Empirical Approach: Creating New Data - Survey, Administrative, Transaction, and Bibliometrics

Wan-Ying Chang, National Science Foundation, National Center for Science and Engineering Statistics
Jodi E. Basner, Clarivate
Maryah Garner, Coleridge Initiative
Jason Owen-Smith, Institute of for Research on Innovation and Science and University of Michigan
Empirical Approach: Creating New Data – Survey, Administrative, Transaction, and Bibliometrics

Wan-Ying Chang¹*, Jodi E. Basner², Maryah Garner³, and Jason Owen-Smith⁴

Affiliations:

1 National Science Foundation, National Center for Science and Engineering Statistics (United States).
2 Clarivate (United States).
3 Coleridge Initiative (United States).
4 Institute for Research on Innovation and Science, Department of Sociology, and Institute for Social Research – University of Michigan (United States).

*Corresponding: Wan-Ying Chang
Email: wchang@nsf.gov

Disclaimer: The views expressed in this document are those of the authors and do not necessarily reflect the views of the National Science Foundation.

1. Introduction

Federal statistical agencies are embracing a new paradigm as they move to combine and use multiple data sources to expand the relevance of information they generate for research and for policy. This shift offers a great opportunity for the community to collaboratively build toward a U.S. Science & Engineering Enterprise Data Network infrastructure [1]. The scope and effectiveness of science, technology, and innovation (STI) indicators – which must “often… tell an end-to-end story on a policy-relevant topic” – increasingly require substantial “layering, linking, or blending…[of]…data” [2]. Current policy-relevant research and reporting on U.S. competitiveness in Science and STEM education is often limited by the need to conduct separate analyses on multiple dis-integrated data sources. Connected data networks will make it possible to go beyond descriptive summaries toward informing policy decisions with deeper evidence-based insights.

The National Center for Science and Engineering Statistics (NCSES) is a leading provider of statistical data on the U.S. science and engineering enterprise. As a principal federal statistical agency, NCSES serves as a clearinghouse for the collection, interpretation, analysis, and dissemination of objective science and engineering data. In recent years, NCSES has explored alternative data sources in conjunction with survey data to bolster the utility of its surveys and avoid imposing further response burden on the survey participants. Several pilot explorations are motivated by requests to match restricted survey data to other data sources to illuminate policy-relevant topics ranging from impact federal research funding has on graduate education, job placement, career progression, and research output as well as scientific mobility indicators, factors impacting scientific workforce diversity, and numerous, persistent gender gaps in publication productivity, earnings and employment. The two surveys that sit in the center of these efforts are the Survey of Earned Doctorates (SED) [3] and the Survey of Doctorate Recipients (SDR) [4], an education survey and a workforce survey, that examine the highly trained and highly invested doctoral population.
In this paper, we present methods used to develop a data network connecting administrative and transactional data on research funding for doctoral students, bibliometric data from scientific publications, and survey data on the individual educational history, demographic characteristics, and post-doctorate employment of STEM PhDs. This new, linked research data provides a rare look at individual profiles of research training, scientific activity, and career development of doctoral graduates in science, engineering, and health fields. The utility of the new research data will be demonstrated with analytical examples built on prior findings of patterns of federal funding of doctoral recipients [1] and the critical role of doctoral research assistantship for women and students of color [5]. The collaborative endeavor for NCSES with the Coleridge Initiative, Institute for Research on Innovation & Science (IRIS) and Clarivate is ongoing to further expand the linked data networks to include PatentsView, federal research grants, and other NSF funded programs.

2. Data Sources

2.1 Survey data

The Survey of Doctorate Recipients (SDR) provides demographic, education, and career history information from individuals with a U.S. research doctoral degree in a science, engineering, or health (SEH) field. The SDR is sponsored by the NCSES within the National Science Foundation (NSF) and by the National Institutes of Health. Conducted biennially since 1973, the SDR is a unique source of information about the educational and occupational achievements and career movement of U.S.-trained doctoral scientists and engineers in the United States and abroad. The sampling frame of SDR is the Doctorate Records File (DRF) constructed from the annual Survey of Earned Doctorates (SED). The SED is an annual census conducted since 1957 of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year. The SED is sponsored by the NCSES and by three other federal agencies: the National Institutes of Health, Department of Education, and National Endowment for the Humanities. The SED collects information on the doctoral recipient's educational history, demographic characteristics, and postgraduation plans. Results are used to assess characteristics of the doctoral population and trends in doctoral education and degrees. Data from the SED were also used to supplement the SDR data.

2. Administrative and Transaction data – IRIS UMETRICS

The UMETRICS data comprises administrative records drawn directly from sponsored projects, procurement, and human resources data systems from that track direct cost expenditures at the level of monthly transactions for nearly 450,000 federal and non-federal sponsored project grants. UMETRICS data includes monthly records for more than 720,000 faculty, staff, students, and post-doctoral trainees employed on those grants. The duration and time frame for the availability of UMETRICS data varies across universities and ranges from three months to 18 years, with an average of 7 years and eight months. The earliest data available begins in 2001; however, the number of reporting universities is sparse in the earlier years, with fewer than half of the universities available before 2013. We have the most coverage for the fiscal years 2014-2017. Each of the 22 universities being examined in this pilot project is represented during this time; 16 universities have full coverage, and five universities are only missing one year. Thus, we focus our report on Ph.D. graduates from SED during the fiscal years 2014-2017.

3. Bibliometric data

The Web of Science™ (WoS) is a widely used source of bibliographic and citation information in over 250 sciences, social sciences, arts, and humanities, covering >21,000 peer-reviewed, high-quality scholarly journals published worldwide, including Open Access journals, Conference proceedings and
book data. The Journal Citation Reports (JCR) offers a systematic, objective means to evaluate the world’s leading journals, with quantifiable, statistical information based on citation data. The Journal Citation Reports help to measure research influence and impact at the journal and category levels and show the relationship between citing and cited journals.

4. Other sources

ResearcherID (RID), generated by Clarivate and linked to the Web of Science, provides a platform for authors within the scholarly research community to manage their publication lists, track their times cited counts and h-index, identify potential collaborators, and avoid author misidentification. Recently ResearcherID (RID) was integrated into Publons, a commercial website launched in 2012, that provides a free service for academics to track, verify, and showcase their peer review and editorial contributions for academic journals. A portion of the researcher profiles are made public.

The Open Researcher and Contributor ID (ORCID) initiative, or ORCID, is a non-profit organization funded through organizational membership and subscription fees. ORCID effectively aggregates information across the research life cycle for researchers, funders, and publishers to enable them to treat data as something they “enter once, re-use often”. “For research institutions, this means validating their researchers’ affiliation(s); for publishers, their works (publications, reviews, datasets, and more); and for funders, their awards and grants.” As with RID, ORCID provides a persistent digital identifier that distinguishes researchers from one another and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between professional activities. The identifiers in RID and ORCID are complementary and information can be exchanged between the two platforms. Just over 435,000 RID profiles, and over 773,000 ORCID profiles will be used in the DAISNG process.

Curriculum Vitae (CV), researcher websites, LinkedIn and ResearchGate (members only) were used as a reference to support the Gold Standard data collection.

3. Linkage methods

1. SED-UMETRICS linkage

We used an "exact match" approach, matching hashed names and hashed birth dates of individuals from the same institution. Relying on hashed linkage techniques provides an essential layer of privacy protection. We carry out the matching procedure in nine sequential steps, relaxing matching rule restrictions with each step. We extract the matches at the end of each step before applying subsequent criteria. The key linking assets are hashed first name, middle name, last name, month and year of birth and institutional affiliation. We begin by matching individuals for whom complete month and year of birth information is available in both SED and UMETRICS. The date of birth criterion is consistent throughout the first three steps. However, the restrictiveness of the name matching criteria varies with each step. As a minimum, we require the matched individual's first and last names to be the same. For a match to occur in steps 4-6 - which are designed to address cases where month of birth information is missing in one or both sources - name variations and year of birth must be the same (and not NA) in both data sets. The same variation in name matching criteria from the previous steps applies here. The pattern continues for steps 7-9 – which addresses cases where both month and year of birth information is missing in one or both datasets. The name variation from the initial steps applies here as well. These steps are for individuals who could not be matched using any date of birth information in the previous steps. The flexibility in the date of birth matching criteria stems from the fact that five institutions in UMETRICS report only the year of birth, and one institution does not report the month or the year of birth. Further, there are observations in SED with the date of birth information missing. Where they are
available, month and year of birth data prove particularly helpful for adjudicating among multiple candidate matches that share common first and last names as is commonly the case with many English, and especially Chinese and Korean name strings.

2. SDR-WoS linkage

Two different stages of machine learning approaches were implemented sequentially to first address survey-to-publication record linking and then publication-to-publication linking (Figure 1). Prior to stage one, publications for two sets of mutually exclusive Gold Standard samples that each included 800 of the 78,320 SDR2015 respondents were manually identified using information provided by SDR2015 respondents in combination with information available on the World Wide Web [6]. The gold standard sets are representative random samples stratifying on demographics, education, and employment characteristics. In stage one, the supervised Gradient Boosting Machine Learning algorithm (GBM) [7] was trained with 800 of the 1,585 SDR2015 Gold Standard respondents (herein referred to as GS1). To mitigate class imbalance effects, stage one tests based on the other 785 Gold Standard respondents (GS2) were structured into subgroups composed of five strongly disambiguating covariates that were selected based on the variable importance ranking of GBM trained with GS1. For each subgroup, the GBM predicted probability threshold for qualifying matched publications was selected to target, if possible, 95% precision. In stage two, the resulting high-confidence publications from Stage 1 were used to seed into and selectively append additional publications from WoS Author Clusters, an independent methodology that disambiguates the more than 171 million WoS publication records.
4. Results

The SED-UMETRICS matched set includes 12,066 graduate students from twenty-one UMETRICS institutions who completed their doctorates during FY 2014-2017 according to the SED. The overall match rate calculated as the proportion of SED doctorate recipients from those universities linked to UMETRICS is 29%. That overall match rate hides significant variation by doctoral field of study ranging from 11% for the non-SEH fields to 48% for the physical sciences (Figure 2(a)). Fields (such as social sciences and non-SEH) that are less reliant on external funding to support doctoral training saw substantially lower match rates because students can only appear in UMETRICS if they were ever employees on a federal or non-federal sponsored project. Marked differences at the intersection of race, gender, and discipline in the granting of doctoral degrees [5] is an important factor contributing to the observed differences in match rate by gender and race. For instance, the match rate for African American doctorate recipients is 18% and that corresponds to a majority (51%) received their degrees in the two SED field areas with the lowest secular match rates.

Linking SED and UMETRICS data provides deeper insight into doctoral student funding than could be had with either source of data alone (Figure 2(b) and (c)). For instance, SED demonstrates that research assistantships are the dominant form of support for doctorate recipients during graduate school but offers little information about the duration of that support or its sources. UMETRICS allows calculation of both those measures but lacks field of degree information. Their combination offers unique insights and novel opportunities.

For instance, Figure 2(c) integrates UMETRICS research support duration information with SED information on sources of support in graduate school for the matched sample. Figure 2(b) similarly breaks out general UMETRICS sources of research support by general SED field areas. It is particularly interesting in document both where fields differ and when they are the same. For instance, 46% of matched social science doctorate recipients rely on non-federal funding sources for support while much smaller percentages of students (5%-18%) in other science and engineering fields rely on support from this source. Students in engineering and health fields are equally reliant on federal sources of funding support, but both differ significantly from physical sciences, a pattern we found surprising and worthy of further exploration.

The final SDR-WoS matched set includes 1,220,565 publications and 1,344,260 SDR respondent-authorships pairs, disambiguated from among more than 160 million pairs, for 57,874 (73.89%) doctorate recipients responded to the 2015 SDR. The overall precision and recall rates are 94.14% and 76.19% respectively. Precision and recall rates were also measured at the subgroup-level to provide a transparent understanding of the resulting dataset's strengths and limitations. Subgroups with high-quality correspondence between survey and publication data resulted in precision and recall rates each greater than 90% (51.5% of the SDR-WoS matched publication dataset). The quality of the results decreases as the availability or quality of the meta-data decreases. For example, for those subgroups where the strongly
disambiguating variables had no high-quality meta-data correspondence, the results demonstrate a need for the identification of alternative methods.

The SDR/SED-WoS linkage provides data to align publishing timeline with doctorate degree award time and other career milestones. The rate of scientific publishing is known to vary by field [8], however across the broad doctoral field of study, those who received fellowship or grant, or research assistantship consistently show higher rates of pre-PhD publishing (Figure 3(a)). The publishing trajectories post-graduation during early career period are closely related to pre-PhD publishing and future employment type (Figure 3(b)). For graduates from the UMETRICS universities, the SED-UMETRICS linkage add rich data on funding duration and graduate research activities to enable a deeper dive into exploring the implication of how funding support can impact research outcomes.
5. Discussions

There is an important trade-off between privacy/confidentiality and linkage quality when connecting data sets like this. We match individuals in the SED and UMETRICS data sets on name, institutions, and date of birth information. However, for privacy protection reasons, the personally identifiable information is hashed for the prospective data sets. To accomplish this we apply the same standardization techniques to information from each dataset and hashing the resulting strings using the same algorithm and seed. The result is that identical cleartext PII strings are transformed into identical hashed strings, which can then be used to for de-identified linkage to integrate the two data sources. Since the personally identifiable information is hashed, our options for linkage are restricted to using an exact matching strategy. We are unable to use a sound-matching or statistical similarity approach. Once the names Jon Doe and John Doe are hashed, we can not identify the similarity of these two names, and thus, we would not identify this individual as a match. This strategy gives us more confidence in the matches we identify; however, the inflexibility of our approach will fail to link individuals who appear in both data sources but with a misspelled first or last name in one. It may lead to us underestimating the number of individuals who receive grant funding while earning a Ph.D.

Similarly, the SDR and WoS linkage was carried out within a closed secure data access environment to safeguard the privacy and confidentiality of the survey respondents. The constraints added to the operational cost and efficiency. The quality of the SDR-WoS results decreases as the availability or quality of the meta-data decreases even with the application of advanced machine learning matching. It was a strategic decision to set different matching criteria for sub-group defined by the most disambiguating features of the GBM model and achieve the target high precision while relaxing the requirement on recall. This allows users of the data set to understand the strengths and limitations and use the information to carefully plan their analysis.

An initial approach of the GBM model building involved using a training data set identified with automated methods rather than using a gold standard sample. The performance was not to expectations and the approach was therefore adjusted to using half of the gold standard data for training and the remaining half for setting match prediction criteria. However, manually obtaining training and testing data is labor intensive and expensive. It took over 3,000 hours to manually obtain publications for 1,600 individuals. Therefore, consideration of ways to develop automated training data sets that perform similar to manually obtained data sets could greatly reduce these efforts.

References


8. Larsen, P. O., Von Ins, M. "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index," Scientometrics, 2010, 84:575-603.