SED/STAR METRICS RECORD LINKAGE

Joshua Tokle, Christina Jones, and Michelle Yin

17 July 2015

American Institutes for Research

Outline

1. Introduction to the problem
2. The data
3. Methodology of record linkage
4. Additional work in progress
5. Conclusion

INTRODUCTION / THE DATA

What are we doing & Why?

What are the effects of National Science Foundation (NSF) funding on PhD outcomes?

We propose that linking university administrative data to the Survey of Earned Doctorates can give an enhanced view of the effects of funding on PhD outcomes.

What is STAR METRICS

STAR METRICS (SM) is a US Interagency program to create a repository of data/tools to assess the impact of federal R&D investments [2, 1]. In reality, the SM data are administrative data collected from participating universities that describe:

Employees Who is being paid on federal grants? When are they paid and what is their occupation?

Vendors What businesses are being paid for goods and services on federal grants? When and where are they being paid and which industry are they from?

Sub-awards and sub-contracts Which organizations are being collaborated with? Where are these organizations located?

Occupational Classification

Occupational classification of job titles is a necessary step in preparing the data before conducting any type of analysis or record linkage. This process involves two steps:

1. Classified occupations based on the person's relationship to the university into five aggregated categories Undergraduate Student, Graduate Student, Postdoctoral Researcher, Staff, Faculty, and Other
2. For those job titles that are assigned the "staff" classification, we then distinguish people based on the work they perform into disaggregated occupational categories Staff Scientist, Research Analyst, Technician, Research Support, Technical Support, Research Administrator, Research Coordinator, Clinicians, Instructional, and Other Staff.

What is SED?

The Survey of Earned Doctorates (SED) is an annual census of all individuals receiving a research doctorate from accredited US institutions [3]. The information collected includes:

- the doctoral recipient’s educational history,
- demographic characteristics,
- and post graduation plans

METHODOLOGY OF RECORD LINKAGE

Probabilistic Record Linkage

In 2014, we performed a pilot study [6] using the Fellegi-Sunter approach [5] on four universities. Main challenges:

- Not many fields in common (name, year sort of)
- No time or budget for clerical review, so
- After initial parameter fitting, we made ad hoc adjustments to linkage parameters

Fellegi-Sunter model

In the simplest version of the model:

- Two sets to be linked A and B where $A \times B = M \cup U$ (matches and nonmatches)
- A binary vector-valued comparison function $\gamma : A \times B \to \{0,1\}^d$
- There are weights m_j such that if g is a particular comparison outcome and z_{jk} is an indicator which is equal to 1, assume

$$P(\gamma = g|M) = \prod_{j=1}^{d} P(\gamma_j = g_j|M) = \prod_{j=1}^{d} m_j^{g_j}(1-m_j)^{1-g_j}$$

  $P(\gamma = g|U)$ is defined similarly
- Weights m_j can be fitted using expectation-maximization (EM) or Bayesian methods.
- Extends naturally to comparison outcomes with more than 2 levels

Fellegi-Sunter model

To determine if $a \in A$ and $b \in B$ represent the same individual, compute the comparison vector $\gamma(a, b) = g$ and take the log-likelihood ratio (match score):

$$\log L = \log P(\gamma = g|M) - \log P(\gamma = g|U)$$

More likely matches will have a higher match score. By studying the data, a threshold is found such that pairs with a match score above the threshold are considered matches.

Model setup

SED and SM fields considered for matching:

- SM has first, middle, and last name fields. Middle name is frequently blank
- SED has `NAMFSTMI` and `NAMELAST` fields. The first is a concatenation of first and last name.
- SED includes `PHDCY`, calendar year of PhD
- We can look at the years an employee appears in SM

Model setup

In the end we performed two linkages which differed in the set of SM records used:

1. Subset of employees who were paid in an occupation coded "graduate student" at least once
2. All SM employees

We used the following field comparisons:

- Jaro-Winkler comparison of last name
- Jaro-Winkler comparison of first names (use regex to extract first word from `NAMFSTMI`)
- Compare `PHDCY` to last year the employee is paid as a graduate student (linkage 1) or last year employee appears in SM (linkage 2)

Model setup

Comparisons have to be discretized for the Fellegi-Sunter model:

- The range of the Jaro-Winkler comparator was broken into 4 levels (copied from BIGMATCH)
- The year comparison was broken into a 3 level comparison. The difference between PHDCY and "last SM year" computed and tested: if equal to 0 or 1, assign level 2; if equal to 2, 3, or -1, assign level 1; otherwise assign level 0.

Model setup

Input data:

- SM data for 4 universities, records dated between 2008 and 2014
- SED data corresponding to the same universities, PHDCY >= 2000
- SM graduate students only, and all SM employees

Blocking: Use university and first character of last name as blocking fields

Finalization: Extract a set of 1-to-1 links with the greatest total match score (linear sum assignment)

In the following slides we consider the links produced by restricting to graduate students.

Case study: University A link counts

Example linkage rate for one university

- Number of unique SM employees: 22,000
- Number paid as graduate student at least once: 10,000
- Last appeared in 2010: 1,200
- Linked to SED: 470
    - Worked less than 6 months (last 2 years): 107
    - Worked between 6 and 12 months (last 2 years): 11
    - Worked more than 12 months (last 2 years): 352
- Number of SED records with PHDCY = 2010: 697

The last category was useful as an independent indicator of “likely graduate student”.

Case study: University D occupations

Some occupations are not being given to PhD candidates.

Occupation	Total	Linked
Adv Oppor Fellow	7	1
Fellow	9	7
Grad Intern/Trainee	9	0
Grad Student	73	27
Graduate Assistant	2	0
Prj Ast - Grader/Reader	3	0
Research Assistant	59	34
Teach Asst Senior	34	22
Teach Asst Standard	66	11
Teaching Assistant	3	0

Value added: analyzing funding source

Out of 2,487 records in the SED 2010 cohort, 711 were linked to SM.

Received Support Type	% of Linked	% of Unlinked
Fellowship	51	50
Grant	24	23
Teaching Asst.	65	65
Research Asst.	**88**	**64**
Other Asst.	4	7
Traineeship	3	5
Other Response	**40**	**61**
No Response	**3**	**6**

“Other Response” includes things like personal savings, family support, and funding from non-US sources.

Issues

- Occupational classification does not map neatly onto educational status (hard to find the PhD students)
- Range breaks in the discretized string comparison may have been too sensitive for our data (problems with short names)
- Some first names correctly contain spaces, taking first word of `NAMFSTMI` not always correct
- Not clear how to turn the graduation year comparison into a discrete value

Conclusion

In this preliminary work we produced a linked data set that could be used to study the effects of federal funding on PhD recipients. From a methodology perspective, we lacked a way to evaluate the quality of our linkage procedure.

ADDITIONAL WORK IN PROGRESS

Phase 2: Pseudo-validated data

- SED asks for birth month and year
- Three STAR METRICS provide employee date of birth

Use these universities to create a “pseudo-validated” set of links for model training and testing.

This is joint work with Wan-Ying Chang (NSF), Wei Cheng (OSU), and Julia Lane (NYU).

Understanding the quality of links

Our ultimate goal is to make a recommendation for ongoing record linkage efforts between SM and SED. To better understand the quality of links, we compare three approaches:

- Rules-based linkage
- Probabilistic linkage
- ML linkage (random forest model)

We can try an ML approach to linkage because we can use the pseudo-validated data as training data.

Occupation analysis

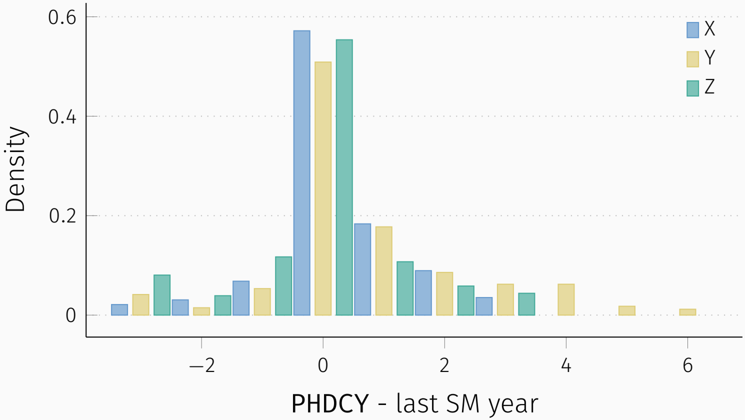
Study of links in the pseudo-validated data set confirms that bucketed occupations are not reliable enough to filter on.

Table: For pseudo-validated links, distribution of bucketed occupation titles (%). Note that the unit of observation is an individual-occupation pair.

Occupation	Uni X	Uni Y	Uni Z
Faculty	8	4	4
Graduate Student	**54**	**71**	**21**
Post Graduate Research	19	12	24
Staff	15	7	**47**
Undergraduate	3	7	2
Other	1	0	2

Analysis of graduate year

Figure: Distribution of PHDCY - last SM year in pseudo-validated data



First name comparison

Breaking the `NAMFSTMI` field in SED on the first space is not reliable (particularly for Chinese names). For example:

- “Joshua Marc” should be split into a first name and middle name
- “Wan Ying” should be considered a first name.

We define new comparators that compare the first name in SM to the entire `NAMFSTMI` field and to the first word of `NAMFSTMI` and returns the better higher score.

Probabilistic matches using DOB

Updated model to create pseudo-validated data:

- Use the new first name comparison
- Compare last name (Jaro-Winkler) and birth month and year (exact)
- Use match score to create labeled training data for a multi-class classifier. We think that a multi-class may do a better job of classifying marginal matches.

Random forest model

We follow Ventura et al. [7] in using random forests [4] in a record linkage context. In a random forest classifier, many regression trees (~ 100) are trained.

- random subsets of predictor variables
- bootstrap samples of labeled cases
- for the final classifier, each tree gets a “vote”
- comparison outcomes must be numeric (not categorical)
- missing values must be handled before training the model

Random forests for record linkage

Each row in the training data contains output from a record comparison plus a class label.

- This is possible because we have a reliable set of links (pseudo-validated data)
- Can input lots of variables and let the model sort them out
- Can fit a multi-class model based on probabilistic match scores (should help correctly classify edge cases)

Random forest for SM/SED

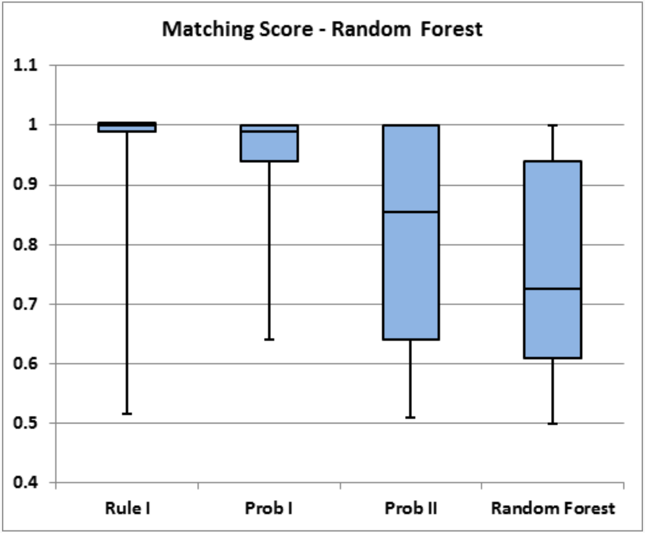
We used the flexibility of this model to try many new variables/comparisons as predictors.

- number days worked in graduate student occupation
- primary and secondary sources of funding (SED)
- number of years in SM and `YRSCOURS`, `YRSDISST` in SED
- PhD field
- first, middle, and last name comparisons, exact and fuzzy (Jaro-Winkler)
- first and last name frequency
- Used naive imputation of missing values (means and modes)

Table: A selection of variables and estimated importance measure

Variable	MeanDecreaseGini
first_full_jw	4883
last_jw	4727
last_name_freq_umetrics	1465
first_name_freq_umetrics	1122
YRSDISST	757
YRSCOURS_minus_grad_years	627
middle_init	510
PHDCY=max_year	266
TUITREMS	140

Initial findings



CONCLUSIONS

Conclusions

- Without a significant amount of clerical review, it’s hard to understand linkage quality.
- Given a set of reliable links that are representative, we can study linkage parameters and attempt to train more sophisticated models.
- Random forests are useful as a tool for understanding variable importance. We think the multi-class approach may improve classification of marginal cases.
- Can we take what we learn from the random forest and achieve the same result with an updated Fellegi-Sunter model?

Questions?

Further Reading I

- Institute for Research on Innovation and Science.
  `http://iris.isr.umich.edu/`.
  Accessed: 2015-07-15.
- STAR METRICS.
  `https://www.starmetrics.nih.gov/`.
  Accessed: 2015-07-15.
- Survey of Earned Doctorates.
  `http://www.nsf.gov/statistics/srvydoctorates/`.
  Accessed: 2015-07-15.
- L. Breiman.
  Random forests.
  Machine learning, 45(1):5–32, 2001.

Further Reading II

- I. Fellegi and A. Sunter.
  A Theory for Record Linkage.
  Journal of the American Statistical Association, 64(328):1183–1210, 1969.

- J. Lane.
  Linking the Survey of Earned Doctorates (SED) to administrative records.
  Technical report, American Institutes for Research, 2015.

Further Reading III

S. L. Ventura, R. Nugent, and E. R. Fuchs.
Seeing the non-stars: (some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records.
Research Policy, Forthcoming.