Automatically assembling a census of an academic field

Allison Morgan
University of Colorado, Boulder

with Samuel Way and Aaron Clauset
Second 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
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

Jane
Professor
jane@example.edu

Mark
Associate Professor
mark@example.edu

Susan
Assistant Professor
susan@example.edu

Cartoons by Jorge Chan; phdcomics.com
Our Approach

Start from department homepage

Courses | Faculty ...

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
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

Start from department homepage

Collect all outward links
Sort links
Pick a link: is this a directory?
If not, try the next likely link

Given the directory URL

Parse tables
Parse lists
Parse divs
Parse articles

HTML structure has been identified

Identify names
Identify titles
Identify webpages
Identify emails

Is their title tenure-track?
If not, remove entry from directory

Directory with every person on the page

For each person

Filtered directory

Cartoons by Jorge Chan; phdcomics.com

Filter non-tenure-track faculty for further analyses

Our Approach
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)

**Reproduces findings of major survey organization:** 12% vs 11% net growth in the number of faculty from the CRA

Computing Research Association: [https://cra.org](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

- Assistant professors
  - New faculty (in 2017 & not in 2011): 853
  - Retained faculty (in 2017 & in 2011): 116 (15%)
  - Departed faculty (in 2011 & not in 2017): 87 (11% of 2011 Assistant professors)
  - Promoted: 527 (65%)
- Associate professors
  - Promoted: 597 (35%)
- Full professors
  - Promoted: 2129 (90%)
  - New faculty (in 2017 & not in 2011): 122

101 (11% of 2011 Associate professors)
Leaky Pipeline

New faculty (in 2017 & not in 2011)

Retained faculty (in 2017 & in 2011)

Departed faculty (in 2011 & not in 2017)

853

87 (11% of 2011 Assistant professors)

527 (65%) promoted

75 (9%) promoted

Assistant professors

Associate professors

Full professors

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

- **New faculty** (in 2017 & not in 2011): 853
- **Retained faculty** (in 2017 & in 2011): 116 (15%)
- **Departed faculty** (in 2011 & not in 2017): 87 (11% of 2011 Assistant professors)
- **Retention**
  - **Assistant professors**
    - Promoted: 527 (65%)
  - **Associate professors**
    - Promoted: 597 (35%)
  - **Full professors**
    - Promoted: 75 (9%)
Leaky Pipeline

- **New faculty** (in 2017 & not in 2011): 853
- **Retained faculty** (in 2017 & in 2011): 116 (15%)
- **Departed faculty** (in 2011 & not in 2017): 87 (11% of 2011 Assistant professors)

**Assistant professors**
- 527 (65%) promoted
- 101 (11% of 2011 Assistant professors)

**Associate professors**
- 933 (55%)
- 167 (10% of 2011 Associate professors)

**Full professors**
- 2129 (90%)
- 229 (10% of 2011 Full professors)

**Promotion**
- 75 (9%) promoted
- 122
Leaky Pipeline

- **New faculty (in 2017 & not in 2011)**: 853
- **Retained faculty (in 2017 & in 2011)**: 116 (15%)
- **Departed faculty (in 2011 & not in 2017)**: 75 (9%) promoted

Assistant professors:
- 87 (11% of 2011 Assistant professors)
- 527 (65%) promoted

Associate professors:
- 167 (10% of 2011 Associate professors)
- 597 (35%) promoted

Full professors:
- 229 (10% of 2011 Full professors)
- 2129 (90%)

Attrition
Leaky Pipeline

Overall attrition for women is slightly higher than men (13.7% vs 12.4%)
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,1,2 Samuel F. Way,1,2 and Aaron Clauset1,2,3,4

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


Prof. Aaron Clauset
PhD Computer Science
aaron.clauset@colorado.edu

Dr. Sam Way
PhD Computer Science
samuel.way@colorado.edu