso	192	237	19	Roma - III	612	556	-10.1
Cassino	80	151	47	Roma - LUISS	89	89	0
Castellanza	19	20	5	Roma - LUMSA	0	15	100
Catania	1953	2002	2.4	Roma - "La Sapienza"	3483	2665	-30.7
Catanzaro	59	55	-7.3	Roma - "Foro Italico"	0	6	100
Chieti e Pescara	613	591	-3.7	Roma - "Tor Vergata"	1238	1341	7.7
Ferrara	688	621	-10.8	Roma - UNINT	8	2	-300
Firenze	2301	2535	9.2	Salerno	633	672	5.8
Foggia	152	143	-6.3	Sassari	369	495	25.5
Genova	1569	1495	-4.9	Siena	1064	1188	10.4
L'Aquila	447	405	-10.4	Siena - Stranieri	38	35	-8.6
Lecce	542	561	3.4	Teramo	143	120	-19.2
Macerata	238	236	-0.8	Torino - Politecnico	991	982	-0.9
Messina	998	989	-0.9	Torino	1936	1746	-10.9
Milano - IULM	48	47	-2.1	Trento	509	540	5.7
Milano - Politecnico	1163	1231	5.5	Trieste	915	802	-14.1
Milano - "Sacro Cuore"	913	933	2.1	Udine	439	387	-13.4
Milano - Bocconi	237	217	-9.2	Urbino	436	486	10.3
Milano	2656	2359	-12.6	Varese	92	83	-10.8
Milano - San Raffaele	11	10	-10	Venezia - "C`a Foscari"	509	474	-7.4
Milano - Bicocca	378	332	-13.9	Venezia - IUAV	204	193	-5.7
Modena e R. Emilia	563	608	7.4	Vercelli	122	129	5.4
Napoli - Seconda	508	1053	51.8	Verona	379	388	2.3
Napoli - "Federico II"	3343	3760	11.1	Viterbo	20	247	91.9
Napoli - "L' Orientale"	201	414	51.4	TOT	51878	52472	1.1

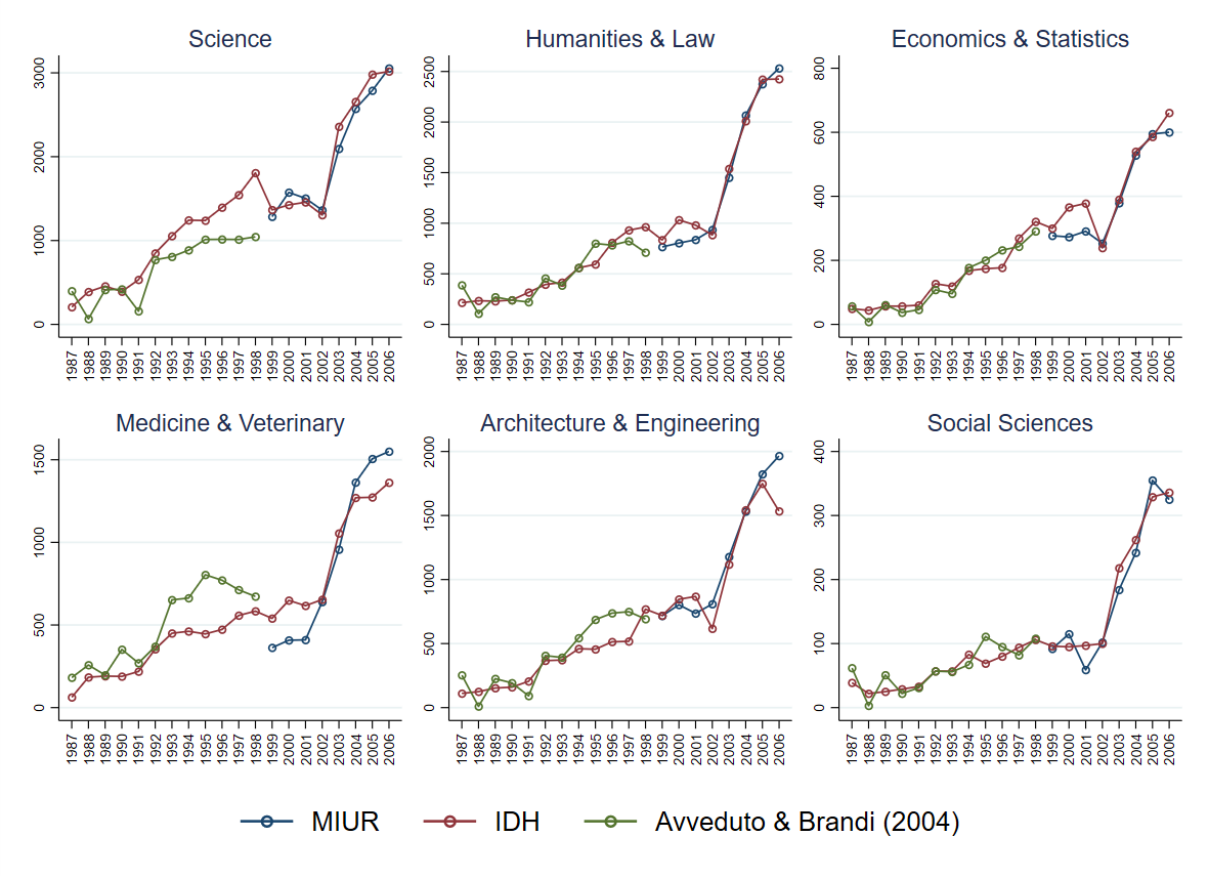

Figure 4: PhD holders (MIUR) and PhD thesis (IDH) by scientific area – 1987-2006

The total number of theses in the IDH database in the period 1986-2006 is 76,093. Missing information required further online searches: 2,022 records do not include university of affiliation and 235 do not indicate the scientific field. Table 3 describes the variables in the IDH database.

Table 3: Description of the variables in the IDH database

Variable	Description
ID	Unique identifier of the thesis
Name	Name of the author of the thesis
Gender	Gender of the author
Publication Year	Year in which the thesis was added to BNCF repository
University	PhD granting university of the thesis
Title	Main title of the thesis
Subtitle	Subtitle of the thesis
Notes	Other information contained on the front page of the thesis (name of doctoral course and PhD cycle)
Other author	Other authors contributing to the thesis (PhD supervisor, PhD coordinator)
MIUR Classification	Classification of thesis scientific area based on MIUR codes
Dewey Decimal Classification	Classification of scientific area thesis using Dewey Decimal bibliographic classes
Macroarea	Macro scientific area of the thesis (Science, Architecture & Engineering, Medicine & Veterinary, Humanities & Law, Economics & Statistics, Social Sciences)

2.2 Italian Academic Careers (AC)

We also collected biographic and career information on academics working in Italian universities.

In 2015, the Italian academic system included 96 universities (30 private, 66 public) and 8 higher education institutions. Each professor working at an Italian university is categorized by academic position (assistant, associate, full professor) and one scientific discipline (SSD). Information on academic position, disciplinary area and university affiliation for all Italian academics in the period 1990-2015 was obtained from MIUR. Figure 5 summarizes the numbers and share of Italian professors by academic position, during this period.

In 1990, there were 42,209 professors active in Italian universities. In the time period considered. 45,795 academics entered the academic system and 33,219 exited, in 2015 there were 54,785 professors working in Italian universities. The maximum number of academics was in 2008, with almost 63,000 researchers, 40% at assistant professor level. After the 2008 recession, numbers of full professors decreased continuously to the end of the period; numbers of associate professors started to increase slightly after 2013 to reach more than 20,000 in 2015.

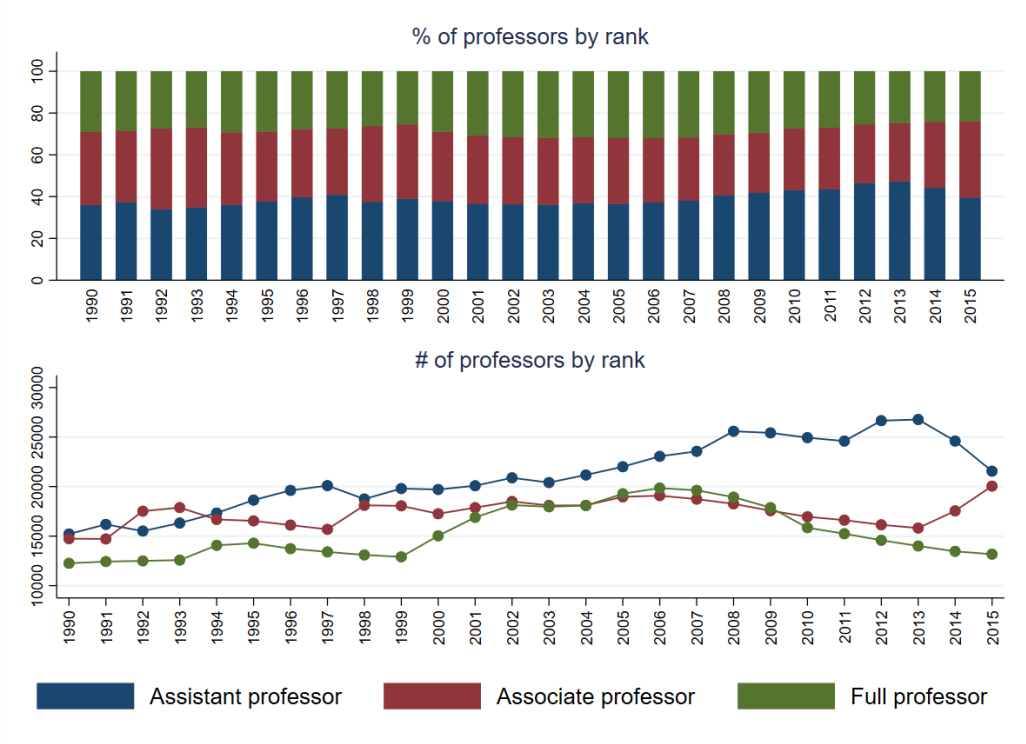

Figure 5: Share and number of professors by academic position (Source: MIUR)

These data have several problems (Checchi et al., 2014). Some are related to use of the same ID to identify an academic who entered the data, then for a period was not included, but then took up another academic position in Italy. There are also some missing data for academic discipline and institution in the first five years of the data (1990-1995). We manually searched on academics' personal pages which identified some of the data.

All professors working in public universities in Italy are tenured civil servants, and those working in private universities are tenured and are included, for administrative purposes, in the Ministry's list, which makes MIUR data representative of the whole population of tenured professors working in Italian public and private academic organizations. Although there are many postdoctoral researchers and researchers working in other types of temporary contract positions in Italian universities, similar to what occurs in other countries, there is no reliable or accessible sets of data on these individuals.

Table 4 presents the variables in the AC database. Notice that information on professor position and affiliation allow for panel data or pooled cross section analysis of academic careers.

Table 4: Description of the variables in the AC dataset

Variable	Description
ID	Unique identifier of the academic
Name	Name of the academic researcher
Gender	Gender of the academic
Birth Year	Year of birth of the academic
Calendar Year	Calendar year of reference (panel dimension: 1990-2015)
SSD	Disciplinary scientific sector of the academic
University	University employing the academic
Department	Department employing the academic
Rank	Rank (assistant, associate or full professor)

3. Matching Doctorate Holders and Academics

In this section, we describe the record linking between the IDH and the AC datasets and provide some basic information on the final matched IDH-AC database, which includes Italian doctorate holders in a tenured academic career in an Italian university.

We performed the matching based on four fields: name, gender, scientific area and year degree was awarded. We ran the first matching based on the researcher's full name (firstname + middlename, e.g. "Anna Maria"), and then re-ran it to account for unmatched records from the first exercise, by considering only first name (i.e., following the previous example, "Anna"). We allowed for some level of mismatch between scientific area and academic position, since having studied for a PhD in one field the academic might be recruited to work in a different, but related field (e.g., a PhD degree in Social Sciences could translate into an academic position in Economics & Statistics, but not in Medicine & Veterinary). We considered a five year time window between date of deposit in the repository and the expected year of dissertation based on researchers' birth year. We describe this in further detail below.

To identify academics with a doctoral degree awarded by an Italian university, we matched academics in the AC dataset with doctoral graduates from Italian universities in the IDH database. The matching procedure was conducted as follows:

- A "narrow" matching based upon academics' and doctoral graduates' full names, gender, scientific area and year of PhD degree award.
- A "broad" matching, to capture academics and theses not matched in the first-step, on first name (i.e. with middle names excluded), gender, discipline and year of PhD degree award.
- A "filtering" procedure to eliminate incongruous academic-doctoral graduate matches.

Figure 6 provides an overview of the record linking approach to construct the IDH-AC database.

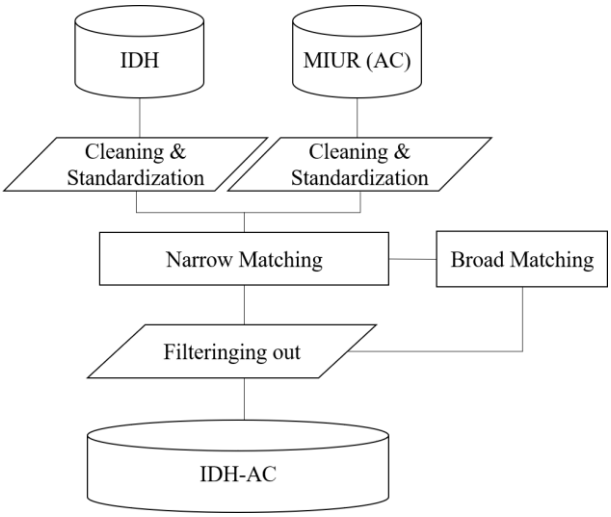

Figure 6: Overview of the data processing and record linkage procedure

First, we took account of the specificities of the Italian language and conducted the following standardization and cleaning procedure on names and surnames:

- accents on the last letters of names were changed into apostrophes; accents on other letters in the name were deleted. For example: "FOÀ Sergio" becomes "FOA' Sergio" and "CALDÉS PINILLA Maria" becomes "CALDES PINILLA Maria".
- names indicated in the IDH dataset by only initials, were corrected to full names based on a manual online search (e.g. "M. Angela" was changed to "Maria Angela").
- All non-alphanumeric characters (e.g., characters "-" between composite names) were removed.

The original "surname+firstname+middlenames" strings were placed in separate fields for each element, i.e. "surname", "firstname+middlenames" and just "firstname".

To take account of possible typos in the names included either of the two datasets, we used fuzzy string matching algorithms[9]. We treated names as raw strings and computed the edit distance, that is, how many operations (inserting/deleting a character, switching two characters next to each other, etc.) had to be applied to the first string to convert it to the second strong. Having specified approve and disapprove levels, we calculated the score of the function using formula (1), where

[9] All algorithms used were implemented using FRIL software, see Jurczyk et al. (2008).

str_A is the first string, str_B is the second string, $e(str_A, str_B)$ is the edit distance between the two strings, a is the approve level and d is the disapprove level. Table 5 presents some examples.

$$score_{str_A,str_B} = \begin{cases} 0, if\ e(str_A, str_B) > d \cdot \max\left[lenght(str_A), lenght(str_B)\right] \\ 1, if\ e(str_A, str_B) > a \cdot \max\left[lenght(str_A), lenght(str_B)\right] \\ \dfrac{d \cdot \max[lenght(str_A), lenght(str_B)] - e(str_A, str_B)}{(d-a) \cdot \max\left[lenght(str_A), lenght(str_B)\right]}, oherwise \end{cases} \quad (1)$$

Table 5: Examples of score values for name matching

str_A	str_B	$e(str_A, str_B)$	a	d	$score$
Falini Andrea	Fallini Andrea	1	0	0.3	0.78
Maliocco Giovanni	Maliocco Giovanna	1	0	0.3	0.82
Piva Paolo	Pizza Paolo	2	0	0.3	0.44
Andrea Verdini	Andera Vedrini	2	0	0.3	0.83

To prevent incongruent matching of a male academic to a female doctorate holder with a similar name (e.g.: Paolo and Paola, or names which could be either masculine or feminine in Italian, such as Andrea), and vice versa, we also accounted for gender.

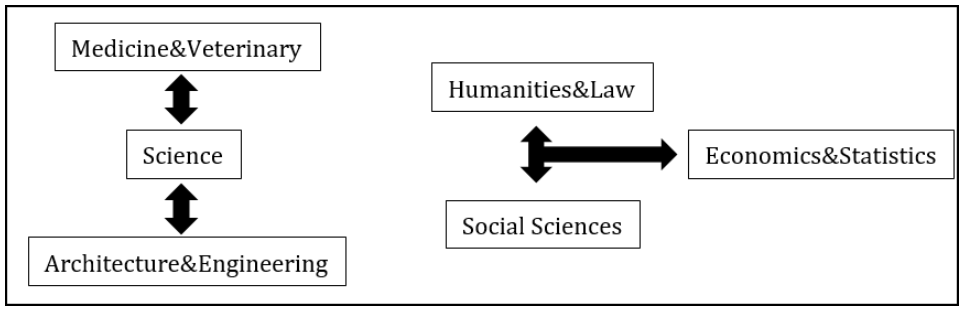

Figure 7: Scientific proximity clusters for the record linkage

We grouped academics and doctorate holders into six disciplines: Medicine & Veterinary, Science (agriculture, biology, chemistry, physics, geology and mathematics), Architecture & Engineering, Humanities & Law, Social Sciences and Economics & Statistics. We grouped the first three and the last three into two "scientific proximity" groups and allowed correspondence between discipline of the PhD thesis and the broad scientific area of the academic, to fall into one of these two groups (Figure 7). In other words, we took account of the fact that studying for a PhD in Economics & Statistics could result in a position in the field of Social Sciences, but not in Medicine & Veterinary for example.

Finally, we took account also of the year of PhD graduation. In IDH we proxy this information using the year of publication of the thesis in the BNCF repository. On the other hand, in AC we do not have the year of PhD of the academics, nor we know whether they actually have a PhD, either from an Italian or a for4eign university. Thus, in this second case, we rely on the birthyear of the academics in AC, in order to compute a "probable" year of PhD graduation. In Italy, individuals enter the education system at the age of 6 and they have 21 years of primary and secondary education before being awarded a doctorate, hence PhD graduation happens nominally at the age of 27. We considered that a researcher could have obtained a PhD in a time window ranging between the 26th and the 32rd year of age. The final score is calculated using formula (2). Table 6 presents some examples: v_1 is the thesis publication year from IDH and v_2 is the "probable" PhD graduation year, computed from the birthyear from AC.

$$score_{v_1,v_2} = \begin{cases} \dfrac{upperBound(v_1) - v_2}{upperBound(v_1) - v_1}, if\ v_2 \in [v_1, upperBound(v_1)] \\ \dfrac{v_2 - lowerBound(v_1)}{v_1 - lowerBound(v_1)}, if\ \ v_2 \in [lowerBound(v_1), v_1] \\ 0, otherwise \end{cases} \quad (2)$$

Table 6: Examples of score values for PhD year matching

v_1	v_2	$lowerBound(v_1)$	$upperBound(v_1)$	$score$
1996	1999	1995	2001	0.5
1996	1997	1995	2001	0.83
1996	1994	1995	2001	0
1996	2001	1995	2001	0.17

We gave each field a weight (see Figure 8), two records were linked if the final confidence score was 81 or higher. If a record from one source was linked to more than one record in the other source, the link with the higher confidence score was retained.

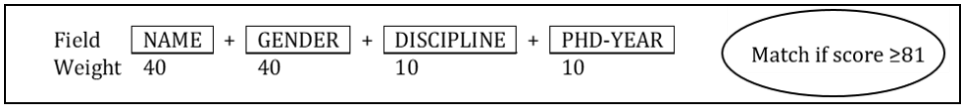

Figure 8: Weight formula for the record linkage

The record linking was performed between the IDH database of doctoral dissertations for all theses deposited in the period 1986-2006 in all disciplines (76,093 theses), and the AC database of academic careers. Since the AC database

does not provide information on PhD education, we consider only researchers who were likely to be awarded a doctoral degree in the period 1986-2006, computing their "probable" PhD graduation year based on their year of birth. In other words, we selected all academics born between 1954, who were 32 years old in 1986, and 1980, who were 26 years old in 2006 (47,734 academics). Here, we take account the scientific field at their first appointment (4.4% of academics changed disciplines in the course of their scientific career).

At the end of record linking procedure we were able to match 27,145 records. If the algorithm identified more than one entry in the IDH database which matched one entry in the AC dataset (or vice versa), we excluded the respective cases from the final dataset. Hence, the final IDH-AC dataset includes 25,412 individuals, and the overall matching yielded 33% of the theses and 53% of the academics considered.

Table 7 presents the number and percentage of linked records per linkage confidence score: only 14% of linked records have a confidence score lower than 90. Appendix B provides additional summary statistics and record linkage results.

Table 7: Number and percentage of the linked record per confidence level

Confidence	Frequency	Percent	Cumulate
81	348	1.3%	1.3%
82	295	1.1%	2.4%
83	1,309	4.8%	7.2%
84	91	0.3%	7.5%
85	448	1.7%	9.2%
86	453	1.7%	10.9%
87	53	0.2%	11.1%
88	750	2.8%	13.8%
89	40	0.1%	14%
90	8,551	31.5%	45.5%
91	507	1.9%	47.4%
92	28	0.1%	47.5%
93	714	2.6%	50.1%
94	4	0%	50.1%
95	3,969	14.6%	64.8%
96	2,658	9.8%	74.6%
98	3,457	12.8%	87.3%
100	3,432	12.7%	100%

3.1 Matching Validation

Comparing the results of the linking procedure, between the two sources was not straightforward. There are no official statistics or data on the percentage of Italian doctoral graduates in a permanent position in Italian academia. We therefore were

forced to rely on ad-hoc studies, carried out by a few universities and research institutions, on the employment outcomes of (some cohorts of) Italian PhD graduates. We wanted to find more information on Italian doctorate holders working in the Italian academic system in order to compare our results. However, some studies investigate only whether graduates from a particular PhD cycle were employed or not (hence, it is not possible to distinguish those who dropped out academia); also, even where sector of employment was reported, they do not always provide information on doctorate holder's country of residence (i.e., we cannot identify PhDs who took up an academic career outside of Italy). Table 8 summarizes some of the most recent studies on this topic. It shows that the 33% resulting from our matching is in line with the findings in these studies on the employment outcomes of Italian PhD graduates which show that around a third of doctoral graduates from an Italian university are employed in Italian academia.

Table 8: Summary of previous studies on Italian PhDs career outcomes

Study	Details	Main findings
Schizzerotto, 2006	University(/ies): Milano, Milano-Bicocca, Trento. Cohort(s): 1998-2005. Methodology: census, CATI, 1179 respondents (66%).	97.1% employed, 58.5% in a public or private university, of which 21.1% are assistant, associate or full professors and 37.4% have a non-permanent position in academia.
Brait et al., 2009	University(/ies): all. Cohort(s): 2004, 2006. Methodology: census, CAWI+CATI.	93.5% employed, 17.3% on a post-doc scholarship or in a temporary position in academia. 6.9% live in a foreign country.
Romano, 2009	University(/ies): Bergamo, Brescia, Milano, Milano-Bicocca, Palermo, Pisa, Pisa-Sant'Anna. Cohort(s): 2005, 2006, 2007. Methodology: census, CAWI, 1758 respondents (49.5%).	85.5% employed,81.6% of them working in Italy. 40% employed in a public or private university, but country and type of contract not specified.
Romano, 2010	University(/ies): Bergamo, Brescia, Milano, Milano-Bicocca, Palermo, Pisa, Pisa-Sant'Anna. Cohort(s): 2007, 2008. Methodology: census, CAWI, 1579 respondents (53.4%).	75.1% employed, 82.4% of them working in Italy. 33% employed in a public or private university, but country and type of contract not specified.
Romano & Himmelmann, 2011	University(/ies): Bergamo, Brescia, Milano, Milano-Bicocca, Palermo, Pavia, Pisa, Pisa-Sant'Anna. Cohort(s): 2008, 2009. Methodology: census, CAWI, 1637 respondents (50.7%).	77.8% employed, 86.8% of them in Italy. 34.6% employed in a public or private university, but country and type of contract not specified.
Romano, 2012	University(/ies): Bergamo, Brescia, Milano, Milano-Bicocca, Palermo, Pavia, Pisa, Pisa-Sant'Anna. Cohort(s): 2009, 2010. Methodology: census, CAWI, 1769	81.9% employed, 90.3% of them in Italy. 32.4% employed oin a public or private university, but country and type of contract not specified.

	respondents (51.1%).	
Sfamurri et al., 2012	University(/ies): Parma. Cohort(s): 2006, 2007, 2008, 2009. Methodology: census, CAWI, 378 respondents (48.2%).	97.3% employed,45.5% in a public or private university. Among them, 19.2% are either assistant, associate or full professors in Italy while 7.6% work in a foreign university.
Bonito, 2013	University(/ies): LUISS. Cohort(s): 2007, 2009. Methodology: census, CATI, 58 respondents (83%).	100% employed including 31% employed in a public or private Italian university, including 20.7% with a non-permanent position in academia.
Giambalvo & Montalbetti, 2013	University(/ies): Bergamo, Brescia, Milano, Milano-Bicocca, Palermo, Pavia, Pisa. Cohort(s): 2010, 2011. Methodology: census, CAWI+CATI, 1938 respondents (60.1%).	87.5% employed, 91.8% in Italy. 29.3% employed in a public or private university, but country and type of contract not specified.
Bergamante et al., 2014	University(/ies): all. Cohort(s): 2006. Methodology: survey, CAWI+CATI, 4879 respondents (48.7%).	92.5% employed, 18% on a post-doc scholarship or in a temporary position in academia. 7.9% live in a foreign country.
Bonini, 2014	University(/ies): Bergamo, Brescia, Milano, Palermo, Pisa, Pavia. Cohort(s): 2011, 2012. Methodology: census, CAWI+CATI, 1537 respondents (55%).	80.3% employed, 92.8% in Italy. 24.8% employed in a public or private university, but country and type of contract not specified.
Gallo et al., 2014	University(/ies): all. Cohort(s): 2008, 2010. Methodology: census, CAWI+CATI.	92.4% employed, 36.5% on a post-doc scholarship or in a temporary position in academia. 12.9% live in a foreign country.

To provide a more precise evaluation of the results of the record linking, we use biographical information retrieved from ORCiD profiles. In fact, we dispose of the ORCiD for 12,553 academics from MIUR data.[10] Among these 12,553, 1,642 profiles include information on PhD training between 1980 and 2003[11] and, thus, are comparable to the results of the algorithm.

We merged the ORCiD profiles with the IDH-AC data and computed the number of true positives, false positives and false negatives, in order to calculate the commonly used metrics to evaluate algorithm results, that it, precision (to evaluate type I errors, or false positives), recall (to evaluate type II errors, or false negatives), accuracy and F1 statistics. Table 9 reports all these evaluation metrics,

[10] This information has been downloaded through Scopus API.

[11] ORCiD profiles include education starting dates; we filtered the years, using 1980 as the lower bound, the first year of PhD training in Italy, and using 2003 as the upper bound, on the assumption that those graduating in 2006 began their doctoral training at least 3 years earlier.

which are all above or around 0.9, demonstrating the high reliability of the matching algorithm.

Table 9: Precision, recall, accuracy and F1 statistics

True positives	False positives	True negatives	False negatives
1263	45	206	128
Precision	Recall	Accuracy	F1
0.966	0.908	0.895	0.936

We are aware that a limitation of ORCiD is that the information is self-reported, so it is possible that some bias or errors might arise if it is used to evaluate the results of the algorithm. Therefore, to further validate our matching algorithm we performed a detailed case analysis using information on academics in physics and chemistry employed at University of Turin and the Polytechnic of Turin, which can be considered appropriate cases.

There were 231 Italian academics working in chemistry and physics in these two Turin universities, in 2015, including 173 who graduated in the period 1986-2006.

The narrow matching exercise resulted in 127 matches and the broad matching resulted in 12 additional academic-doctoral graduate pairs. The filtering out exercise resulted in 139 unique (i.e., one-academic-to-one-doctoral graduate) matches.

We manually checked the CVs of the remaining academics:

- 17 did not have a PhD degree;
- 7 had been awarded a PhD by a foreign university;
- 4 had a PhD degree awarded by a private Italian institution with no legal obligation to deposit doctoral dissertations in the BNCF repository;
- 4 received their doctorate from an Italian university, but their theses were not in the BNCF repository;[12]
- 2 were not matched by the algorithm.

Thus the matching strategy correctly identified 93% of the academics awarded a PhD degree by an Italian university (139 matched by the algorithm over 173-17-7=149 total academics with an Italian PhD degree).

At the end of the matching process, the resulting IDH-AC database combined information from the two sources described in section 2. It includes 338,840

[12] This is due to the small imbalance between date of thesis defence and date of deposit in the BNCF (see Section 2.1).

observations, corresponding to 25,412 individuals, observed over 13 years, on average, between 1990 and 2015. Table 10 reports variable name, description and source in IDH-AC database.

Table 10: Name, description and source of the variables in IDH-AC database

Variable	Description	Source
ID Thesis	Unique identifier of the thesis	IDH
Name	Name of the academic researcher	IDH
Gender	Gender	IDH
Publication Year	Year in which the thesis was added to BNCF repository	IDH
PhD University	PhD granting university	IDH
Title	Main title of the thesis	IDH
Subtitle	Subtitle of the thesis	IDH
Notes	Other information contained on the front page of the thesis (title of the doctoral course and PhD cycle)	IDH
Other author	Other author responsible for the thesis (namely the PhD supervisor or the PhD coordinator)	IDH
MIUR Classification	Classification of the scientific area of the thesis using MIUR disciplinary codes	IDH
Dewey Classification	Classification of the scientific area of the thesis using the Dewey Decimal bibliographical classes	IDH
Macroarea	Macro scientific area of the thesis (Science, Architecture & Engineering, Medicine & Veterinary, Humanities & Law, Economics & Statistics, Social Sciences)	IDH
ID Prof.	Unique identifier of the academic	AC
Birth Year	Birth year of the academic	AC
Calendar Year	Calendar year of reference (panel dimension: 1990-2015)	AC
SSD	Disciplinary scientific sector of the academic	AC
University	University of employment of the academic	AC
Department	Department of employment of the academic	AC
Rank	Rank of the academic (assistant, associate or full professor)	AC

4. Italian Doctorate Holders' Academic Careers in Italy: Stylized facts from the IDH-AC database

In this section, we conduct a preliminary analysis of the main stylized facts related to the Italian academic job market for PhD degree holders. We use the IDH-AC database to analyse a small set of questions related to the labour market outcomes of doctoral graduates in Italy. We focus on the characteristics studied most

frequently in the literature on academic careers: gender, promotion, inbreeding and mobility.

Much of the research in the economics of science focuses on the determinants of scientific productivity, but tends to ignore careers and mobility - perhaps because both are assumed to be linked closely to productivity (Allison & Long, 1990; Long et al., 1993). However, although it might be expected that academic career advancement is based on scientific merit, there is evidence that this is not the case. In many countries, merit is not the only driver of promotion. Long et al. (1993) show that seniority and gender are equally if not more important; similarly, following the seminal work of Crane (1965, 1970), it has been shown that hiring can rely on the prestige effect and favour graduates from top institutions. There can also be an inbreeding effect, resulting in the institution hiring from among its own graduates (see Horta et al., 2009).

Early work on academic career paths refer to the Anglo-Saxon countries, which have a very specific academic labour market (Long et al., 1979, 1993; Long & Fox, 1995). Since the institutional set-up is different in continental Europe, many studies focused on the specific case of EU countries (see Musselin 2004, 2005 for an international analysis).

For instance, Combes et al. (2008) focus on publication profiles and network connections in the procedure for hiring economists in France; Heining et al. (2007) investigate time-to-entry and time-to-promotion in the academic labour market for economists in Germany; Cruz-Castro & Sanz-Menéndez (2010) examine the relationship between scientific performance and rewards, for academic scientists in Spain. There are a few studies that focus on the Italian academic labour market: (Pezzoni et al., 2012) compare the career advancement of academics in physics and chemistry in France and Italy; and Carrozza & Minucci (2014) investigate the mobility and career decisions of early-stage researchers in the field of social sciences and humanities in Italy, based on in-depth interviews .

Most of this body of work focuses on a specific scientific field or cohort of researchers, in order to study their labour market outcomes. Systematic information on academic careers by subject are scarce; data are available only from ad-hoc surveys. This highlights the advantages of the IDH-AC dataset: i) it provides micro-level data on the population of researchers in all disciplines; and ii) follows their careers since award of the PhD degree for up to 30 years. This makes IDH-AC a valuable and rare source of information for researchers in economics and sociology of science.

4.1 Gender gap in academic careers

In their seminal study, Long et al. (1993) highlight the gender gap and the lower chances of female scientists achieving a high rank position in a university. The

study of gender differences in career paths has attracted the attention of both sociologists and economists (see, e.g,, Ginther & Kahn, 2004; Kahn, 1993; Levin & Stephan, 1998; McDowell et al., 2001).

Table 11: Share of female researchers by rank and scientific area

Scientific area	PhD	Assistant Prof.	Associate Prof.	Full Prof.
Arch&Eng	31.6	30.6	24.3	12.3
Econ&Stat	44.7	47.9	40.8	20.6
Hum&Law	54.4	53.9	48.9	30.9
Med&Vet	58.1	55.4	39.4	25.2
Science	50.1	51.5	41.4	21.6
SocSci	47.4	48	37.5	27.4
Total	48.6	48.4	39	22.3

Notes: Column "PhD" draws from IDH database.

Table 11 column 1 shows that the total share of female PhD degree holders in the whole IDH database is 48.6%. Women represented more than half of the academics in Medicine & Veterinary and Humanities & Law and Science while Architecture & Engineering is the scientific field with the biggest gender imbalance. Columns 2 to 4 show the share of women by academic rank in the IDH-AC database: assistant professor, associate professor and full professor. The gender gap increases with higher academic ranking, that is, for associate and full professor positions, while gender composition by scientific area among assistant professors is largely the same. The gap is even more pronounced for Architecture & Engineering (and STEMM[13] fields in general), especially at full professor level; while in social sciences and humanities, in particular Humanities & Law, the gender gap is smaller.

Table 12: Share of PhDs who pursued an academic career in Italy, by gender and scientific area

Scientific area	M	F
Architecture & Engineering	42.6	31.1
Economics & Statistics	52.2	41.8
Humanities & Law	41.3	31.1
Medicine & Veterinary	31.5	19.2
Science	38	28.5
Social Sciences	47.1	34.6
Total	40.6	29.2

[13] Science, Technology, Engineering, Mathematics and Medicine.

Table 11 shows that almost half (48.6%) of doctoral graduates in the IDH database are females; we looked at the gender distribution by scientific field and career stage. It is possible that there are gender differences in choice of academic career across different disciplines. Table 12 shows the share of male and female PhD degree holders who pursued an academic career in Italy, overall and for each scientific area. For instance, among all male PhD degree holders, 40.6% are employed in an Italian university compared to 29.2% for females. The gap between these two shares is around 10 percentage points across all fields, and is highest in Medicine & Veterinary and Social Sciences (respectively 12.3 and 12.5 percentage points). While we cannot exclude the possibility that some female doctoral graduates continued their academic careers abroad, Table 12 shows that female doctoral graduates are less likely to pursue an academic career in the Italian system.

4.2 Seniority and promotion

In academic jobs, career progress takes time: seniority is rewarded by promotion (Sanz-Menéndez et al., 2013). The time spent by a scientist in a given academic rank has been shown to be one of the most important factors determining the chances of promotion (Long et al., 1993; Modena et al., 1999), either directly (more senior researchers stand a higher chance of being promoted, ceteris paribus) or indirectly, based on scientific production (more senior scientists have longer to accumulate publications, which can drive promotion). Since the IDH-AC database provides information on researchers since their PhD training, we can calculate both time-to-entry (from PhD degree award to first appointment as assistant professor) and time-to-promotion (to associate and full professor) in Italian academia.

Table 13 - Time to reach different career stages by scientific area

Scientific area	PhD - Assistant Prof.	Assistant Prof.- Associate Prof.	Associate Prof. - Full Prof.
Arch&Eng	4.00	7.62	6.89
Econ&Stat	3.48	6.86	6.04
Hum&Law	4.73	7.21	5.73
Med&Vet	5.07	8.59	7.49
Science	4.63	8.89	7.16
SocSci	5.44	7.03	6.90
Total	4.50	7.90	6.56

Table 13 column 1 shows the average length of the postdoctoral period, that is, the number of years between doctoral award to first appointment as assistant professor. On average, it takes four years for a new doctoral graduate researcher

to obtain a permanent position in Italian academia. The waiting time is longer (5+ years) for Medicine & Veterinary and Social Sciences fields, while in Economics & Statistics and Architecture & Engineering areas the post-doc period can be around three years. Columns 3 and 4 show that the time to be appointed associate or full professor differs significantly across scientific areas. In social science and humanities fields, promotion is faster than in STEMM fields, in general, and in Economics & Statistics which has the shortest times for entry and promotion. Table 14 shows that the average age of academics in the field of Economics & Statistics is younger compared to other disciplines, at every given career stage.

Table 14 - Age at career stage by scientific area

Scientific area	PhD	Assistant Prof.	Associate Prof.	Full Prof.
Arch&Eng	30.8	31.9	37.4	44.0
Econ&Stat	29.9	30.9	36.2	41.9
Hum&Law	31.2	33.2	38.4	43.8
Med&Vet	33.9	34.5	40.8	47.4
Science	31.0	31.7	38.0	44.8
SocSci	31.5	35	40.0	46.4
Total	31.0	32.3	37.9	44.1

Since the IDH-AC database covers academic careers from 1986 to 2015, we can study how entry and promotion evolve during the period of observation. Figure 9 presents the average differences in years between time to entry and promotion for two cohorts of doctorate holders: those who graduated in 1986-1996 and those who graduated in 1997-2006. It can be seen that Humanities & Law, Medicine & Veterinary and Social Science graduates achieve tenure more quickly. This applies, also, to promotion from assistant to associate professor position. Architecture & Engineering is the only scientific field where the average time related to each career stage was takes relatively longer for the first cohort considered.

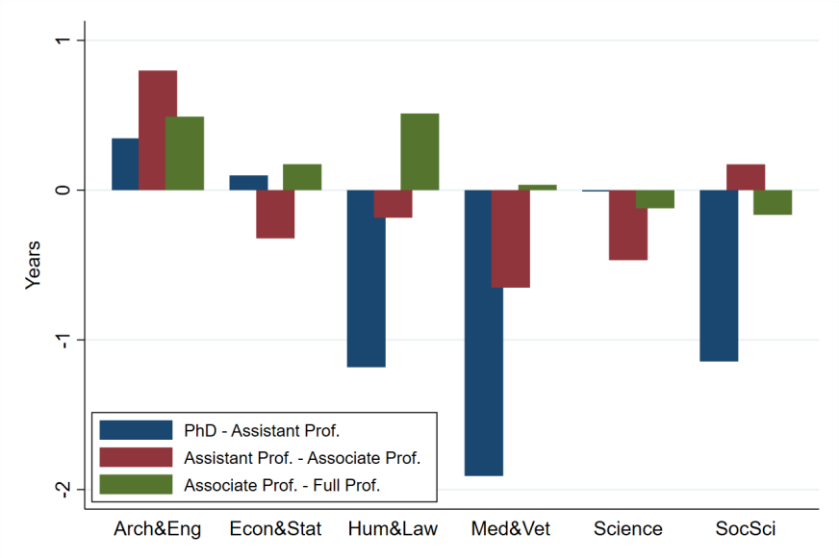

Figure 9: Average difference for time-to entry and promotion between 1986-96 and 1997-06 cohorts of doctorate holders

4.3 Inbreeding

Academic inbreeding was common in the US until the late 1970s (Hargens & Farr, 1973; Hargens & Hagstrom, 1967), and remains substantial in many countries in Europe, at least at the beginning of the academic career (Horta, 2013; Horta et al., 2009). Godechot (2016) shows that, in France, during the 1980s, inbred doctoral graduates were 17 times more likely to be hired by the awarding university than outbred graduates. IDH-AC provides information on the period since researchers' doctorate degree award, which allows us to identify Alma Mater and explore the effect of inbreeding at different career stages.

Table 15 – Share of inbred academics by rank and scientific area

Scientific area	Assistant Prof.	Associate Prof.	Full Prof.
Arch&Eng	58.0	51.0	41.0
Econ&Stat	34.0	27.1	23.0
Hum&Law	33.0	26.3	20.1
Med&Vet	53.8	45.2	32.4
Science	57.8	51.1	39.2
SocSci	34.6	24.4	27.4
Total	48.6	41.0	31.3

Table 15 column 1 shows that almost half of Italian PhDs are hired as assistant professors in their Alma Mater and that this is more common in STEMM fields, particularly Architecture & Engineering and Science. The overall percentage of inbred faculty at associate and full professor levels is lower (41% and 31.3% respectively), but remains more common in STEMM fields. Among social science

and humanities, Social Sciences mostly accounts for the highest percentage of inbred academics, at every career stage, with the exception of Economics & Statistics and associate professor positions.

We explore the dynamics of inbreeding in Figure 10, which shows the share of inbred academics at every career stage by scientific area and doctoral training cohort (1986-96 and 1997-06). We see that, for every discipline, inbred shares for assistant and associate professor positions are higher among academics in the most recent PhD cohort. The reverse applies to the share of inbred full professors: this is stable or lower for the most recent cohort, for among all scientific fields except Social Science, where it increases from 23% to 38% from the first to the second period.

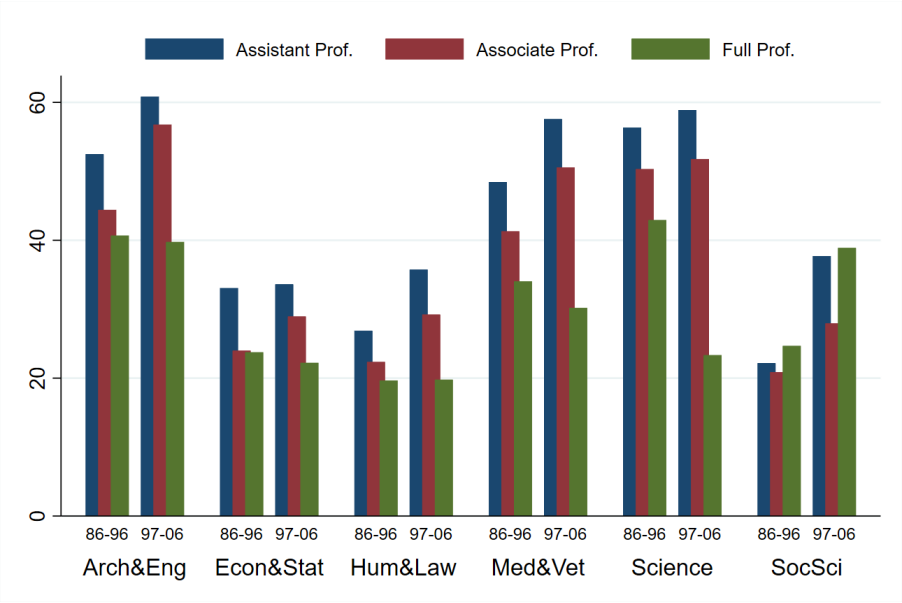

Figure 10: Share of inbred academics by rank and area in 1986-96 and 1997-06 cohorts of doctorate holders

4.4 Job Mobility

Job mobility is considered an important characteristic because it enhances knowledge circulation and contributes to the well-functioning of the academic market (Geuna, 2015). However, the effect of mobility on career advancement seems to depend on the structure of the academic system. Mobility can have a positive effect on the academic career by providing access to a bigger network of individuals, which can increase scientific productivity (Jonkers, 2011). However, it can have a negative effect. Mobile researchers, especially in case of young scientists in their post-doctoral period, might experience difficulty integrating in the local environment (Cruz-Castro & Sanz-Menéndez, 2010; Melin, 2005).

Table 16 - Number of changes of affiliation since PhD degree award

Changes of affiliation since doctorate	Freq.	Percent	Cum.
0	10,851	42.7	42.7
1	11,509	45.3	88.0
2	2,413	9.5	97.5
3	520	2.0	99.5
4	98	0.4	99.9
5	19	0.1	99.9
6	2	0.0	100
Total	25,412	100	

Table 16 show the number of changes of affiliation since award of the doctoral degree, for researchers in the IDH-AC database. Only 12% of the academics are is "highly mobile", that is, more than one affiliation during the course of the academic career; 45.3% moved once while 42.7% are considered "immobile" who stayed at the same university throughout their careers.

Table 17: Share of academics by number of changes of affiliation since PhD and scientific area

Scientific area	Changes of affiliation		
	0	1	≥2
Architecture & Engineering	48.7	38.9	12.4
Economics & Statistics	25.8	52.6	21.5
Humanities & Law	26.8	53.7	19.5
Medicine & Veterinary	48.1	42.3	9.6
Science	49.0	37.7	13.3
Social Sciences	30.0	56.5	13.5
Total	40.9	44.2	14.9

Table 17 presents the spread of these categories among scientific areas. Column 1 shows a higher percentage of immobile academics in STEMM fields (nearly 50%) compared to ca. 26%-30% of researchers in social science and humanities scientific areas. The higher mobility of social science and humanities fields is confirmed in columns 2 and 3, where we can see that more than 50% of the academics in these scientific areas changed affiliation at least once during their career. Academics in Economics & Statistics and Humanities & Law are the most mobile: 21.5% and 19.5%, respectively, moved twice or more since the PhD.

Table 18 explores the career stage related to mobility. Column 1 shows that the most common mobility, across all disciplines but especially for social sciences and

humanities, is the move from the PhD degree awarding institution to the university of first appointment.[14] Columns 2 to 4 show that, on average, 20.5% of the academics are mobile at the assistant professor level, 8.7% at associate professor level and 7.8% at full professor level. These percentages are higher for Economics & Statistics and Humanities & Law, which confirm that they are the scientific fields with the highest incidence of mobility among academics across all career stages.

Table 18: Share of mobile academics by rank and scientific area

Scientific area	PhD	Assistant Prof.	Associate Prof.	Full Prof.
Arch&Eng	43.0	21.6	7.1	5.0
Econ&Stat	67.2	43.7	15.3	10.1
Hum&Law	68.8	27.7	11.6	11.2
Med&Vet	47.5	8.7	6.1	7.6
Science	43.3	15.6	6.4	5.6
SocSci	66.8	13.5	10.1	6.2
Total	52.9	20.5	8.7	7.8

Table 19: Share of mobile academics promoted by rank and scientific area

Scientific area	Mobile Assistant Prof. promoted	Mobile Associate Prof. promoted
Architecture & Engineering	71.4	31.8
Economics & Statistics	63.4	38.0
Humanities & Law	62.0	38.9
Medicine & Veterinary	51.2	32.3
Science	54.1	38.3
Social Sciences	50.0	33.3
Total	60.7	36.3

Table 19 shows whether mobility is associated to promotion. Column 1 shows the share of assistant professors who were promoted to associate professor in a different university. Column 2 shows whether promotion to full professor was associated to a change of university. Mobility is related to promotion to associate professor positions in 60.7% of cases, particularly in Architecture & Engineering (71.4%), while only 36.3% of the newly appointed full professors come from a

[14] The difference with the share of inbred Assistant Prof. (Table 15) is due to the few cases where the doctoral graduate was promoted directly to associate professor.

different university. This share is slightly higher for Humanities & Law, where it reaches almost 39%.

5. Conclusions

Doctorate holders are central actors in the creation and dissemination of innovative scientific knowledge. To analyse their career progression requires knowledge of the functioning of the academic labour market and the returns to the academic system of the public resources invested in third level education.

In this work, we reviewed the relevant evidence, focusing mostly on methodological issues. We found that the scarce information on Italy stems, mostly, from ad hoc surveys, covering a small number of institutions, years and scientific fields. The available data do not allow robust analysis of the supply of PhD degree holders and their careers in the Italian academic system. We constructed a novel comprehensive database, which allows the investigation of several issues.

We provided some examples by producing descriptive statistics and conducting an exploratory analysis based on this new database. The IDH-AC database contains information from IDH, a pilot dataset of information on Italian doctorate holders, extracted from the BNCF repository, and the AC dataset, which includes ministry data on academics recruited to work in Italian universities. We discussed several technical issues related to name disambiguation and record linking which are crucial one for ensuring good quality data. Although we present a stylized and preliminary representation of Italian academic careers, it confirms the database potential and usefulness for research in the field of economics of science.

References

Allison, P. D., & Long, J. S. (1990). Departmental Effects on Scientific Productivity. *American Sociological Review*, *55*(4), 469–478. https://doi.org/10.2307/2095801

Auriol, L., Misu, M., & Freeman, R. A. (2013). *Careers of Doctorate Holders*. https://www.oecd-ilibrary.org/content/paper/5k43nxgs289w-en

Bergamante, F., Canal, T., & Gualtieri, V. (2014). *Non sempre mobili: I risultati dell'indagine Isfol sulla mobilità geografica dei dottori di ricerca*. Isfol. http://isfoloa.isfol.it/xmlui/handle/123456789/1381

Bonini, F. (2014). *Consorzio Stella: Indagine sui dottori di ricerca 2011-2012 a un anno dal conseguimento del titolo: Il percorso formativo e i suoi esiti occupazionali e sociali*. Cineca.

Bonito, B. (2013). *Indagine sull'inserimento professionale dei dottori di ricerca*. LUISS Guido Carli.

Brait, F., De Vitiis, C., Petrillo, R., Russo, M., Strozza, M., & Ungaro, P. (2009). *L'indagine sui dottori di ricerca: Un'esperienza pilota* (No. 10; Documenti ISTAT).

Brandi, M. C., & Avveduto, S. (2004). Le migrazioni qualificate in Italia. *Studi Emigrazione*, *41*(156), 797–829.

Carrozza, C., & Minucci, S. (2014). Keep on Movin'? Research Mobility's Meanings for Italian Early-Stage Researchers. *Higher Education Policy*, *27*(4), 489–508. https://doi.org/10.1057/hep.2014.23

Checchi, D., De Fraja, G., & Verzillo, S. (2014). *Publish or Perish: An Analysis of the Academic Job Market in Italy*. University of Nottingham, School of Economics. https://EconPapers.repec.org/RePEc:not:notecp:14/04

Cipollone, P. E., & Avveduto, S. (2004). *La mobilità delle intelligenze in Europa. Internazionalizzazione della formazione e dottorato di ricerca*. Franco Angeli.

Combes, P.-P., Linnemer, L., & Visser, M. (2008). Publish or peer-rich? The role of skills and networks in hiring economics professors. *Labour Economics*, *15*(3), 423–441. https://doi.org/10.1016/j.labeco.2007.04.003

Crane, D. (1965). Scientists at Major and Minor Universities: A Study of Productivity and Recognition. *American Sociological Review*, *30*(5), 699–714. https://doi.org/10.2307/2091138

Crane, D. (1970). The Academic Marketplace Revisited: A Study of Faculty Mobility Using the Cartter Ratings. *American Journal of Sociology*, *75*(6), 953–964. https://doi.org/10.1086/224848

Cruz-Castro, L., & Sanz-Menéndez, L. (2010). Mobility versus job stability: Assessing tenure and productivity outcomes. *Research Policy*, *39*(1), 27–38. https://doi.org/10.1016/j.respol.2009.11.008

European Commission. (2014). *Innovation Union competitiveness report 2013 :Commission staff working document.* Directorate General for Research and Innovation. https://data.europa.eu/doi/10.2777/44320

Gallo, F., Arcaleni, E., & Cutillo, A. (2014). *L'inserimento professionale dei dottori di ricerca*. ISTAT.

Geuna, A. (2015). *Global Mobility of Research Scientists*. Elsevier. https://doi.org/10.1016/C2014-0-00191-8

Giambalvo, O., & Montalbetti, C. A. (2013). *Consorzio Stella: Indagine sui dottori di ricerca 2010-2011 a un anno dal conseguimento del titolo: Il percorso formativo e i suoi esiti occupazionali e sociali*. Cineca.

Ginther, D. K., & Kahn, S. (2004). Women in Economics: Moving Up or Falling Off the Academic Career Ladder? *Journal of Economic Perspectives*, *18*(3), 193–214. https://doi.org/10.1257/0895330042162386

Godechot, O. (2016). The chance of influence: A natural experiment on the role of social capital in faculty recruitment. *Social Networks*, *46*, 60–75. https://doi.org/10.1016/j.socnet.2016.02.002

Hargens, L. L., & Farr, G. M. (1973). An Examination of Recent Hypotheses About Institutional Inbreeding. *American Journal of Sociology*, *78*(6), 1381–1402. https://doi.org/10.1086/225470

Hargens, L. L., & Hagstrom, W. O. (1967). Sponsored and Contest Mobility of American Academic Scientists. *Sociology of Education*, *40*(1), 24–38. https://doi.org/10.2307/2112185

Heining, J., Jerger, J., & Lingens, J. (2007). *Success in the Academic Labour Market for Economics—The German Experience* (University of Regensburg Working Papers in Business, Economics and Management Information Systems No. 422). University of Regensburg, Department of Economics. https://EconPapers.repec.org/RePEc:bay:rdwiwi:803

Horta, H. (2013). Deepening our understanding of academic inbreeding effects on research information exchange and scientific output: New insights for academic based research. *Higher Education*, *65*(4), 487–510. https://doi.org/10.1007/s10734-012-9559-7

Horta, H., Veloso, F. M., & Grediaga, R. (2009). Navel Gazing: Academic Inbreeding and Scientific Productivity. *Management Science*, *56*(3), 414–429. https://doi.org/10.1287/mnsc.1090.1109

Jonkers, K. (2011). Mobility, productivity, gender and career development of Argentinean life scientists. *Research Evaluation*, *20*(5), 411–421. https://doi.org/10.3152/095820211X13176484436177

Jurczyk, P., Lu, J. J., Xiong, L., Cragan, J. D., & Correa, A. (2008). FRIL: A Tool for Comparative Record Linkage. *AMIA Annual Symposium Proceedings*, *2008*, 440.

Kahn, S. (1993). Gender Differences in Academic Career Paths of Economists. *The American Economic Review*, *83*(2), 52–56.

Levin, S. G., & Stephan, P. E. (1998). Gender Differences in the Rewards to Publishing in Academe: Science in the 1970s. *Sex Roles*, *38*(11), 1049–1064. https://doi.org/10.1023/A:1018882711314

Long, J. S., Allison, P. D., & McGinnis, R. (1979). Entrance into the Academic Career. *American Sociological Review*, *44*(5), 816–830. https://doi.org/10.2307/2094529

Long, J. S., Allison, P. D., & McGinnis, R. (1993). Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity. *American Sociological Review*, *58*(5), 703–722. https://doi.org/10.2307/2096282

Long, J. S., & Fox, M. F. (1995). Scientific Careers: Universalism and Particularism. *Annual Review of Sociology*, *21*, 45–71.

McDowell, J. M., Singell, L. D., & Ziliak, J. P. (2001). Gender and Promotion in the Economics Profession. *Industrial and Labor Relations Review*, *54*(2), 224–244. https://doi.org/10.2307/2696008

Melin, G. (2005). The dark side of mobility: Negative experiences of doing a postdoc period abroad. *Research Evaluation*, *14*(3), 229–237. https://doi.org/10.3152/147154405781776102

Modena, M. G., Lalla, M., & Molinari, R. (1999). Determinants of career structure and advancement among Italian cardiologists. An example of segregation and discrimination against women? *European Heart Journal*, *20*(17), 1276–1284. https://doi.org/10.1053/euhj.1999.1579

Musselin, C. (2004). Towards a European Academic Labour Market? Some Lessons Drawn from Empirical Studies on Academic Mobility. *Higher Education*, *48*(1), 55–78. https://doi.org/10.1023/B:HIGH.0000033770.24848.41

Musselin, C. (2005). European academic labor markets in transition. *Higher Education*, *49*(1), 135–154. https://doi.org/10.1007/s10734-004-2918-2

Pezzoni, M., Sterzi, V., & Lissoni, F. (2012). Career progress in centralized academic systems: Social capital and institutions in France and Italy. *Research Policy*, *41*(4), 704–719. https://doi.org/10.1016/j.respol.2011.12.009

Romano, M. F. (2009). *Indagine sui dottori di ricerca 2005-2007—Il percorso formativo e i suoi esiti occupazionali e sociali*. CILEA.

Romano, M. F. (2010). *Indagine sui dottori di ricerca 2007-2008 a un anno dal conseguimento del titolo. Il percorso formativo ed i suoi esiti occupazionali e sociali*. CILEA.

Romano, M. F. (2012). *Indagine sui dottori di ricerca 2009-2010 a un anno dal conseguimento del titolo. Il percorso formativo e i suoi esiti occupazionali e sociali*. CILEA.

Romano, M. F., & Himmelmann, M. (2011). *Indagine sui dottori di ricerca 2008-2009 a un anno dal conseguimento del titolo. Il percorso formativo e i suoi esiti occupazionali e sociali*. CILEA.

Sanz-Menéndez, L., Cruz-Castro, L., & Alva, K. (2013, August 10). *Time to tenure in Spanish universities: An event history analysis*. PloS One; PLoS One. https://doi.org/10.1371/journal.pone.0077028

Schizzerotto, A. (2006). *Gli esiti occupazionali dei dottori di ricerca*. Il mestiere della ricerca: un impegno per la cultura e per l'innovazione, Università degli Studi di Milano.

Sfamurri, C., Bergamaschi, F., Incarnato, S., & Melloni, R. (2012). *Indagine sulla Condizione Occupazionale dei Dottori di Ricerca*. Università degli Studi di Parma.

Appendix A: Retrieving and Cleaning Information from the BNCF Archive

BNCF data are selected by librarians for the purpose of archiving and classifying the publications, thus, bibliographic information can be assumed to have an overall high degree of accuracy. Figure A1 provides an example of the detailed codification, in a spreadsheet, of the XML tags for a thesis stored in the BNCF repository.

Figure A1: XML tags of a thesis stored in the BNCF Opac

XML structure	Spreadsheet structure
<rec>	
<cf t="001">BNI0000005</cf>	001: Record id
<df t="200" il="1" i2=" ">	
<sf c="a">^La %malattia di Gaucher</sf>	200-a: Title
<sf c="e">efficacia della terapia sostitutiva enzimatica in due pazienti affetti dalla forma di tipo 1.</sf> <sf c="e">tesi di dottorato di ricerca: scienze pediatriche</sf>	200-e: Subtitle(s)
<sf c="f">Francesco Nigro</sf>	
<sf c="g">tutore: A. Fiumara</sf> <sf c="g">coordinatore: Giustiniano Reitano</sf> <sf c="g">Università degli studi di Catania</sf>	200-g: Responsibility
</df>	
<df t="210" il=" " i2=" "> <sf c="d">2003.</sf> </df>	210: Publication year
<df t="300" il=" " i2=" "> <sf c="a">8. ciclo</sf> </df> <df t="300" il=" " i2=" "> <sf c="a">A. a. 1995-1996</sf> </df>	300: Notes
<df t="676" il=" " i2=" "> <sf c="a">618.92</sf> <sf c="v">21</sf> <sf c="9">PEDIATRIA</sf> </df>	676: Dewey classification
<df t="686" il=" " i2=" "> <sf c="a">MED/38</sf> <sf c="9">Pediatria generale e specialistica</sf> </df>	686: Miur classification
<df t="700" il=" " i2="1"> <sf c="a">Nigro</sf> <sf c="b">, Francesco</sf> <sf c="3">BNIV000008</sf> </df>	700: Author name
<df t="702" il=" " i2="1"> <sf c="a">Fiumara</sf> <sf c="b">, Agata</sf> <sf c="3">BNIV000007</sf> </df> <df t="702" il=" " i2="1"> <sf c="a">Reitano</sf> <sf c="b">. Giustiniano</sf> <sf c="f"><1931- ></sf> <sf c="3">SBLV148695</sf> </df>	702: Other author(s)
<df t="712" il="0" i2="2"> <sf c="a">Università degli studi</sf> <sf c="c"><Catania></sf> <sf c="3">CFIV008644</sf> </df>	712: University affiliation
</rec>	

Nevertheless, we needed to do some data cleaning. The cleaning of the raw data obtained from the BNCF was conducted in the following steps:

- the original information obtained from the BNCF was parsed into several fields on a spreadsheet, to reduce the noise accompanying unnecessary strings of code;
- individuals' names were standardized. For example, name information was coded in UTF-8 and divided by commas to identify first and last names;
- parsed data on publication year and university of affiliation were cleaned by checking for spelling mistakes/typos and disambiguating university names. All variants or misspelled university names were standardized and assigned to the corresponding institution. If the field corresponding to the PhD granting institution in the XML was empty, the information was searched for and, if possible, retrieved from another field;
- by considering the MIUR and Dewey Decimal classifications. we grouped the theses into six broad scientific areas: Medicine & Veterinary, Science (agriculture, biology, chemistry, physics, geology and mathematics), Architecture & Engineering, Humanities & Law, Social Sciences and Economics & Statistics. If any of these fields were empty in the XML, we compared the information doctoral course title to the official Cineca repository of doctoral courses offered by Italian universities, which associates each course to a macro-scientific area;
- We identified and removed duplicate records, including those that had different identifiers, but the same author name, publication year, university affiliation, Dewey Decimal and MIUR classification combination;
- Finally, we excluded all records with incomplete name information (missing first name or surname).

Appendix B: Summary statistics for the record linking

Appendix B provides the summary statistics for the database of doctoral graduates from Italian universities who pursued an academic career in Italian academia.

Table 20 and Figure 11 present the share of thesis-academic links in the total number of theses. We observe that the share of doctorate holders employed in an Italian university decreases over time. This is because the more recent graduates had less time to obtain a position in academia and, also, because recent reforms to the higher education sector have introduced new, non-permanent positions which extend the post-doc period. Also, Medicine & Veterinary graduates tend to be less interested in a job in academia,[15] whereas Economics & Statistics and Social Sciences graduates are most likely to pursue an academic career.

Table 21 and Figure 12 present the share of academic-thesis links found among all academics. It is interesting that, with the exception of Medicine & Veterinary for the reasons already discussed (see fn. 8), Economics & Statistics is the scientific field with the lowest share of professors with a PhD degree awarded by an Italian university (57%) while Architecture & Engineering accounts for the highest share of professors with an Italian PhD degree (68%).

[15] Anecdotally, it is a common practice in this field to give scholarships to newly graduated physicians to try to retain them while they wait to obtain a permanent position. During this period, they do not perform research; rather they work as junior doctors.

Table 20: Percentage of BNCF thesis linked to a MIUR academic by scientific field and PhD year (1986-2006)

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Tot
Hum&Law	0.59	0.54	0.46	0.57	0.59	0.60	0.53	0.53	0.53	0.46	0.47	0.44	0.43	0.41	0.39	0.40	0.35	0.34	0.30	0.23	0.20	0.36
Science	0.58	0.53	0.54	0.54	0.57	0.52	0.49	0.43	0.40	0.35	0.36	0.34	0.31	0.31	0.28	0.31	0.28	0.26	0.22	0.18	0.16	0.30
Econ&Stat	0.44	0.68	0.66	0.71	0.74	0.68	0.57	0.55	0.69	0.53	0.53	0.52	0.46	0.51	0.47	0.42	0.45	0.48	0.42	0.35	0.33	0.46
Med&Vet	0.58	0.41	0.51	0.44	0.46	0.35	0.38	0.34	0.27	0.29	0.25	0.23	0.26	0.20	0.24	0.25	0.25	0.21	0.16	0.16	0.15	0.23
Arch&Eng	0.85	0.53	0.65	0.60	0.67	0.55	0.54	0.55	0.48	0.49	0.47	0.44	0.39	0.42	0.37	0.38	0.34	0.29	0.26	0.19	0.17	0.35
SocSci	0.67	0.59	0.65	0.59	0.74	0.60	0.60	0.54	0.52	0.64	0.37	0.48	0.52	0.41	0.51	0.49	0.32	0.42	0.37	0.20	0.28	0.40
Tot	0.65	0.53	0.54	0.55	0.58	0.53	0.50	0.46	0.44	0.41	0.40	0.38	0.36	0.36	0.34	0.35	0.31	0.29	0.25	0.20	0.19	0.32

Table 21: Share of MIUR academics linked to a BNCF thesis by scientific field and expected PhD year (1986-2006)

	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Tot
Hum&Law	0.46	0.46	0.51	0.51	0.52	0.62	0.61	0.64	0.66	0.64	0.65	0.68	0.67	0.68	0.67	0.69	0.64	0.61	0.59	0.40	0.34	0.60
Science	0.43	0.50	0.52	0.56	0.61	0.62	0.67	0.68	0.71	0.76	0.71	0.73	0.73	0.74	0.72	0.70	0.73	0.68	0.61	0.49	0.30	0.64
Econ&Stat	0.30	0.28	0.34	0.45	0.40	0.53	0.53	0.62	0.65	0.62	0.71	0.60	0.73	0.70	0.69	0.68	0.63	0.63	0.53	0.48	0.34	0.57
Med&Vet	0.30	0.24	0.33	0.35	0.35	0.36	0.41	0.45	0.50	0.52	0.47	0.46	0.43	0.42	0.32	0.33	0.27	0.28	0.23	0.19	0.15	0.36
Arch&Eng	0.59	0.50	0.55	0.60	0.65	0.69	0.72	0.78	0.75	0.76	0.78	0.76	0.78	0.74	0.77	0.78	0.71	0.69	0.60	0.49	0.34	0.68
SocSci	0.54	0.36	0.70	0.71	0.71	0.65	0.74	0.72	0.68	0.70	0.70	0.63	0.78	0.63	0.70	0.70	0.71	0.53	0.55	0.52	0.40	0.66
Tot	0.46	0.46	0.51	0.51	0.52	0.62	0.61	0.64	0.66	0.64	0.65	0.68	0.67	0.68	0.67	0.69	0.64	0.61	0.59	0.40	0.34	0.60

Table 22: PhD holders (MIUR), PhD thesis (IDH) and differences by scientific area (I:IDH, M:MIUR)

	1987			1988			1989			1990			1991			1992			1993			1994			1995			1996		
	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ
Science	323	399	19	303	66	-359	341	412	17	588	420	-40	264	158	-67	750	770	3	944	807	-17	993	886	-12	928	1015	9	1222	1013	-21
Hum&Law	266	388	31	205	107	-92	243	271	10	296	240	-23	196	221	11	391	456	14	387	385	-1	386	561	31	436	783	44	850	824	-3
Econ&Stat	66	57	-16	35	8	-338	65	61	-7	61	37	-65	37	46	20	124	108	-15	128	96	-33	137	177	23	144	232	38	163	243	33
Med&Vet	99	182	46	144	258	44	149	197	24	236	352	33	123	270	54	264	370	29	412	652	37	329	663	50	334	771	57	550	713	23
Arch&Eng	190	254	25	89	10	-790	160	226	29	183	193	5	101	92	-10	353	406	13	346	392	12	465	544	15	459	738	38	681	749	9
SocSci	47	62	24	20	3	-567	27	51	47	31	22	-41	15	31	52	58	57	-2	56	56	0	62	67	7	67	95	29	81	82	1
TOT	991	1342	26	796	452	-76	985	1218	19	1395	1264	-10	736	818	10	1940	2167	10	2273	2388	5	2372	2898	18	2368	3634	35	3547	3624	2

	1997			1998			1999			2000			2001			2002			2003			2004			2005			2006		
	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ	I	M	%Δ
Science	1533	1045	-47	1723	1013	-70	1189	1285	7	1438	1574	9	1520	1504	-1	842	1363	38	1824	2095	13	2334	2573	9	3221	2789	-15	3029	3055	1
Hum&law	1055	711	-48	1056	800	-32	793	766	-4	998	804	-24	1047	837	-25	547	936	42	1181	1451	19	1668	2067	19	2483	2376	-5	2433	2532	4
Econ&Stat	262	291	10	304	200	-52	269	277	3	351	273	-29	430	291	-48	187	253	26	307	379	19	454	528	14	604	595	-2	661	600	-10
Med&Vet	707	673	-5	606	804	25	485	363	-34	574	408	-41	691	410	-69	392	640	39	786	957	18	1165	1363	15	1474	1506	2	1366	1550	12
Arch&Eng	827	692	-20	1054	687	-53	788	717	-10	916	803	-14	915	736	-24	510	809	37	1051	1178	11	1463	1532	5	1735	1823	5	1539	1966	22
SocSci	104	108	4	106	111	5	88	92	4	92	115	20	114	59	-93	38	102	63	174	184	5	229	242	5	331	355	7	336	325	-3
TOT	4488	3520	-28	4849	3615	-34	3612	3500	-3	4369	3977	-10	4717	3837	-23	2516	4103	39	5323	6244	15	7313	8305	12	9848	9444	-4	9364	10028	7

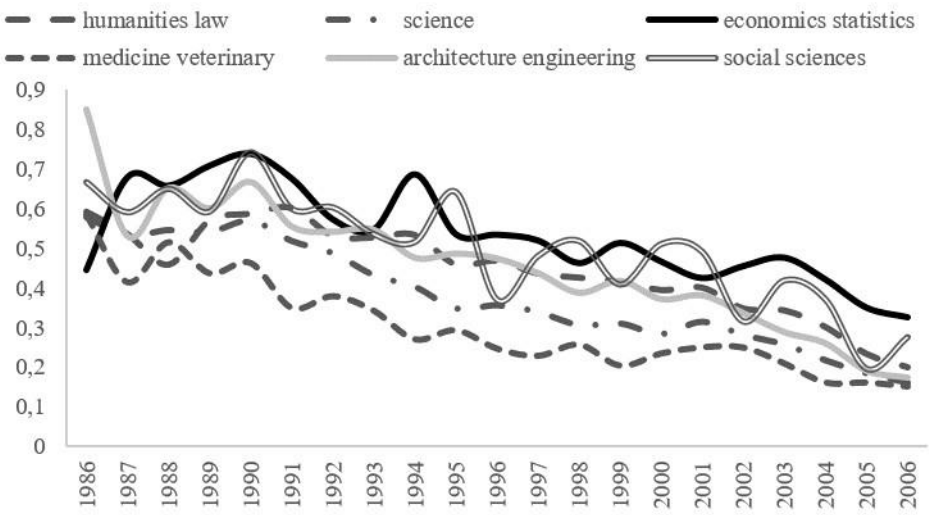

Figure 11: BNCF thesis linked to a MIUR academic by scientific field and PhD year (%)

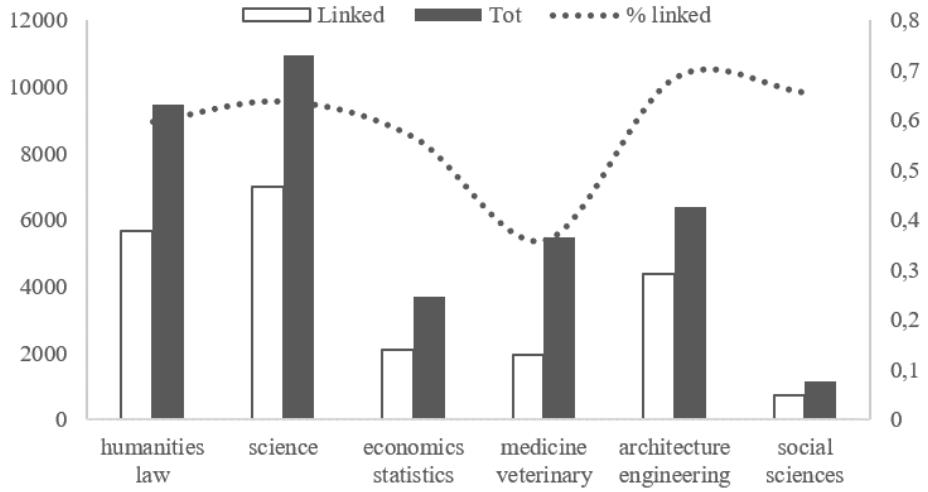

Figure 12: MIUR academics linked to BNCF thesis by scientific field (86-06)