Network Structure and Dynamics of the Scientific Workforce

Aaron Clauset
@aaronclauset
Computer Science Dept. & BioFrontiers Institute
University of Colorado, Boulder
External Faculty, Santa Fe Institute
faculty

[1] these awesome cartoons copyright Jorge Cham, via PhDCcomics.com
• faculty research agendas shape what discoveries are made
• faculty train students and postdocs
• faculty have long careers
• faculty are numerous

faculty play a special role in the scientific workforce

[1] these awesome cartoons copyright Jorge Cham, via PhDComics.com
but:

- who hires whose graduates as faculty?
- what does the "system" of faculty production look like?
- what predicts faculty placement?
- where are there inequalities in this system?
- what are their consequences? what drives them?
who hires whose graduates as faculty?
COLLECT ALL THE DATA
collect all the data

complete, hand-curated data for 19,000 tenure-track faculty across 461 departments in
  • Computer Science (205 depts)
  • Business (112)
  • History (144)
roughly 4000 hours of manual data collection

[2] all data from public sources, mainly faculty CVs and homepages
collect all the data

complete, hand-curated data for 19,000 tenure-track faculty across 461 departments in
- Computer Science (205 depts)
- Business (112)
- History (144)

roughly 4000 hours of manual data collection

[2] all data from public sources, mainly faculty CVs and homepages
<table>
<thead>
<tr>
<th></th>
<th>Computer Science</th>
<th>Business</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>institutions</td>
<td>205</td>
<td>112</td>
<td>144</td>
</tr>
<tr>
<td>tenure-track faculty</td>
<td>5032</td>
<td>9336</td>
<td>4556</td>
</tr>
<tr>
<td>mean size</td>
<td>25</td>
<td>83</td>
<td>32</td>
</tr>
</tbody>
</table>

\[ \sum = 18,924 \]
<table>
<thead>
<tr>
<th></th>
<th>Computer Science</th>
<th>Business</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>institutions</td>
<td>205</td>
<td>112</td>
<td>144</td>
</tr>
<tr>
<td>tenure-track faculty</td>
<td>5032</td>
<td>9336</td>
<td>4556</td>
</tr>
<tr>
<td>mean size</td>
<td>25</td>
<td>83</td>
<td>32</td>
</tr>
<tr>
<td>Full Professors</td>
<td>2400 (48%)</td>
<td>4294 (46%)</td>
<td>2097 (46%)</td>
</tr>
<tr>
<td>Associate Prof.</td>
<td>1772 (35%)</td>
<td>2521 (27%)</td>
<td>1611 (35%)</td>
</tr>
<tr>
<td>Assistant Prof.</td>
<td>860 (17%)</td>
<td>2521 (27%)</td>
<td>848 (19%)</td>
</tr>
<tr>
<td>female</td>
<td>15%</td>
<td>22%</td>
<td>36%</td>
</tr>
</tbody>
</table>
collect all the data

<table>
<thead>
<tr>
<th></th>
<th>Computer Science</th>
<th>Business</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>institutions</td>
<td>205</td>
<td>112</td>
<td>144</td>
</tr>
<tr>
<td>tenure-track faculty</td>
<td>5032</td>
<td>9336</td>
<td>4556</td>
</tr>
<tr>
<td>mean size</td>
<td>25</td>
<td>83</td>
<td>32</td>
</tr>
<tr>
<td>Full Professors</td>
<td>2400 (48%)</td>
<td>4294 (46%)</td>
<td>2097 (46%)</td>
</tr>
<tr>
<td>Associate Prof.</td>
<td>1772 (35%)</td>
<td>2521 (27%)</td>
<td>1611 (35%)</td>
</tr>
<tr>
<td>Assistant Prof.</td>
<td>860 (17%)</td>
<td>2521 (27%)</td>
<td>848 (19%)</td>
</tr>
<tr>
<td>female</td>
<td>15%</td>
<td>22%</td>
<td>36%</td>
</tr>
<tr>
<td>PhDs in-sample</td>
<td>87%</td>
<td>84%</td>
<td>89%</td>
</tr>
</tbody>
</table>

∑ = 18,924

equally closed hiring systems
faculty market is a network

- vertices are PhD-granting universities
- consumers $\leftrightarrow$ producers
- $v$ hires from $u$, add an edge $u \rightarrow v$
faculty market is a network

- vertices are PhD-granting universities
- consumers $\leftrightarrow$ producers
- $v$ hires from $u$, add an edge $u \rightarrow v$

[1] actual exchanges of 267 faculty among 10 elite CS departments, from our 2011 data, without self-hires
explore the data yourself:
http://danlarremore.com/faculty/
huge inequalities in faculty production
huge inequalities in faculty production

Gini coefficients (production)
- 0.69, 0.62, 0.72

50% of faculty from
- 18, 16, 8 universities

net producers $k_{out}/k_{in} > 1$
- 24%, 36%, 18%

1-10 producers vs.
- 11-20 : 1.6, 2.1, 3.0x more
- 21-30 : 3.1, 2.3, 5.6x more

[2] U.S. Income Gini coefficient = 0.45
a prestige hierarchy

• difficult to talk about inequalities in academia without talking about rankings
• let’s extract a data-driven ranking from the network
a prestige hierarchy
• select permutation (a ranking) $\pi$ that minimizes the number of "rank violations" : edges $(u, v)$ where $\pi_v < \pi_u$

• higher-ranked nodes have greater "placement power"

• equivalent to minimum feedback arc set problem (NP-hard)

[1] these "MVR"s have a deep history in social theory for extracting dominance or prestige hierarchies from data, especially in animal behavior
[2] MFAS: find the set of arcs of minimum cardinality whose removal converts a directed graph $G$ into a directed acyclic graph
[3] there are many equivalent MVRs for our network. we sample these using a zero-temp MCMC, and average across them to obtain $\langle \pi \rangle$
• given an ordering $\pi$ with $\psi(\pi, A)$ rank violations on network $A$

• repeat *ad infinitum*: choose a pair $(u, v)$, swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain $\pi'$, compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$

• for instance:
• given an ordering $\pi$ with $\psi(\pi, A)$ rank violations on network $A$

• repeat ad infinitum: choose a pair $(u, v)$, swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain $\pi'$, compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$

• for instance:
• given an ordering $\pi$ with $\psi(\pi, A)$ rank violations on network $A$

• repeat *ad infinitum*: choose a pair $(u, v)$, swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain $\pi'$, compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$

• for instance:
a prestige hierarchy

• what do these prestige hierarchies look like?
• what do they tell us about the structure of faculty hiring?
• what predicts placement?
prestige rankings correlate with USNews and NRC

- here, prestige $\pi$ quantifies placement power
- uncertainty increases as prestige decreases
- similar results, but different orderings for Business and History

[1] uncertainties derived from the set of MVRs
most placements are down the hierarchy
most placements are down the hierarchy

- down : 88%, 86%, 91%
- up : 12%, 14%, 9%
- $\langle \Delta \pi \rangle = 47, 27, 42$ steps down
- CS: top 15% of departments produce 68% of their own faculty and hire 7% from outside
top 25% of departments

what predicts placement?

• compare 10 single features:
  
  prestige
  US News rank
  NRC rank
  out-degree
  in-degree
  out/in degree
  eigenvector centrality
  harmonic centrality
  closeness centrality
  random
what predicts placement?

- prestige best *single* predictor in all 3 fields
- order of other features varies by field
- AUCs all below 0.67 = plenty of room for improvement
prestige correlates with network position

• core and periphery
prestige correlates with network position

- core and periphery
prestige correlates with network position

- core and periphery, homeland and colonies

- prestige is *influence*, via doctoral placement, over research agendas, research communities, and departmental norms across the discipline
inequality and prestige hierarchies

- prestige is influence, via doctoral placement
- faculty flow out of core, into periphery ("the colonies")
- small fraction stay inside core
- only ~10% of hires flow "upstream"
inequality and prestige hierarchies

- prestige is influence, via doctoral placement
- faculty flow out of core, into periphery ("the colonies")
- small fraction stay inside core
- only ~10% of hires flow "upstream"

future work

- how to measure cultural influence of core departments?
- what is different about "upstream" hires?
- what role for other inequalities : gender, ethnicity/race, SES, neighborhood effects, productivity, etc.?
inequality and prestige hierarchies

- prestige is influence, via doctoral placement
- faculty flow out of core, into periphery ("the colonies")
- small fraction stay inside core
- only \( \sim 10\% \) of hires flow "upstream"

future work

- how to measure cultural influence of core departments?
- what is different about "upstream" hires?
- what role for other inequalities: gender, ethnicity/race, SES, neighborhood effects, productivity, etc.?
women are *dramatically* under-represented in computing

18% of bachelors degrees
20% of doctoral degrees
20% of industry positions
15% of CS faculty positions

[2] Reliably industry-wide estimates are hard to come by, but see http://cnet.co/1GZh268
what role does gender play in CS faculty hiring?
what role does gender play in CS faculty hiring?

- learn a model of CS faculty hiring for all placements 1970-2010 via logistic regression (adapted for non-independence of hires)

- consider 6 factors:
  - gender
  - postdoc training
  - changing in geographic region
  - doctoral prestige
  - prestige difference
  - scholarly productivity
what role does gender play in CS faculty hiring?

- learn a model of CS faculty hiring for all placements 1970-2010 via logistic regression (adapted for non-independence of hires)

- consider 6 factors:
  - gender
  - postdoc training
  - changing in geographic region
  - doctoral prestige
  - prestige difference
  - scholarly productivity
Faculty productivity data

- Publication records (titles, dates, etc.) for 95.1% of sampled faculty
- Mean number of pubs prior to first faculty appointment is 11.3 (but distribution has a heavy tail)
- And, mean varies by subfield
• use a topic model to learn subfields from all pre-hire paper titles:

for each subfield, we tabulated a distribution over paper counts, weighted by each faculty’s inferred emphasis in that field

for each faculty, we computed a \( z \)-score for their productivity \( \rho_i \) relative to their subfield mixture
fitting the model
fitting the model

- we learn weights $\tilde{w}$ for our covariates by minimizing placement error (with L1 regularization)

$$err = \frac{1}{m} \sum_{i=1}^{m} [O(u_i) - M(u_i)]^2 + \lambda \sum_k |\tilde{w}_k|$$

- add covariates one-by-one, in greedy fashion, with gender added last
fitting the model

- we learn weights \( \vec{w} \) for our covariates by minimizing placement error (with L1 regularization)

\[
\text{err} = \frac{1}{m} \sum_{i=1}^{m} [O(u_i) - M(u_i)]^2 + \lambda \sum_{k} |\vec{w}_k|
\]

- add covariates one-by-one, in greedy fashion, with gender added last

[1] fitting done by Nelder-Mead
• adding gender does not improve placement accuracy

• three possibilities:
   1. gender is irrelevant to hiring decisions
   2. we modeled gender’s effect incorrectly
   3. gender’s effect is included in the other variables [see footnotes]

[1] since 2002, men postdoc at equal rates to women (about 28%). prior to 2002, women postdoc’d at greater rates than men (28% vs. 16%)
[2] post-2002, women with postdocs are as productive as men without postdocs; men with postdocs are significantly more productive than women with postdocs
institutions and individuals

• use our learned model to simulate hiring patterns for each institution over 1970-2010

• compare actual vs. expected number of female hires
institutions and individuals

• use our learned model to simulate hiring patterns for each institution over 1970-2010
• compare actual vs. expected number of female hires

• for top 50 institutions, an oscillation: an interference effect? from non-independence of hiring and, two distinct pools of candidates

[1] the errors our model makes are interesting, and non-uniform. for instance, men and women tend to exceed the model’s expectations at similar rates; but, for under-performing individuals, men tend to fall short of expectations by a wider margin. furthermore, people with postdoc training tend to exceed the model’s expectations, but women with postdoc experience tend to exceed expectations by a wider margin than do men
gender parity
gender parity

- where is the CS faculty gender ratio going?
- from our 40 years of observable data, the trend is toward parity
- ratio increases by 0.43% per year

[1] 1992 was a genuine outlier, with 21 of 59 (36%) hires being women; mean of 3 years before and after were 10% and 18%
gender parity

• where is the CS faculty gender ratio going?
• from our 40 years of observable data, the trend is toward parity
• ratio increases by 0.43% per year
• linear extrapolation: gender parity in new hires around 2075

[1] 1992 was a genuine outlier, with 21 of 59 (36%) hires being women; mean of 3 years before and after were 10% and 18%
gender parity

- where is the CS faculty gender ratio going?
- from our 40 years of observable data, the trend is toward parity
- ratio increases by 0.43% per year
- linear extrapolation: gender parity in new hires around 2075

[1] mean gap between PhD and faculty gender ratios over 40 years is only 1%, which was large in the 70s, but is more modest in the 2000s
the scientific workforce
the scientific workforce

- common inequalities and hierarchical structure across disciplines
- prestige is cultural influence, via doctoral placement
- prestige predicts placement
- dominant core-periphery structure (homeland vs. colonies)
- other fields? interdisciplinary work? upstream hires?
the scientific workforce

• common inequalities and hierarchical structure across disciplines
• prestige is cultural influence, via doctoral placement
• prestige predicts* placement
• dominant core-periphery structure (homeland vs. colonies)
• other fields? interdisciplinary work? upstream hires?

gender’s role in faculty hiring

• no systematic effect alone
• correlates with productivity, postdocs, geography (which are effects)
• interference effect in hiring
• "core" departments drive gender ratios everywhere
• what about other inequalities? other fields?

[1] women with a postdoc are as productive as men without a postdoc; since 2002, men/women postdoc at similar rates (28%), implying that most female applicants today will appear less productive than most male applicants
proving the gender balance in computer science would
represented in the professoriate, making up only 15% of
degrees and 20% of doctorates in 2011,

1

http://cnet.co/1GZh268

larremore@santafe.edu

offer from

successful institutions implies a more positive collective assessment of
educational and research outcomes. When institutions are unequally
collection of such pairwise assessments, a discipline

more prestigious institutions in an effort to bolster their own prestige
dominance may be neglected, leaving social prestige, in which less
social prestige mechanisms (cial hierarchy, which may emerge from either physical dominance or
network (Fig. 1) represents a collective assessment:

assessment: when an institution

ciencies of faculty hiring across disciplines is lacking.
tories of individual scholars (School of Public Health, Boston, MA 02115, USA.

*Corresponding author. E-mail: aaron.clauset@colorado.edu

Dynamics, Harvard School of Public Health, Boston, MA 02115, USA.

2

the lesser the correlation between prestige and merit.
are irrelevant, then prestige is equi

equal that encompasses differences in both scholastic merit and non-

Faculty hiring is a ubiquitous featur

common and steeply hierarchical structure that reflects pro

nearly 19,000 regular faculty in three disparate discipline

Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.

Department of Computer Science, University of Colorado, Boulder, CO 80309 USA.

5

et al.

—

orts both to support equal opportunities and

inequality in faculty hiring networks and provide new insights to the underrepresentation of women

up the rankings of universities, suggesting that the e

ences do exist, e.g., in scholarly productivity, postdoctoral training rates, and in career movements

features, the addition of gender did not significantly reduce modeling error. However, gender di

institutions and (ii) the scholarly productivity of the candidates. After including these, and other

hiring outcomes and scholarly productivity for 2659 tenure-track faculty across 205 Ph.D.-granting

I. INTRODUCTION

1,2,3

who hires whose graduates as faculty

10.1126/sciadv.1400005

Copyright 2015 by the authors, some rights reserved; exclusive license to the publisher granted by the American Association for the Advancement of Science. Distributed under a Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC).

Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks

Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks

Samuel F. Way,1,* Daniel B. Larremore,2,† and Aaron Clauset1,3,2,‡
1 Department of Computer Science, University of Colorado, Boulder CO, 80309 USA
2 Santa Fe Institute, Santa Fe NM, 87501 USA
3 BioFrontiers Institute, University of Colorado, Boulder CO, 80303 USA

WWW (2016) [arxiv:1602.00795]


Funding:
• Kauffman Foundation
• University of Colorado Boulder