

RE.ST.EM.

(Research Students' Employability)

How to trace and hear from doctoral alumni

IoE CDE Seed Corn (from Track1 “Student Employability”) –
status of progress

July 7th 2021

PI: Giulio Marini, dr (g.marini@ucl.ac.uk)

Co-PI: Tatiana Fumasoli, dr

Presentation is about:

- Possible sources/approaches to the topic; Main large spots in the topic
- RESTEM project in the making:
 - Stage 0 – Knowing the population
 - Stage 1 – getting a sample
 - Stage 2 – grabbing contacts
 - Stage 3 – Shipping a questionnaire
 - Stage 4 – Dissemination
 - Stage 5 – Reporting, applying for grants

Current knowledge/sources

- Currently, HESA collects data about Doctoral leavers, but
 - it does not trace those who exit EU
 - Does not deepen knowledge on doctoral experience as predictor of future success
- ONS Labour Force might investigate some aspects (e.g., *overeducation*)
- Even other sources do not pair the increasing relevance of the topic
 - By and large we have poor knowledge of what may help PhD holders getting a good job
 - Especially for nationals like the Chinese one, the percentage of PhD-holders non working in the UK/EU is (still) very large
 - We have small, if not null, knowledge about internationals PhDs working in other systems (non-academic or non-EU based)

This is what we have from HESA (unrounded omitted)

| Row Labels | 2011/12 | 2012/13 | 2013/14 | 2014/15 | 2015/16 | 2016/17 | Grand Total |
|--|--------------|--------------|--------------|--------------|--------------|--------------|---------------|
| Non-European Union | | | | | | | 2362 |
| Domiciled in: Other European Union | | | | | | | 201 |
| Domiciled in: United Kingdom | | | | | | | 2161 |
| Not known/stateless | | | | | | | 1052 |
| Domiciled in: Other European Union | | | | | | | 148 |
| Domiciled in: United Kingdom | | | | | | | 904 |
| Other European Union | | | | | | | 11847 |
| Domiciled in: Other European Union | | | | | | | 8844 |
| Domiciled in: United Kingdom | | | | | | | 3003 |
| United Kingdom | | | | | | | 38803 |
| Domiciled in: Other European Union | | | | | | | 318 |
| Domiciled in: United Kingdom | | | | | | | 38485 |
| Grand Total | 7700 | 8995 | 8808 | 9323 | 9520 | 9718 | 54064 |
| "Real Total" as per Ethos – British Library | 18416 | 20305 | 20135 | 21062 | 21314 | 22308 | 123540 |
| Percentage of coverage from HESA | 41.8% | 44.3% | 43.7% | 44.3% | 44.7% | 43.6% | 43.8% |

source: HESA Destination of leavers respondents Full Person Equivalent (FPE) v1

A relevant dataset built with ORCID ([Bohannon, 2017](#))

First 20 destinations for ORCID profiles who got a PhD in GB (up to 2016); N=31294; for 2011 year of PhD attainment N=1218

| | | |
|----------------------|-------------|-------------|
| GB | 16480 | 52.7% |
| <i>missing</i> | 6835 | 21.8% |
| US | 1071 | 3.4% |
| AU | 688 | 2.2% |
| PT | 474 | 1.5% |
| MY | 452 | 1.4% |
| CN | 268 | 0.9% |
| BR | 265 | 0.8% |
| IT | 220 | 0.7% |
| IE | 216 | 0.7% |
| DE | 208 | 0.7% |
| CA | 201 | 0.6% |
| SA | 198 | 0.6% |
| ES | 189 | 0.6% |
| MX | 166 | 0.5% |
| IR | 157 | 0.5% |
| NZ | 138 | 0.4% |
| FR | 136 | 0.4% |
| TR | 136 | 0.4% |
| GR | 125 | 0.4% |
| <i>Commonwealth*</i> | <i>2230</i> | <i>7.1%</i> |
| <i>EU*</i> | <i>2166</i> | <i>6.9%</i> |

Bohannon, 2017:

This notebook processes the 2.8 million profiles in the [ORCID public data](#) (16 December 2016) into a data set of the 741,867 profiles that have at least one listed education or employment affiliation. This outputs two files:

- ORCID_migrations_2016_12_16.csv (222 MB)
- ORCID_migrations_2016_12_16_by_person.csv (31 MB)

Today:

11,818,924 ORCID IDs and counting. [See more...](#)

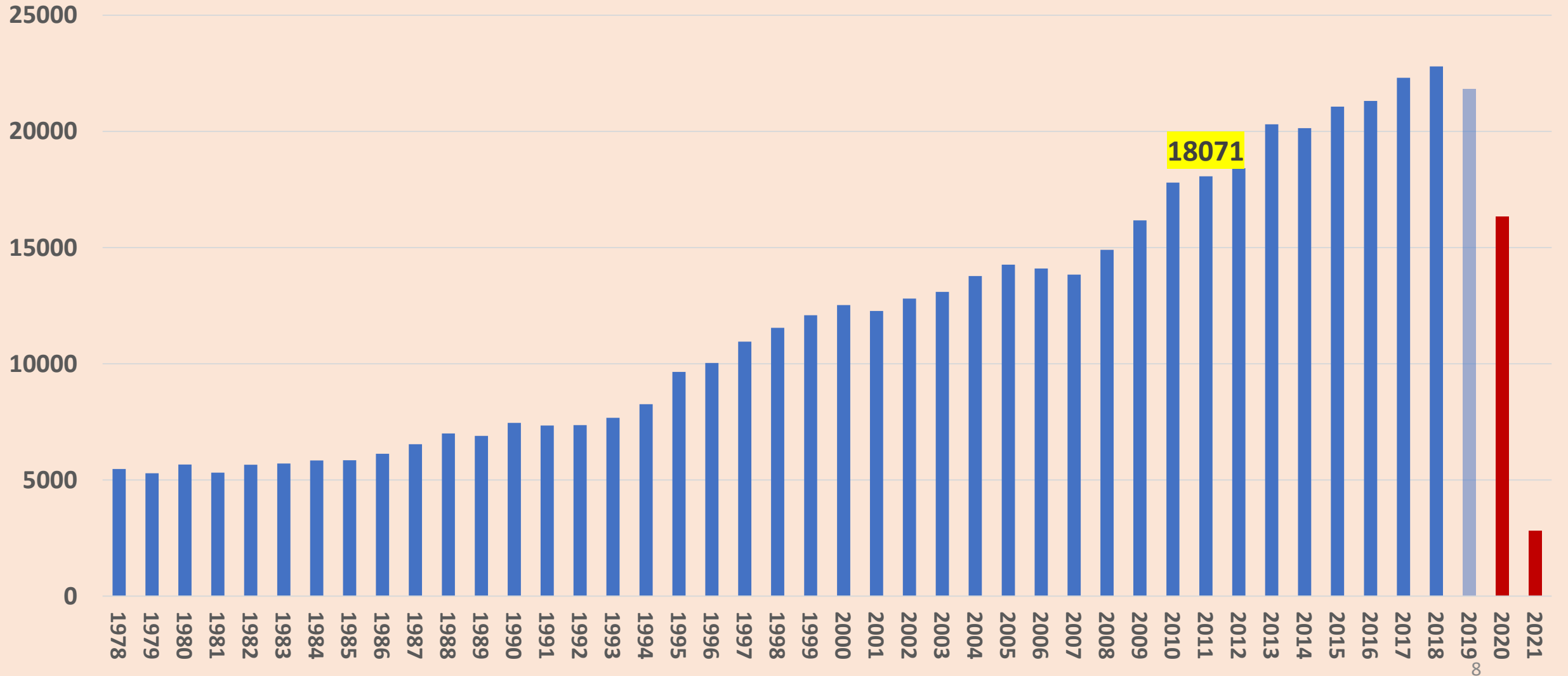
* does not include GB itself

Project stages, quick overview

1. Collection of data about individuals who got a PhD in any UK HEI (Ethos British Library doctoral theses repositories)
2. Searching those persons via (sample trials):
 - a) Google Scholar
 - b) Scopus
 - c) Clarivate Web of Science
 - d) ORCID
 - e) LinkedIn
 - f) ResearchGate
 - g) Other possibilities
3. Developing a questionnaire to be sent to most current email address (either academic retrievable via publications or LinkedIn ones)

Stage 0 – How is this universe?

PhD deposited from GB&NI universities, time series



First stage

- Searching from Ethos PhD Theses UK Library we searched the most common Chinese last names in 2011 (10 years ago) [N \approx 1000-]
- The 100 most frequent Chinese last names cover a large part of the entire Han population (using not tones grabs even more)
- Few of these last names coincides with non-Chinese ones
 - “Song” which is also Korean,
 - some others from Malesia, Indonesia, Singapore,
 - Taiwan or Hong Kong SAR (still ethnically Chinese, but different legacy in higher education)
 - few Westerners such as “Long” or “Day” (out of searching “Dai”),
 - ... but first names help in disentangling

A First original dataset

- Grabbing from [Ethos British Library](#) we built an original dataset with: name, surname, PhD title and awarding institution (and year of PhD attainment)
- (Expansion/refinement of searches by surname/years/disciplinary domains/etc/ are possible)

Advanced search

[Search tips](#)

| | |
|---------------------------------------|--------------------|
| <input type="text" value="Chen"/> | Author's last name |
| AND <input type="text" value="2012"/> | Year of award |
| AND <input type="text"/> | Any word |

[+ Add Term](#) [GO](#)

- Limit search to items available for immediate download
- Include restricted or embargoed items

Search results - 54 records

Can't find the UK doctoral thesis you want? Click [here](#) to ask our experts

Results page: ⏪ ⏩ 6 of 6 Sort by: Relevance

[The DNA-binding specificity of forkhead transcription factors](#)

Author: Chen, Xi
Awarding Body: University of Manchester
Awarded: 2012



Example of primary search

Example from row data

| Title | Surname, Name | Awarding institution | Year of attainment |
|--|-----------------|----------------------------------|--------------------|
| Probabilistic retrieval models : relationships, context-specific application, selection and implementation | Wang, Jun | Queen Mary, University of London | 2011 |
| Numerical study of turbulent flow in eccentric annular pipe | Wang, Hengliang | Imperial College London | 2011 |
| A deformation based approach to structural steel design | Wang, Facheng | Imperial College London | 2011 |
| Bacterial biofilms and biomineralisation on titanium | Wang, Anqi | University of Birmingham | 2011 |
| Identification of molecular cloning of trypsin inhibitor form the skins of oriental ranid frogs | Wang, Min | Queen's University Belfast | 2011 |
| The effects of individual differences on mobile phone users' operational behaviour | Wang, Wen-Chia | Brunel University | 2011 |
| English language identity on the Internet established by the second language learners in higher education and its application in second language learning and teaching | Wang, Yi | Liverpool John Moores University | 2011 |
| Ensemble diversity for class imbalance learning | Wang, Shuo | University of Birmingham | 2011 |
| Lay participation in China | Wang, Zhuoyo | University of Birmingham | 2011 |

First 10 universities awarding PhD to Chinese nationals in 2011

| University | N | % |
|---|-----------|-------------|
| University of Cambridge | 84 | 8.43 |
| University of Edinburgh | 52 | 5.22 |
| University of Nottingham | 48 | 4.82 |
| University of Manchester | 46 | 4.62 |
| Imperial College London | 44 | 4.42 |
| University of Oxford | 41 | 4.12 |
| University of Bristol | 39 | 3.92 |
| University College London (University of London) | 37 | 3.71 |
| University of Birmingham | 29 | 2.91 |
| Loughborough University | 28 | 2.81 |

Second stage is...

- Assessing degree of efficiency of each possible solution in terms of individuating people (= current email addresses)
- Grabbing email addresses
- Google Scholar allows to match all UK-bred doctoral holders
- A minority of these people have a Google Scholar profile, resulting in having the following problems:
 - I am losing people (not all) working outside academic or public research institutes
 - A minority of the minority is supposed to be the final set of valid answers out of a survey

Example of leaver (1)

The image shows a screenshot of a Google Scholar search result. At the top, the Google Scholar logo is on the left, and a search bar contains the text "Ensemble diversity for class imbalance learning" with a magnifying glass icon on the right. Below the search bar, the word "Articoli" is displayed with a blue arrow icon. The main content area shows a search result for the paper "Ensemble diversity for class imbalance learning" by S Wang, 2011, from etheses.bham.ac.uk. The author's name "S Wang" is circled in blue. The abstract of the paper is visible, starting with "This thesis studies the diversity issue of classification ensembles for class imbalance learning problems. Class imbalance learning refers to learning from imbalanced data sets, in which some classes of examples (minority) are highly under-represented comparing to other classes (majority). The very skewed class distribution degrades the learning ability of many traditional machine learning methods, especially in the recognition of examples from the minority classes, which are often deemed to be more important and interesting. Although ...". Below the abstract, there are icons for a star, a document, and a link, followed by the text "Citato da 18 Articoli correlati Tutte e 5 le versioni". At the bottom of the result, there is a link to "Visualizzazione del risultato migliore di questa ricerca. Mostra tutti i risultati". On the left side of the search results, there are filters for "In qualsiasi momento" (with sub-options: Dal 2021, Dal 2020, Dal 2017, Intervallo specifico...), "Ordina per pertinenza" (with sub-option: Ordina per data), and "Qualsiasi lingua" (with sub-option: Pagine in Italiano).

Google Scholar

"Ensemble diversity for class imbalance learning"

Articoli

In qualsiasi momento
Dal 2021
Dal 2020
Dal 2017
Intervallo specifico...

Ordina per pertinenza
Ordina per data

Qualsiasi lingua
Pagine in Italiano

Ensemble diversity for class imbalance learning
S Wang 2011 - etheses.bham.ac.uk

This thesis studies the diversity issue of classification ensembles for class imbalance learning problems. Class imbalance learning refers to learning from imbalanced data sets, in which some classes of examples (minority) are highly under-represented comparing to other classes (majority). The very skewed class distribution degrades the learning ability of many traditional machine learning methods, especially in the recognition of examples from the minority classes, which are often deemed to be more important and interesting. Although ...


☆ 📄 Citato da 18 Articoli correlati Tutte e 5 le versioni 🔗

Visualizzazione del risultato migliore di questa ricerca. [Mostra tutti i risultati](#)

Example of found leaver (2)

← → ↻ 🏠 <https://scholar.google.com/citations?user=kFBS6asAAAAJ&hl=it&oi=sra> ⋮ 🛡️ ☆

≡ Google Scholar



Shuo Wang

School of Computer Science, [The University of Birmingham](#), UK
Email verificata su cs.bham.ac.uk - [Home page](#)

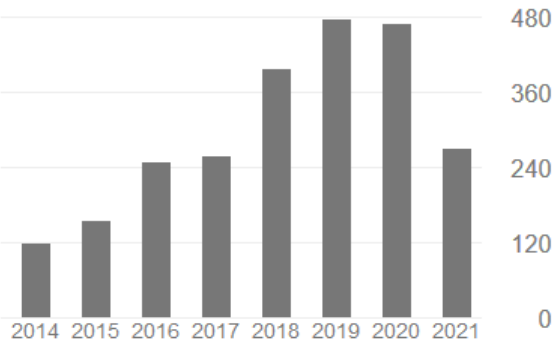
[Machine learning](#) [data mining](#) [ensemble learning](#) [class imbalance learning](#)
[software engineering](#)

✉️ SEGUI

| TITOLO | CITATA DA | ANNO |
|--|-----------|------|
| Multiclass Imbalance Problems: Analysis and Potential Solutions S Wang, X Yao Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 1-12 | 449 | 2012 |
| Using class imbalance learning for software defect prediction S Wang, X Yao IEEE Transactions on Reliability 62 (2), 434-443 | 422 | 2013 |
| Diversity analysis on imbalanced data sets by using ensemble models S Wang, X Yao Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on | 410 | 2009 |

Citata da [VISUALIZZA TUTTO](#)

| | Tutte | Dal 2016 |
|-----------|-------|----------|
| Citazioni | 2593 | 2119 |
| Indice H | 19 | 18 |
| i10-index | 25 | 22 |



First original descriptive evidence (potentially ongoing)

| Country | N |
|----------------|----------|
| GB | 29 |
| CN | 24 |
| US | 19 |
| MY | 7 |
| HK (SAR) | 2 |
| AU | 2 |
| DE | 1 |
| IE | 1 |
| NL | 1 |
| SG | 1 |
| TW | 1 |
| <i>unknown</i> | 1 |

Among these people 6 are working in companies (AstraZeneca, Amazon, apparently related to medicine or biology – probably because they publish nevertheless)

Third stage is about to be:

- Finalising a questionnaire that takes into account some literature in the topic
- Deploying questionnaire over UCL Opinio
- Sending invitations
- Main dimensions are:
 - Section 1. Motivation to become PhD
 - Section 2. Doctoral Education
 - Section 3. Post-doctoral career
 - Section 4. Current position

Fourth: some dissemination

- Apparently we can use the budget only for dissemination/travels.
- Current negotiations to present first results at:
 - USI – RISIS network
 - LERU network
 - Others to come

Fifth Stage: Open challenges

- **Finish ethical review for sampling**
- From Doctoral affiliation attempting current affiliation: thorough but inefficient
- From current affiliation going backwards: by far more efficient, but loose (but good to have data from first point)
- If attempting wider harvesting:
 - affiliations within publications can be more powerful, but it grabs *mobility* (not necessarily first affiliation is doctoral leavers' alma mater)
 - Any source has its own limitations
- Pursue sampling to refine questionnaire itself (this is a pilot)
- Writing report for IoE CDE / CHES / QSS
- Applying for larger grant(s)!