What drives the productivity of scientific labor?

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

network data science

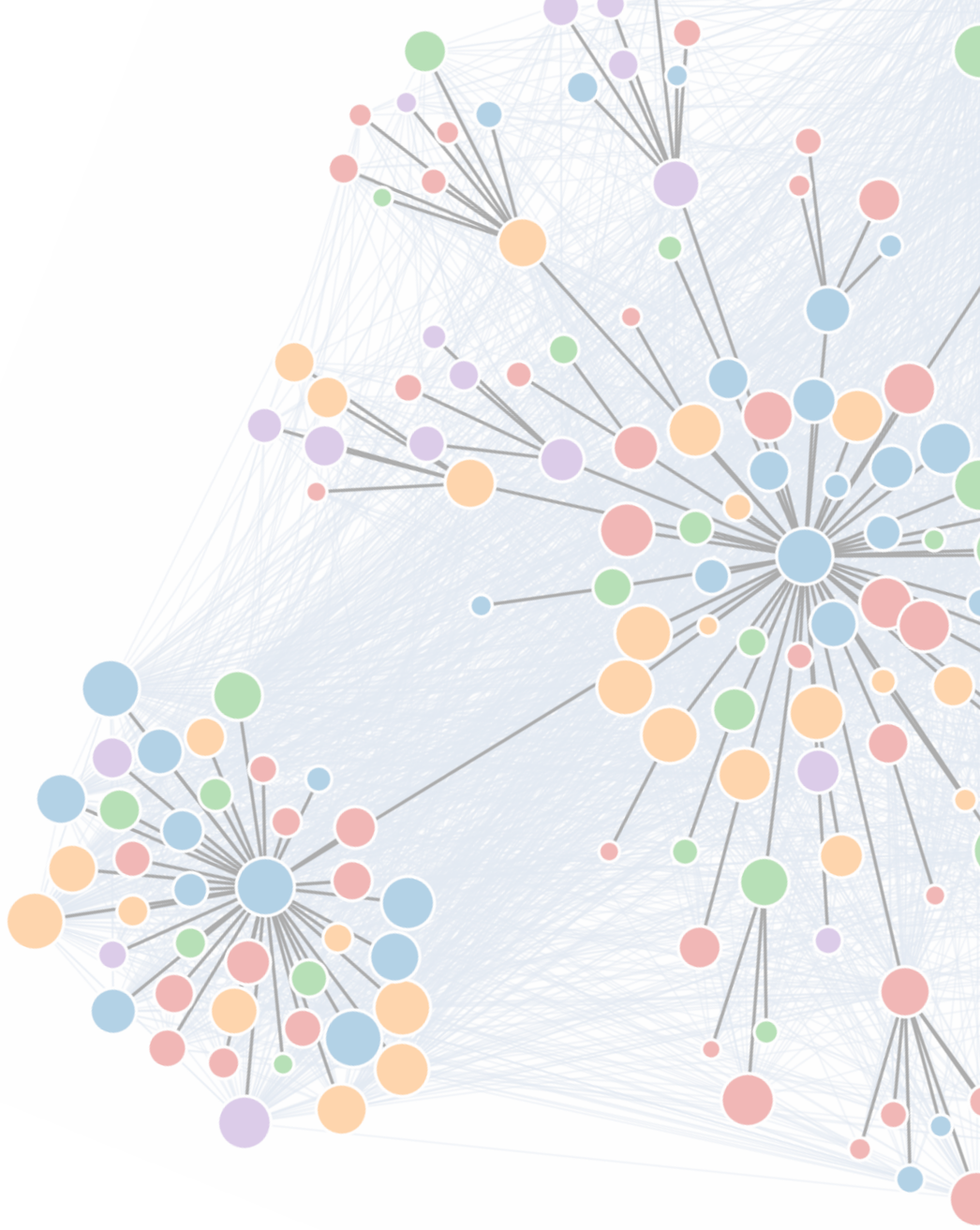
University of Colorado **Boulder**

Network Analysis and Modeling

This graduate-level course will examine modern techniques for analyzing and modeling the structure and dynamics of complex networks. The focus will be on statistical algorithms and methods, and both lectures and assignments will emphasize model interpretability and understanding the processes that generate real data. Applications will be drawn from computational biology and computational social science. No biological or social science training is required. (Note: this is not a scientific computing course, but there will be plenty of computing for science.)

Full lectures notes online (~150 pages in PDF)

https://aaronclauset.github.io/courses/5352/



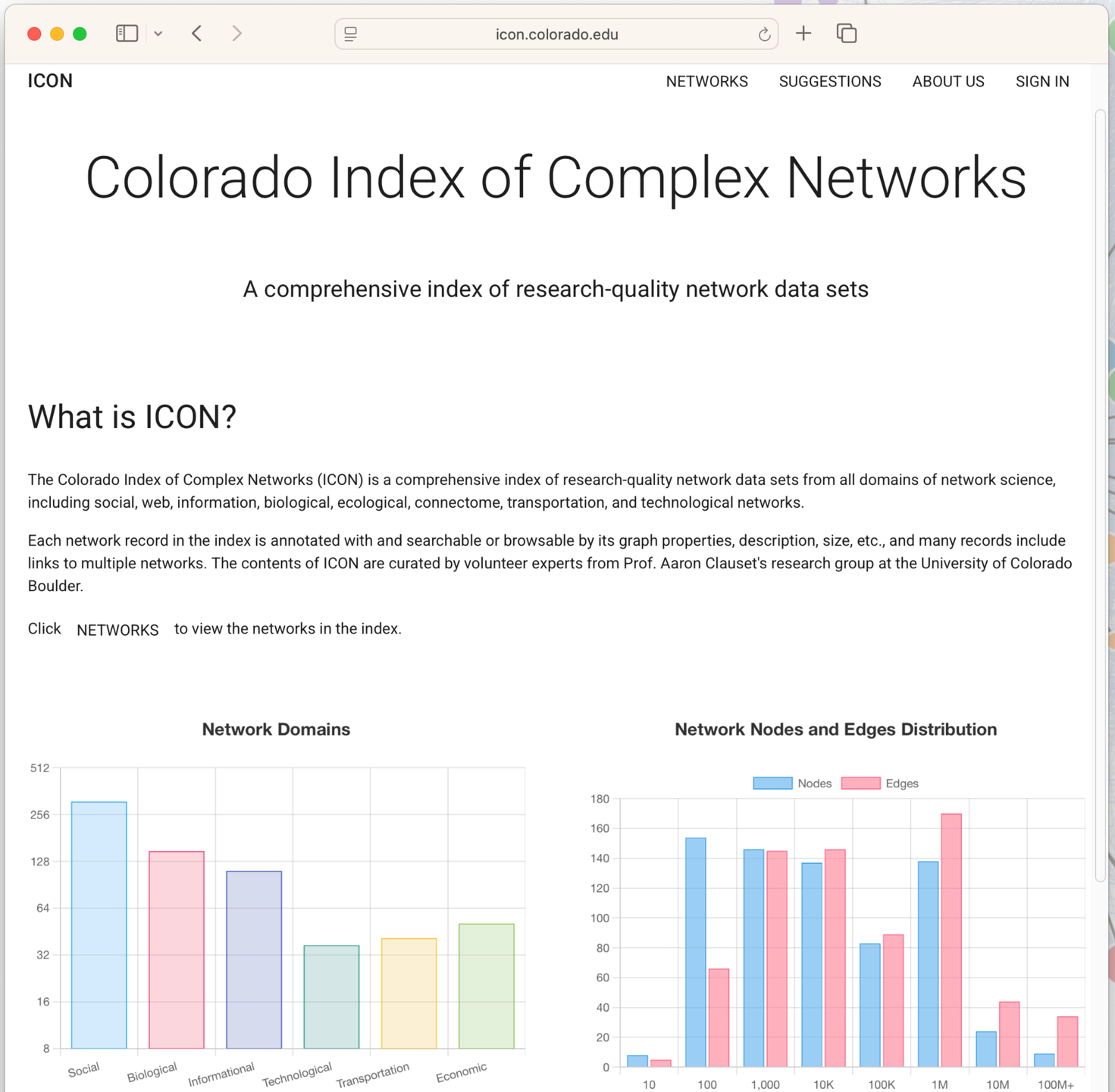
Network editors & packages

NetworkX [python]
igraph [python, R, c++]
graph-tool [python, c++]
GraphLab [python, c++]

UCI-Net
NodeXL
Gephi
Pajek
Network Workbench
Cytoscape
yEd graph editor
Graphviz

Network data sets

Colorado Index of Complex Networks
icon.colorado.edu



network effects in scientific labor

- *networks* mediate most scientific activities:

  scientific training, hiring, collaboration, teaching, attention, peer review, etc.

  networks act like a form of *unequally distributed social capital*

- a productive collaborator → increases your productivity
- a prominent collaborator → increases your prominence

PhD student
Aaron Clauset
(New Mexico)

Professor
Mark Newman
(Michigan)

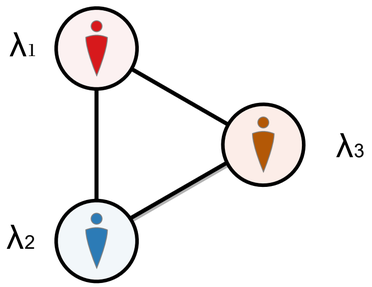
how much does *who* you work with impact *your* productivity and prominence?

this requires marginalizing out the network's effects on a node's scholarly output

partial figure of 379 network scientists in 2006 and their collaborations, from Newman "Finding community structure in networks using the eigenvectors of matrices" (2006)

untangling the network's effects

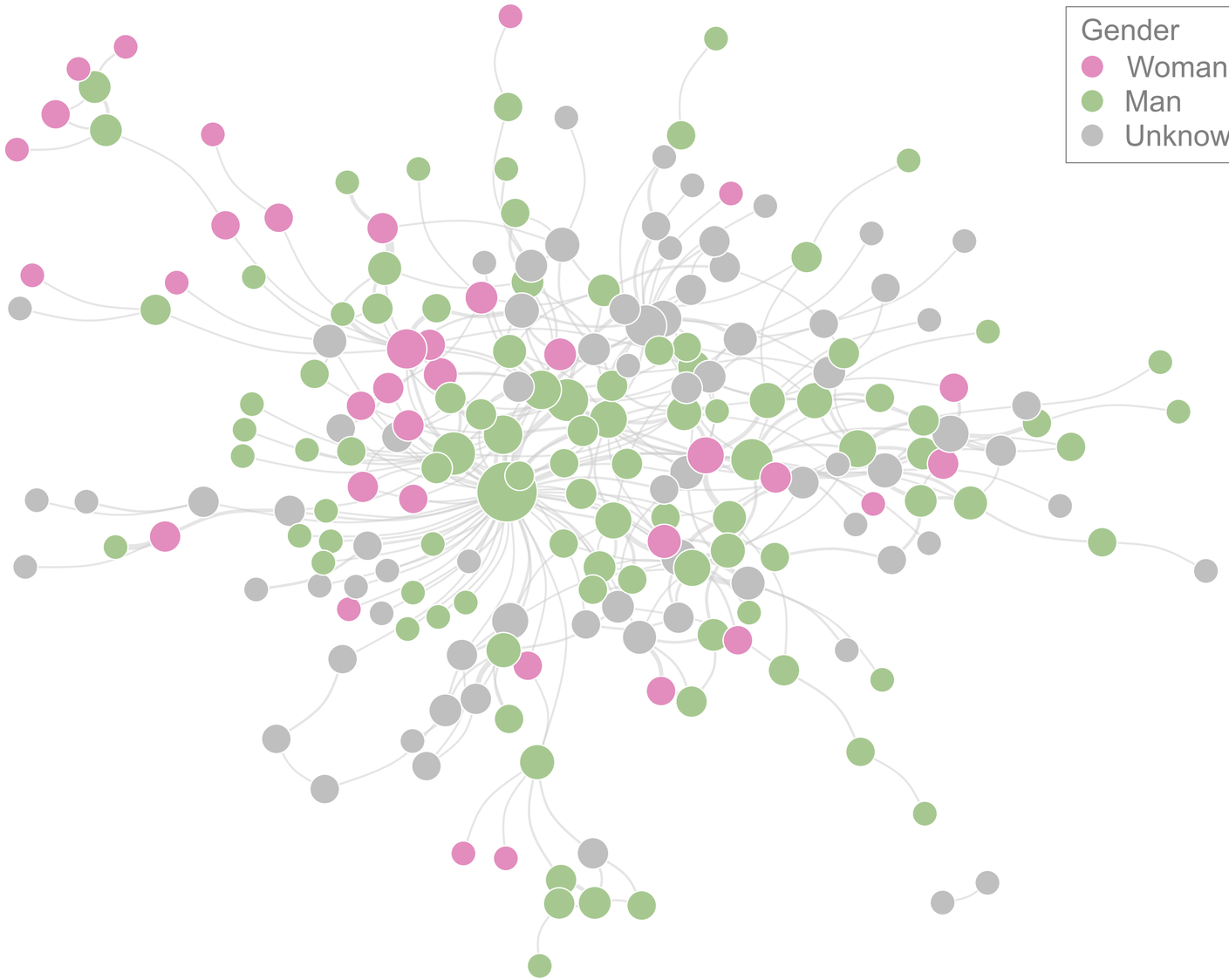
Coauthorship network



how to estimate individual productivity,

net of coauthors' own individual productivity?

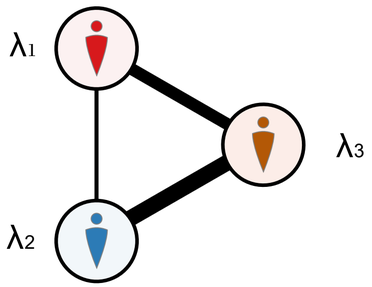
for example…



Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

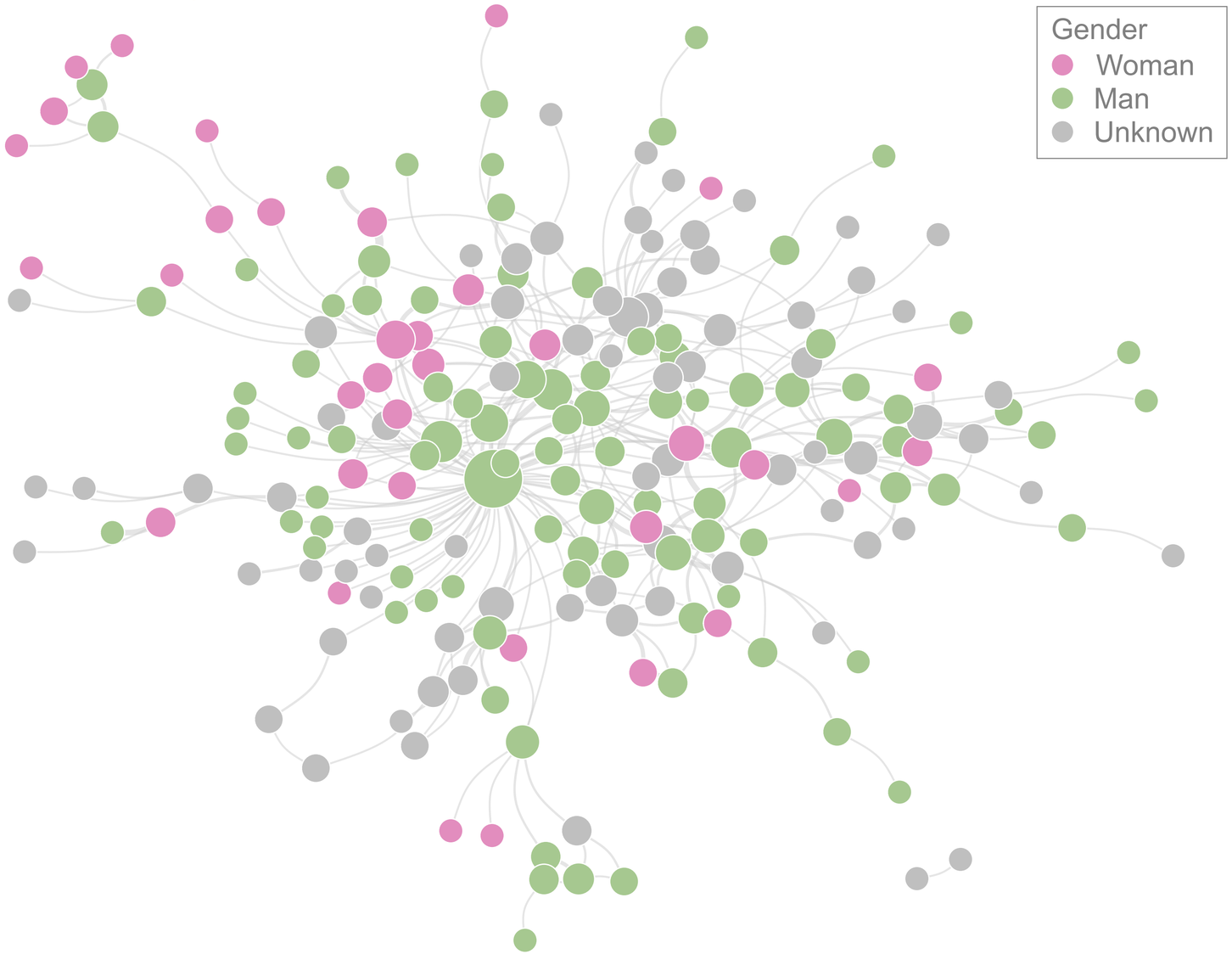
Coauthorship network



Pairwise productivity

Author pair	Papers	Time
	1	2
	3	2
	4	3
(i,j)	N_{ij}	t_{ij}

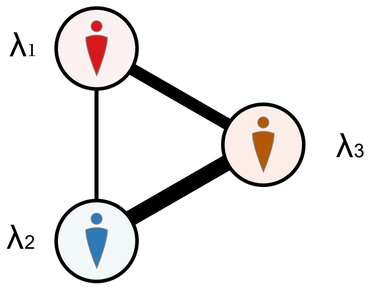
who is the most individually productive?



Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network



Pairwise productivity

Author pair	Papers	Time
	1	2
	3	2
	4	3
(i,j)	N_{ij}	t_{ij}

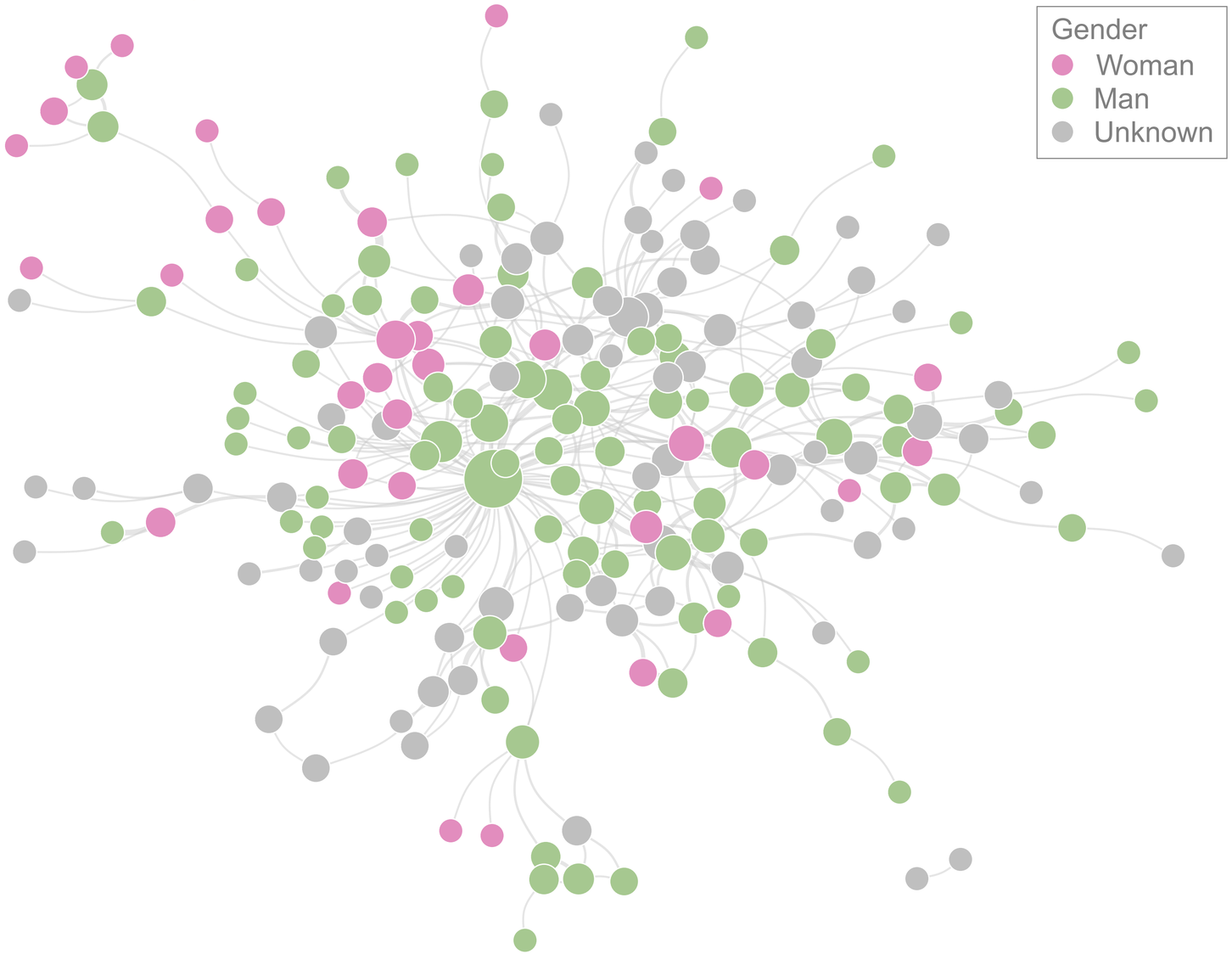
probabilistic model of *pairwise* productivity

number (i,j)-coauthored papers

$$\Pr(N_{ij}, t_{ij} \mid \lambda_i, \lambda_j)$$

given time period

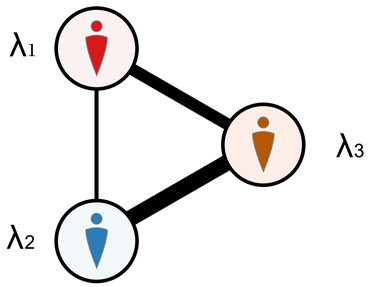
individual productivities



Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network



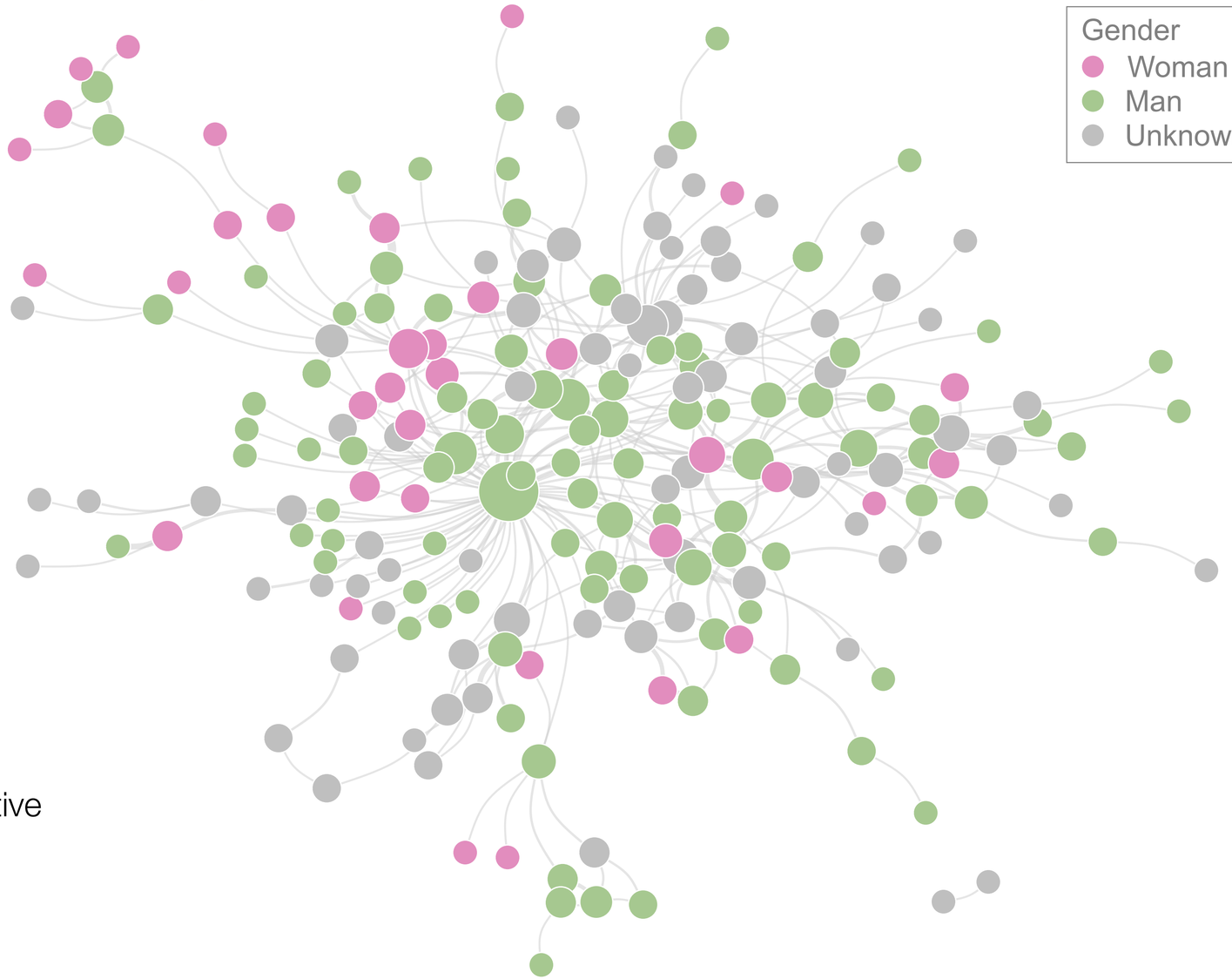
Pairwise productivity

Author pair	Papers	Time
	1	2
	3	2
	4	3
(i,j)	N_{ij}	t_{ij}

probabilistic model of *pairwise* productivity

$$\Pr(N_{ij}, t_{ij} \mid \lambda_i, \lambda_j) = \text{Poisson}([\lambda_i + \lambda_j]t_{ij})$$

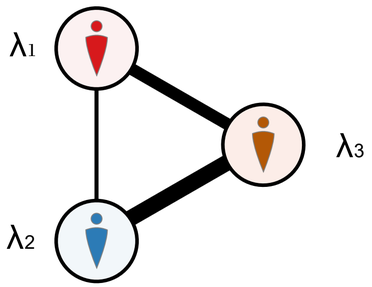
productivity is additive



Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

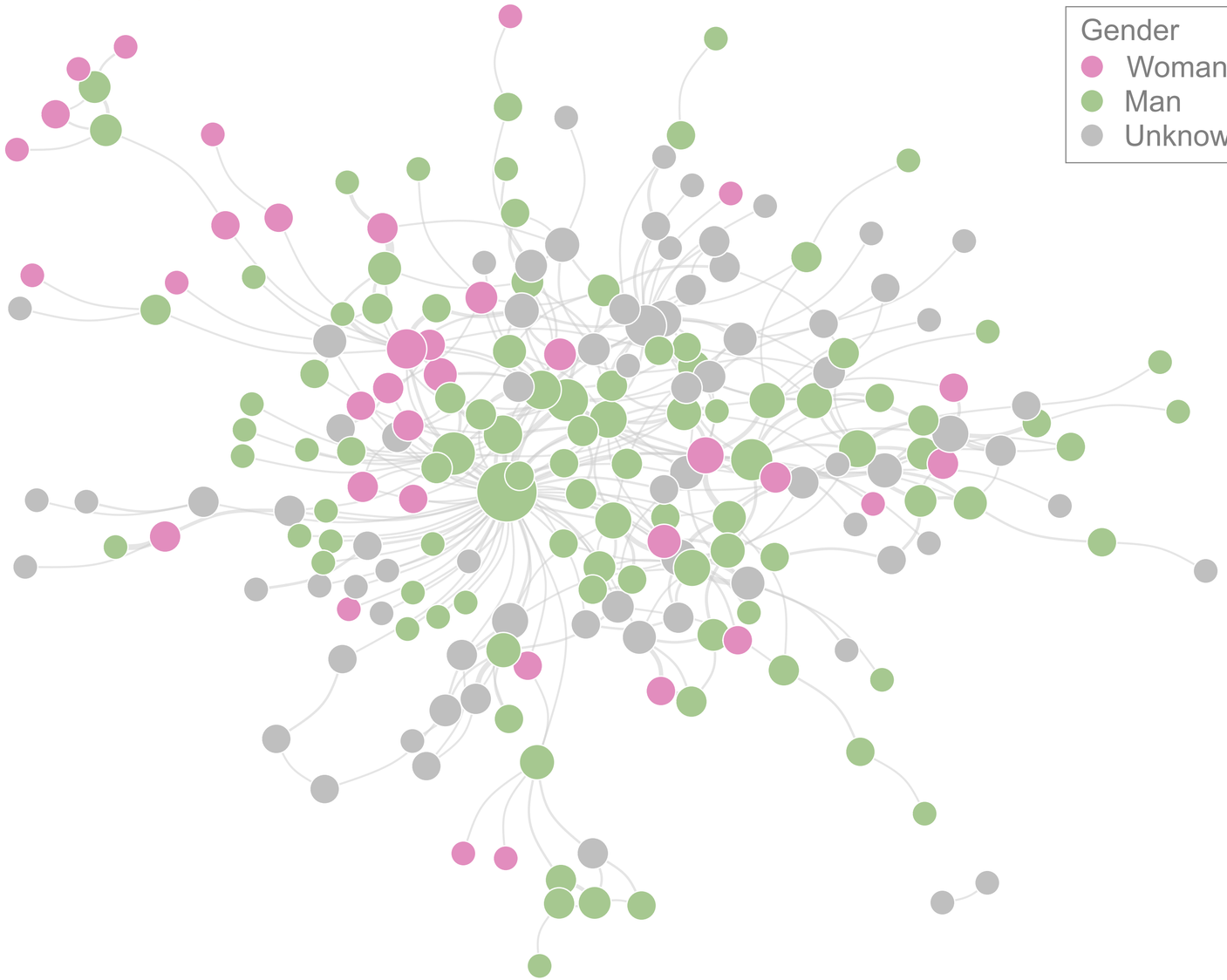
untangling the network's effects

Coauthorship network



Individual researcher metrics

Author	Papers	Estimate
	4	
	5	
	7	
i	N_i	

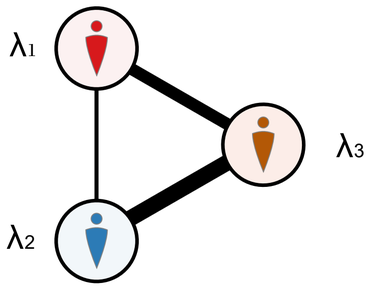


$$\Pr(N_{ij}, t_{ij} \mid \lambda_i, \lambda_j) = \text{Poisson}([\lambda_i + \lambda_j]t_{ij})$$

Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network

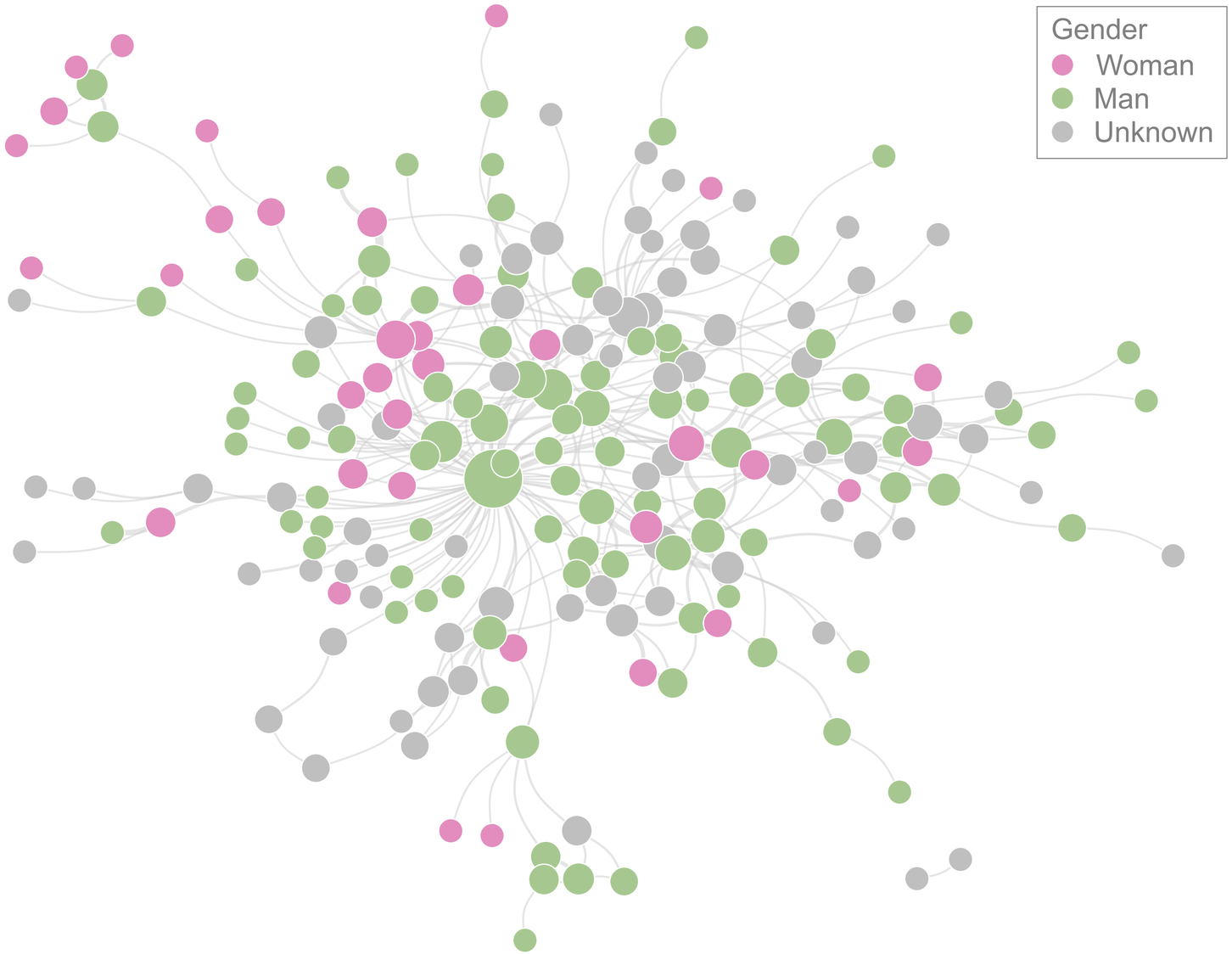


Individual researcher metrics

Author	Papers	Estimate
	4	$\lambda_1 = 0.33$
	5	$\lambda_2 = 0.17$
	7	$\lambda_3 = 1.17$
i	N_i	

who is the most individually productive?

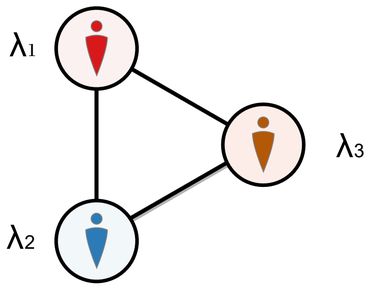
$$\Pr(N_{ij}, t_{ij} \mid \lambda_i, \lambda_j) = \text{Poisson}([\lambda_i + \lambda_j]t_{ij})$$



Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

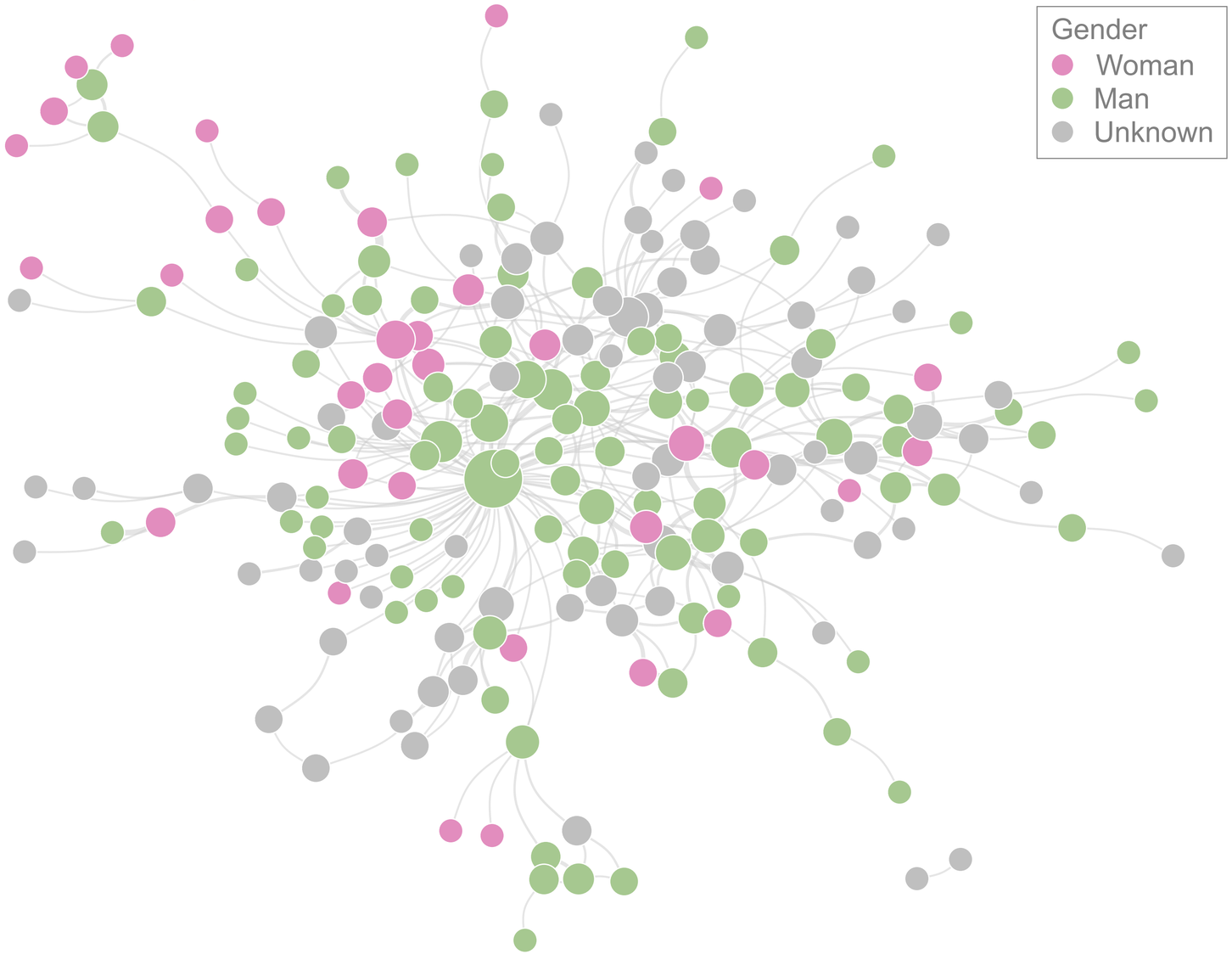
Coauthorship network



how to estimate individual prominence,

defined as number of "high impact" papers?

for example…

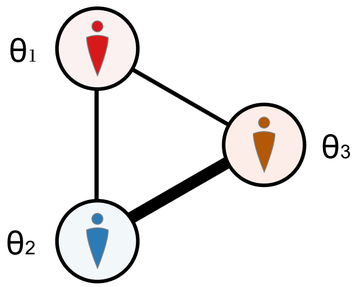


A "high impact" paper is one in the top 8% of citations received among papers published in that field in that same year

Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

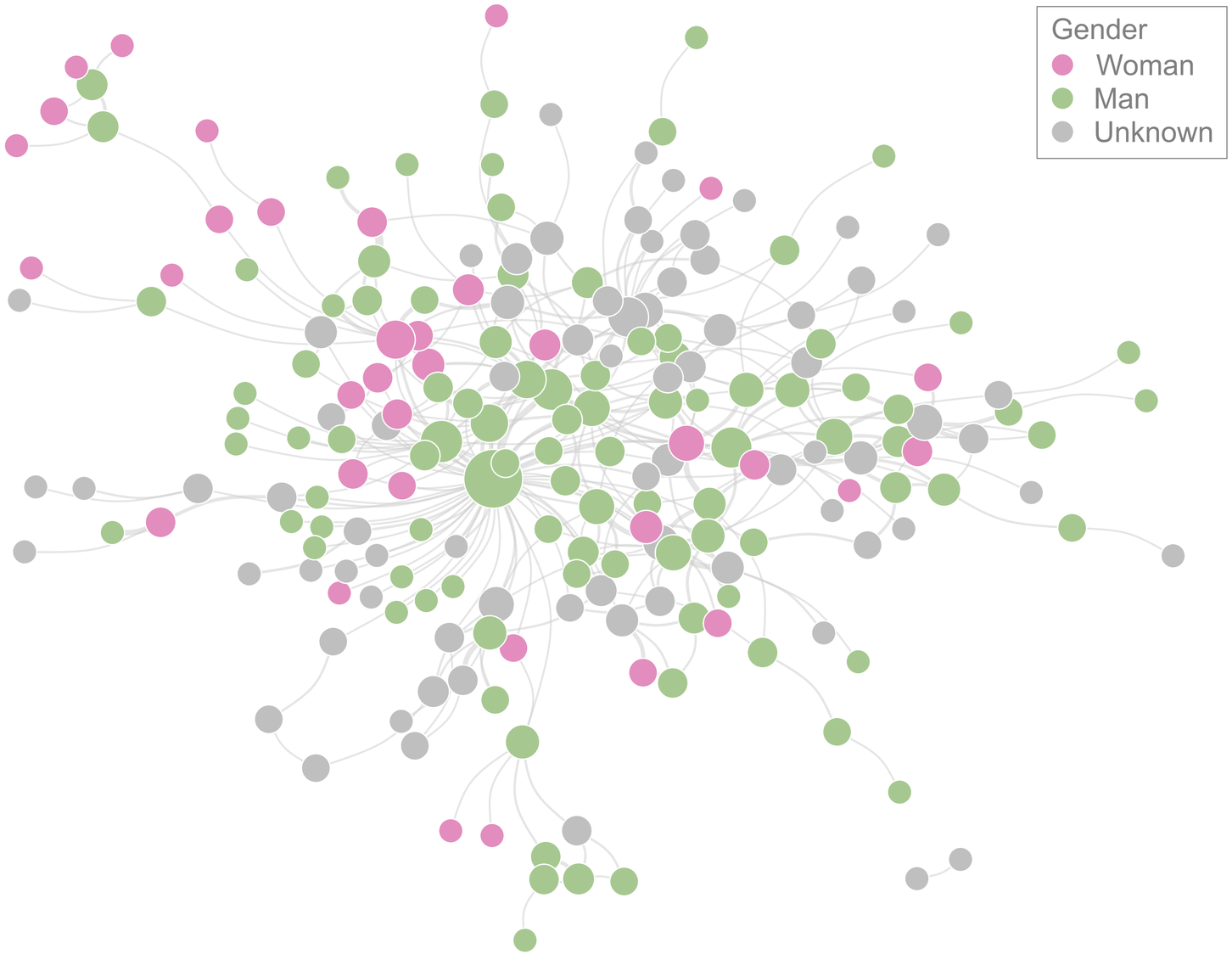
Coauthorship network



Pairwise impact

Author pair	Papers	Hit papers
	1	1
	3	1
	4	3
(i,j)	N_{ij}	m_{ij}

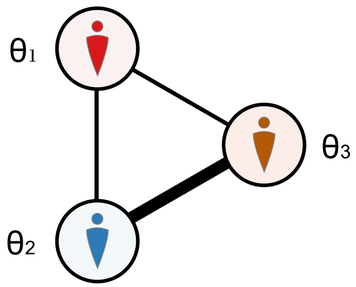
who is the most individually prominent?



A "hit" paper is one in the top 8% of citations received among papers published in that field in that same year
Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network



Pairwise impact

Author pair	Papers	Hit papers
	1	1
	3	1
	4	3
(i,j)	N_{ij}	m_{ij}

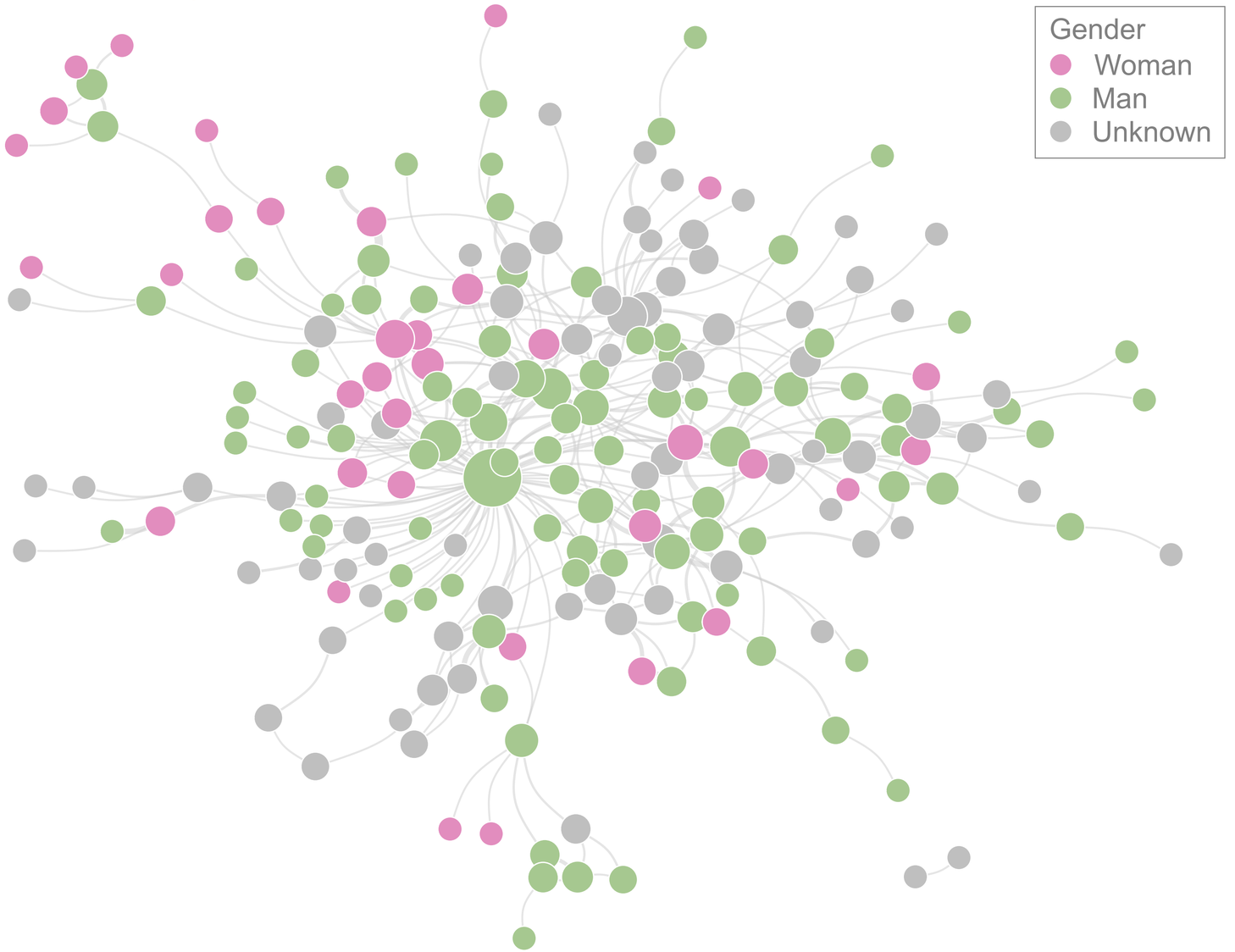
probabilistic model of *pairwise* prominence

number (i,j)-coauthored papers

$$\Pr(N_{ij}, m_{ij} \mid \theta_i, \theta_j) =$$

individual prominences

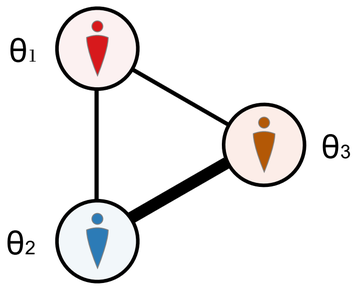
number of "high impact" papers



A "hit" paper is one in the top 8% of citations received among papers published in that field in that same year
Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network



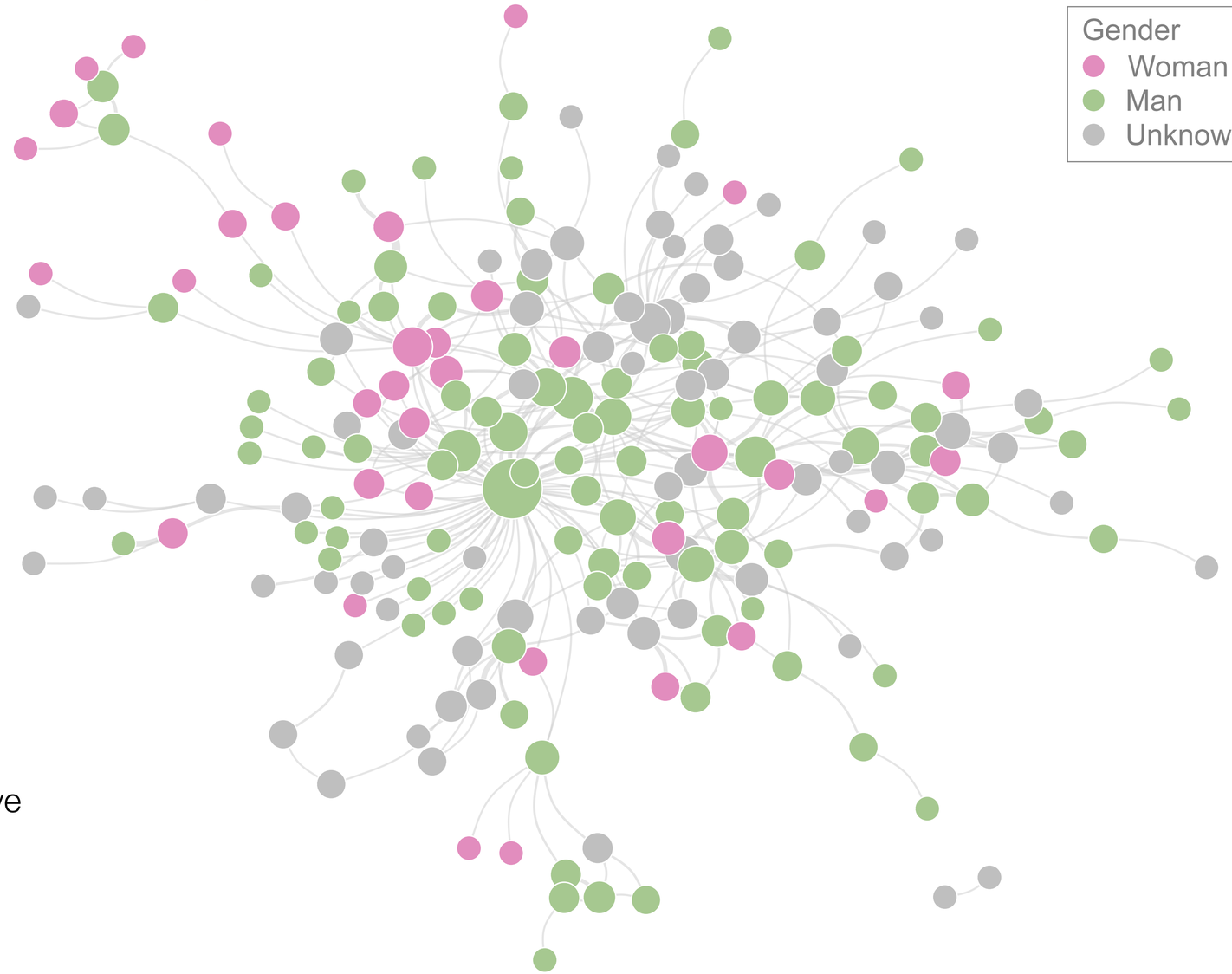
Pairwise impact

Author pair	Papers	Hit papers
	1	1
	3	1
	4	3
(i,j)	N_{ij}	m_{ij}

probabilistic model of *pairwise* prominence

$$\Pr(N_{ij}, m_{ij} \,|\, \theta_i, \theta_j) = \text{Binomial}(N_{ij}, [\theta_i + \theta_j])$$

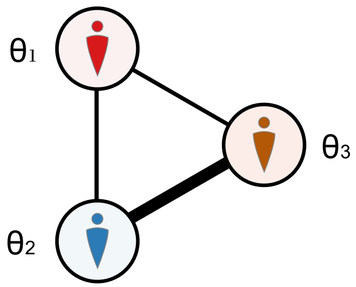
prominence is additive



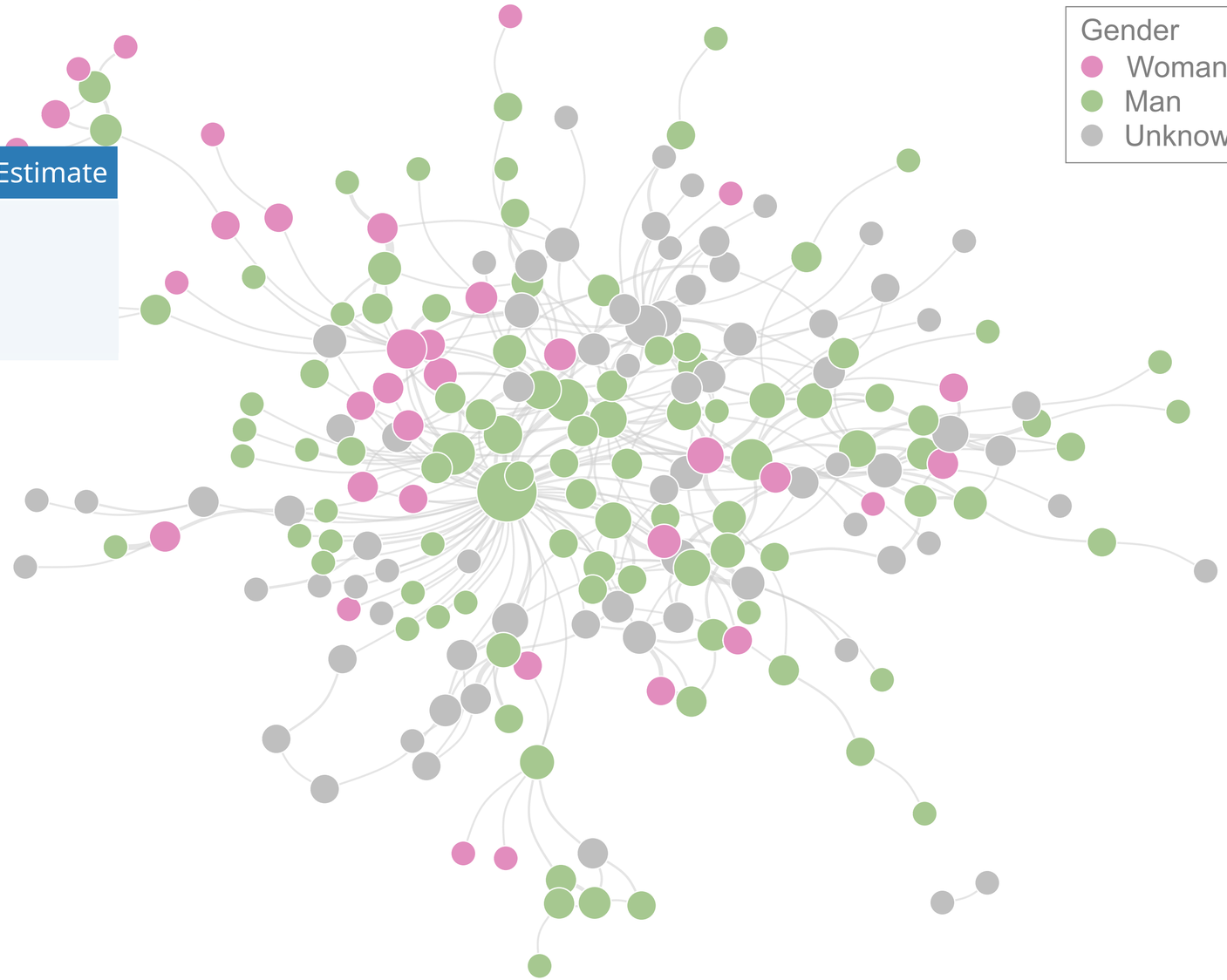
A "hit" paper is one in the top 8% of citations received among papers published in that field in that same year
Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network



Author	Papers	Hit papers	Estimate	Estimate
	4	2	$\lambda_1 = 0.33$	
	5	4	$\lambda_2 = 0.17$	
	7	4	$\lambda_3 = 1.17$	
i	N_i	m_i		

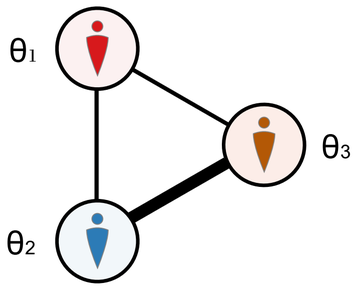


$$\Pr(N_{ij}, m_{ij} \,|\, \theta_i, \theta_j) = \text{Binomial}(N_{ij}, [\theta_i + \theta_j])$$

A "hit" paper is one in the top 8% of citations received among papers published in that field in that same year
Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Coauthorship network



Author	Papers	Hit papers	Estimate	Estimate
	4	2	λ1 = 0.33	θ1 = 0.29
	5	4	λ2 = 0.17	θ2 = 0.71
	7	4	λ3 = 1.17	θ3 = 0.04
i	N_i	m_i		

middle productivity λ, middle prominence θ

low productivity λ, high prominence θ

high productivity λ, low prominence θ

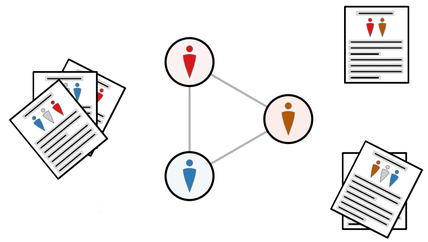
A "hit" paper is one in the top 8% of citations received among papers published in that field in that same year
Li et al., "Untangling the network effects of productivity and prominence among scientists" (2022)

untangling the network's effects

Microsoft Academic Graph (MAG) 1950—2019

Microsoft Research

- 198,202 mid-career researchers, with 10+ papers by 15th year of publishing history
- spanning 6 STEM fields (biology, chemistry, CS, math, medicine, physics)
- analyze only first-last author pairs (mitigates middle-author effects)
- 'high impact' paper = in upper 8% for given year-field



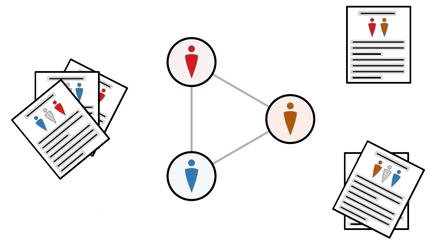
"field" is given by MAG's level-0 classification — there are 19 of these, 6 of which we select
prestige measure must be international, and production of highly cited papers generally correlates well with other national measures of prestige — define "elite" as top 10 institutions under z-score of total highly cited papers over time period
prior to estimation, prune all subtrees of collaboration network — model parameters are non-identifiable on trees; robustness analysis indicates prestige (lower) but no gender correlation with dropped authors

untangling the network's effects

Microsoft Academic Graph (MAG) 1950—2019

Microsoft Research

- 198,202 mid-career researchers, with 10+ papers by 15th year of publishing history
- spanning 6 STEM fields (biology, chemistry, CS, math, medicine, physics)
- analyze only first-last author pairs (mitigates middle-author effects)
- 'high impact' paper = in upper 8% for given year-field



Estimate individual (λ_i, θ_i) parameters for each scientist

- estimate using network 1950—*T*, for variable *T* in [1975,2017]
- bootstrapped convex optimization, then record mean λ_i and θ_i
- investigate how (λ_i, θ_i) covary with gender, prestige, etc.

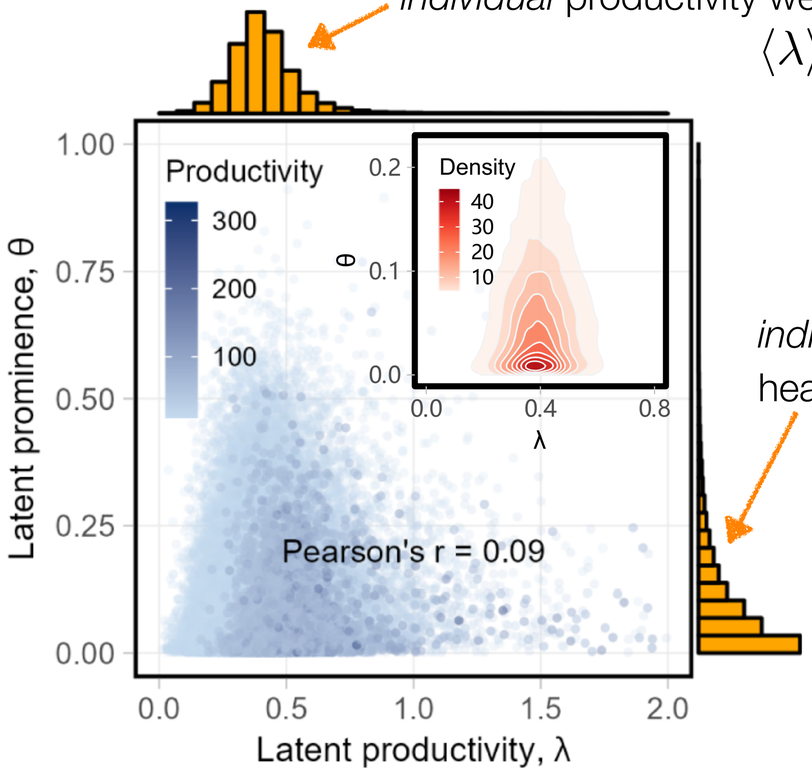
"field" is given by MAG's level-0 classification — there are 19 of these, 6 of which we select
prestige measure must be international, and production of highly cited papers generally correlates well with other national measures of prestige — define "elite" as top 10 institutions under z-score of total highly cited papers over time period
prior to estimation, prune all subtrees of collaboration network — model parameters are non-identifiable on trees; robustness analysis indicates prestige (lower) but no gender correlation with dropped authors

model checking

- applied to 198,202 mid-career STEM researchers 1975-2017

individual productivity well-concentrated ✅

$\langle \lambda \rangle = 0.42$ papers / year



individual prominence highly skewed ✅

heavy-tailed $\Pr(\theta)$

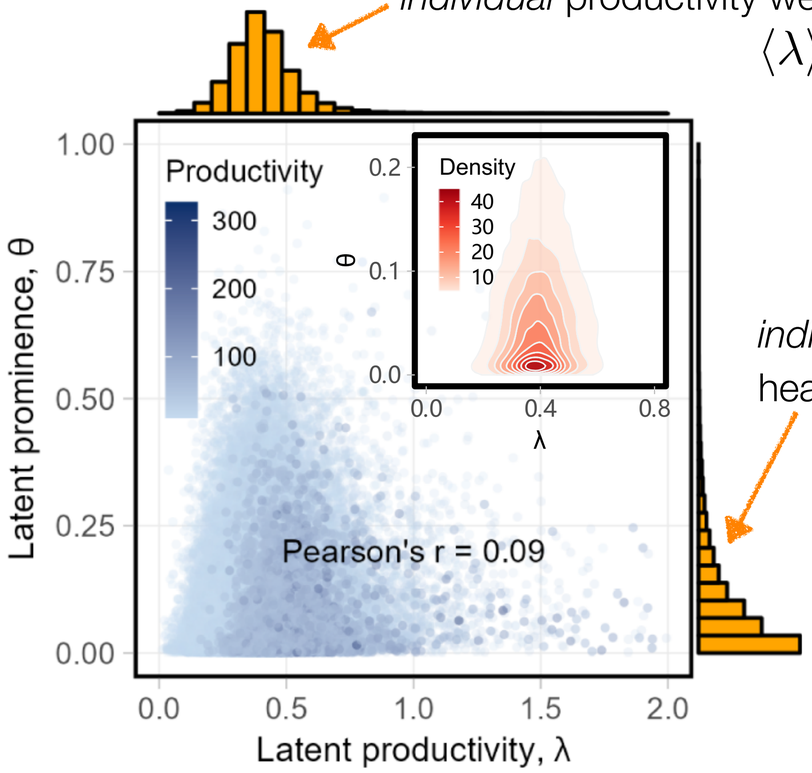
dots omitted if $\hat{\theta} < 10^{-3}$ (researchers with no highly cited papers)

model checking

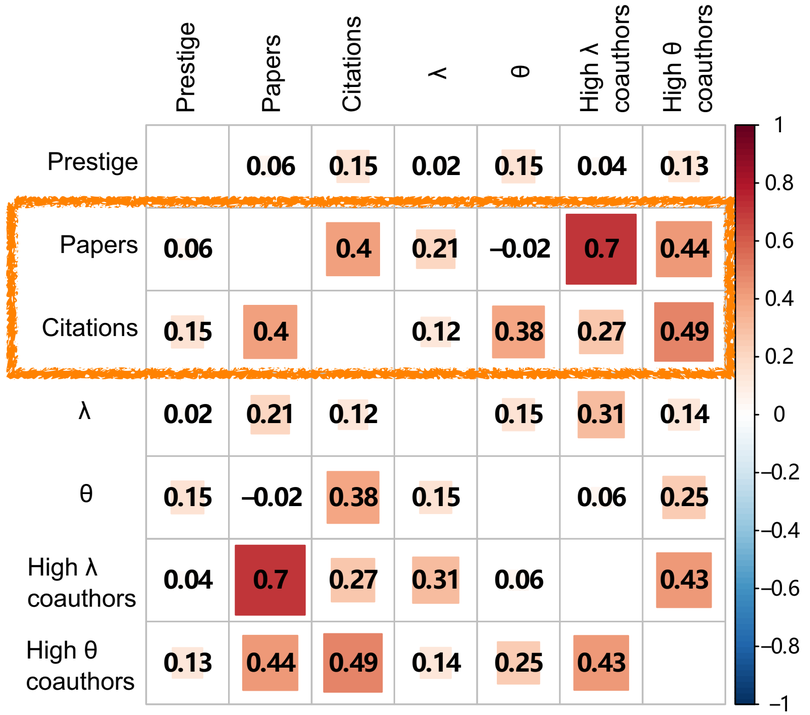
- applied to 198,202 mid-career STEM researchers 1975-2017

individual productivity well-concentrated

$\langle \lambda \rangle = 0.42$ papers / year



individual prominence highly skewed heavy-tailed $\Pr(\theta)$



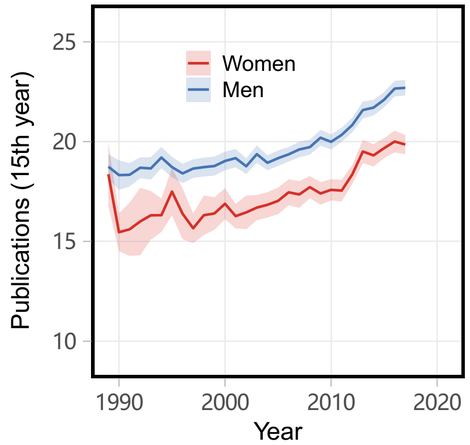
✅ my paper count = highly correlated with high λ coauthors

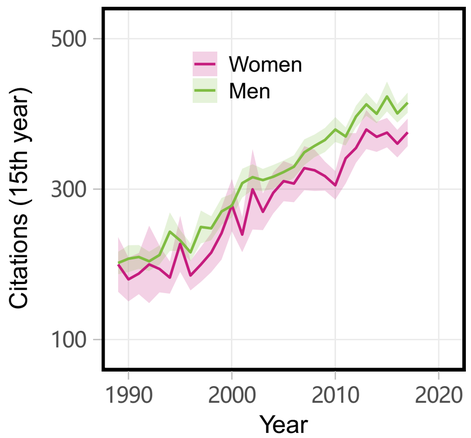
✅ my citations = well correlated with high λ & θ coauthors

dots omitted if $\hat{\theta} < 10^{-3}$ (researchers with no highly cited papers)

gender vs. productivity & prominence

- past work : men publish more papers than women & receive more citations
  - compare (λ_i, θ_i) over time, for men and women





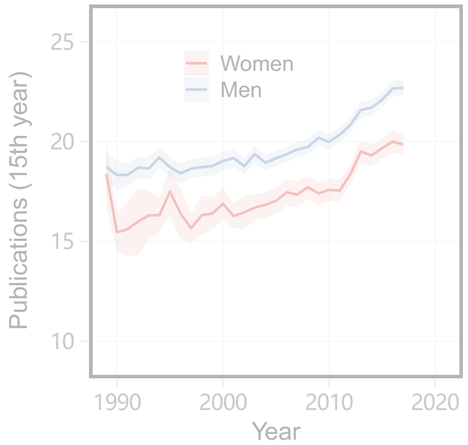
substantial gendered differences in "raw" productivity and prominence

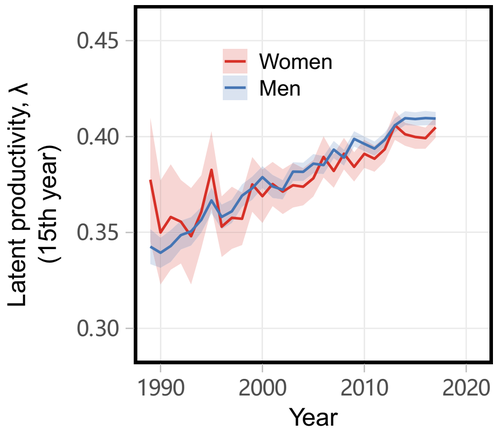
- men much more productive
- men slightly more prominent

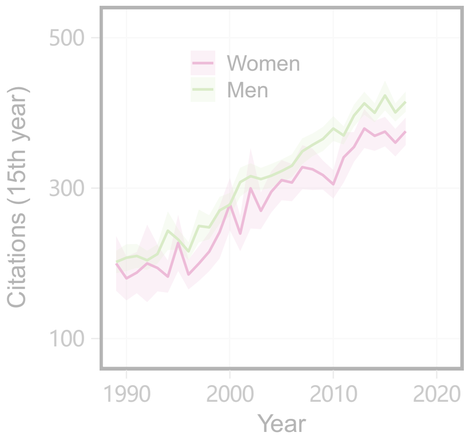
shaded areas are 95% confidence intervals

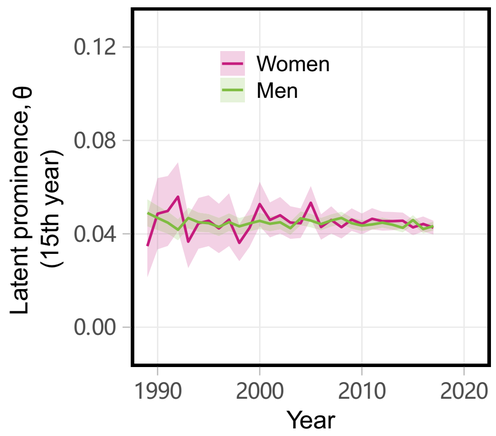
gender vs. productivity & prominence

- past work : men publish more papers than women & receive more citations
  - compare (λ_i, θ_i) over time, for men and women → their networks are different









but: latent productivity and prominence is *not gendered* ✅

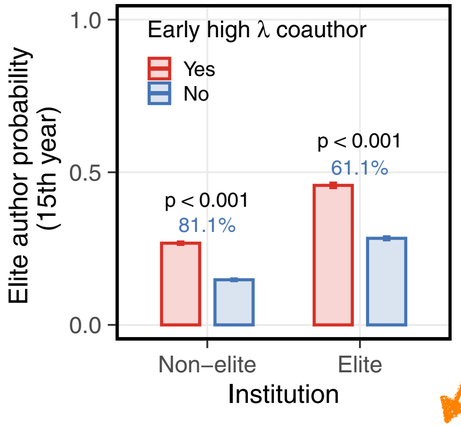
- size and composition of collaboration networks is gendered
- latent productivities increase steadily
- latent prominence stable over time

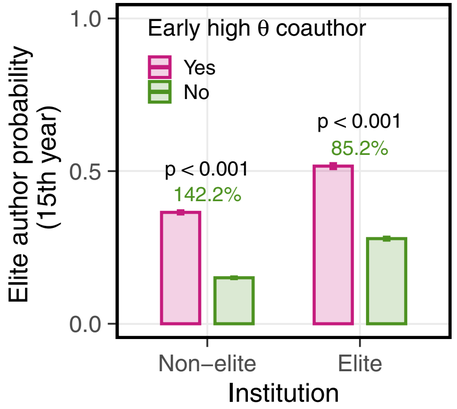
*not causal, but implies effects of known gendered causal factors on productivity (eg, parenthood) may operate by reshaping collaborating networks

see Morgan et al. "The unequal impact of parenthood in academia." (2021)

effects of elite collaborators

- how much does an early-career collaboration with an elite senior researcher influence you?
  - elite senior researches with high λ or high θ





early high-λ or early high-θ collaborator substantially increases likelihood of high prominence in mid-career

- much more common at elite institutions
- *effect appears at non-elite institutions too* ✅

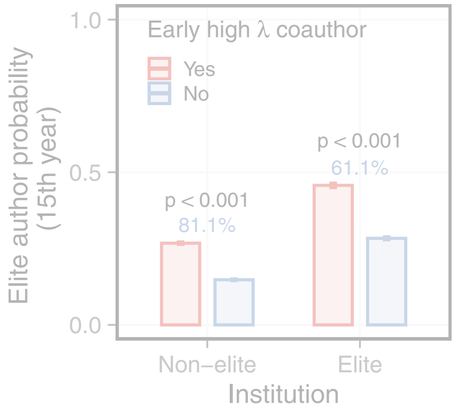
Pr. of high-λ early collab = 0.18 (elite) & 0.15 (non-elite)

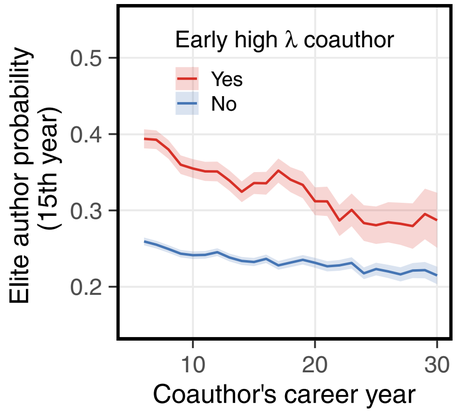
Pr. of high-θ early collab = 0.14 (elite) & 0.07 (non-elite)

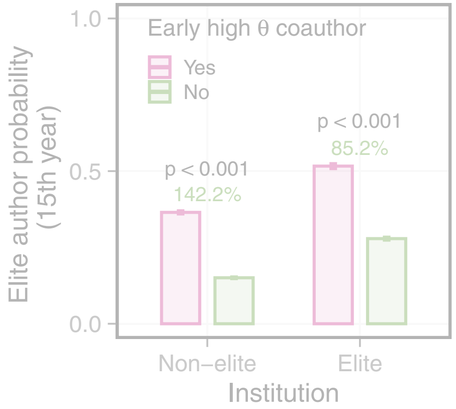
elite institutions = top 10 by *z*-score of high impact papers
early-career collaboration = within first 5 years of publishing history
"elite author" = upper 5% of citations among authors in a given field-year
shaded areas are 95% confidence intervals

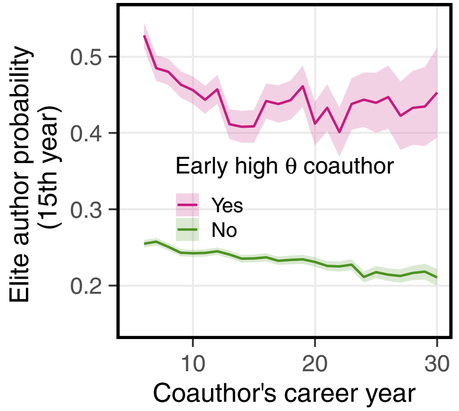
effects of elite collaborators

- how much does an early-career collaboration with an elite senior researcher influence you?
  - elite senior researches with high λ or high θ









the 'benefits' are substantial regardless of coauthors career age

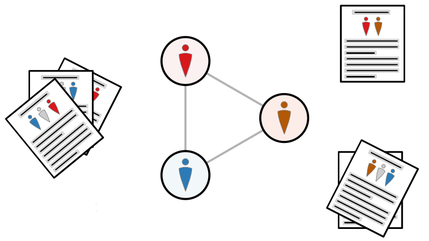
- slight decrease for most senior coauthors
- collaboration networks act like a partially transferrable form of *social capital* in science ✅

elite institutions = top 10 by *z*-score of high impact papers
early-career collaboration = within first 5 years of publishing history
"elite author" = upper 5% of citations among authors in a given field-year
shaded areas are 95% confidence intervals

how important is who you work with?

networks act like unequally distributed social capital in science

- *they mediate our scientific attention, evaluation, and collaboration*



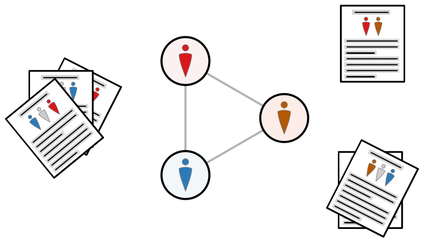
differences in collaboration networks can explain

- gendered differences in productivity & prominence →
  - no difference in *individual* productivity & prominence (λ_i, θ_i)
  - but men do publish more and are more prominent, implying the difference is due to their networks
  - hence, should not compare *unadjusted* measures of productivity & prominence (the network confounds)

how important is who you work with?

networks act like unequally distributed social capital in science

- *they mediate our scientific attention, evaluation, and collaboration*



differences in collaboration networks can explain

- gendered differences in productivity & prominence
- early-career productivity & prominence

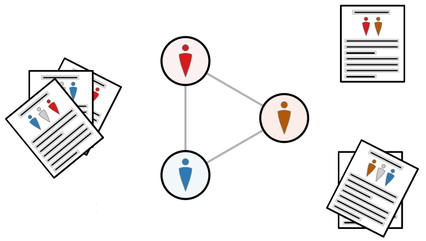
→
- elite senior coauthors bequeath some of their networks to junior coauthors — inter-generational transfers
- this effect is independent of institutional prestige
- but junior scholars have greater access to elite coauthors at elite institutions

study limitations:
none of these analyses are causal, although they do suggest specific mechanisms that can be tested
no data on race/ethnicity in our bibliographic analyses, although literature suggests under-represented minorities may have similar or larger network differences as women
we focus only on STEM fields, which tend to have strong collaboration norms; unclear what results might be for fields with different collaboration norms

how important is who you work with?

networks act like unequally distributed social capital in science

- *they mediate our scientific attention, evaluation, and collaboration*



differences in collaboration networks can explain

- gendered differences in productivity & prominence
- early-career productivity & prominence
- what else?

can we intervene in these networks to mitigate inequalities? 🤔

- funds for new collaborations, eg, after parenthood?
- early-career fellowships to work with elite senior coauthors? →

Edge interventions can mitigate demographic and prestige disparities in the Computer Science coauthorship network

Kate Barnes
kathryn.barnes@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Mia Ellis-Einhorn
melliseinh@gmail.com
Haverford College
Haverford, Pennsylvania, USA

Carolina Chávez-Ruelas
carolina.chavezruelas@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Nayera Hasan
nhasan1@haverford.edu
Haverford College
Haverford, Pennsylvania, USA

Mohammad Fanous
mfanous@haverford.edu
Haverford College
Haverford, Pennsylvania, USA

Blair D. Sullivan
sullivan@cs.utah.edu
The University of Utah
Salt Lake City, Utah, USA

Sorelle Friedler
sorelle@cs.haverford.edu
Haverford College
Haverford, Pennsylvania, USA

Aaron Clauset
aaron.clauset@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Preprint, arxiv:2506.04435 (2025)

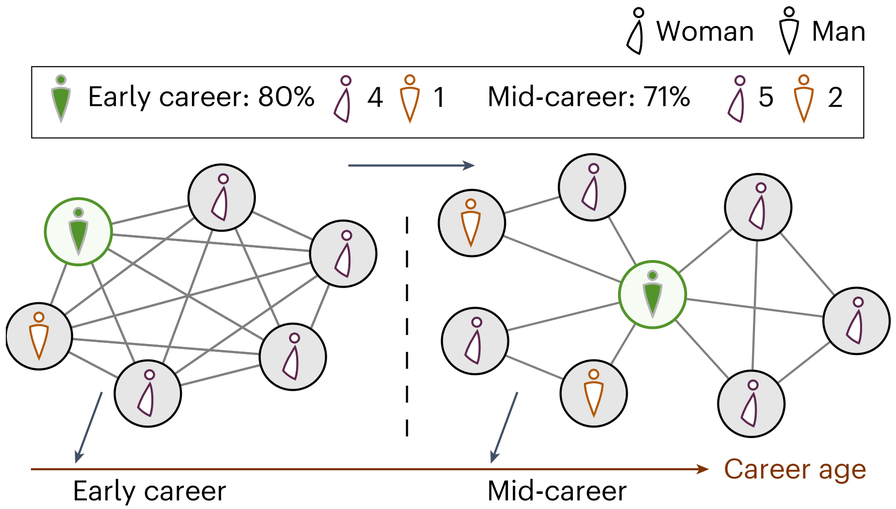
study limitations:
- none of these analyses are causal, although they do suggest specific mechanisms that can be tested
- no data on race/ethnicity in our bibliographic analyses, although literature suggests under-represented minorities may have similar or larger network differences as women
- we focus only on STEM fields, which tend to have strong collaboration norms; unclear what results might be for fields with different collaboration norms

gender and racial diversity socialization

- how does the *diversity* of early-career teams influence later-career collaboration networks?

  homophily : women prefer women, men prefer men

  what if preferences can be socialized?



man who trains with high density of women → man who advises high density of early-career women

Article
https://doi.org/10.1038/s43588-
Gender and racial diversity socialization in science
Received: 4 December 2024
Accepted: 20 March 2025
Weihua Li[1,2,3,4,5,6], Hongwei Zheng[7] ✉, Jennie E. Brand[8] & Aaron Clauset[9,10,11] ✉
Nature Computational Science (2025)

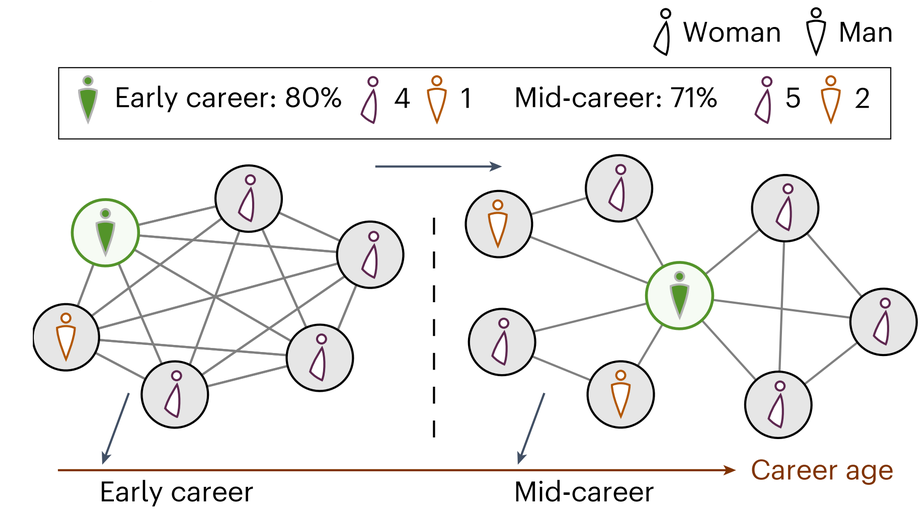
results here for 17,917 advisors and all their advisees, spanning STEM fields, social sciences, and humanities; similar results for coauthorships via Microsoft Academic Graph 1950-2021
diversity socialization measured relative to a subfield- and trend-adjusted null model, ie, the effects we measure are beyond what we'd expect for field-level differences in representation and demographic trends of increasing diversity

gender and racial diversity socialization

- how does the *diversity* of early-career teams influence later-career collaboration networks?

homophily : women prefer women, men prefer men

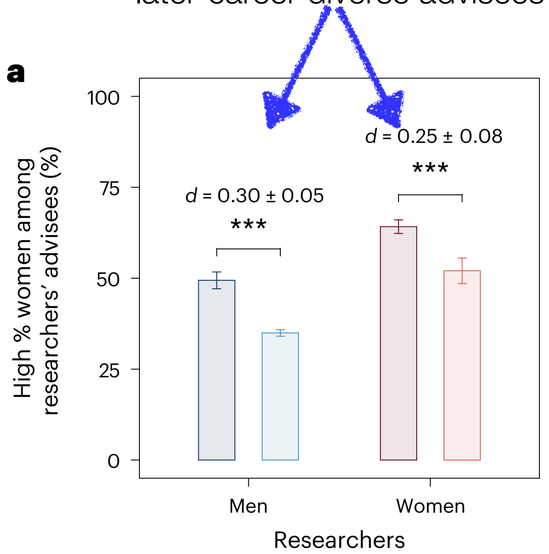
what if preferences can be socialized?



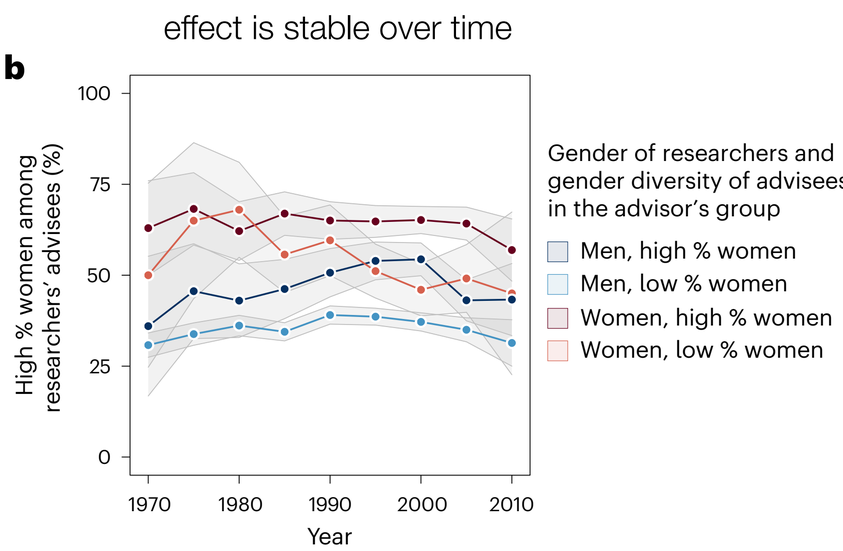
man who trains with high density of women → man who advises high density of early-career women

Article
https://doi.org/10.1038/s43588-
Gender and racial diversity socialization in science
Received: 4 December 2024
Accepted: 20 March 2025
Weihua Li[1,2,3,4,5,6], Hongwei Zheng[7] ✉, Jennie E. Brand[8] & Aaron Clauset[9,10,11] ✉
Nature Computational Science (2025)

diversity socialization leads to later-career diverse advisees



effect is stable over time



homophily is not destiny — diversity is a learnable preference

results here for 17,917 advisors and all their advisees, spanning STEM fields, social sciences, and humanities; similar results for coauthorships via Microsoft Academic Graph 1950-2021
diversity socialization measured relative to a subfield- and trend-adjusted null model, ie, the effects we measure are beyond what we'd expect for field-level differences in representation and demographic trends of increasing diversity

the scientific ecosystem

no scientist is an island → productivity is driven by environmental & network effects

- where you train *doesn't* seem to matter (conditioned on getting a faculty job)
- where you work *does* matter : prestige → available labor → more scientific output

- prestige shapes your working environment — like a slope underneath the whole ecosystem
- no results on the *value* of different scientific outputs, only their total volume
- many unobserved confounds: quality of job talk, fundability of ideas, quality of writing, etc.

the scientific ecosystem

no scientist is an island → productivity is driven by environmental & network effects

- where you train *doesn't* seem to matter (conditioned on getting a faculty job)
- where you work *does* matter : prestige → available labor → more scientific output

productivity & prominence are network effects → who you work with shapes your output

- net of their coauthors, men & women *equal* in their *individual* productivity & prominence
- early-career researchers can *inherit* collaboration networks → like generational wealth transfer

- prestige shapes who you *can* work with (environment)
- no results about scientific *value,* only the volume and the attention outputs receive
- comparing individual researchers is unfair without accounting for different coauthor network effects

the scientific ecosystem

no scientist is an island → productivity is driven by environmental & network effects

- where you train *doesn't* seem to matter (conditioned on getting a faculty job)
- where you work *does* matter : prestige → available labor → more scientific output

productivity & prominence are network effects → who you work with shapes your output

- net of their coauthors, men & women *equal* in their *individual* productivity & prominence
- early-career researchers can *inherit* collaboration networks → like generational wealth transfer

ecosystem metaphor is rich → what other environmental or network effects?

- how should we intervene in an *ecosystem*?
- what might accelerate scientific discovery?

- The undervaluing of elite women in physics
- Gender and racial diversity socialization in science
- Epistemic inequality in the diffusion of scientific ideas

references & collaborators

Productivity, prominence, and the effects of academic environment

Samuel F. Way[a,1], Allison C. Morgan[a], Daniel B. Larremore[a,b,2], and Aaron Clauset[a,b,c,1,2]

[a]Department of Computer Science, University of Colorado, Boulder, CO, USA; [b]BioFrontiers Institute, University of Colorado, Boulder, CO, USA; [c]Santa Fe Institute, Santa Fe, NM, USA

PNAS 116(22), 10729-10733 (2019)

Quantifying hierarchy and dynamics in US faculty hiring and retention

https://doi.org/10.1038/s41586-022-05222-x K. Hunter Wapman[1]✉, Sam Zhang[2], Aaron Clauset[1,3,4] & Daniel B. Larremore[1,3]✉

Nature 610, 120–127 (2022)

Labor advantages drive the greater productivity of faculty at elite universities

Sam Zhang[1]*, K. Hunter Wapman[2], Daniel B. Larremore[2,3], Aaron Clauset[2,3,4]*

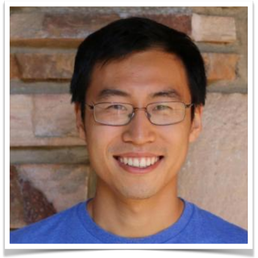
Science Advances 8, eabq7056 (2022)

Untangling the network effects of productivity and prominence among scientists

Weihua Li[1,2,3,4]✉, Sam Zhang[5], Zhiming Zheng[1,2,3,4], Skyler J. Cranmer[6] & Aaron Clauset[7,8,9]

Nature Communications 13, 4907 (2022)

Dr. Samuel F Way
(now: Spotify)



Sam Zhang
(Colorado)

Dr. K. Hunter Wapman
(Colorado)

Prof. Weihua Li
(Beihang)

Prof. Zhiming Zhang
(Beihang)

Dr. Allison Morgan
(now: Code for America)

Prof. Skyler Cranmer
(Ohio State)

Prof. Daniel Larremore
(Colorado)

Funding: