A Feasibility Study Linking the Survey of Earned Doctorates to UMETRICS and ProQuest

Workshop on the Use of Alternative and Multiple Data Sources for Federal Statistics

December 16, 2015

Wan-Ying Chang (NSF/NCSES)
Julia Lane (NYU)
Joshua Tokle, Christina Jones, Ahmad Emad (AIR)
Background

In 2013, the Survey of Graduate Students and Postdoctorates in Science and Engineering reported that federal grants are the primary source of financial support for 17% of all full-time graduate students. It is the third largest major source of support after institutional support (42%) and self support (35%).

Doctoral students’ attrition rate in the U.S. has been at 57% across all disciplines. Excluding personal factors, research indicates that the type of financial support and the level of students’ academic integration are crucial factors to doctoral completion rates.

The UMETRICS project extended the federal STAR METRICS effort and obtained records of wage payment made from federal and non-federal grants to university employees. The transactional data can be enhanced by linkages to other sources and used to study the influence of research experiences to the outcome of graduate students.
Making Connections

Grant Experiences

UMETRICS
University Grant Transactions

Outcomes
SED
Doctorate Recipients’ Post-graduation Plans

Outcomes
ProQuest
PhD & Master Dissertations & Theses
Research Questions

1. How well can doctorate recipients be linked to UMETRICS and ProQuest?

2. Can grant transactional data be used to identify features related to likelihood of completing a doctoral degree?

3. Do the grant experiences influence the employment choice of doctorate recipients?
Data Elements

- **UMETRICS**
  - Employee (paid on fed or non-fed grants) transactions: names, job titles, pay period dates, award numbers
  - Award transactions: funding agency, title and abstract

- **Survey of Earned Doctorates**
  All research doctorates from U.S. institutions: names, educational history, demographics, sources of financial support, and post-graduation plans

- **ProQuest**
  Abstract and full text PDFs of graduate works: degree awarded, institution, names of authors and advisors, subject of dissertation
Methods

I. Machine learning record linkage

II. Use big data tools to explore grant profiles

III. Evaluate outcomes of graduate students
Challenges with Transactional Data

Time coverage and job titles (used to code occupations) varies by universities

Range of Transaction Data

Transactions by Occupation Classes

- Other
- Faculty
- Postgraduate Research
- Graduate Student
- Undergraduate
Record Linkage Approaches

- **Traditional methods**
  - Deterministic matching (rule-based)
  - Probabilistic matching (Fellegi-Sunter model)

- **Machine learning methods**
  Pseudo-validated links based on richer data from a subset of universities were used as training data to build random forest models for predicting matching status
SED – UMETRICS Linkage Results

Doctorate Recipients Matched

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Match</td>
<td>95.41</td>
<td>22.33</td>
</tr>
<tr>
<td>Probabilistic Match</td>
<td>86.90</td>
<td>78.41</td>
</tr>
<tr>
<td>Pseudo-validated</td>
<td>89.56</td>
<td>89.92</td>
</tr>
<tr>
<td>Random Forests</td>
<td>93.44</td>
<td>80.83</td>
</tr>
</tbody>
</table>

- Precision = % linked records that are true matches
- Recall = % true matches that are linked by the algorithm

Estimated using gold standard data
Visualizing Individual Grant Profiles

- UMETRICS transactions enhanced by SED
- Useful for data verification and cleaning
Grant Support Duration

- 15% received support from the start
- Others, on average, waited for 1 year and 9 months

- 68% showed a gap before the degree time
- Mean gap length = 1 year 2 months
Funding Agencies

- Top funding agencies differ by university
- Linked cases have longer support
Unsupervised Random Forests Clustering

Find hidden structure

- Construct a RF predictor to distinguish unlabeled observed data from synthetic data
- Use the RF predictor to define dissimilarity between pairs of unlabeled observed data
- Perform multidimensional scaling
- Run a clustering algorithm
- Apply the variable importance measures to identify discriminant features
Unsupervised Random Forests Clustering

- The unsupervised RF yielded three clusters nicely corresponding to medium (69%), low (39%), and high (82%) levels of SED linkage.

- Variable importance analysis suggests when the complete grant profiles are available, the longer profiles are more likely to be linked to SED.
Postgraduation Plans and Grant Experiences

Simple logistic regression shows that the linkage indicator contributes in predicting the propensity of taking a postdoc position or working primarily in research and development.

**Response = POSTDOC**

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Wald</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth year</td>
<td>1</td>
<td>0.39</td>
<td>0.5339</td>
</tr>
<tr>
<td>Race category</td>
<td>5</td>
<td>3.56</td>
<td>0.614</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>0.08</td>
<td>0.7758</td>
</tr>
<tr>
<td>Broad field</td>
<td>7</td>
<td>388.30</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>U.S. citizenship</td>
<td>2</td>
<td>14.39</td>
<td>0.0007</td>
</tr>
<tr>
<td>Parents’ education</td>
<td>3</td>
<td>2.62</td>
<td>0.4542</td>
</tr>
<tr>
<td>Graduate debt</td>
<td>1</td>
<td>0.12</td>
<td>0.7328</td>
</tr>
<tr>
<td>Married</td>
<td>2</td>
<td>2.50</td>
<td>0.2863</td>
</tr>
<tr>
<td>Stay in U.S.</td>
<td>2</td>
<td>2.47</td>
<td>0.2914</td>
</tr>
<tr>
<td>Tuition waiver</td>
<td>1</td>
<td>2.59</td>
<td>0.1076</td>
</tr>
<tr>
<td>Research Asst</td>
<td>1</td>
<td>0.01</td>
<td>0.915</td>
</tr>
<tr>
<td>UMETRICS link</td>
<td>1</td>
<td>16.89</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

**Response = R&D**

<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Wald</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth year</td>
<td>1</td>
<td>34.66</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Race category</td>
<td>5</td>
<td>7.97</td>
<td>0.1581</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>3.86</td>
<td>0.0493</td>
</tr>
<tr>
<td>Broad field</td>
<td>7</td>
<td>162.69</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>U.S. citizenship</td>
<td>2</td>
<td>8.32</td>
<td>0.0156</td>
</tr>
<tr>
<td>Parents’ education</td>
<td>3</td>
<td>5.45</td>
<td>0.1414</td>
</tr>
<tr>
<td>Graduate debt</td>
<td>1</td>
<td>0.05</td>
<td>0.8284</td>
</tr>
<tr>
<td>Married</td>
<td>2</td>
<td>0.85</td>
<td>0.6522</td>
</tr>
<tr>
<td>Stay in U.S.</td>
<td>2</td>
<td>1.73</td>
<td>0.4216</td>
</tr>
<tr>
<td>Tuition waiver</td>
<td>1</td>
<td>7.55</td>
<td>0.006</td>
</tr>
<tr>
<td>Research Asst</td>
<td>1</td>
<td>17.59</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>UMETRICS_link</td>
<td>1</td>
<td>22.37</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Challenges and Promises

- Wide range of data elements including longitudinal patterns, numerical and text summaries needs a wide range of tools to be explored as a whole

- Differences in time coverage, job title codes, and non-fed grant descriptions among universities call for careful interpretations of analysis

- When combined, the data provide rare information on graduate training for studying educational and career pathways of graduate students

- Can be used to evaluate existing survey responses and to improve survey contents
Please direct questions and comments to…

Wan-Ying Chang
wchang@nsf.gov

Thank you!