Automatically assembling a census of an academic field

Allison Morgan, Samuel Way, Aaron Clauset
University of Colorado Boulder
About Me

Third year PhD Student in CS at CU Boulder

Collaborators and I study the “sociology of science”

Interested in computational methods to study under-representation in academia

RESEARCH ARTICLE

NETWORK SCIENCES

Systematic inequality and hierarchy in faculty hiring networks

Aaron Clauset,1,2,3*, Samuel Arbesman,4 Daniel B. Larremore5,6


The misleading narrative of the canonical faculty productivity trajectory

Samuel F. Way*,1, Allison C. Morgan*, Aaron Clauset1,2,3, and Daniel B. Larremore4,5,6

Proceedings of the National Academy of Sciences Oct 2017, 201702121
Motivation

Much of the sociology of science studies small samples of the academic workforce at a single point in time.

Can we build a tool to efficiently collect the employment information of all faculty across institutions, across time?
Challenge

Every department contains a public directory of its faculty

With the same information: names, titles, email addresses, and webpages

But, information is distributed and not well structured
Our Approach

Start from department homepage

Navigate to its faculty directory

Identify the directory's HTML structure & extract faculty information

faculty_name: Jane
  title: Professor
  website: ...
  email: ...

Filter non-tenure-track faculty for further analyses

  title: Assistant Professor
  title: Research Professor
  title: Full Professor
  title: Instructor

Cartoons by Jorge Chan; phdcomics.com
Our Approach

(i) Navigate to the directory
(ii) Identify the HTML structure of the directory
(iii) Identify faculty members
(iv) Sample the relevant faculty members
Navigation

From a department homepage, sort all outgoing links by keywords:

[“professor”, "faculty", “people", "directory", “personnel", “staff” … ]

For more than half of departments, this heuristic results in the shortest path.

Navigation

To stop at directories, we use a random forest classifier trained on all directory pages, and a sample of non-directory pages.

Important features: [“NAME”, “TITLE”, “EMAIL”, “PHONE”, “website”, “profile”, “office”, “interest”]

Average accuracy is 82%*

* To avoid skipping directory pages, we parse any page which has a likelihood of being a directory > 0. Results in perfect recall, at the expense of precision.
Summary of Engineering Results

**Fast:** average <1 minute vs ~8 hours to produce a single department’s faculty directory

**Accurate:** 99% recall (nearly all tenure-track faculty are retrieved) and precision (few non-tenure-track faculty are retrieved)*

**Comparable to findings of major survey organization:** 16% vs 11% net growth in the number of faculty from the CRA

*Manually checked against a third of departments; Computing Research Association: https://cra.org
So what can we do with this tool?

We investigate the “leaky pipeline”: women leave STEM at various career stages, resulting in their under-representation at the faculty level.

Journal of Animal Science, 74(11), 2843-2848, 1996

PloS ONE, 11(7), e0157447, 2016
Leaky Pipeline

Three stages of tenure-track

- **New faculty (in 2017 & not in 2011)**
  - 1038
  - **117** (13%)
  - Departed faculty (in 2011 & not in 2017)
    - 108 (12% of 2011 Assistant professors)

- **Assistant professors**
  - 567 (65%)

- **Associate professors**
  - 160
  - 203 (11% of 2011 Associate professors)
  - 635 (35%)

- **Full professors**
  - 194
  - 78 (9%)
  - 2220 (89%)
  - 271 (11% of 2011 Full professors)
Leaky Pipeline

Arrows represent the flow from tenure-track stage in 2011 to 2017

New faculty (in 2017 & not in 2011) 1038
Retained faculty (in 2017 & in 2011) 117 (13%)
Departed faculty (in 2011 & not in 2017) 108 (12% of 2011 Assistant professors)

Assistant professors

Associate professors

Full professors

567 (65%)

78 (9%)
Leaky Pipeline

- New faculty (in 2017 & not in 2011)
  - Assistant professors
  - 117 (13%)
- Retained faculty (in 2017 & in 2011)
  - Associate professors
  - 992 (54%)
- Departed faculty (in 2011 & not in 2017)
  - Full professors
  - 2220 (89%)

Retention
Leaky Pipeline

- New faculty (in 2017 & not in 2011)
  - Assistant professors: 567 (65%)
  - Associate professors: 635 (35%)
  - Full professors
- Retained faculty (in 2017 & in 2011)
- Departed faculty (in 2011 & not in 2017)

Promotion
Leaky Pipeline

New faculty (in 2017 & not in 2011)

Retained faculty (in 2017 & in 2011)

Departed faculty (in 2011 & not in 2017)

- **Assistant professors**
  - 108 (12% of 2011 Assistant professors)

- **Associate professors**
  - 203 (11% of 2011 Associate professors)

- **Full professors**
  - 271 (11% of 2011 Full professors)

Attrition
Leaky Pipeline

Overall attrition for women is slightly higher than men (15.5% vs 14.3%)
Future Work

Expand support to other academic fields

Use the Internet Archive to collect the historical data

Cartoons by Jorge Chan; phdcomics.com
Thanks!

Automatically assembling a full census of an academic field

Allison C. Morgan,1,* Samuel F. Way,1,† and Aaron Clauset1,2,3,‡

1Department of Computer Science, University of Colorado, Boulder, CO, USA
2BioFrontiers Institute, University of Colorado, Boulder, CO, USA
3Santa Fe Institute, Santa Fe, NM, USA