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)

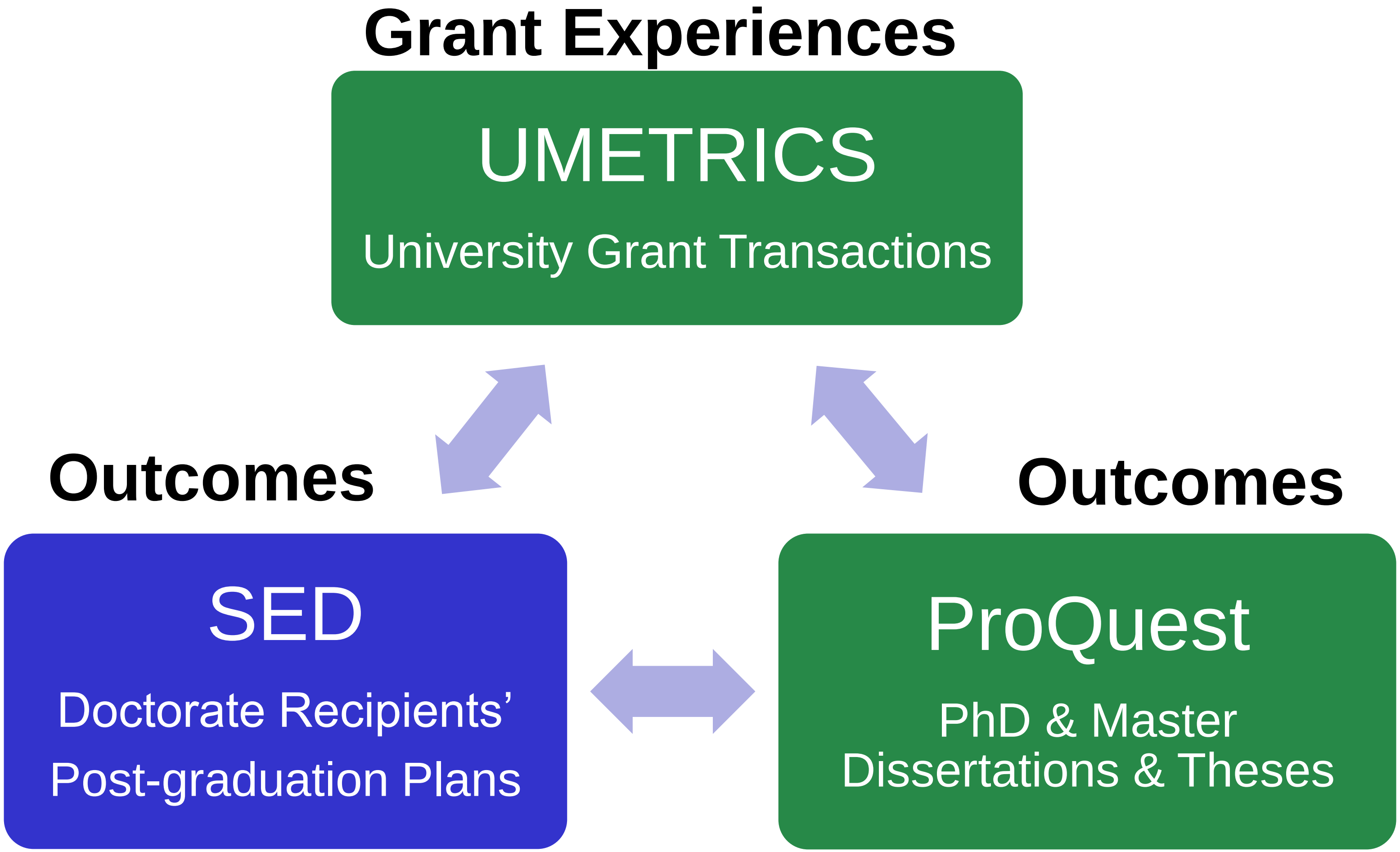
Background

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

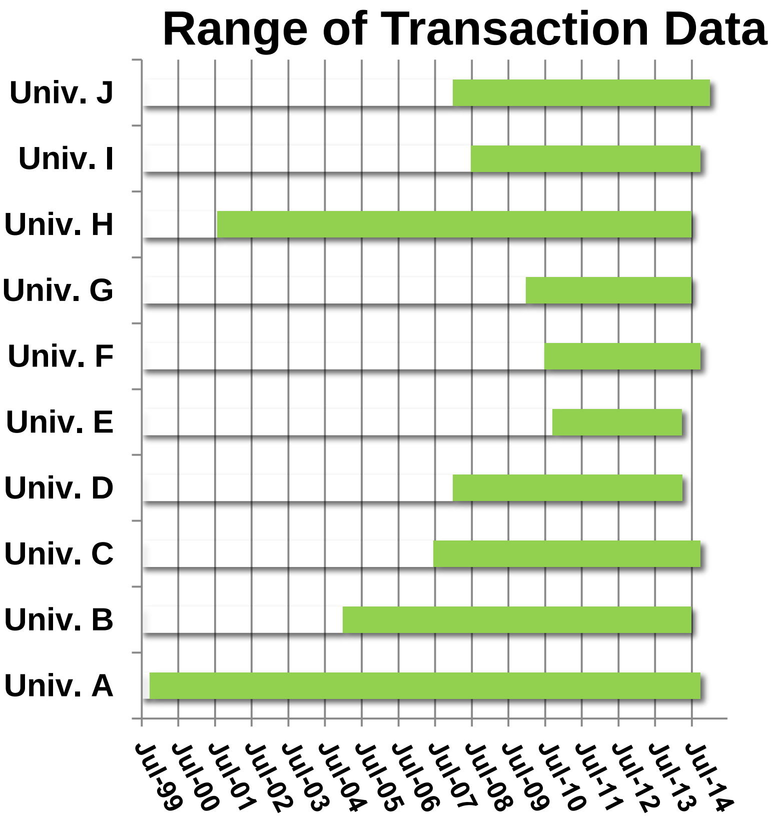
I. Machine learning record linkage

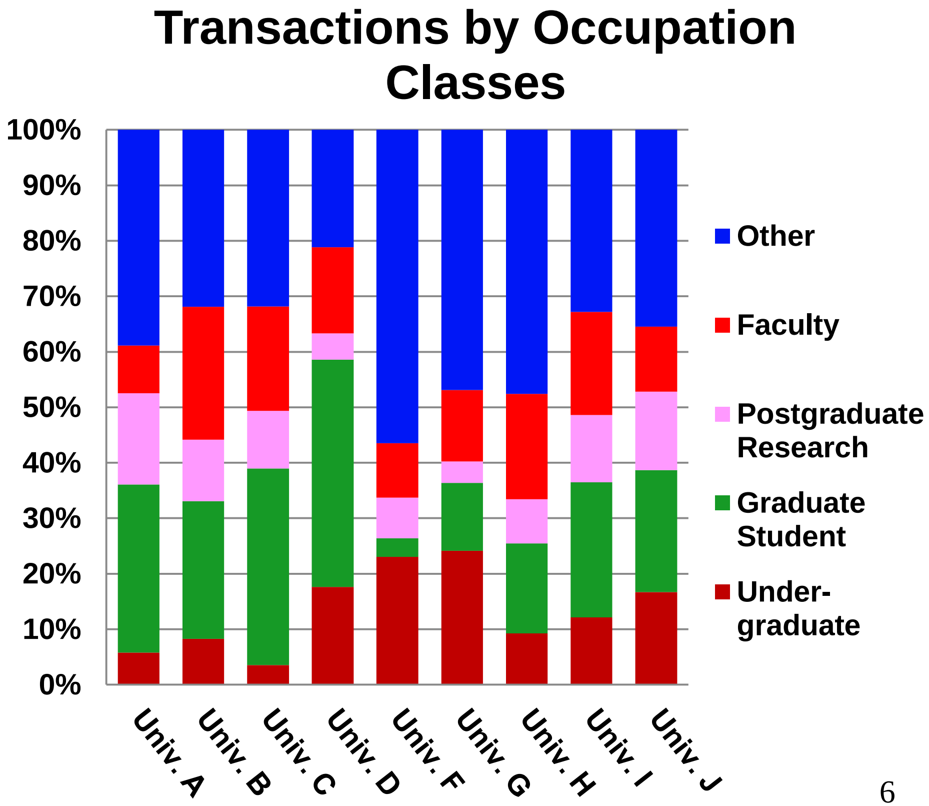
II. Use big data tools to explore grant profiles

III. Evaluate outcomes of graduate students

Challenges with Transactional Data

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



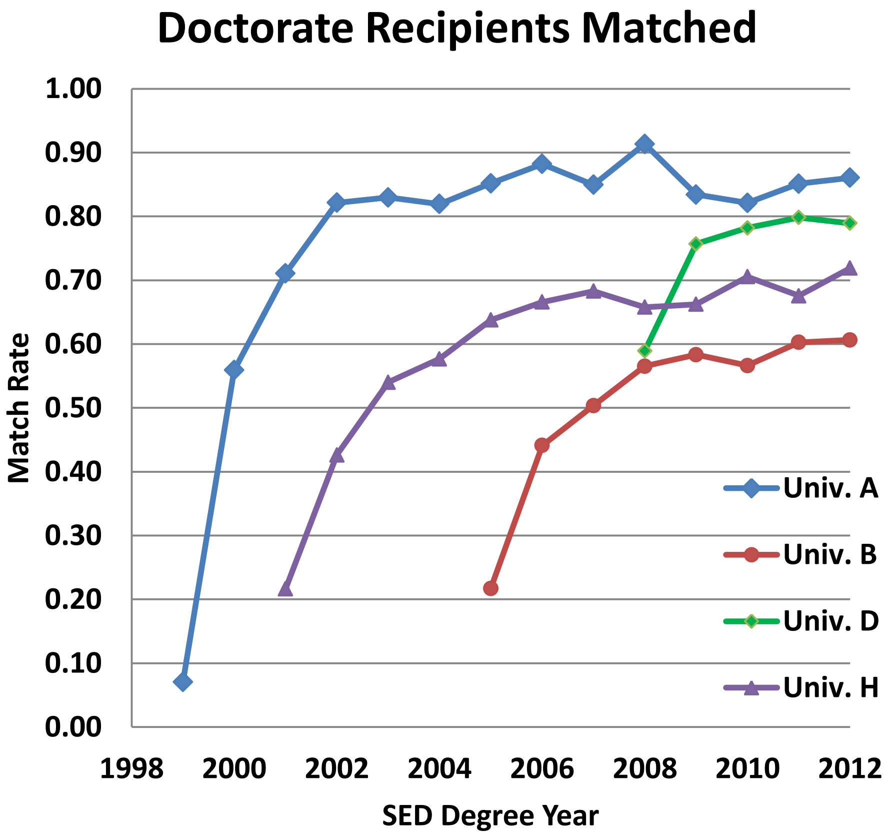


Record Linkage Approaches

- 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



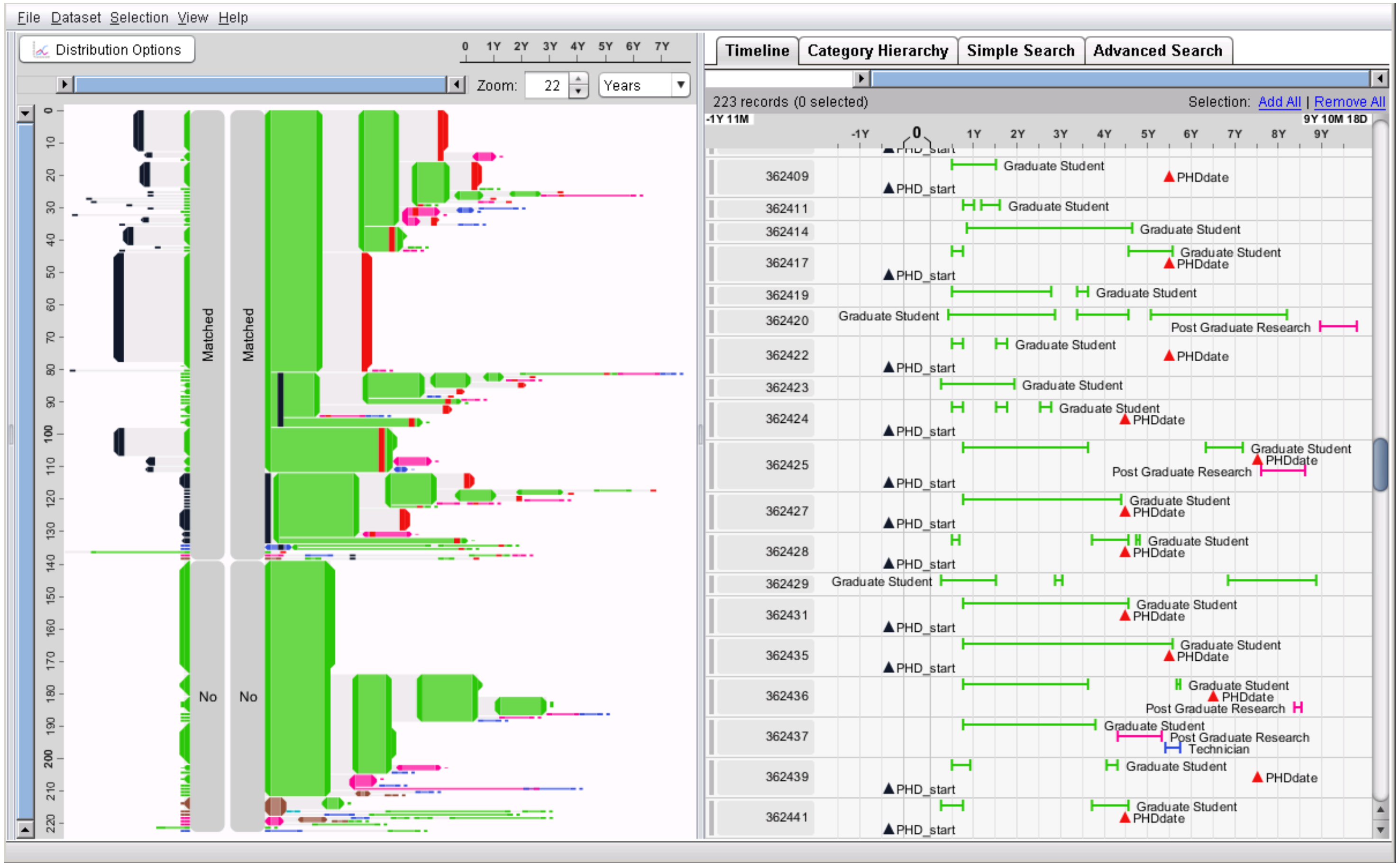
Method	Precision	Recall
Exact Match	95.41	22.33
Probabilistic Match	86.90	78.41
Pseudo-validated	89.56	89.92
Random Forests	93.44	80.83

- 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



Zoom: 22 Years

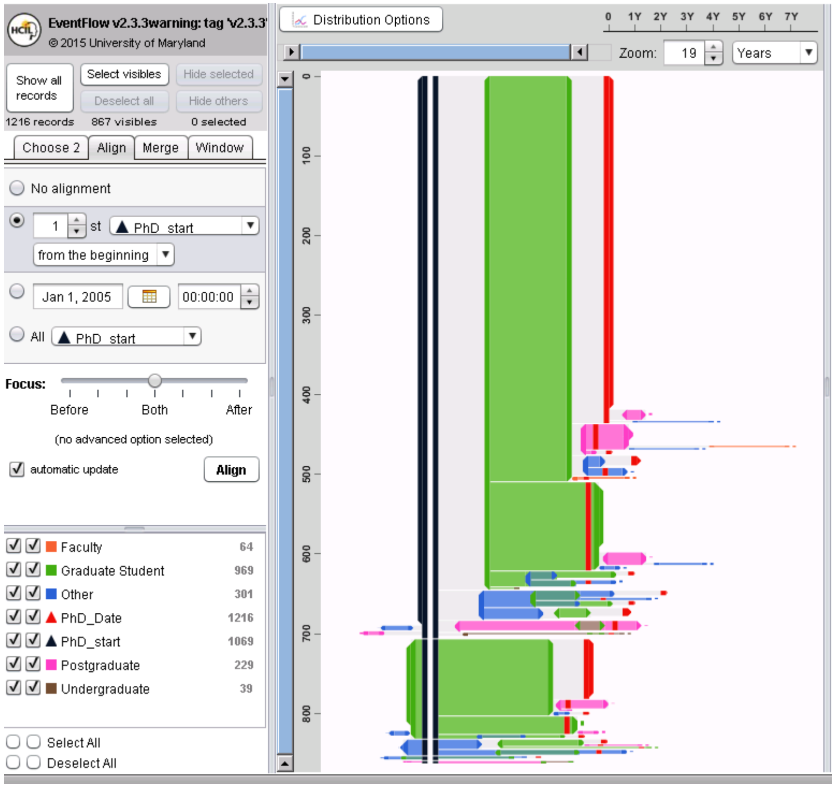
223 records (0 selected)

File Dataset Selection View Help
Distribution Options
0 1Y 2Y 3Y 4Y 5Y 6Y 7Y
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220
Matched
Matched
No
No
Timeline
Category Hierarchy
Simple Search
Advanced Search
Selection: Add All | Remove All
-1Y 11M
9Y 10M 18D
-1Y 0 1Y 2Y 3Y 4Y 5Y 6Y 7Y 8Y 9Y
362409 Graduate Student PHD_start PHDdate
362411 Graduate Student
362414 Graduate Student
362417 Graduate Student PHDdate PHD_start
362419 Graduate Student
362420 Graduate Student Post Graduate Research
362422 Graduate Student PHD_start PHDdate
362423 Graduate Student
362424 Graduate Student PHDdate PHD_start
362425 Graduate Student PHDdate Post Graduate Research PHD_start
362427 Graduate Student PHDdate PHD_start
362428 Graduate Student PHDdate PHD_start
362429 Graduate Student
362431 Graduate Student PHDdate PHD_start
362435 Graduate Student PHDdate PHD_start
362436 Graduate Student PHDdate Post Graduate Research
362437 Graduate Student Post Graduate Research Technician
362439 Graduate Student PHD_start PHDdate
362441 Graduate Student PHDdate PHD_start

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



EventFlow v2.3.3warning: tag 'v2.3.3'
© 2015 University of Maryland

Show all records | Select visibles | Hide selected | Deselect all | Hide others

1216 records 867 visibles 0 selected

Choose 2 | Align | Merge | Window

○ No alignment

● 1 st ▲ PhD_start
from the beginning

○ Jan 1, 2005 00:00:00

○ All ▲ PhD_start

Focus: Before Both After

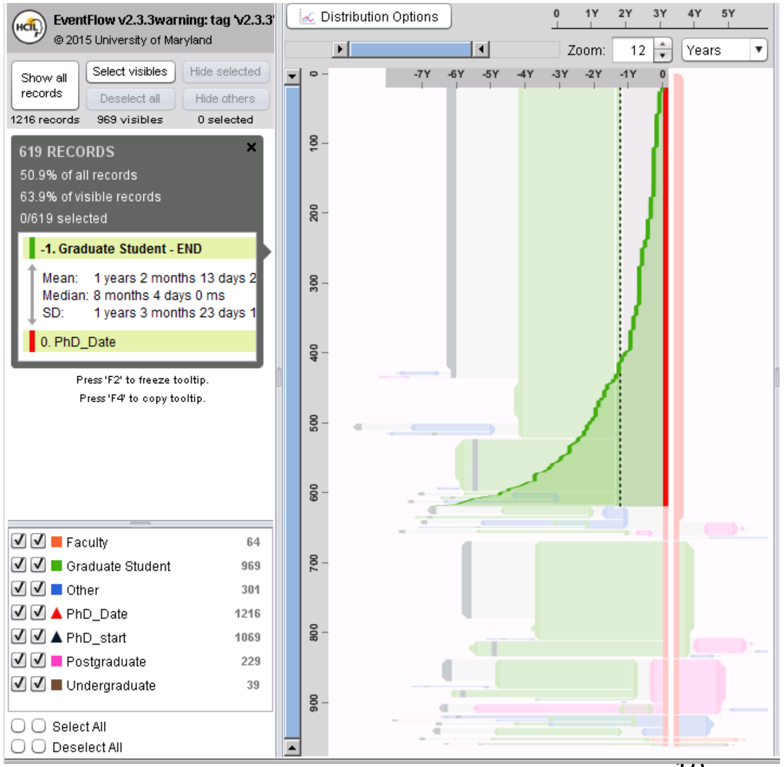
(no advanced option selected)

☑ automatic update Align

☑	☑	🟧 Faculty	64
☑	☑	🟩 Graduate Student	969
☑	☑	🟦 Other	301
☑	☑	▲ PhD_Date	1216
☑	☑	▲ PhD_start	1069
☑	☑	🟪 Postgraduate	229
☑	☑	🟫 Undergraduate	39

○ ○ Select All
○ ○ Deselect All

Distribution Options
Zoom: 19 Years



EventFlow v2.3.3warning: tag 'v2.3.3'
© 2015 University of Maryland

Show all records | Select visibles | Hide selected | Deselect all | Hide others

1216 records 969 visibles 0 selected

619 RECORDS
50.9% of all records
63.9% of visible records
0/619 selected

-1. Graduate Student - END
Mean: 1 years 2 months 13 days 2
Median: 8 months 4 days 0 ms
SD: 1 years 3 months 23 days 1
0. PhD_Date

Press 'F2' to freeze tooltip.
Press 'F4' to copy tooltip.

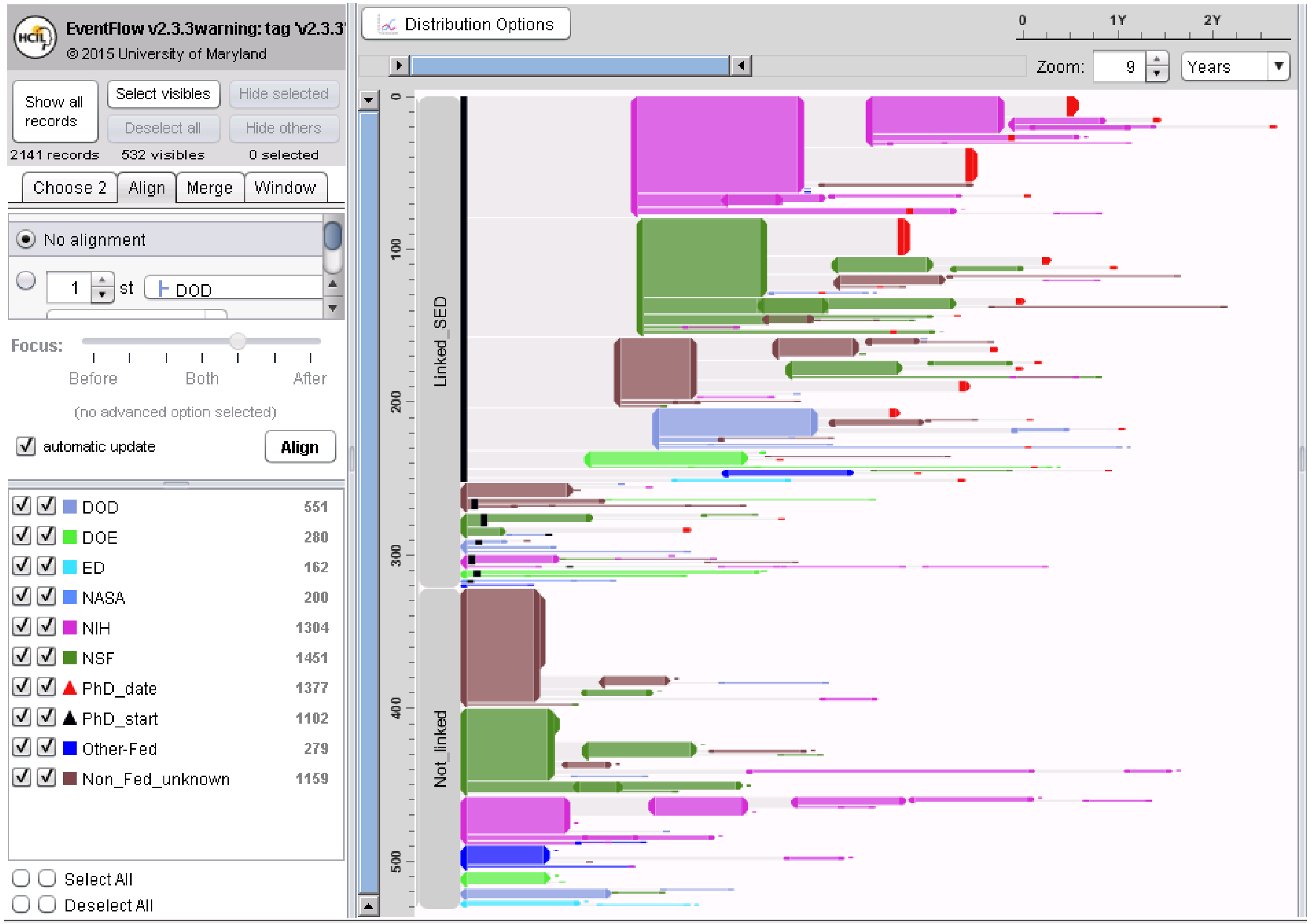
☑	☑	🟧 Faculty	64
☑	☑	🟩 Graduate Student	969
☑	☑	🟦 Other	301
☑	☑	▲ PhD_Date	1216
☑	☑	▲ PhD_start	1069
☑	☑	🟪 Postgraduate	229
☑	☑	🟫 Undergraduate	39

○ ○ Select All
○ ○ Deselect All

Distribution Options
Zoom: 12 Years

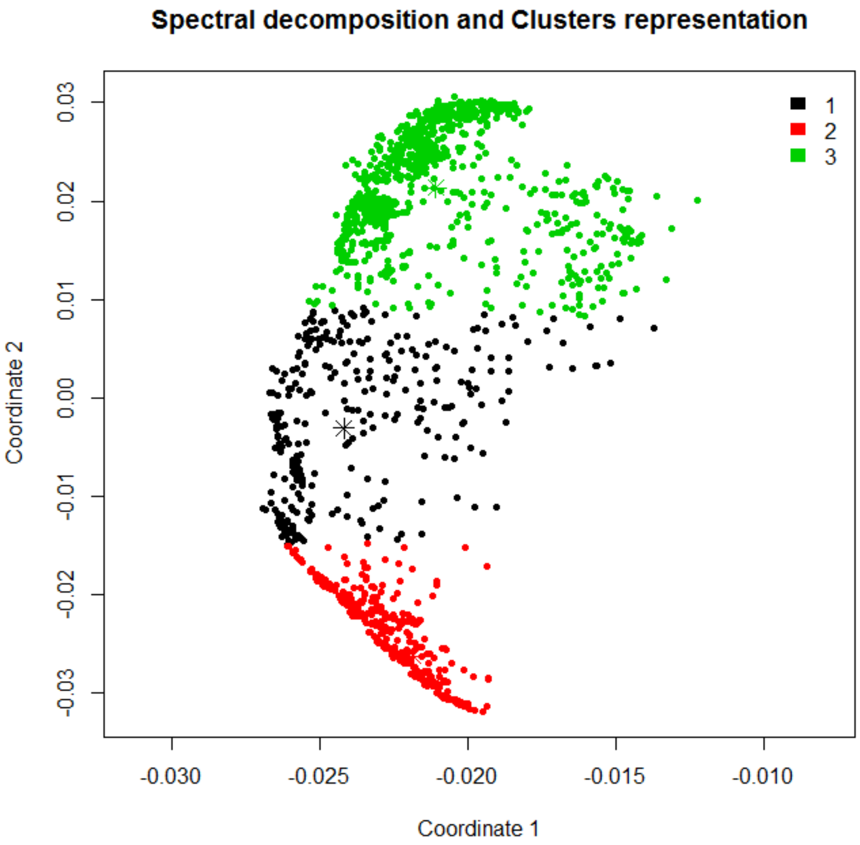
Funding Agencies

- Top funding agencies differ by university
- Linked cases have longer support



☑ ☑	DOD	551
☑ ☑	DOE	280
☑ ☑	ED	162
☑ ☑	NASA	200
☑ ☑	NIH	1304
☑ ☑	NSF	1451
☑ ☑	▲ PhD_date	1377
☑ ☑	▲ PhD_start	1102
☑ ☑	Other-Fed	279
☑ ☑	Non_Fed_unknown	1159

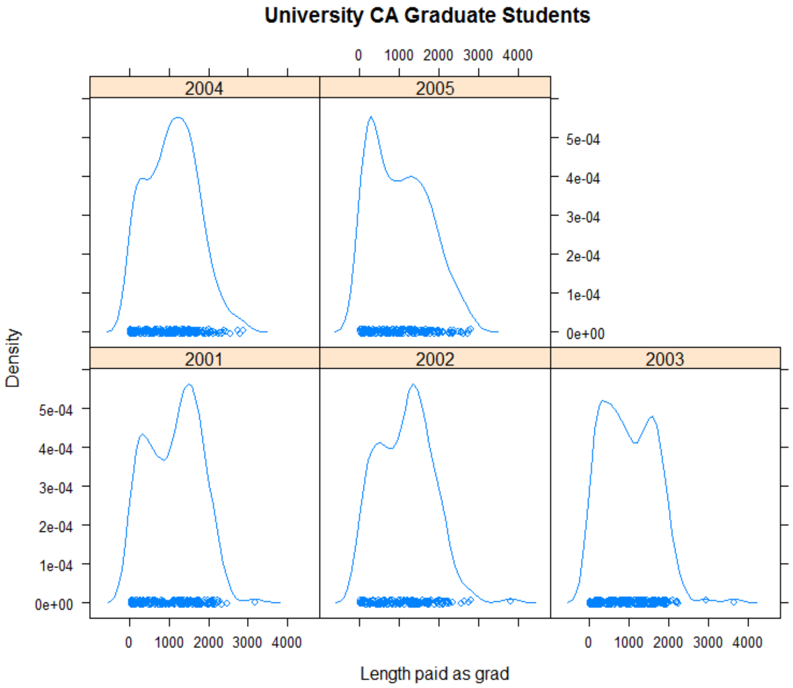
Unsupervised Random Forests Clustering



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

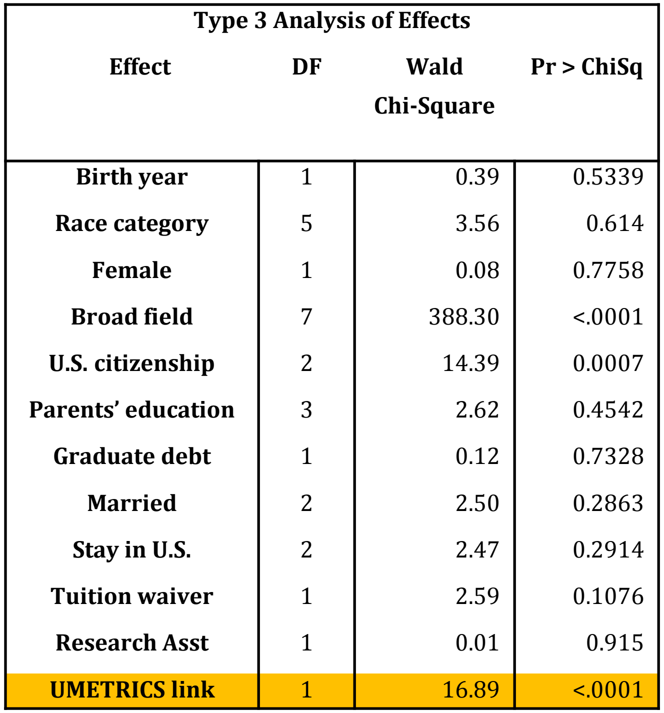


- The unsupervised RF yielded three clusters nicely corresponding to medium (69%), low (39%), and high (82%) levels of SED linkage
- Variable importance analysis suggests when the complete grant profiles are available, the longer profiles are more likely to be linked to SED

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

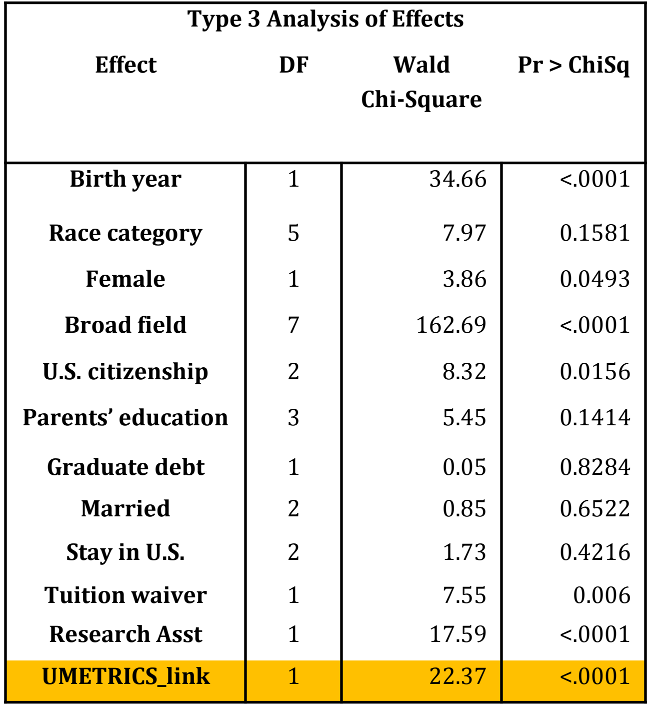
Response= POSTDOC



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Birth year	1	0.39	0.5339
Race category	5	3.56	0.614
Female	1	0.08	0.7758
Broad field	7	388.30	<.0001
U.S. citizenship	2	14.39	0.0007
Parents' education	3	2.62	0.4542
Graduate debt	1	0.12	0.7328
Married	2	2.50	0.2863
Stay in U.S.	2	2.47	0.2914
Tuition waiver	1	2.59	0.1076
Research Asst	1	0.01	0.915
UMETRICS link	1	16.89	<.0001

Response = R&D



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Birth year	1	34.66	<.0001
Race category	5	7.97	0.1581
Female	1	3.86	0.0493
Broad field	7	162.69	<.0001
U.S. citizenship	2	8.32	0.0156
Parents' education	3	5.45	0.1414
Graduate debt	1	0.05	0.8284
Married	2	0.85	0.6522
Stay in U.S.	2	1.73	0.4216
Tuition waiver	1	7.55	0.006
Research Asst	1	17.59	<.0001
UMETRICS_link	1	22.37	<.0001

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!