The effects of specialization in research trajectories

Nicolas Robinson-Garcia

May 7, 2019

Main research question and objectives

How having specific scientific profiles affect individuals' research careers?

- Identification of scientific profiles
- Comparisons between profiles and length of scientific careers, productivity, impact, gender... (to be discussed)

Rationale

1. Construction of predicting models of contributorship based on bibliometric indicators
2. Validation of such models
3. Prediction of contributorship for a set of individuals for their complete publication history
4. Profiling of individuals
5. Comparisons

Methodological notes

- Steps 1 and 2 use Plos contributorship data (old dataset)
- Steps 3-5 use complete publication history of a selection of authors
- Analysis is done by fields using high level NOWT Classification of seven subject categories (based on paper references):
  - Medical & Life Sciences
  - Natural Sciences
  - Multidisciplinary
  - Social & Behavioral Sciences
  - Law, Arts & Humanities
  - Engineering
  - Language, Information & Communication

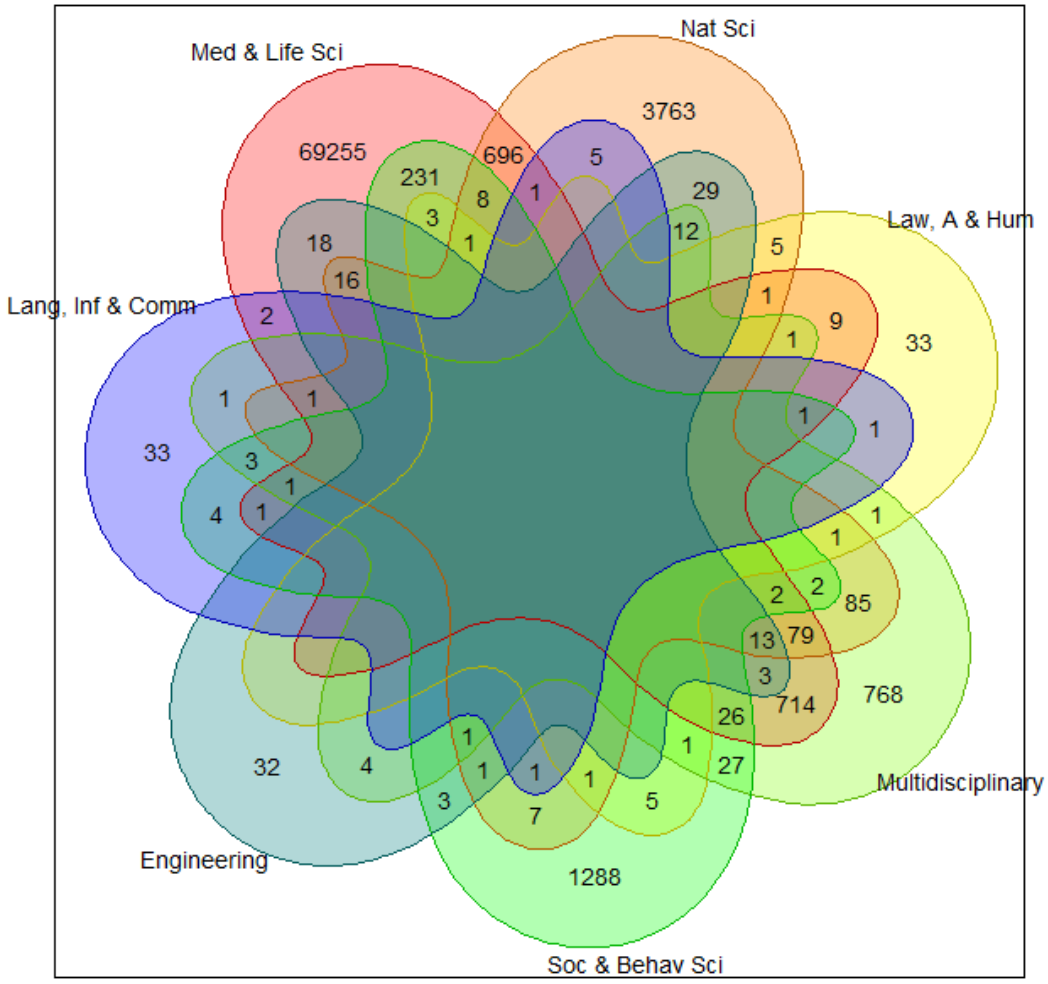
Variables

- Author position
- Number of authors
- Number of pubs. of an author at the time of the publication
- Academic age at the time of the publication
- Document type: Article or review
- Contribution types as dummy variables
- Number of contribution per author

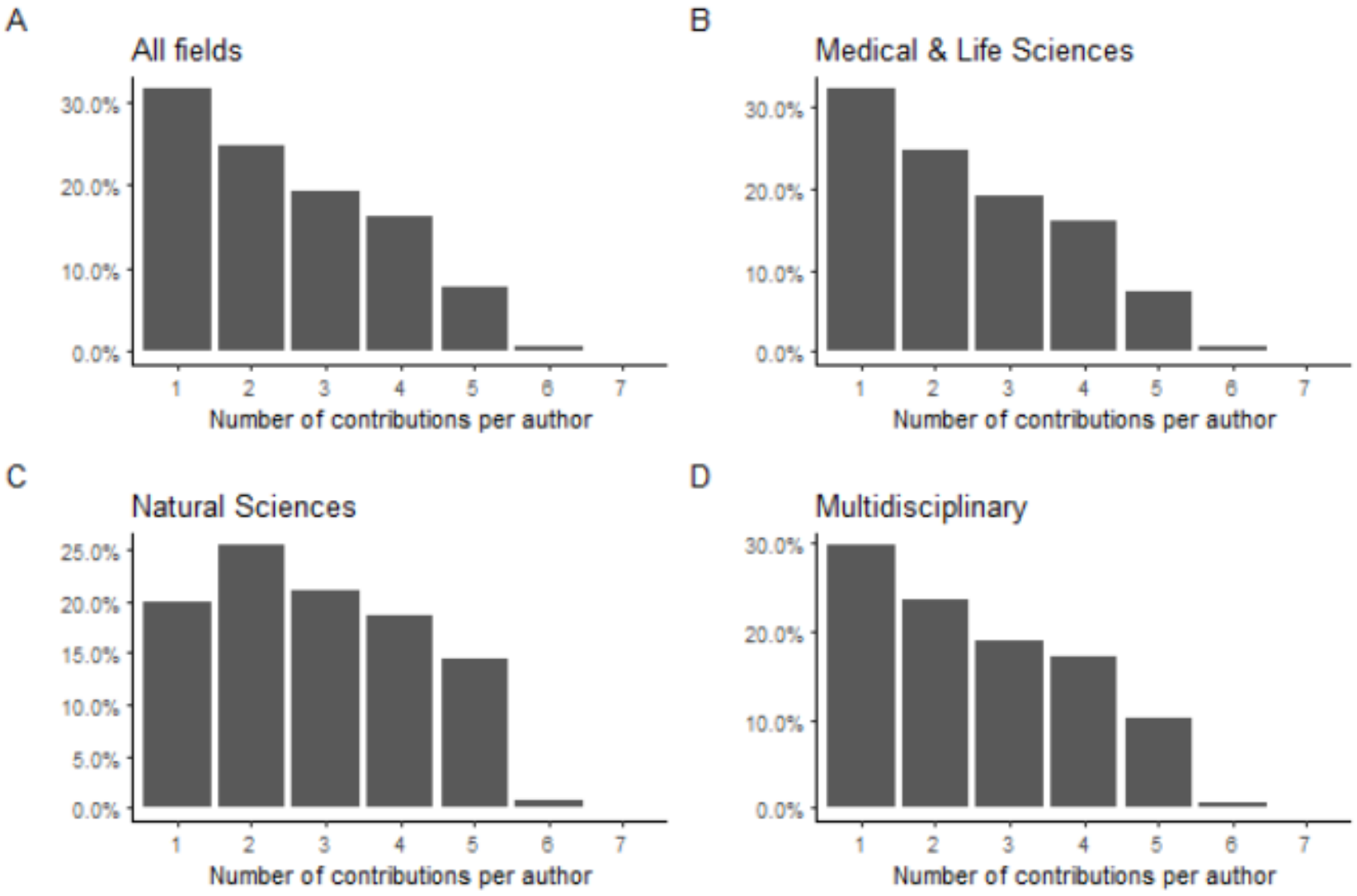
Some descriptives: General counts

Fields	Publications	Authors
Med & Life Sci	71083	348710
Natural Sci	4731	19994
Multidisciplinary	1746	9054
Soc & Behav Sci	1618	5538
Engineering	132	547
Law, A & Hum	64	291
Lang, Inform & Comm	55	181

Some descriptives: Publication overlap between fields



Some descriptives: Differences on distribution of labor



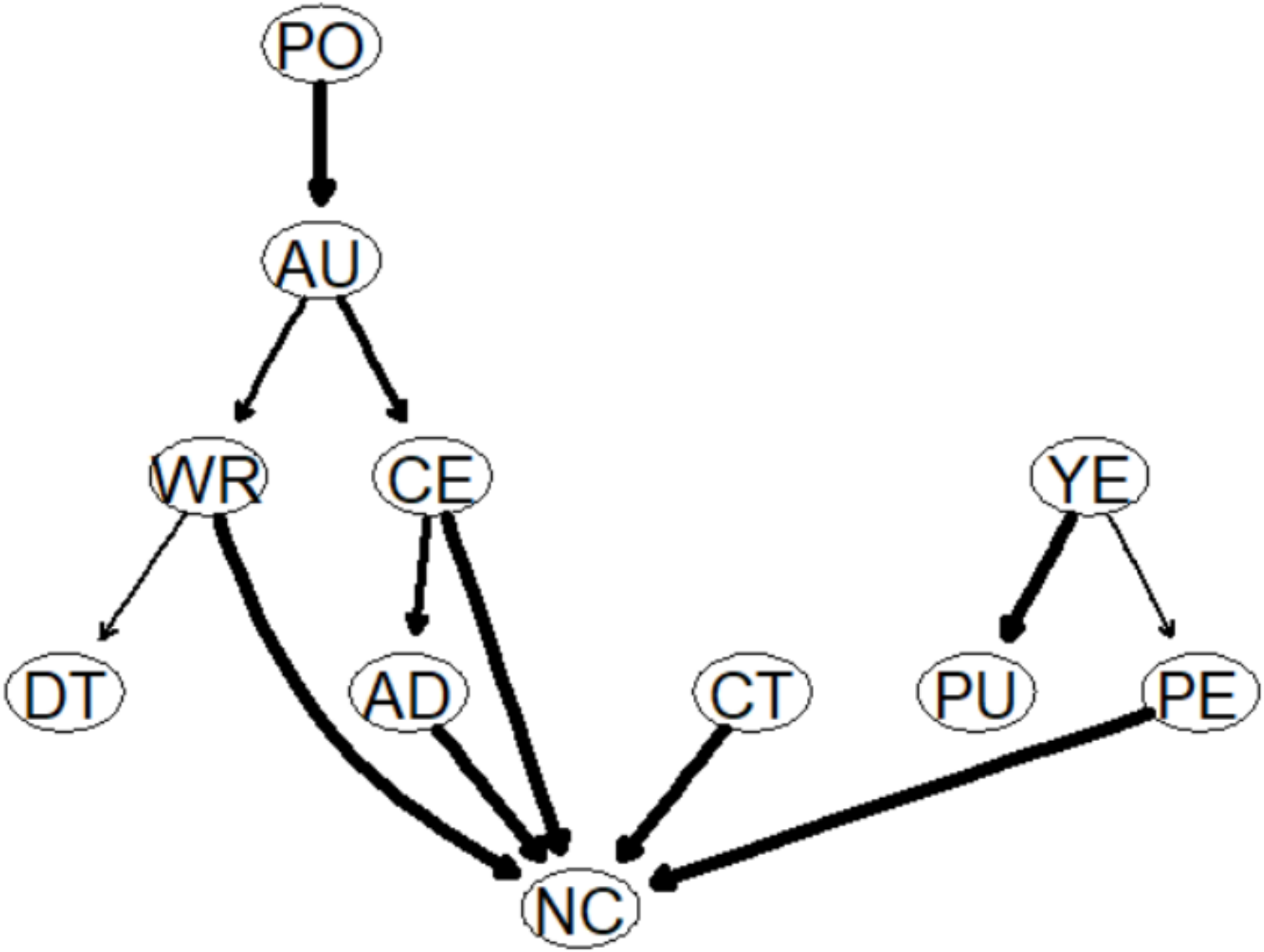
Some descriptives: Differences on distribution of labor

Step 1. Bayesian networks

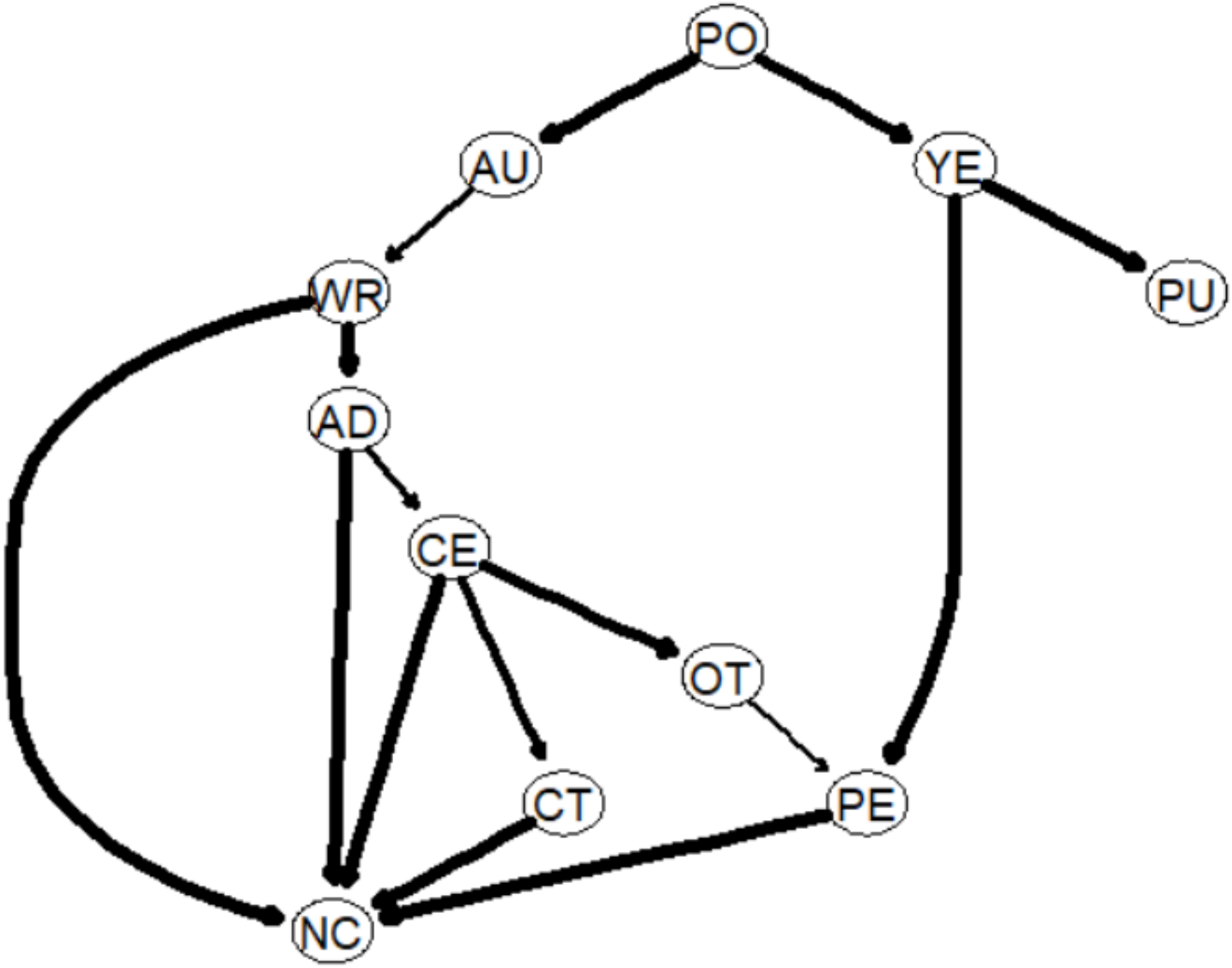
Machine learning technique aiming at identifying relation between variables

- No *a priori* assumptions on the relation of variables is made.
- The algorithm is run separately for each field
- We use bootstrapping and run the algorithm 50 times per field to maintain only stronger linkes between variables
- The method allows cross-validation and estimation of classification errors

Some examples: Law, Arts & Humanities



Some examples: Engineering



Cross-validation: Classification errors

Mean class. error	Eng	A & Hum	Lang	Med & Life	Mult	Nat S
Wrote paper	0.04	0.11	0	0.02	0.02	0.01
Analyzed data	0.04	0.07	0.05	0.02	0.01	0.00
Conceived exp.	0.05	0.06	0.08	0.01	0.01	0.00
Performed exp.	0.05	0.09	0.10	0.02	0.01	0.01
Contributed	0.04	0.10	0.08	0.02	0.02	0.01
Approved				0.03		0.01
Other	0.07	0.14	0.05	0.01	0.01	0.01

Next steps

1. Create *ideal* table conditioning bibliometric variables to show the *expected* contribution of authors by field, career, experience and position.
2. Select a set of authors and predict contribution for their complete publication history.
3. Profiling of authors: Archetypal analysis? Clustering techniques?
4. Comparisons between profiles based on (for starters) career length, productivity and impact

Target journal: PNAS

Authors: Nicolas, Rodrigo, Cassidy, Vincent and Tina