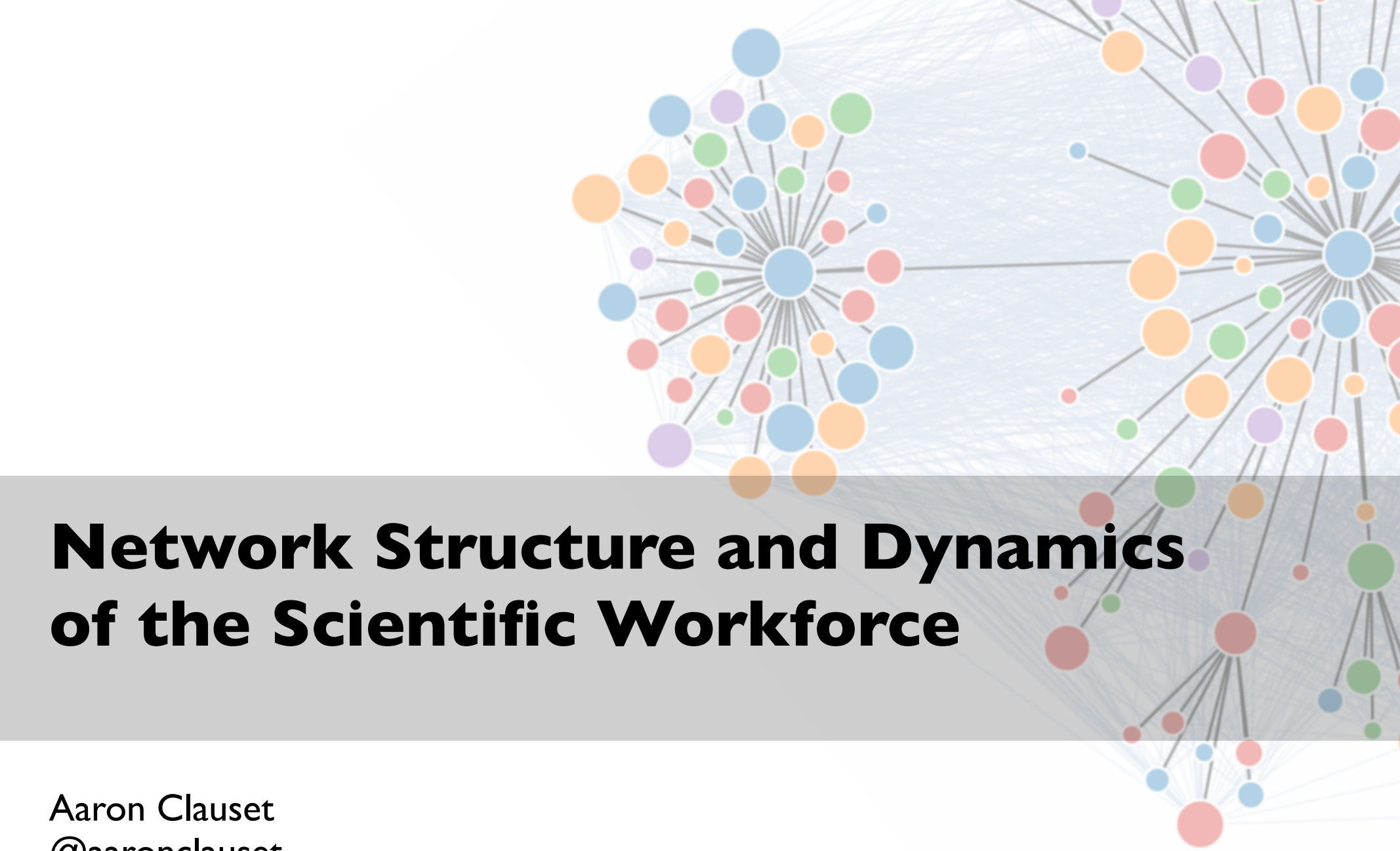
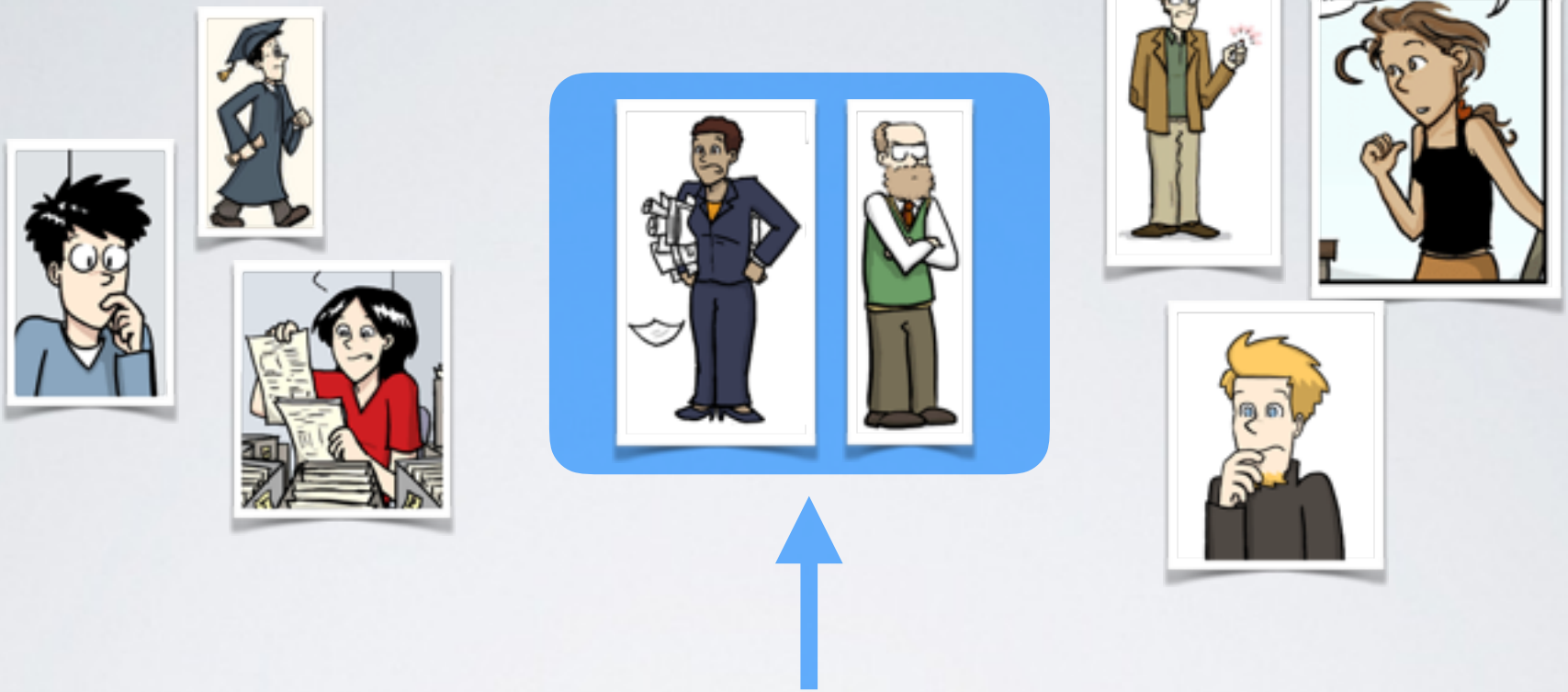


Network Structure and Dynamics of the Scientific Workforce

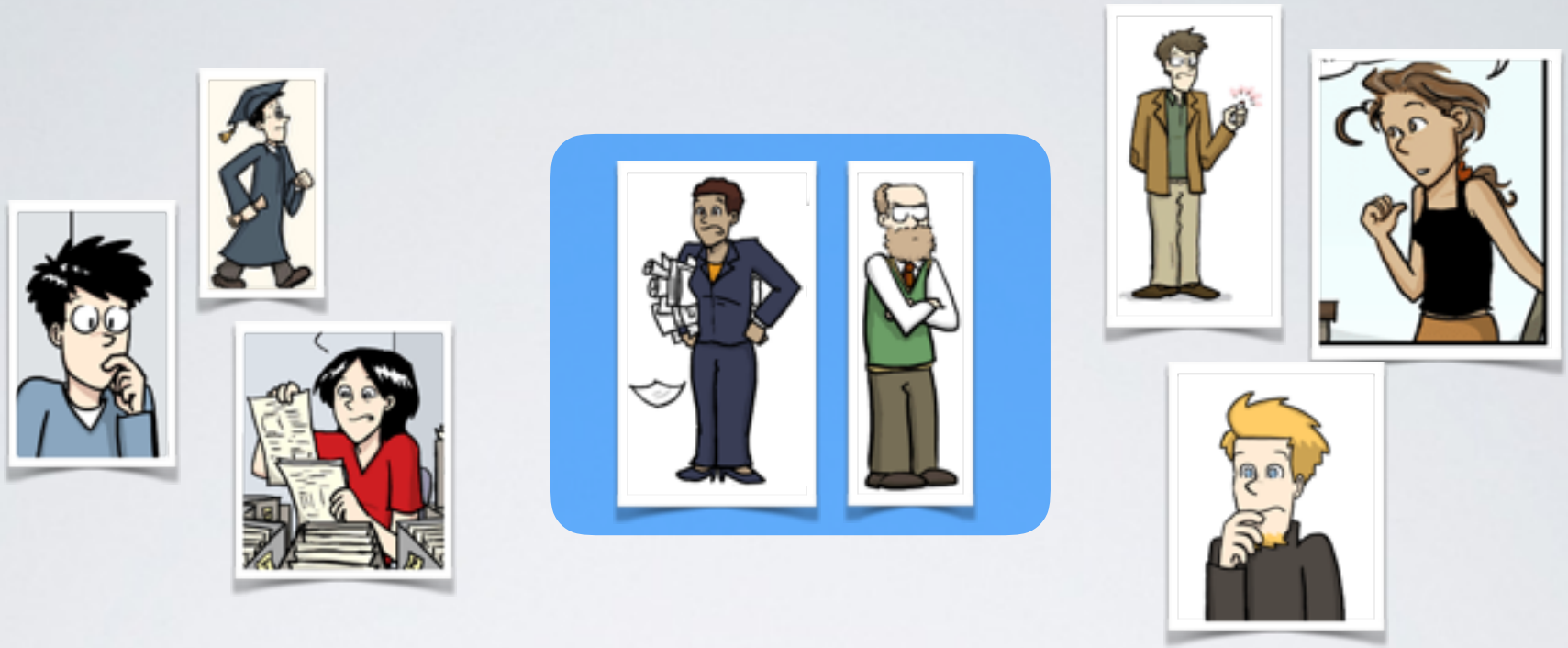
Aaron Clauset
@aaronclauset
Computer Science Dept. & BioFrontiers Institute
University of Colorado, Boulder
External Faculty, Santa Fe Institute

10 May 2016



faculty

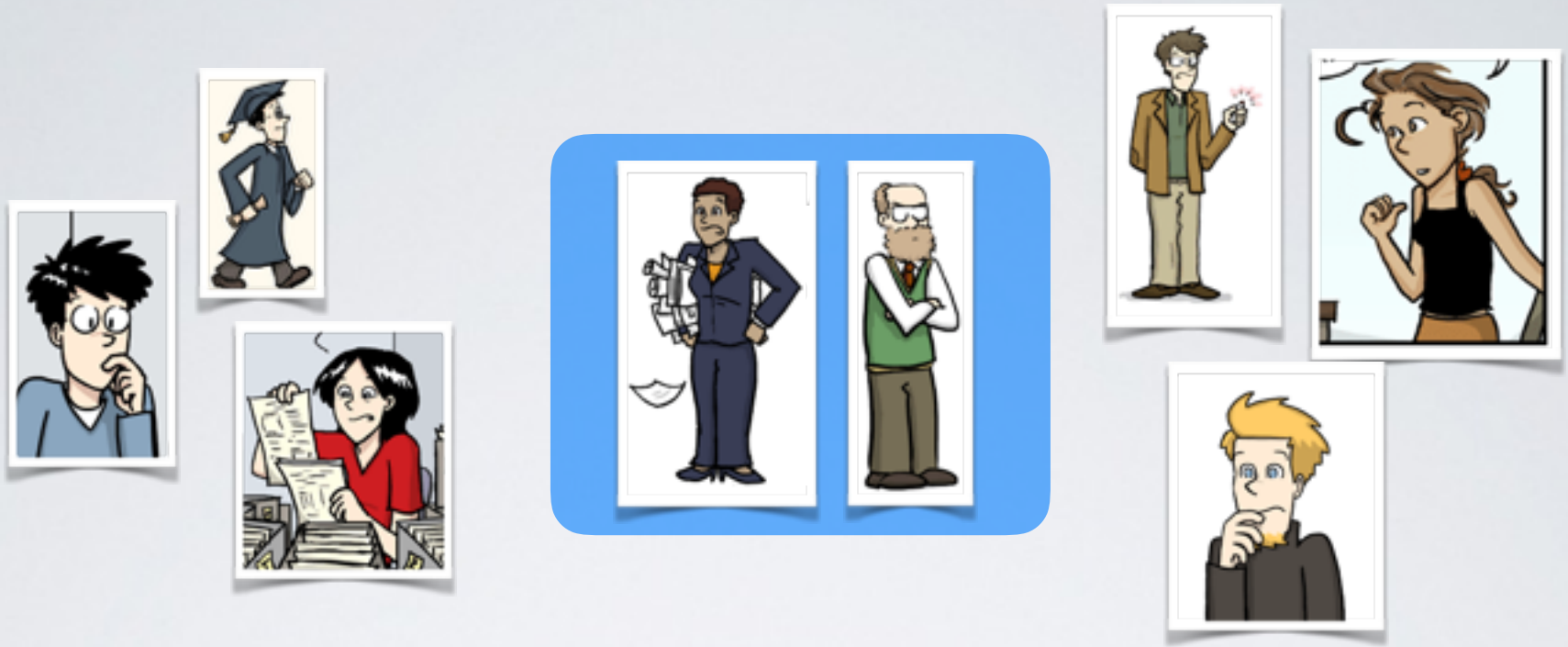
[1] these awesome cartoons copyright Jorge Cham, via PhDComics.com



- 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?

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

who hires whose graduates as faculty?


COLLECT
ALL THE DATA
memegenerator.net

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

```
>>> record 1059
# facultyName : James H. Martin
# email       : [redacted]
# sex         : M
# department  : Computer Science
# place       : University of Colorado, Boulder
# current     : Full Professor
# [Education]
# degree      : BS
# place       : Columbia University
# field       : Computer Science
# years       : ????-1981
# [Education]
# degree      : PhD
# place       : UC Berkeley
# field       : Computer Science
# years       : ????-1988
# [Faculty]
# rank        : Assistant Professor
# place       : University of Colorado, Boulder
# years       : 1989-1995
# [Faculty]
# rank        : Associate Professor
# place       : University of Colorado, Boulder
# years       : 1995-2007
# [Faculty]
# rank        : Full Professor
# place       : University of Colorado, Boulder
# years       : 2007-2011
# recordDate  : 7/4/2011
```

[1] CS data from 2011, Business schools from 2012, History from 2013
[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

```
>>> record 1059
# facultyName : James H. Martin
# email       : [redacted]
# sex         : M
# department  : Computer Science
# place       : University of Colorado, Boulder
# current     : Full Professor
# [Education]
# degree      : BS
# place       : Columbia University
# field       : Computer Science
# years       : ????-1981
# [Education]
# degree      : PhD
# place       : UC Berkeley
# field       : Computer Science
# years       : ????-1988
# [Faculty]
# rank        : Assistant Professor
# place       : University of Colorado, Boulder
# years       : 1989-1995
# [Faculty]
# rank        : Associate Professor
# place       : University of Colorado, Boulder
# years       : 1995-2007
# [Faculty]
# rank        : Full Professor
# place       : University of Colorado, Boulder
# years       : 2007-2011
# recordDate  : 7/4/2011
```

[1] CS data from 2011, Business schools from 2012, History from 2013
[2] all data from public sources, mainly faculty CVs and homepages

collect all the data

	Computer Science	Business	History
institutions	205	112	144
tenure-track faculty	5032	9336	4556
mean size	25	83	32

$\sum = 18,924$

collect all the data

	Computer Science	Business	History
institutions	205	112	144
tenure-track faculty	5032	9336	4556
mean size	25	83	32
Full Professors	2400 (48%)	4294 (46%)	2097 (46%)
Associate Prof.	1772 (35%)	2521 (27%)	1611 (35%)
Assistant Prof.	860 (17%)	2521 (27%)	848 (19%)
female	15%	22%	36%

$\sum = 18,924$

collect all the data

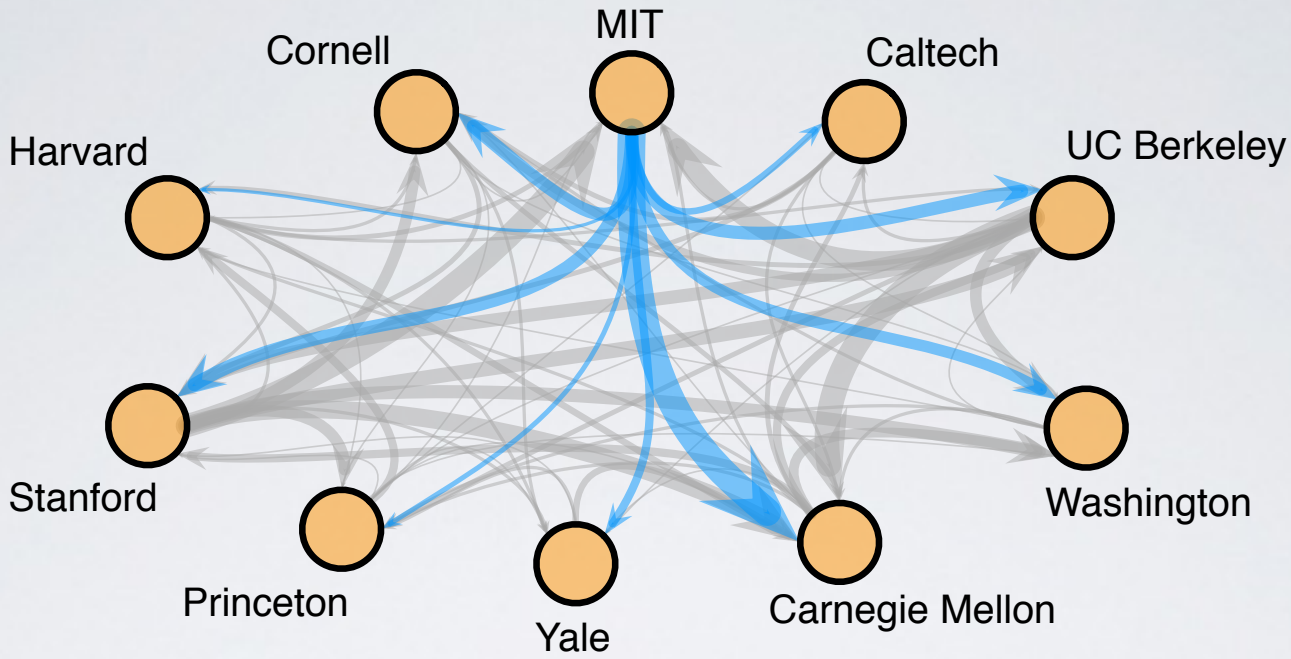
	Computer Science	Business	History
institutions	205	112	144
tenure-track faculty	5032	9336	4556
mean size	25	83	32
Full Professors	2400 (48%)	4294 (46%)	2097 (46%)
Associate Prof.	1772 (35%)	2521 (27%)	1611 (35%)
Assistant Prof.	860 (17%)	2521 (27%)	848 (19%)
female	15%	22%	36%
PhDs in-sample	87%	84%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

$\sum = 18,924$

nearly closed hiring systems

faculty market is a *network*

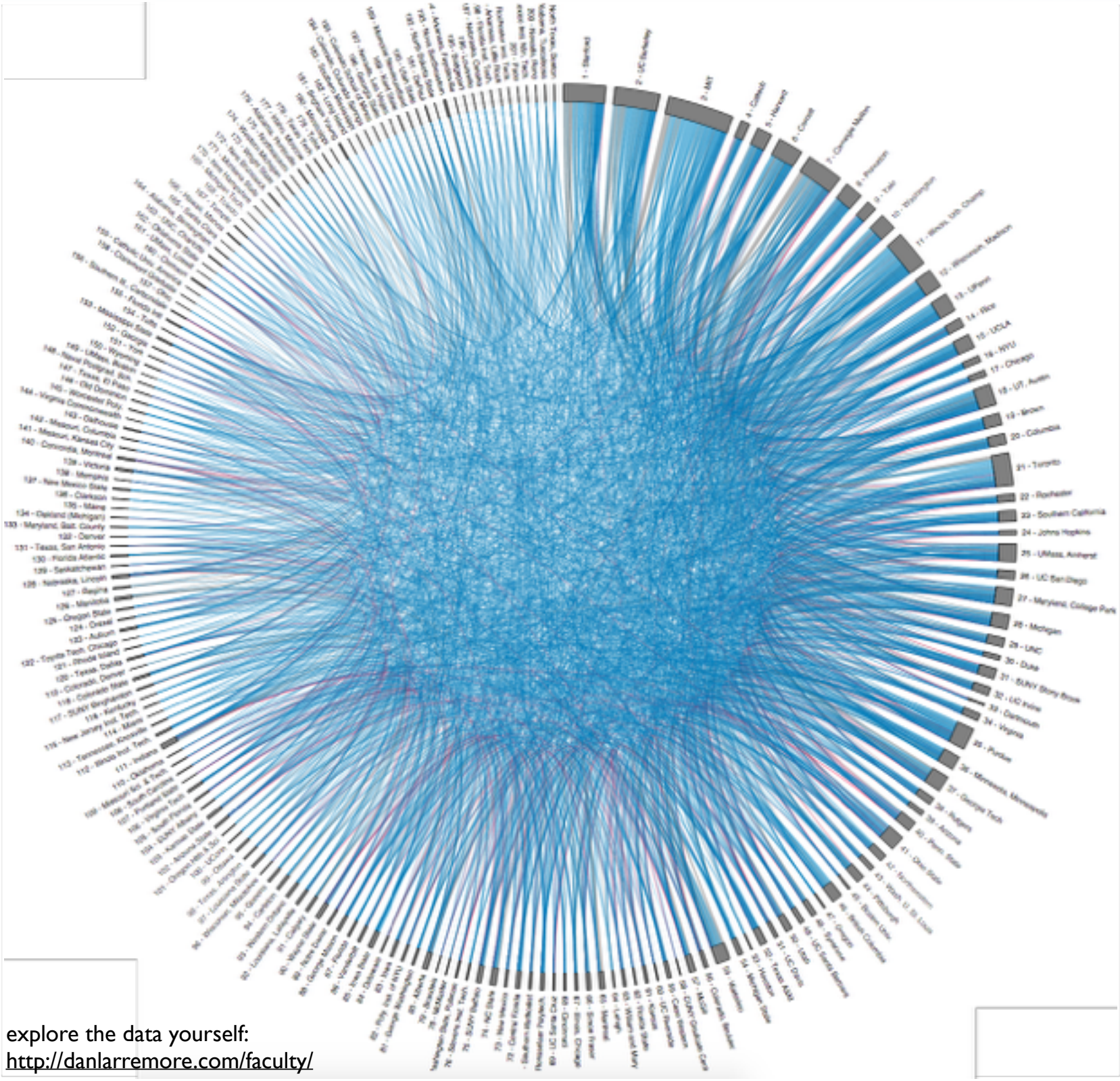
- vertices are PhD-granting universities
- consumers $\leftrightarrow$ producers
- v hires from u, add an edge $u \rightarrow v$



faculty market is a *network*

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

[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

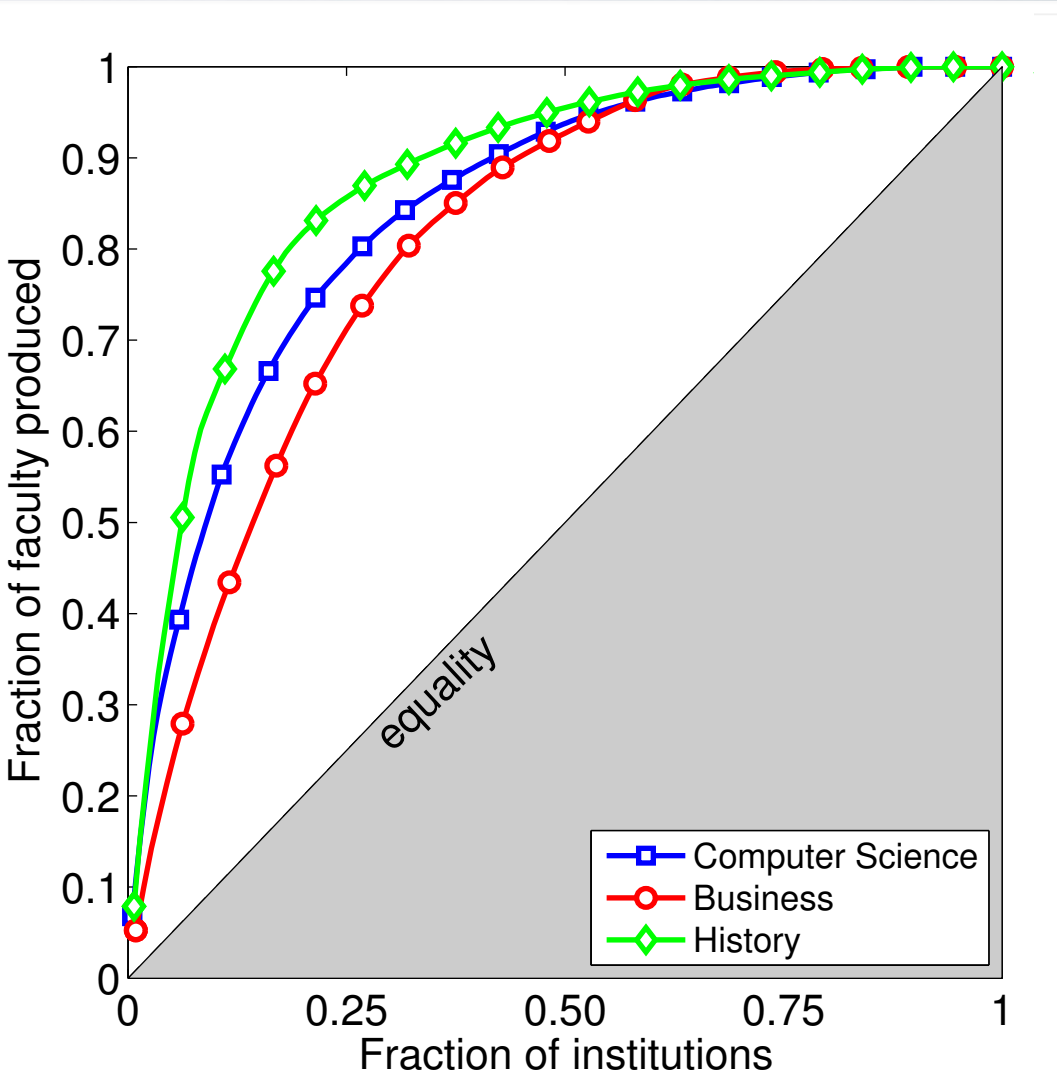
- 18, 16, 8 universities

net producers $k_{\text{out}}/k_{\text{in}} > 1$

- 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

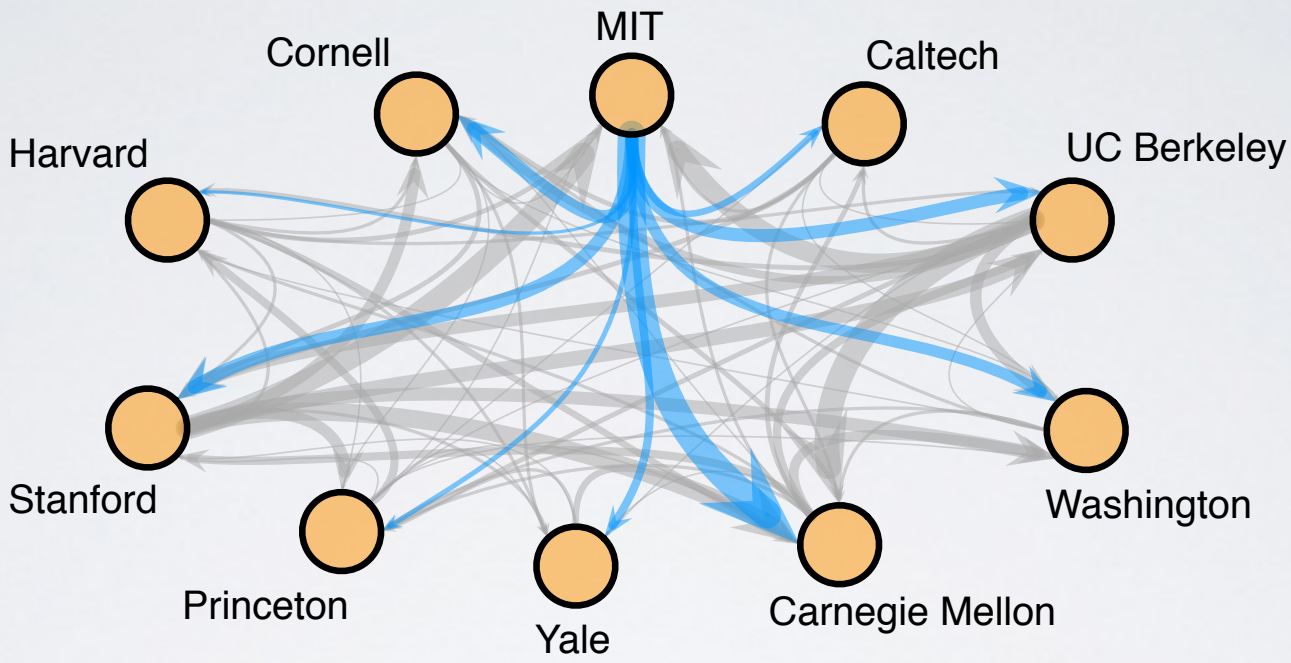


[1] order: CS, Business, History
[2] U.S. Income Gini coefficient = 0.45

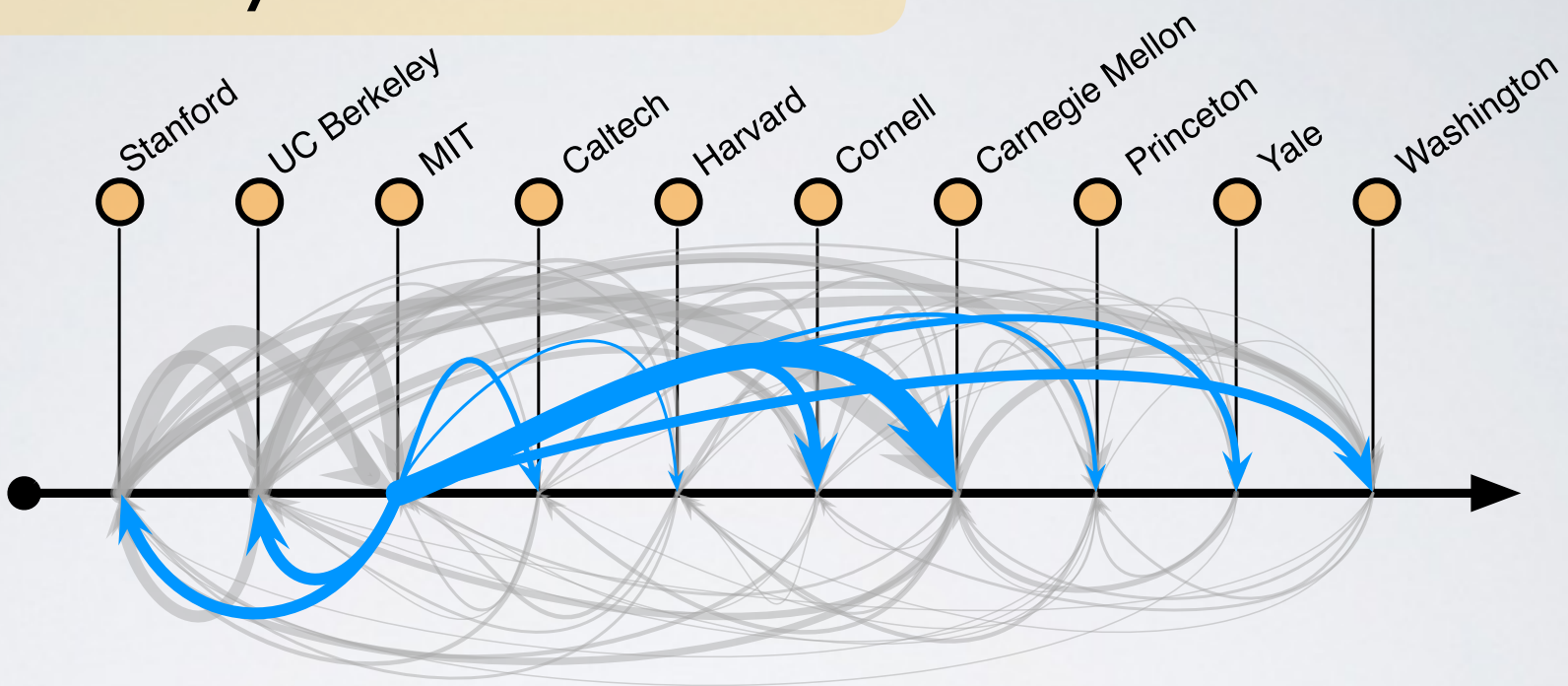
a prestige hierarchy

- difficult to talk about inequalities in academia without talking about rankings
- let's extract a data-driven ranking from the network

a prestige hierarchy



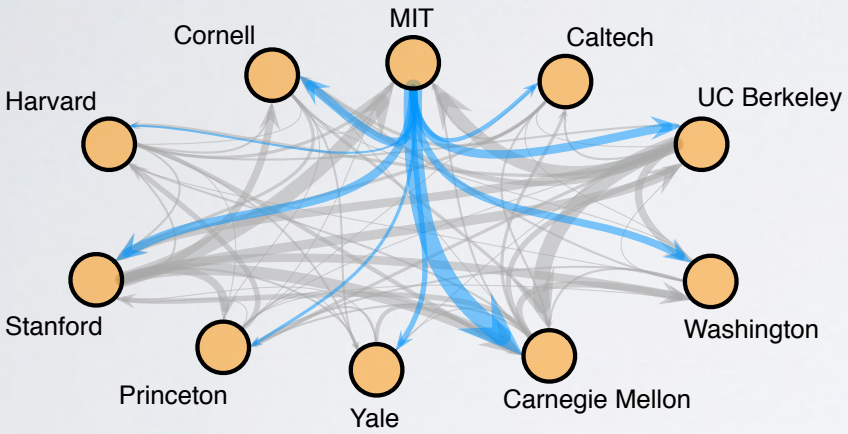
a prestige hierarchy



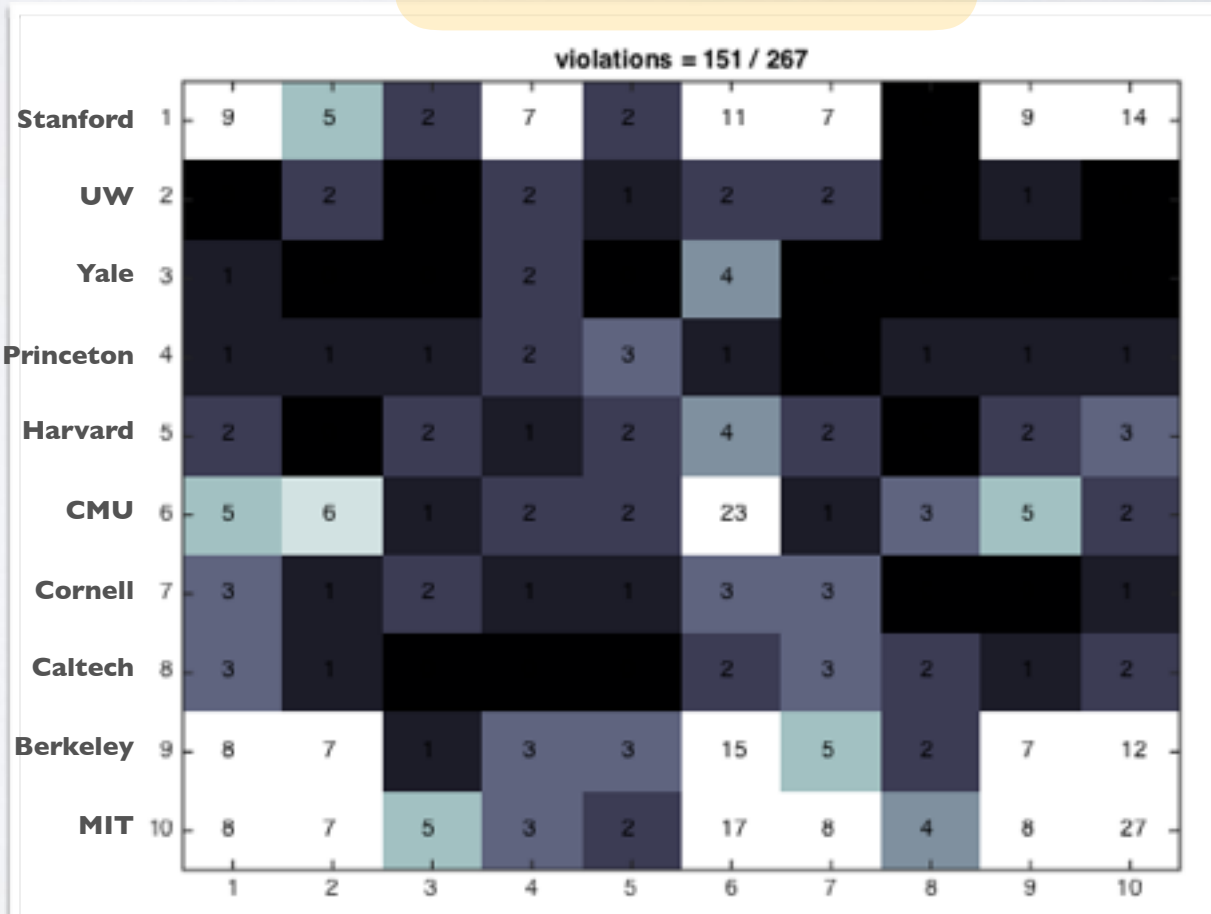
- select permutation (a ranking) π that minimizes the number of "rank violations" : edges (u, v) where $\pi_v < \pi_u$
- higher-ranked nodes have greater "placement power"
- equivalent to *minimum feedback arc set* problem (NP-hard)

[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 $\langle \pi \rangle$

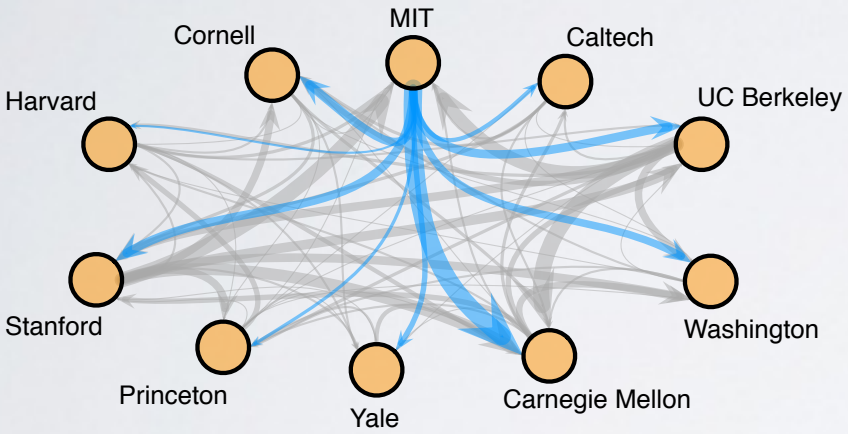
- given an ordering π with $\psi(\pi, A)$ rank violations on network A
- repeat *ad infinitum*: choose a pair (u, v), swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain π', compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$
- for instance:



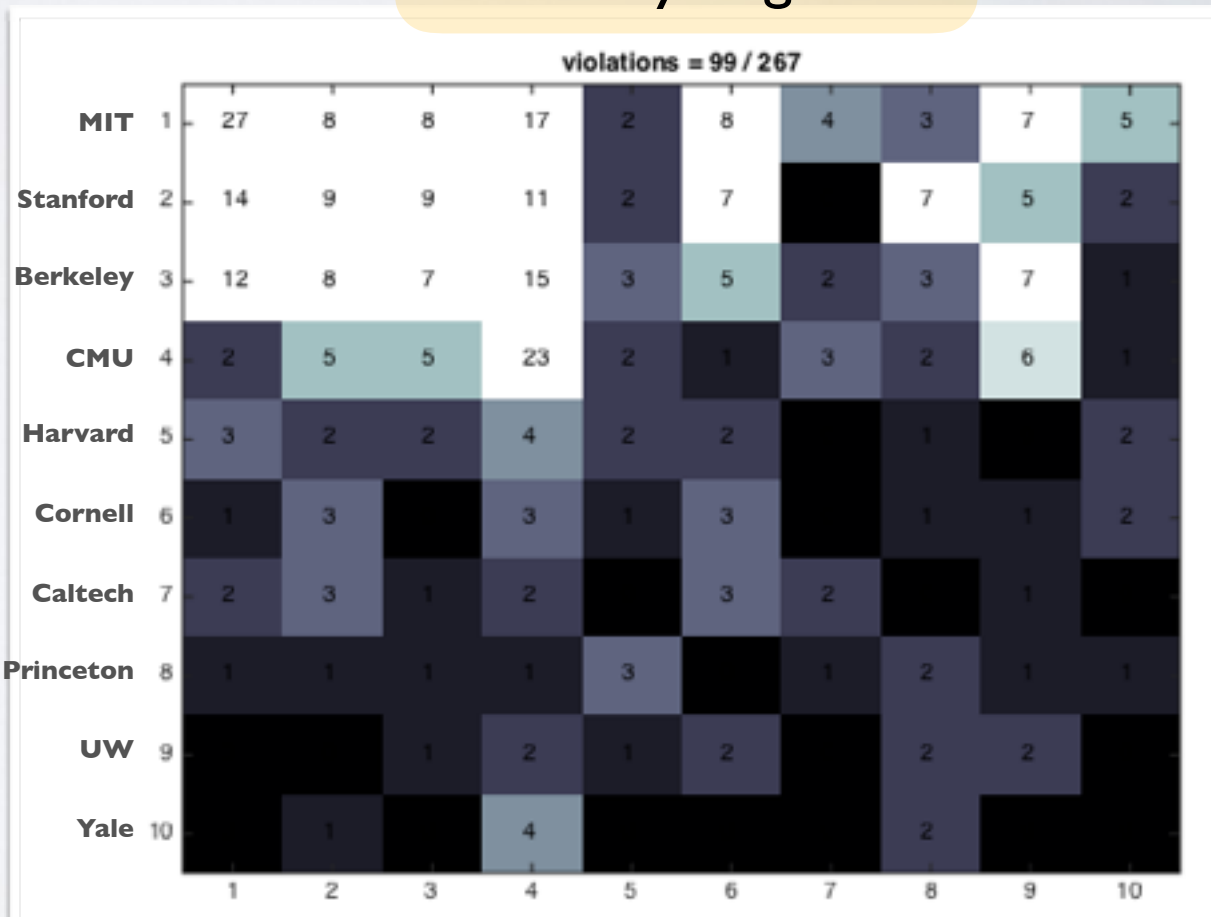
random



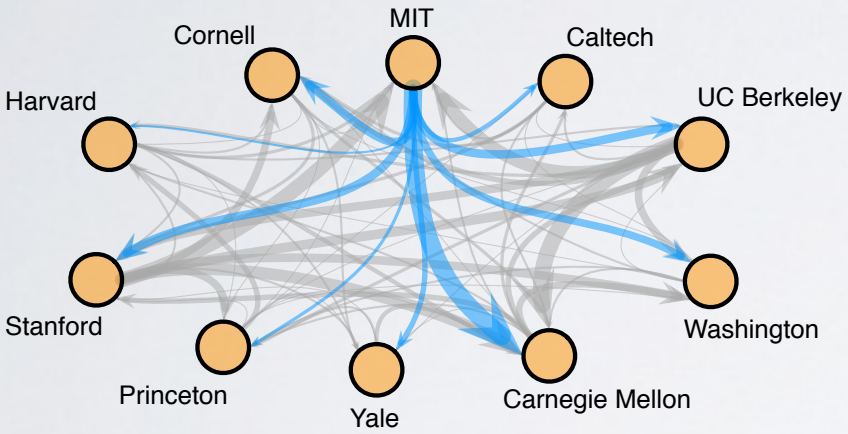
- given an ordering π with $\psi(\pi, A)$ rank violations on network A
- repeat *ad infinitum*: choose a pair (u, v), swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain π', compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$
- for instance:



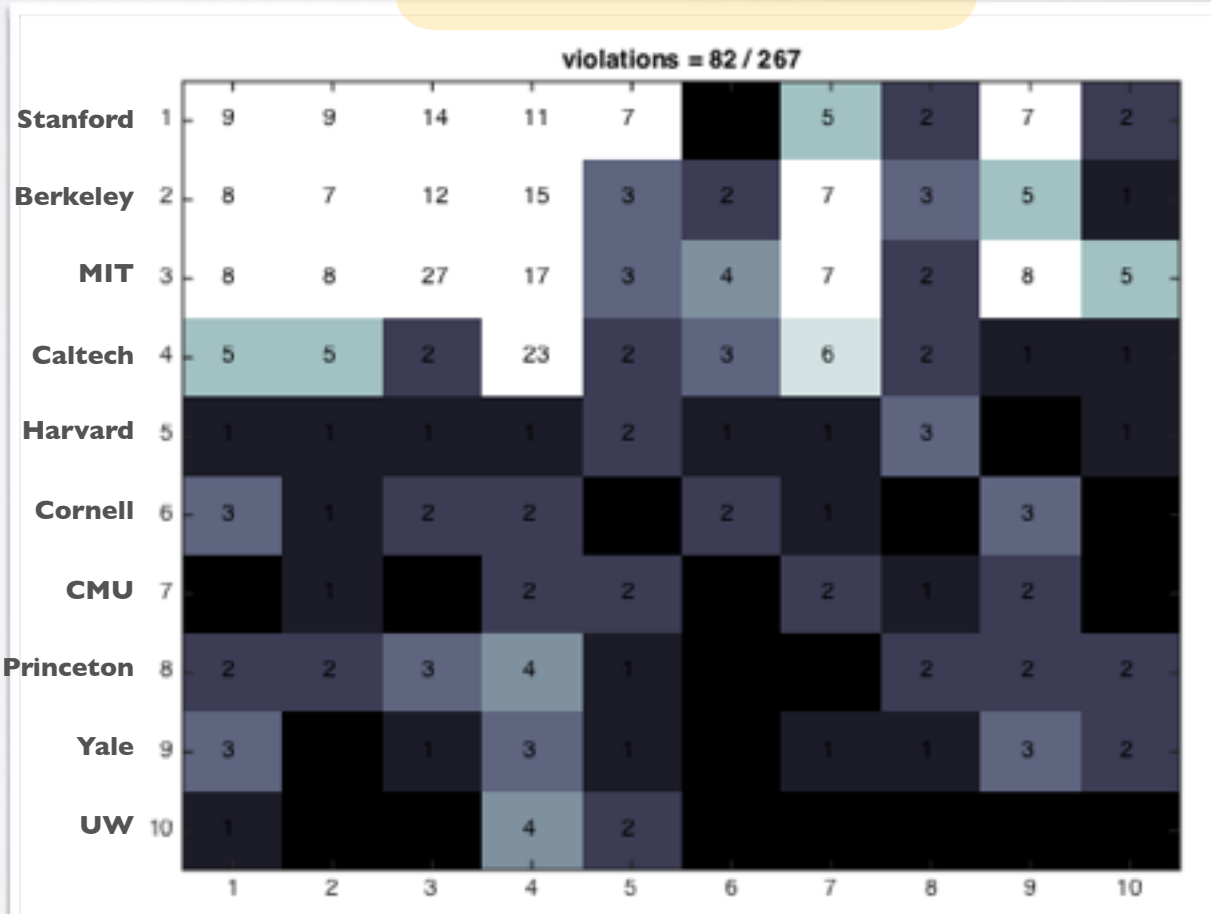
sort by degree



- given an ordering π 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 π', compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$
- for instance:

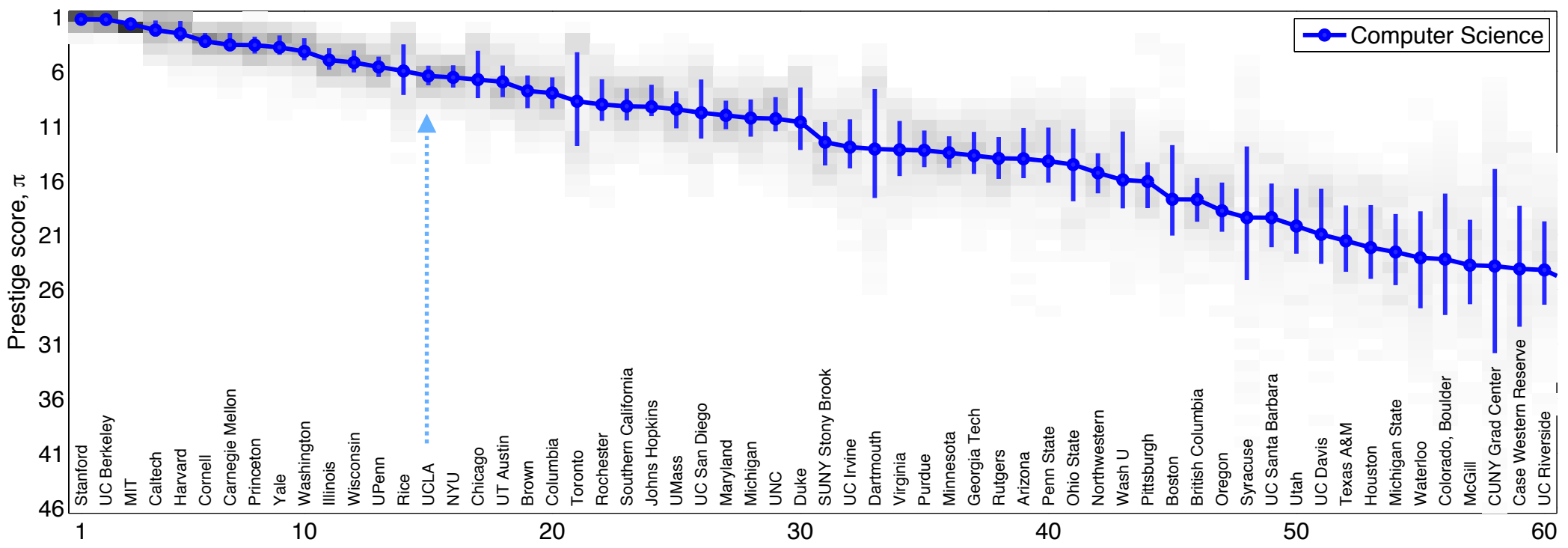


MVR



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* π quantifies *placement power*
- uncertainty increases as prestige decreases
- similar results, but different orderings for Business and History

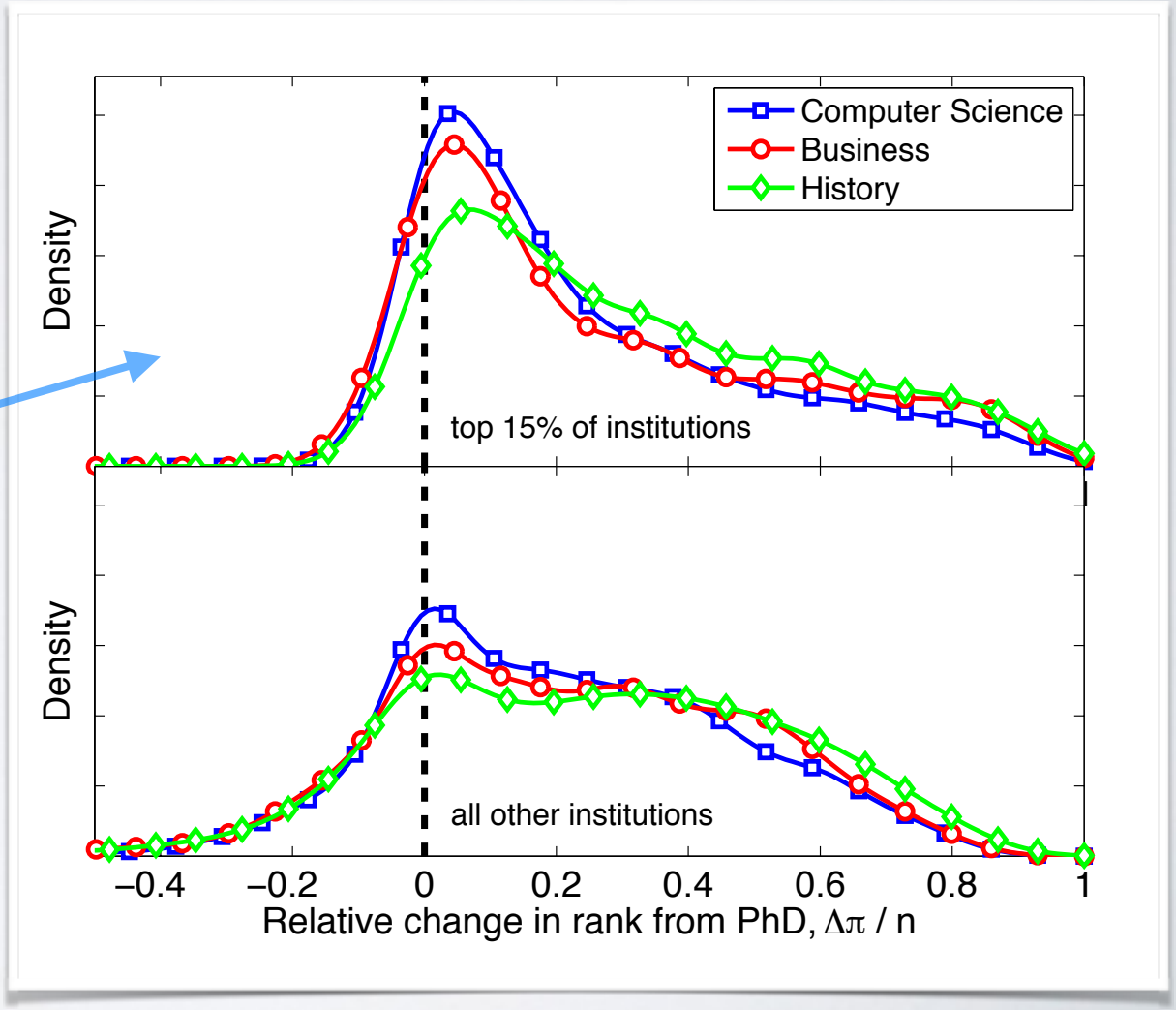
[1] uncertainties derived from the set of MVRs

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



[1] order: CS, Business, History

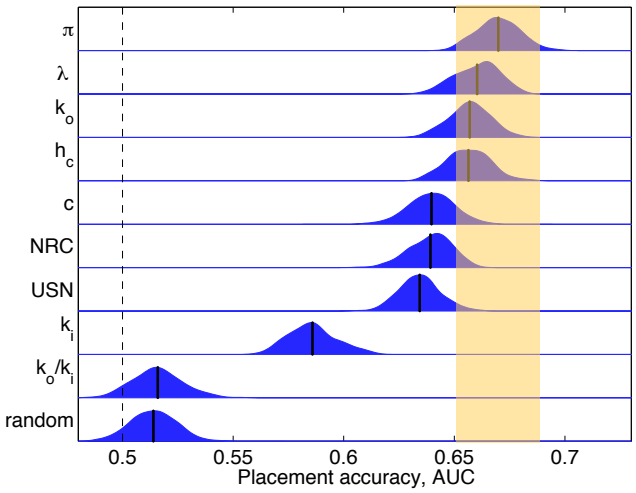
what predicts placement?

- compare 10 single features:

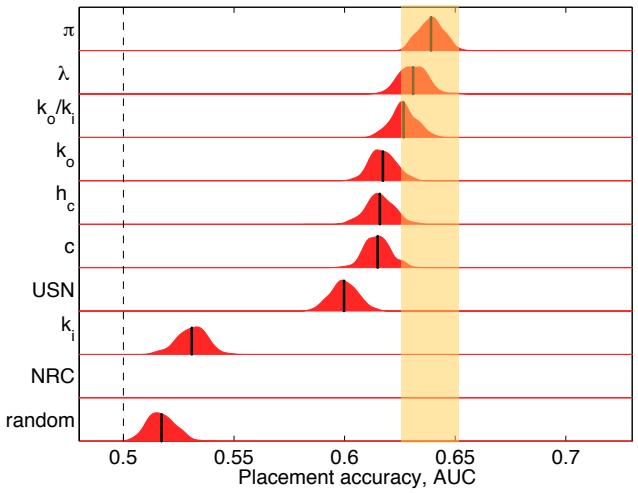
  prestige
  US News rank
  NRC rank
  out-degree
  in-degree
  out/in degree
  eigenvector centrality
  harmonic centrality
  closeness centrality
  random

what predicts placement?

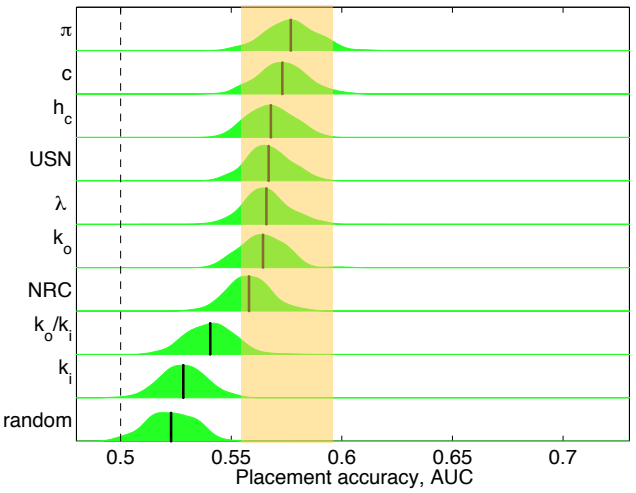
- prestige best *single* predictor in all 3 fields
- order of other features varies by field
- AUCs all below 0.67 = plenty of room for improvement



CS



Business



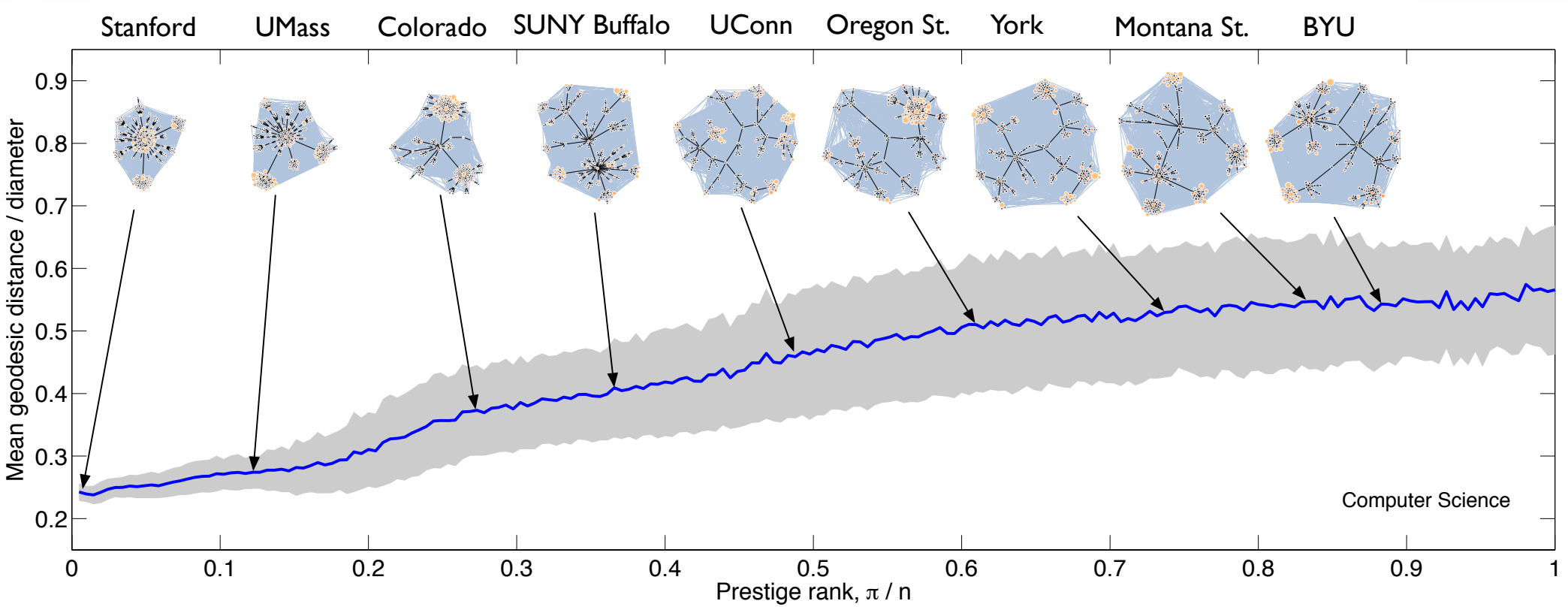
History

prestige correlates with network position

- core and periphery

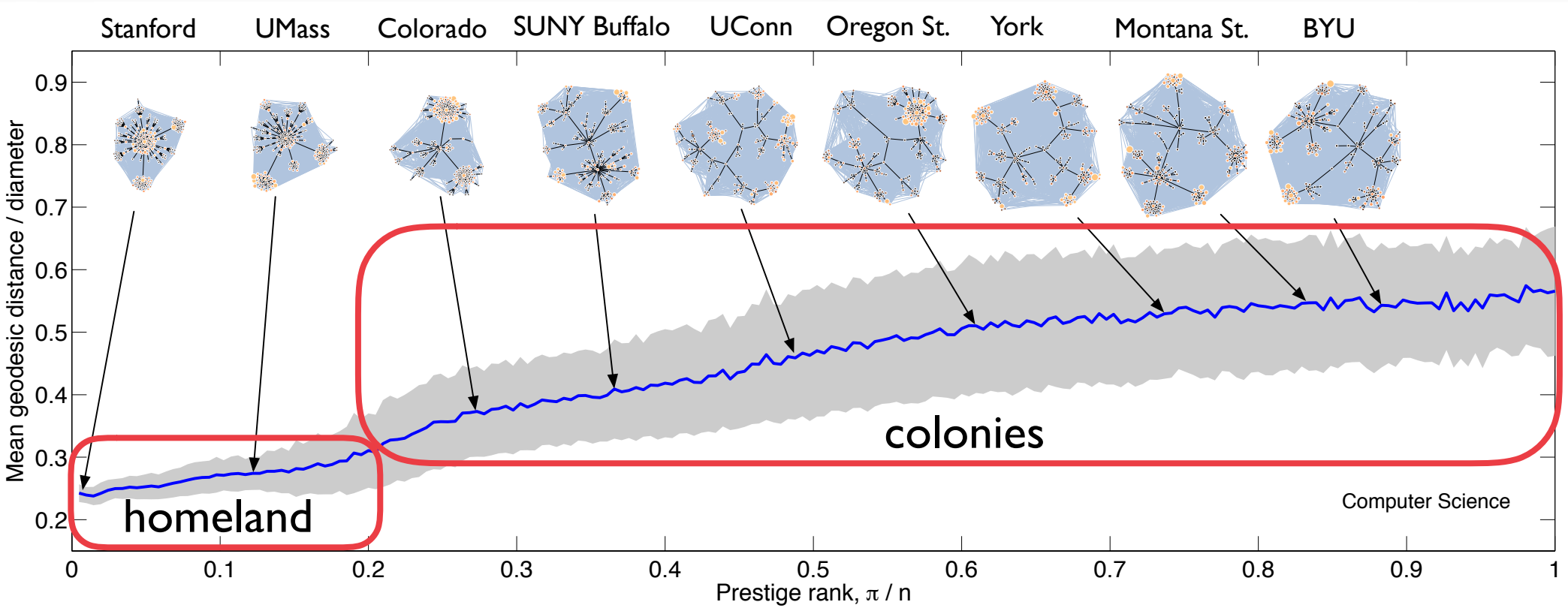
prestige correlates with network position

- 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 ~10% of hires flow "upstream"

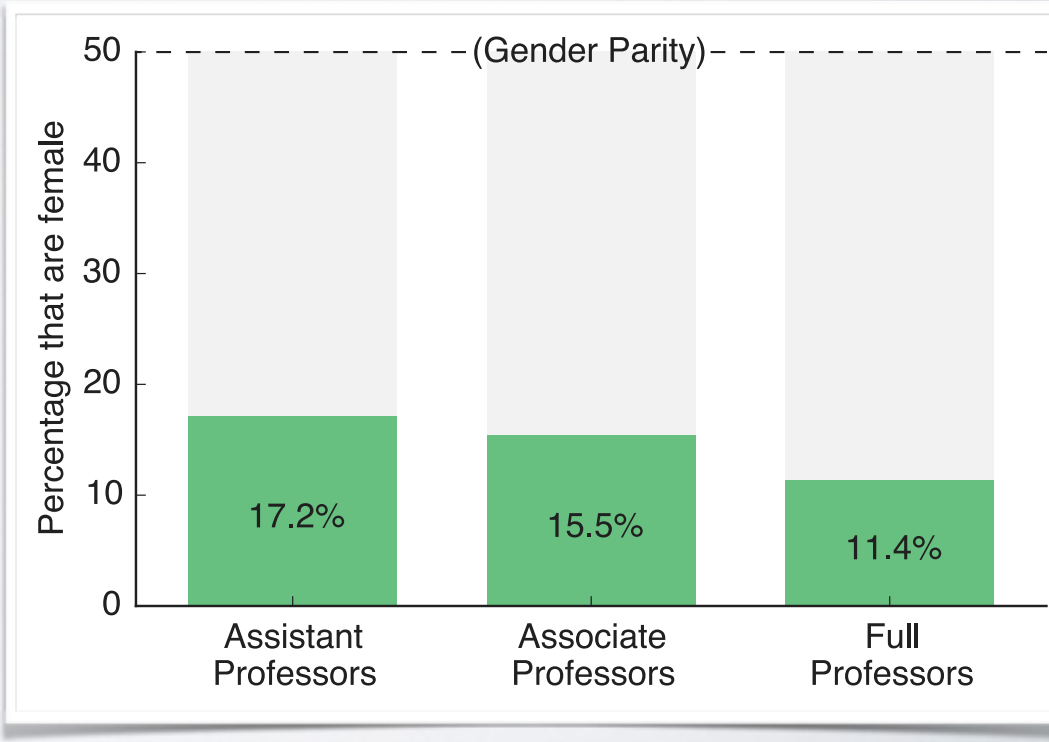
future work

- 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



[1] National Center for Education Statistics (2011 data)
[2] Reliably industry-wide estimates are hard to come by, but see http://cnet.co/1GZh268
[3] Clauset, Arbesman, Larremore, Science Advances 1(1), e1400005 (2015)

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)

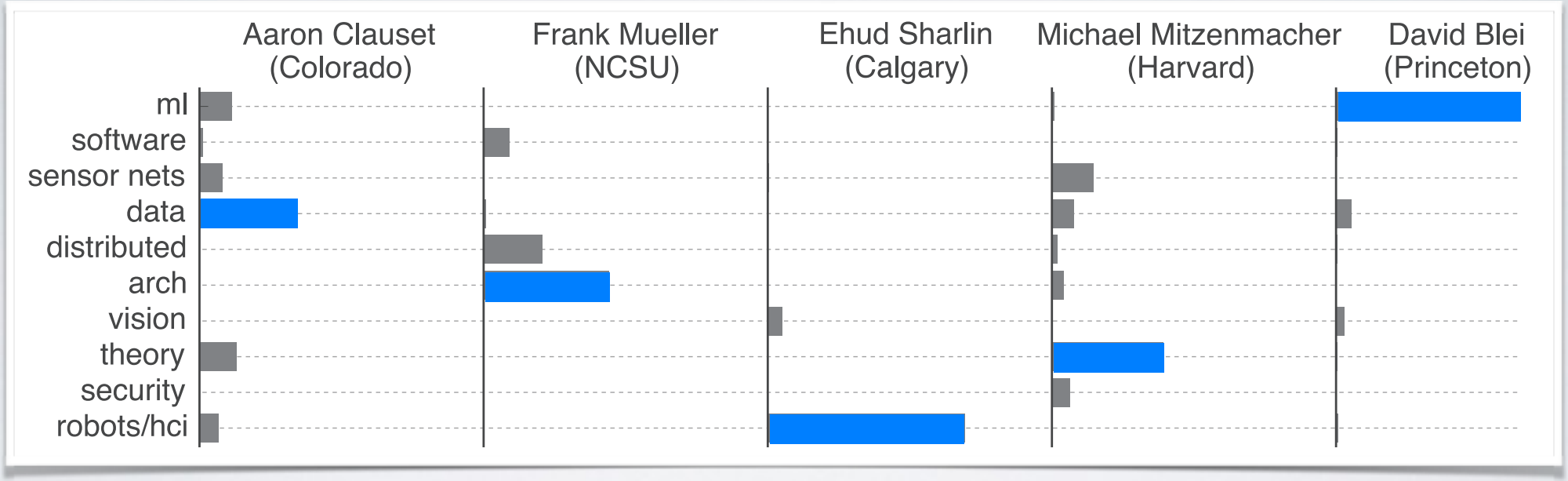
- consider 6 factors:

  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 ρ_i relative to their subfield mixture

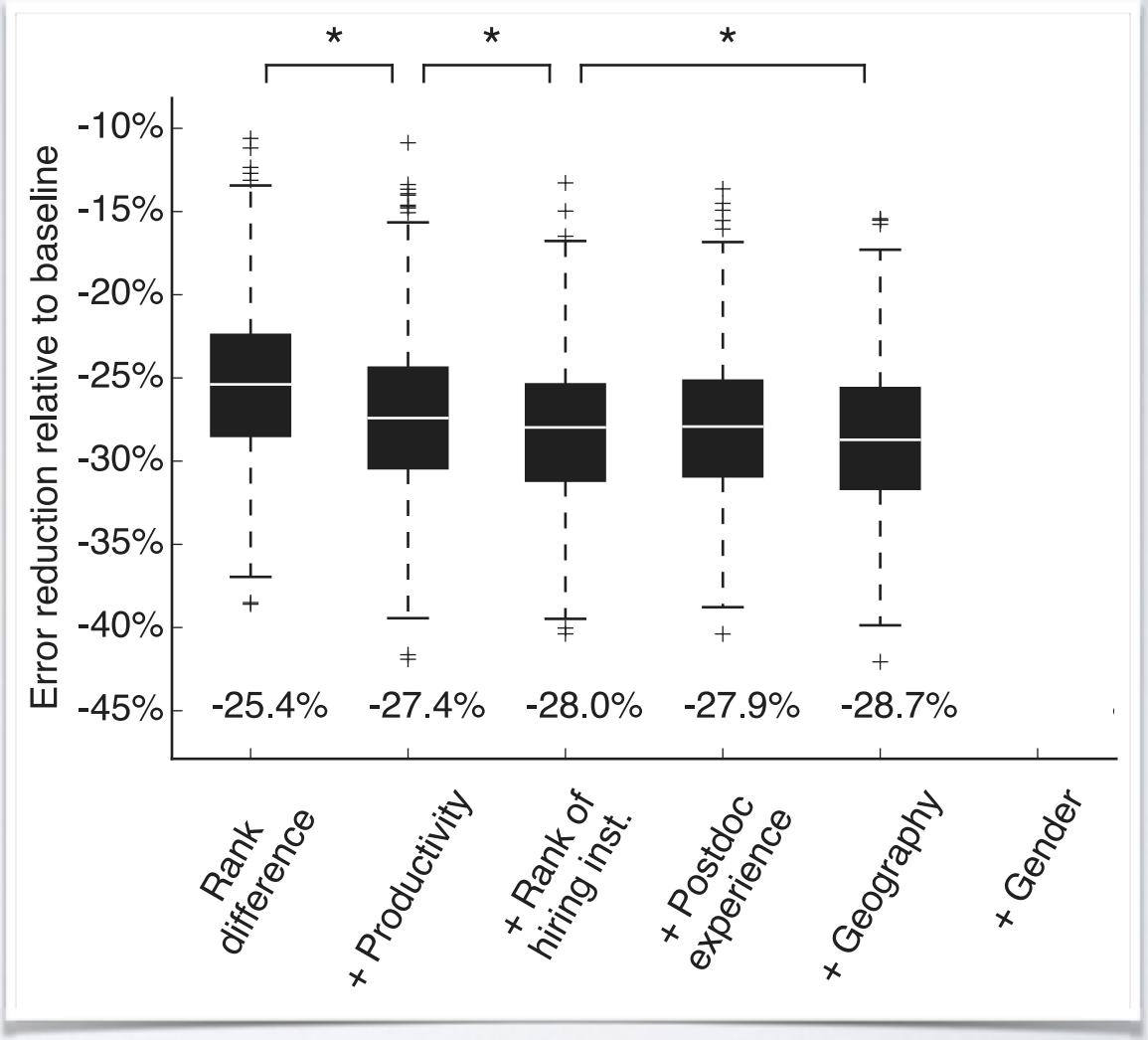
fitting the model

fitting the model

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

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

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



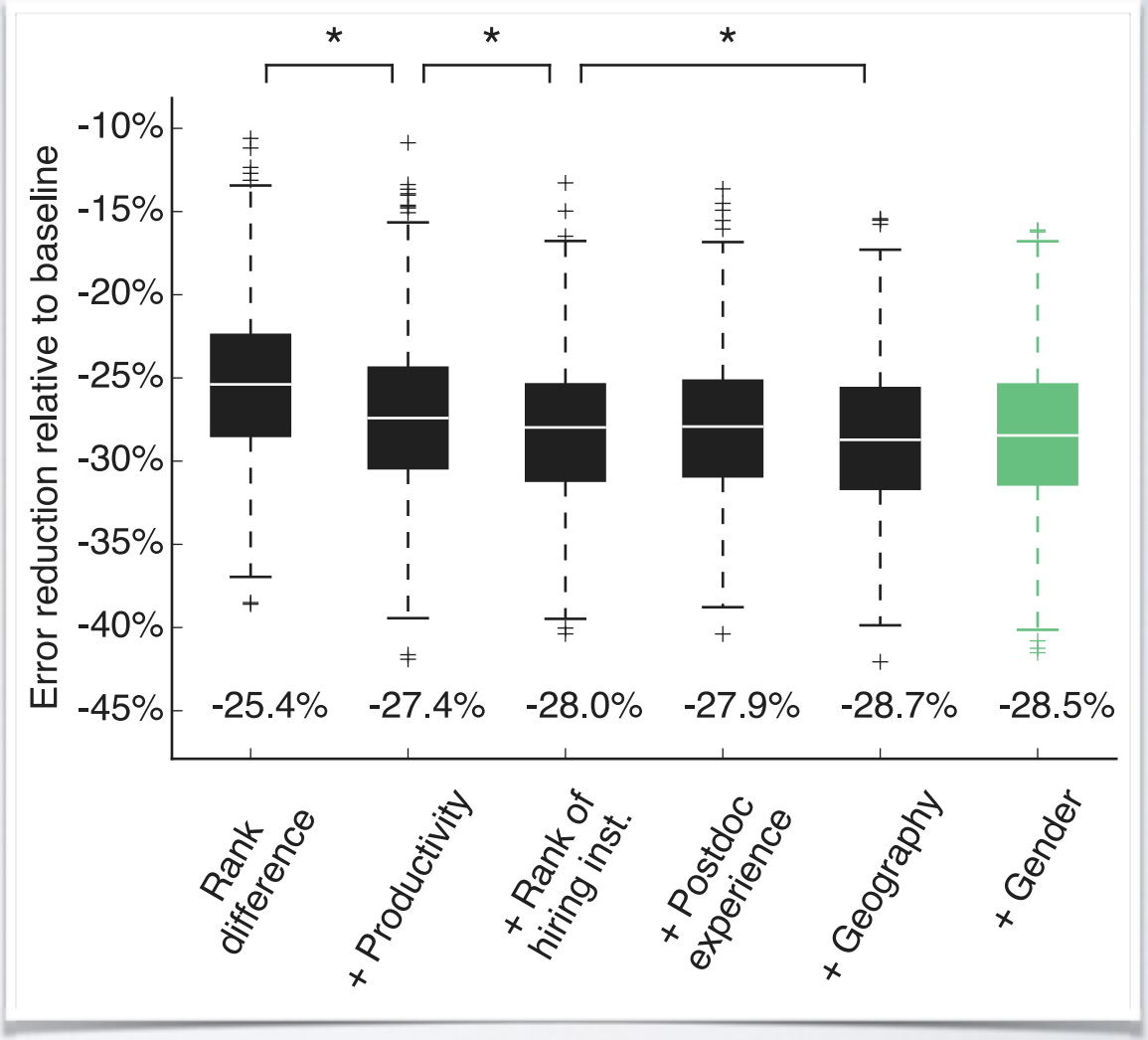
[1] fitting done by Nelder-Mead

fitting the model

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

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

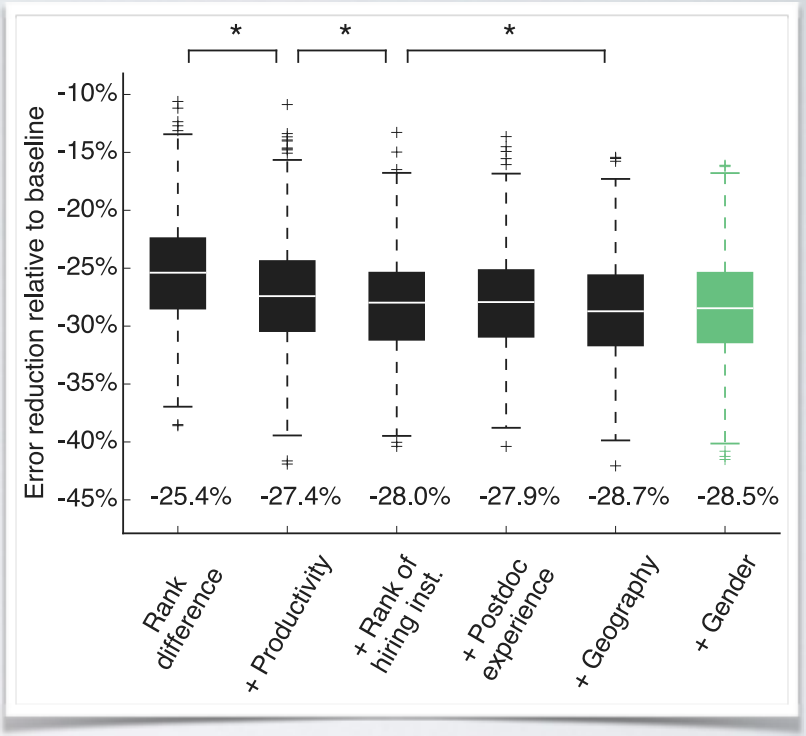
- add covariates one-by-one, in greedy fashion, with gender added last



[1] fitting done by Nelder-Mead

system-level results

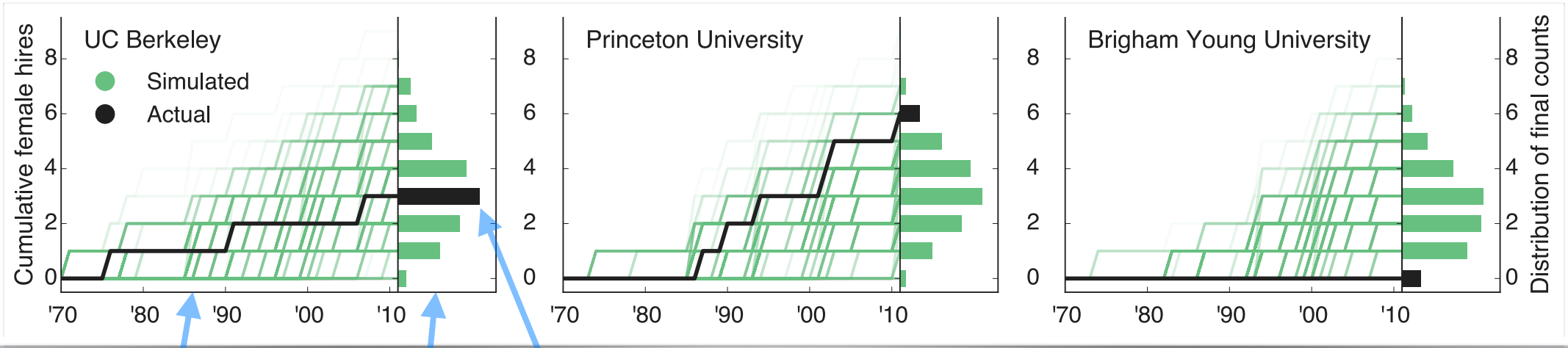
- adding gender *does not improve* placement accuracy
- three possibilities:
  1. gender is irrelevant to hiring decisions
  2. we modeled gender's effect incorrectly
  3. gender's effect is included in the other variables [see footnotes]



[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



1000 simulated histories of hiring

simulated distribution of hires

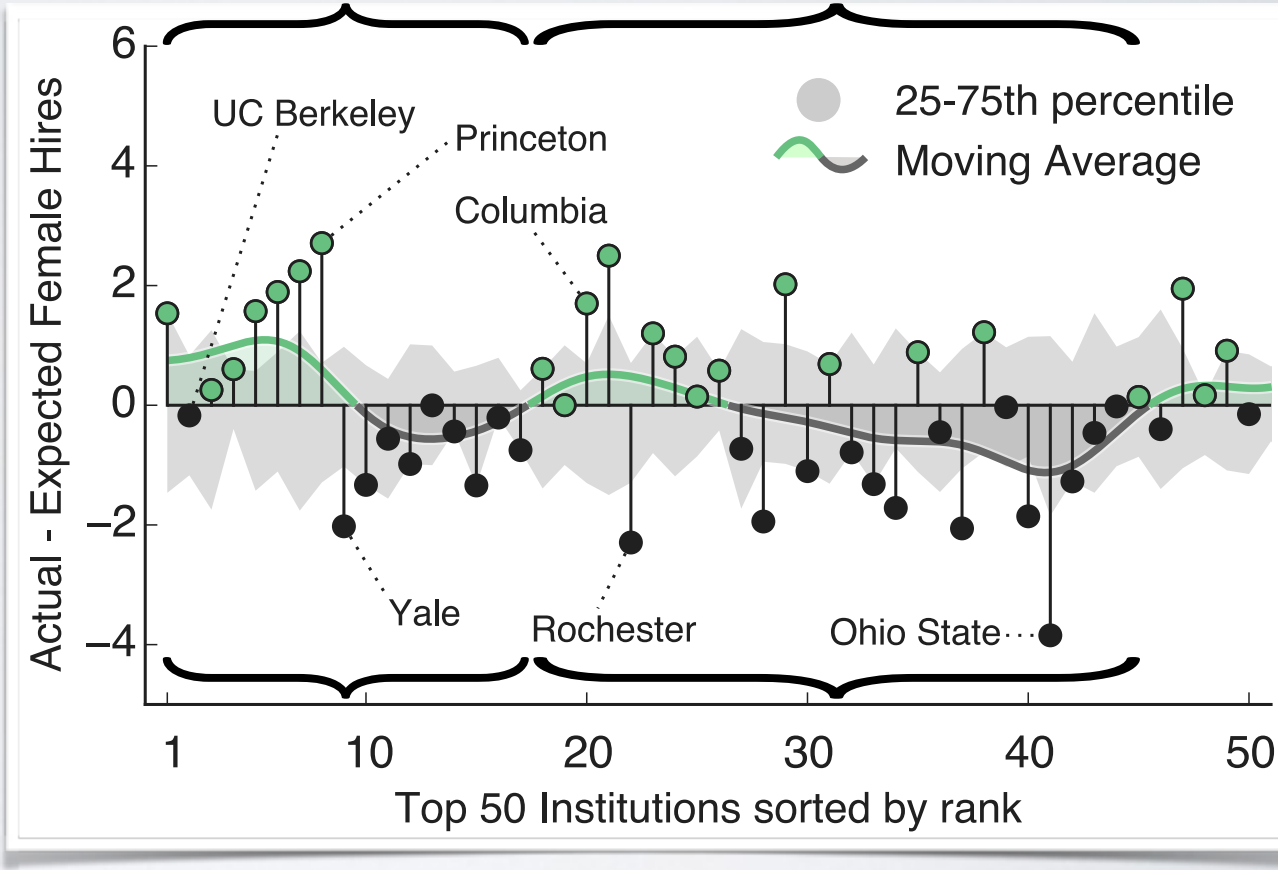
observed count in 2011

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

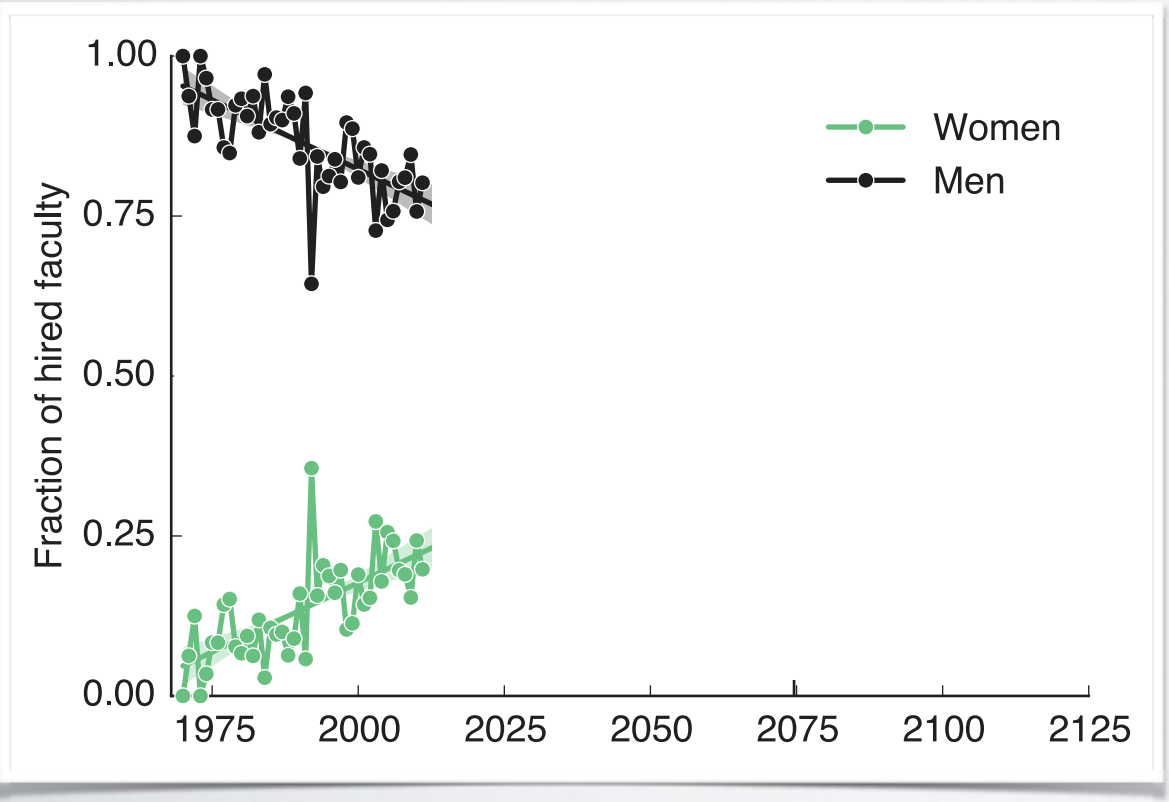


[1] the errors our model makes are interesting, and non-uniform. for instance, men and women tend to exceed the model's expectations at similar rates; but, for under-performing individuals, men tend to fall short of expectations by a wider margin. furthermore, people with postdoc training tend to exceed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

gender parity

gender parity

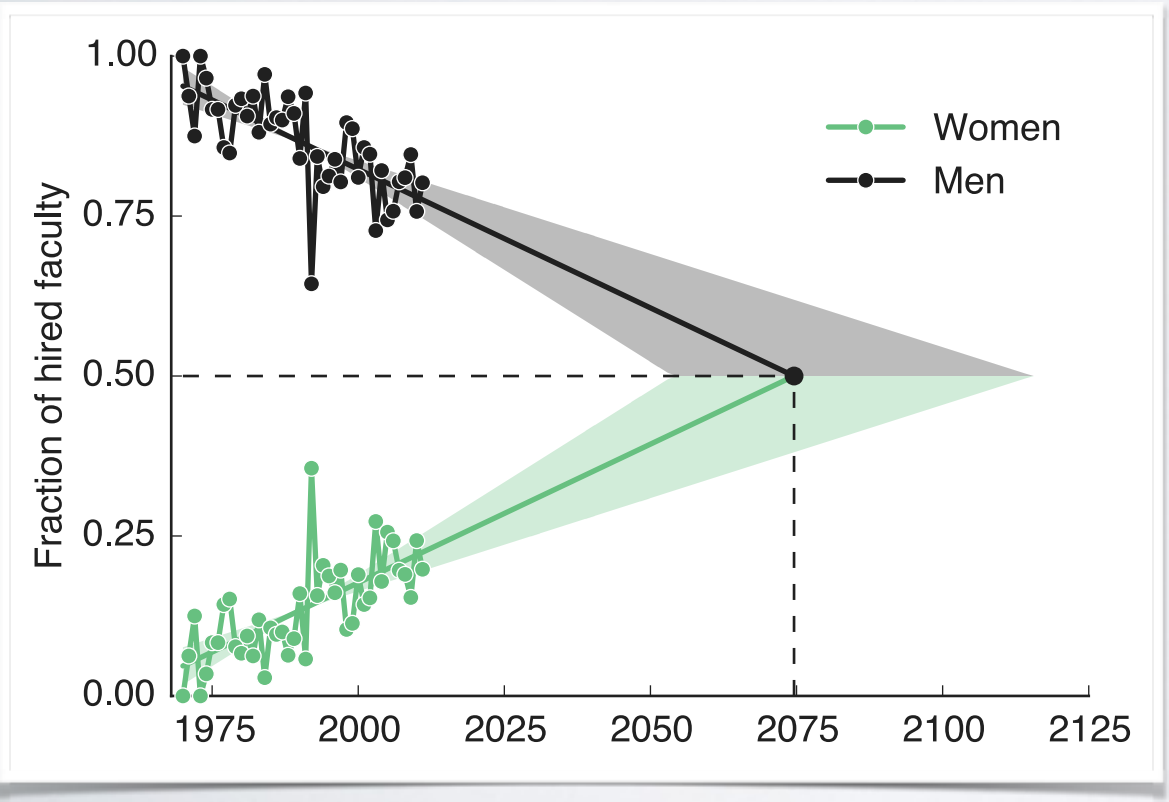
- where is the CS faculty gender ratio going?
- from our 40 years of observable data, the trend is toward parity
- ratio increases by 0.43% per year



[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*

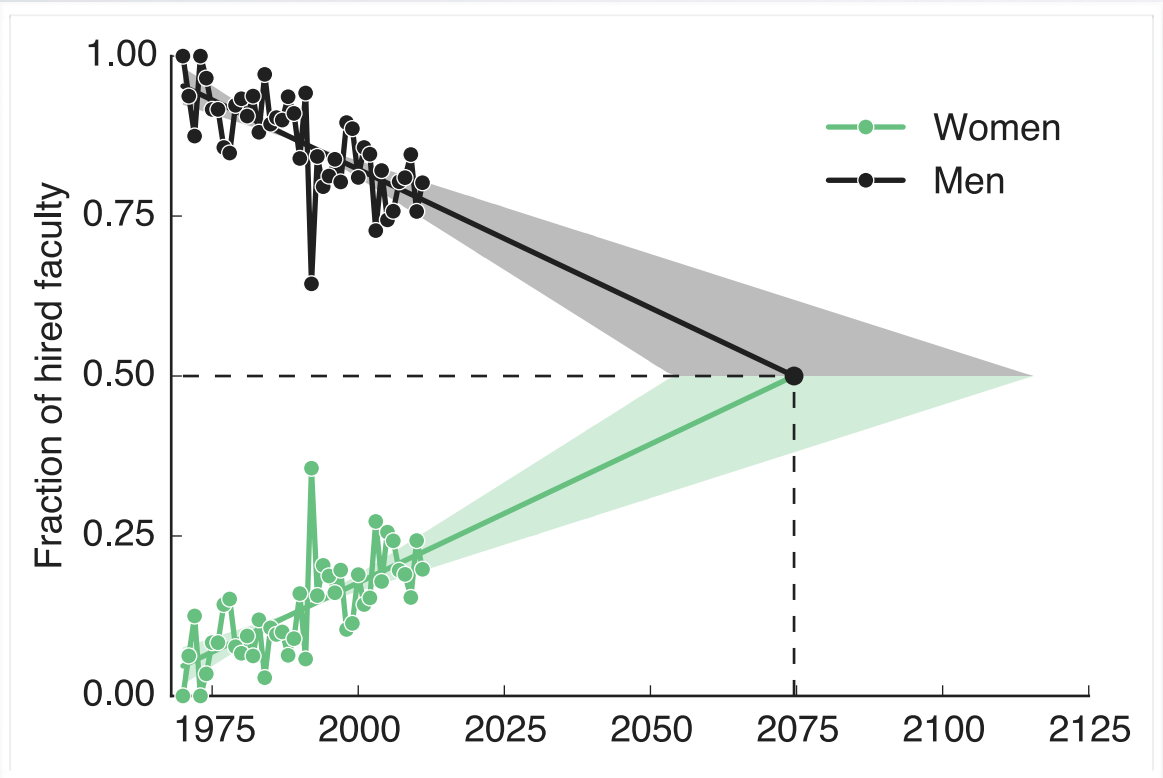


can we change this?!

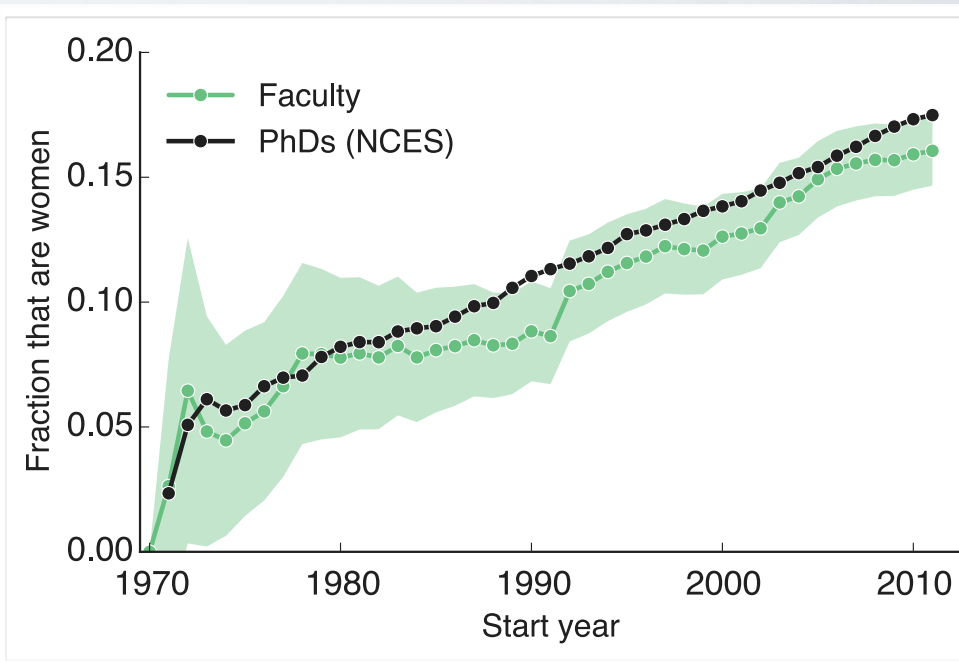
[1] 1992 was a genuine outlier, with 21 of 59 (36%) hires being women; mean of 3 years before and after were 10% and 18%

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*



can we change this?!



[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

Sam Way
(Colorado)

Dan Larremore
(Santa Fe)

Sam Arbesman
(Lux Capital)

RESEARCH ARTICLE

NETWORK SCIENCES

Systematic inequality and hierarchy in faculty hiring networks

Aaron Clauset,[1,2,3]* Samuel Arbesman,[4] Daniel B. Larremore[5,6]

10.1126/sciadv.1400005

Science Advances 1(1), e1400005 (2015)

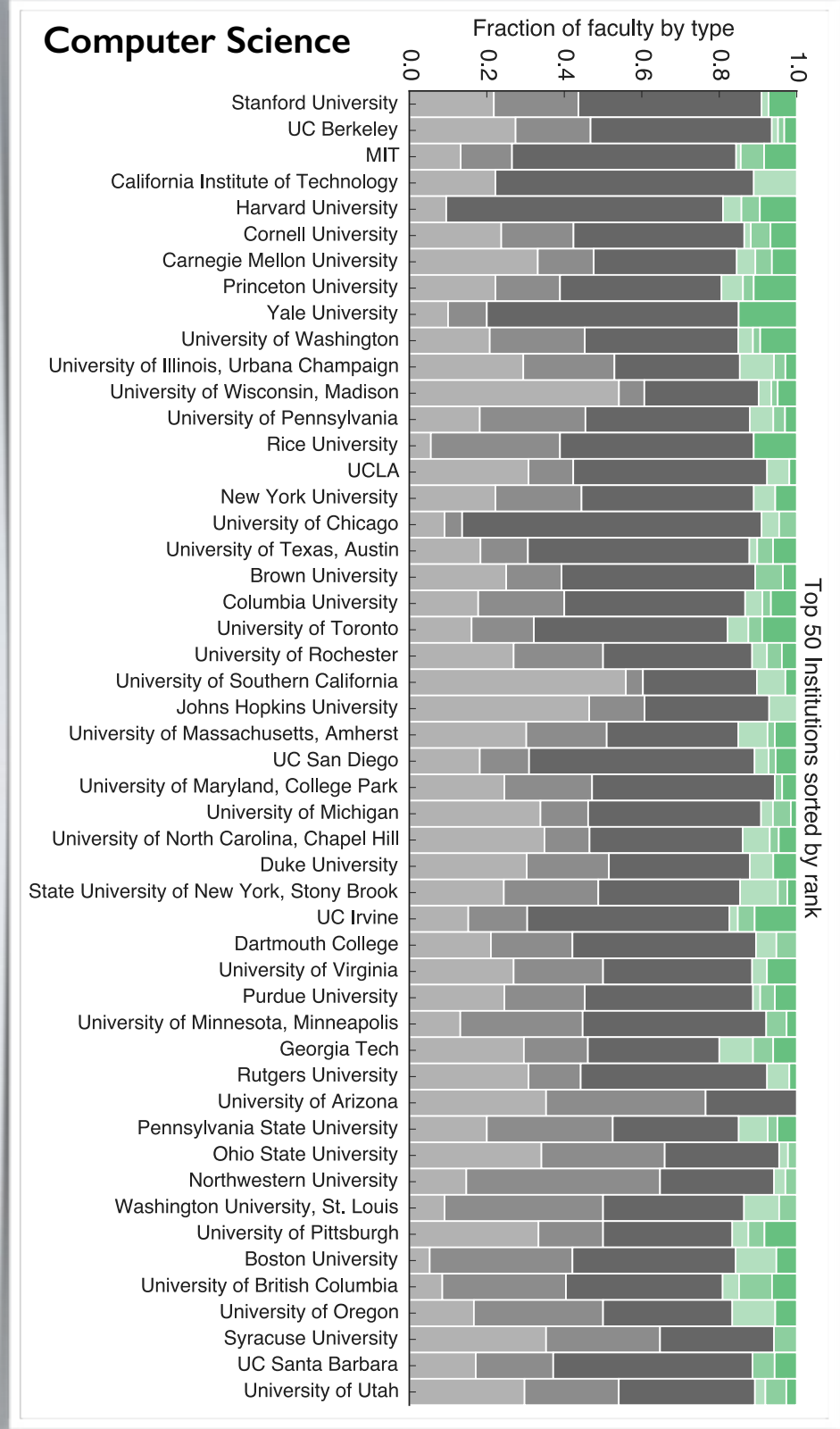
Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks

Samuel F. Way,[1,*] Daniel B. Larremore,[2,†] and Aaron Clauset[1,3,2,‡]

[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


Computer Science
Fraction of faculty by type
0.0
0.2
0.4
0.6
0.8
1.0
Stanford University
UC Berkeley
MIT
California Institute of Technology
Harvard University
Cornell University
Carnegie Mellon University
Princeton University
Yale University
University of Washington
University of Illinois, Urbana Champaign
University of Wisconsin, Madison
University of Pennsylvania
Rice University
UCLA
New York University
University of Chicago
University of Texas, Austin
Brown University
Columbia University
University of Toronto
University of Rochester
University of Southern California
Johns Hopkins University
University of Massachusetts, Amherst
UC San Diego
University of Maryland, College Park
University of Michigan
University of North Carolina, Chapel Hill
Duke University
State University of New York, Stony Brook
UC Irvine
Dartmouth College
University of Virginia
Purdue University
University of Minnesota, Minneapolis
Georgia Tech
Rutgers University
University of Arizona
Pennsylvania State University
Ohio State University
Northwestern University
Washington University, St. Louis
University of Pittsburgh
Boston University
University of British Columbia
University of Oregon
Syracuse University
UC Santa Barbara
University of Utah
Top 50 Institutions sorted by rank


Female Full Professors
Female Associate Professors
Female Assistant Professors
Male Full Professors
Male Associate Professors
Male Assistant Professors