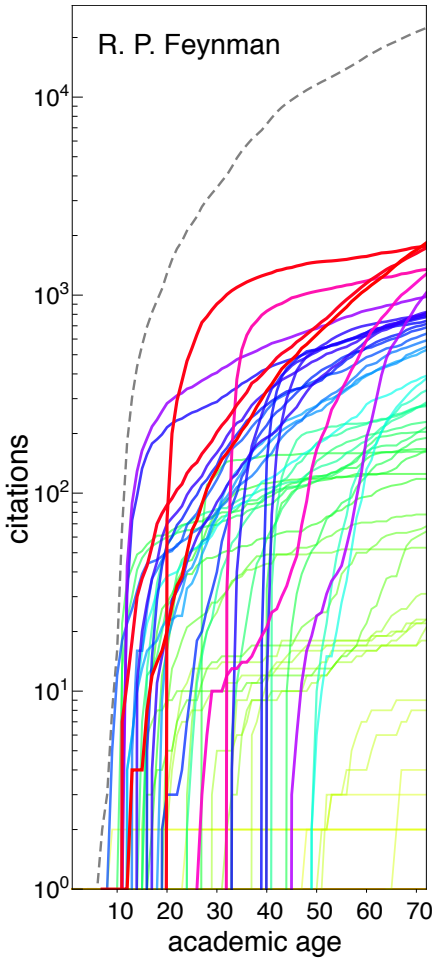
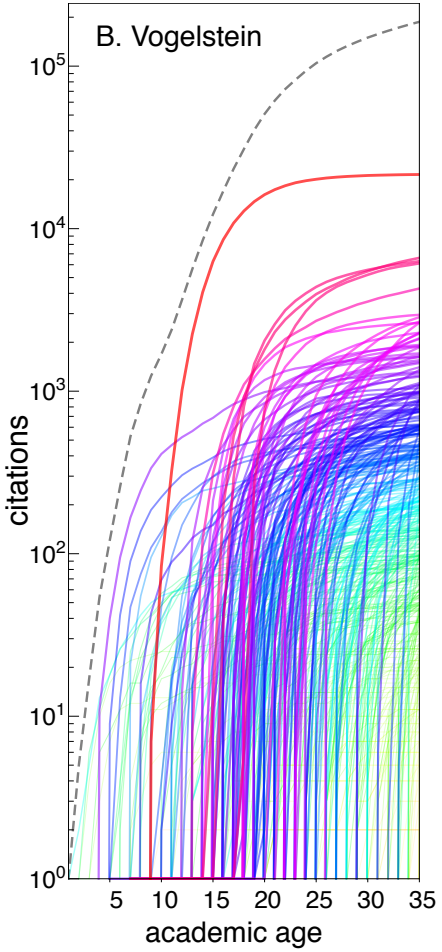
The computational social science of academic career growth



IMT INSTITUTE FOR ADVANCED STUDIES LUCCA

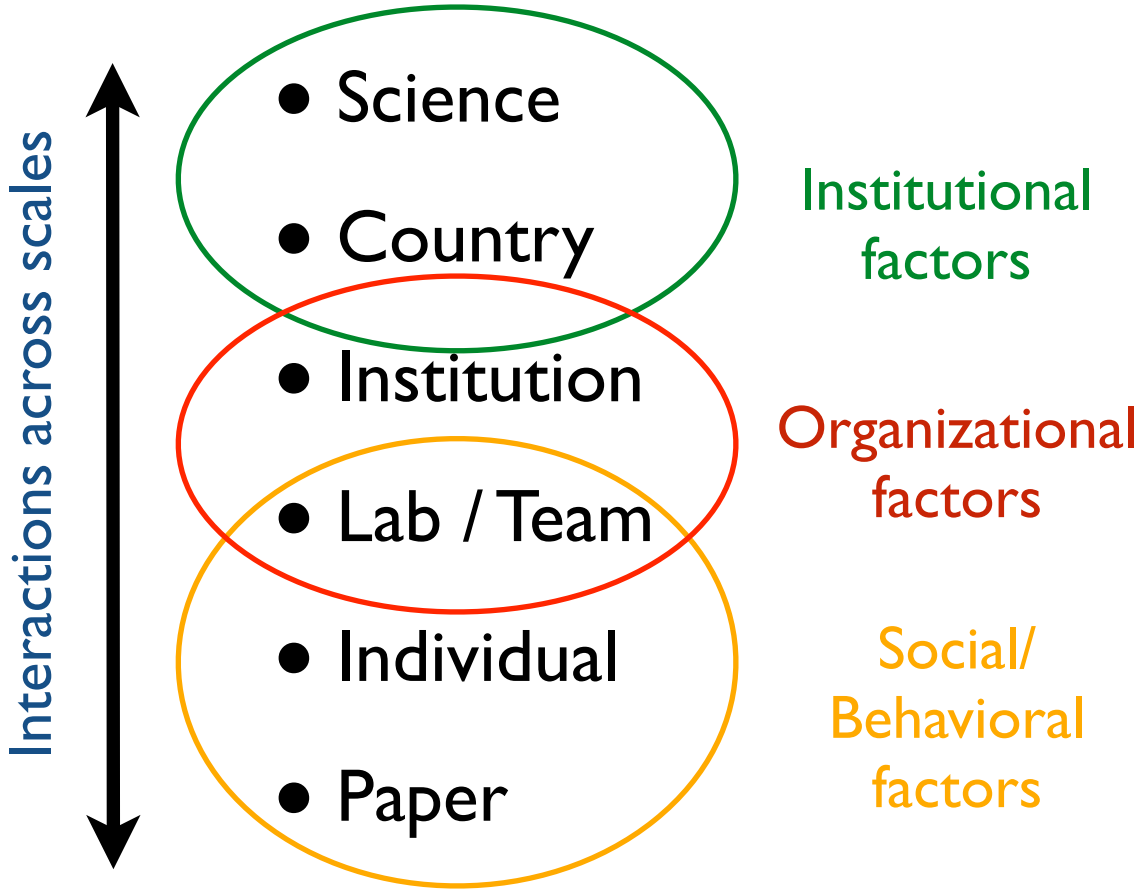
Alexander M. Petersen



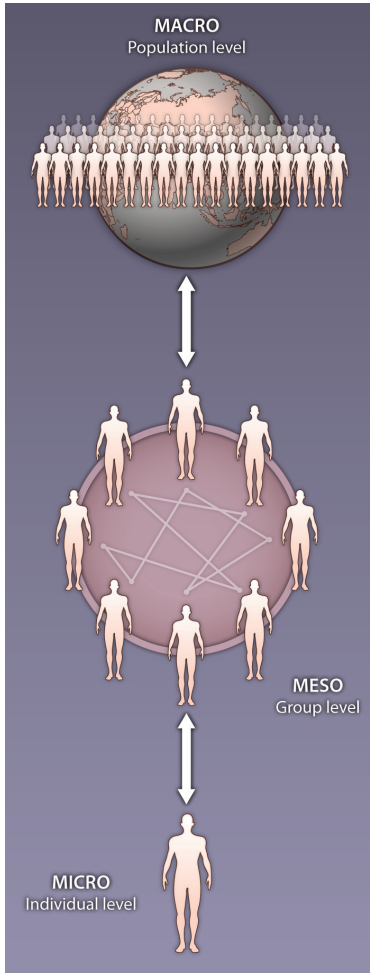
Science is a multi-scale system with emergent complexity

Science of Science

Practical Question: how to measure scientific output and impact at various scales while accounting for systemic heterogeneity

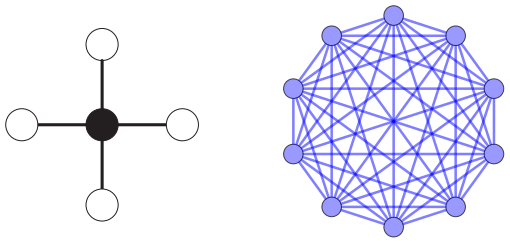


K. Börner, et al. A multi-level systems perspective for the science of team science. Sci. Transl. Med. 2, 49cm24 (2010).



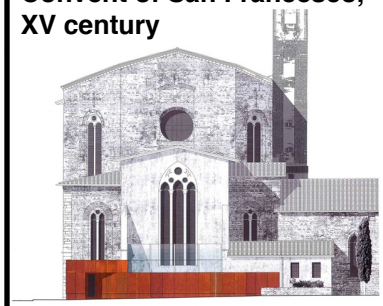
Paradigm shifts

Limited complexity in small knowledge networks

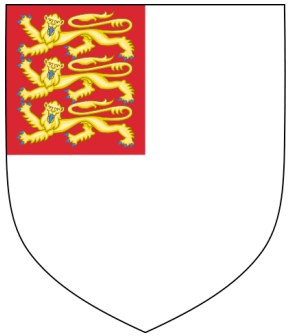


Early scholarly societies, e.g. national societies, scholastic monasteries, noble courts

Convent of San Francesco, XV century

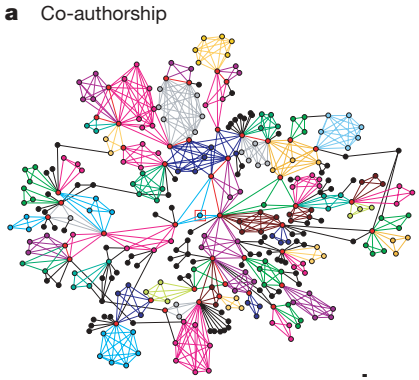


The Royal Society of London for Improving Natural Knowledge, Established 1660



growth and increasing *organizational complexity*

Emergent complexity in large knowledge networks



G. Palla, A.-L. Barabasi, T. Vicsek. Quantifying social group evolution. Nature 446, 664-667 (2007)

S. Wuchty, B. F. Jones, B. Uzzi. The increasing dominance of teams in production of knowledge. Science 316, 1036-9 (2007)

Harvard University

Urban property
210 acres (85 ha) (Main campus)
21 acres (8.5 ha) (Medical campus)
360 acres (150 ha) (Allston campus)
4,500 acres (1,800 ha) (other holdings)

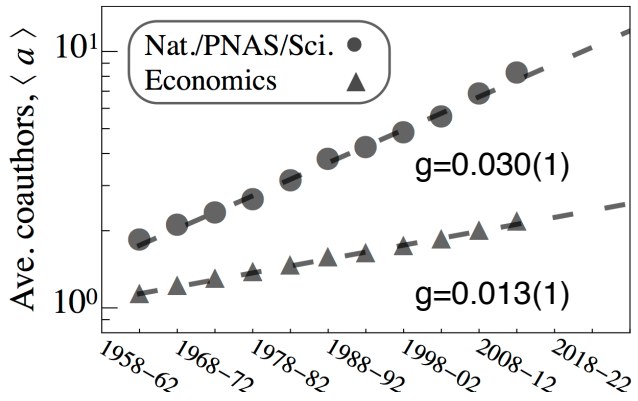
Academic staff
2,100

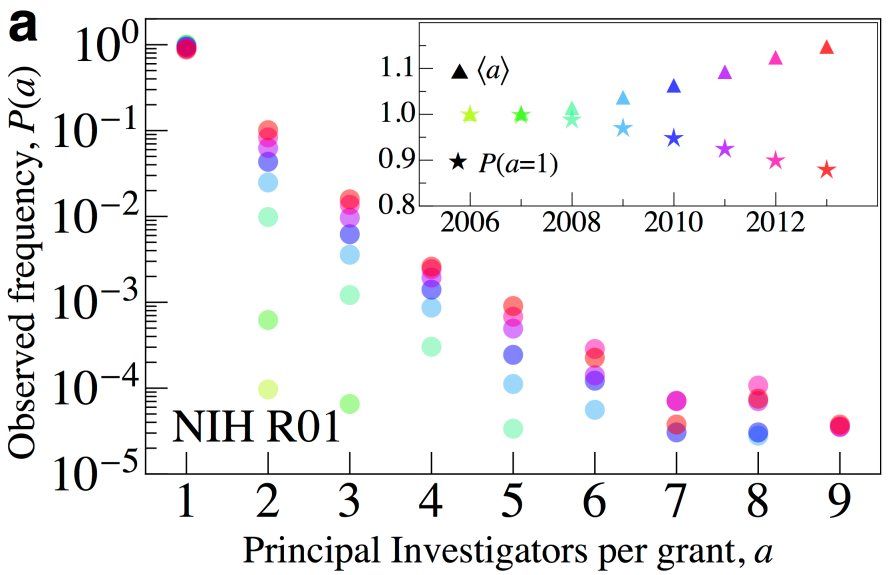
Admin. staff
2,500 non-medical
11,000 medical

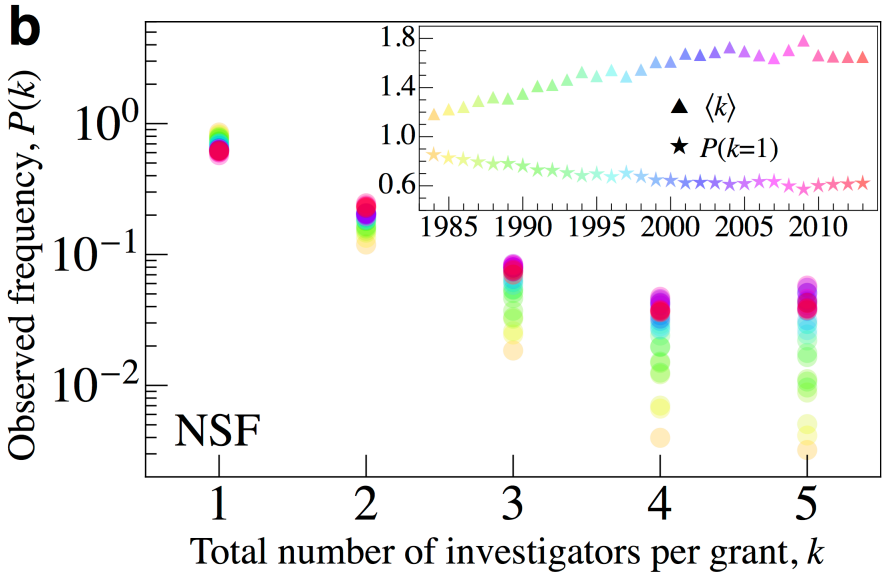
Endowment
US$30 billion (2012) (Large-cap company, e.g. same market capitalization as Enel and Mitsubishi)

How might paradigm shifts in science affect science careers?

Access to new opportunities increasingly dependent on the embedding within teams / organizational units







Macro (institutions)

- Exponential growth of Science
- Economics of research universities and govt. funding
- Increasing role of teams (division of labor) in science with implications on allocation of resources

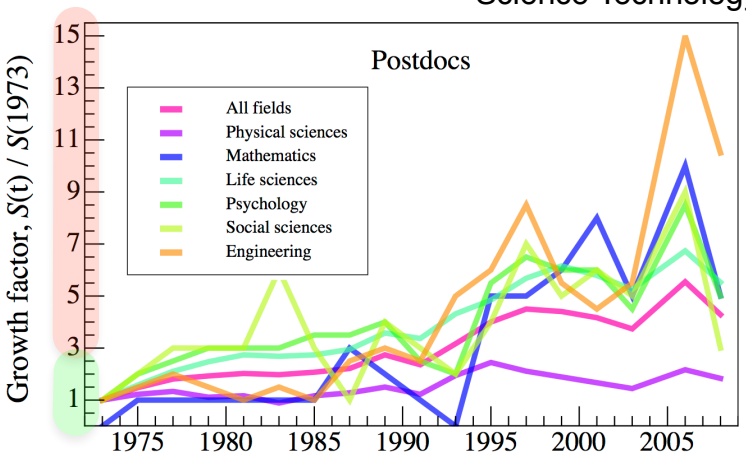
Micro (individual careers)

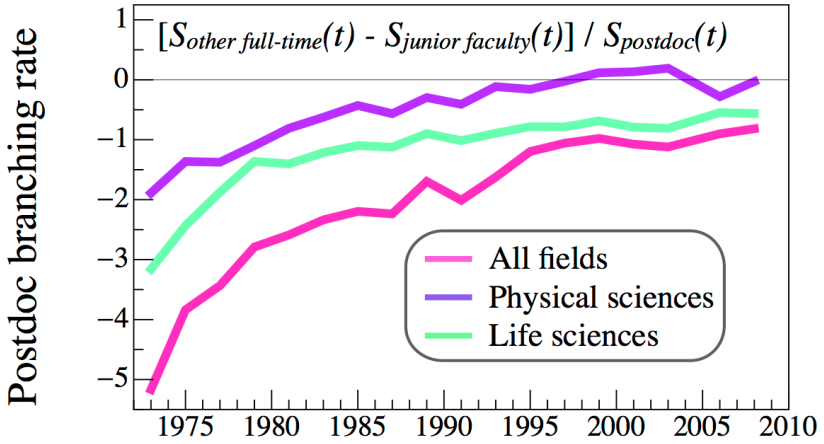
- Growth of careers
- Collaboration patterns within careers
- Competition
- Issues of ethics (rules of the game)

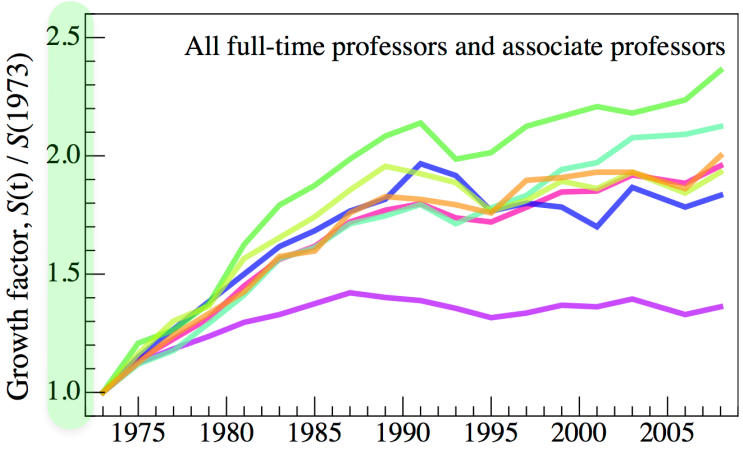
Increased competition for limited resources

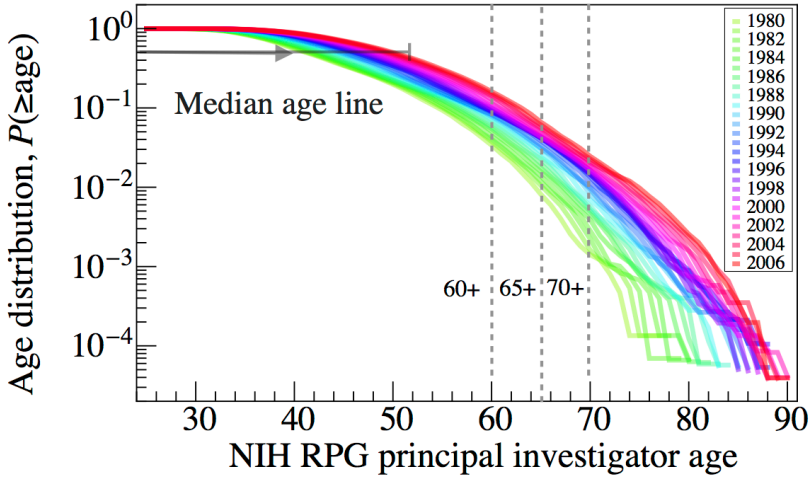
- Bottle-neck in the tenure track model: redirection of PhDs into postdocs and non-tenure track personnel
- Demographic shifts: aging, globalization, and brain drain (e.g. 2004 Euro expansion)

Science Technology & Medicine Faculties (STEM), USA









Ethical scandals reveal the price of success

"...one survey estimated that almost 7% of students in US universities have used prescription stimulants [Adderall and Ritalin] in this way, and that on some campuses, up to 25% of students had used them in the past year. These students are early adopters of a trend that is likely to grow, and indications suggest that they're not alone."

Towards responsible use of cognitive-enhancing drugs by the healthy

Society must respond to the growing demand for cognitive enhancement. That response must start by rejecting the idea that 'enhancement' is a dirty word, argue **Henry Greely and colleagues**.

NATURE|Vol 456|11 December 2008

Professor's little helper

The use of cognitive-enhancing drugs by both ill and healthy individuals raises ethical questions that should not be ignored, argue **Barbara Sahakian** and **Sharon Morein-Zamir**.

NATURE|Vol 450|20/27 December 2007

"One in five respondents said they had used drugs for non-medical reasons to stimulate their focus, concentration or memory. Use did not differ greatly across age-groups..., which will surprise some. "

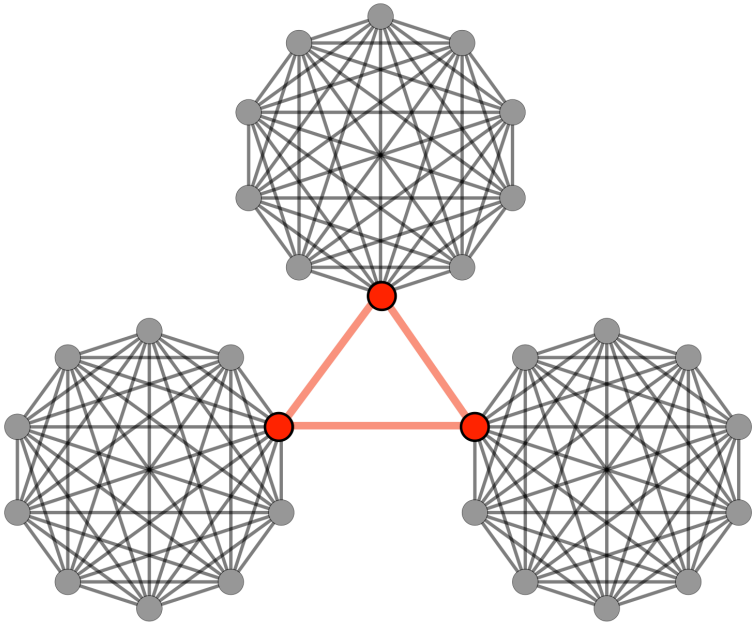
NATURE|Vol 452|10 April 2008

Poll results: look who's doping

In January, *Nature* launched an informal survey into readers' use of cognition-enhancing drugs. **Brendan Maher** has waded through the results and found large-scale use and a mix of attitudes towards the drugs.

Team Ethics: Credit distribution in large team science

The reward system in science developed during a period when teams were relatively small. Hence, there is an inherent difficulty in distributing fairly sliced credits in large modular teams comprised of *heterogenous* members



$a = 30, N = 138$

2008-2012

NEJM (Medicine), $P(\geq 30) = 0.065$
PRL (Physics), $P(\geq 30) = 0.040$
Cell (Biology), $P(\geq 30) = 0.017$

Cutting the "credit pie" fairly:
Who gets credit? "Who's on first"?

Citation (impact) credit:
- Is it shared equally amongst *a* coauthors?

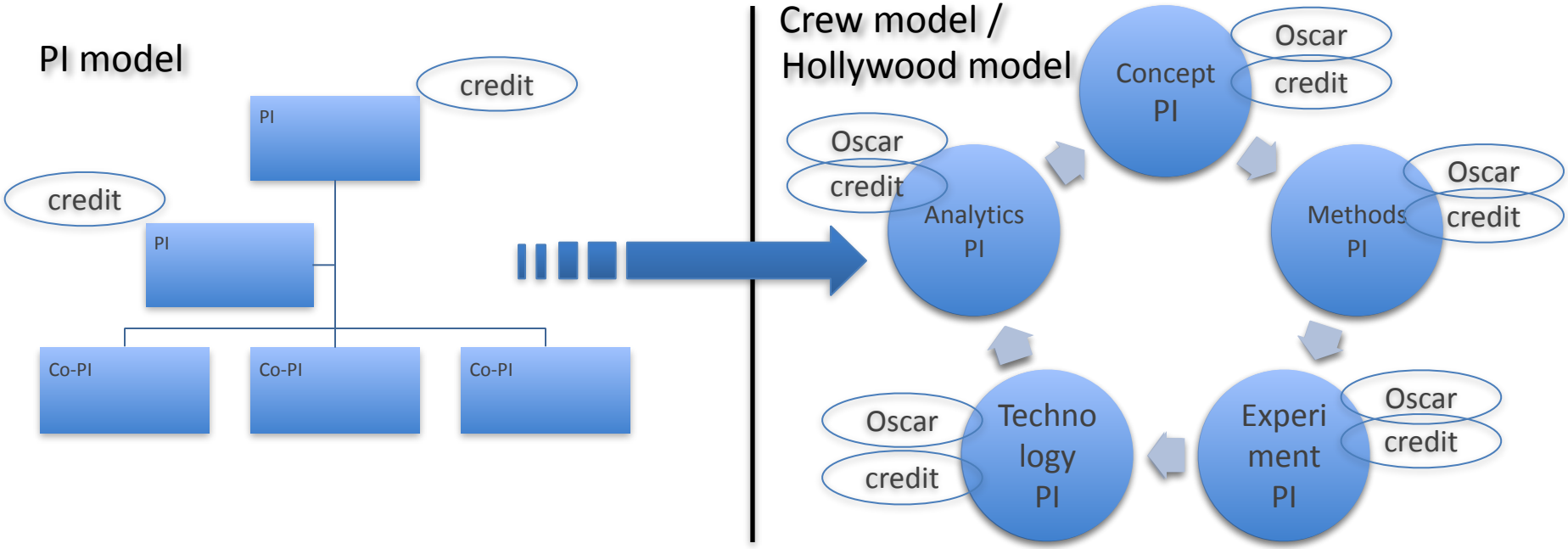
Fraud/Retraction anti-credit:
- can impact all *a* coauthors
- If credit is shared equally then should blame also?

~ factor of 20 increase in retractions from 2000 - 2010
The retraction penalty: Evidence from the web of science.
Lu SF, Jin GZ, Uzzi B, Jones B. Scientific Reports 3, 3146 (2013).

Redesigning the credit system in science?

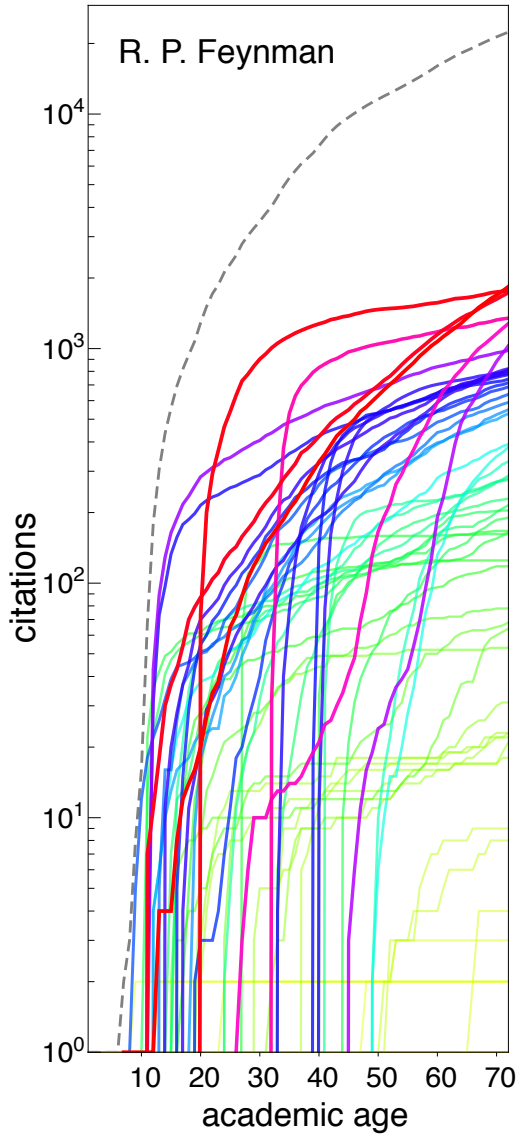
Adoption of career models from communities that embraced a team structure (e.g., filmmaking)

- PI model ➡ crew model
- uni-polar reward system ➡ multi-polar reward system



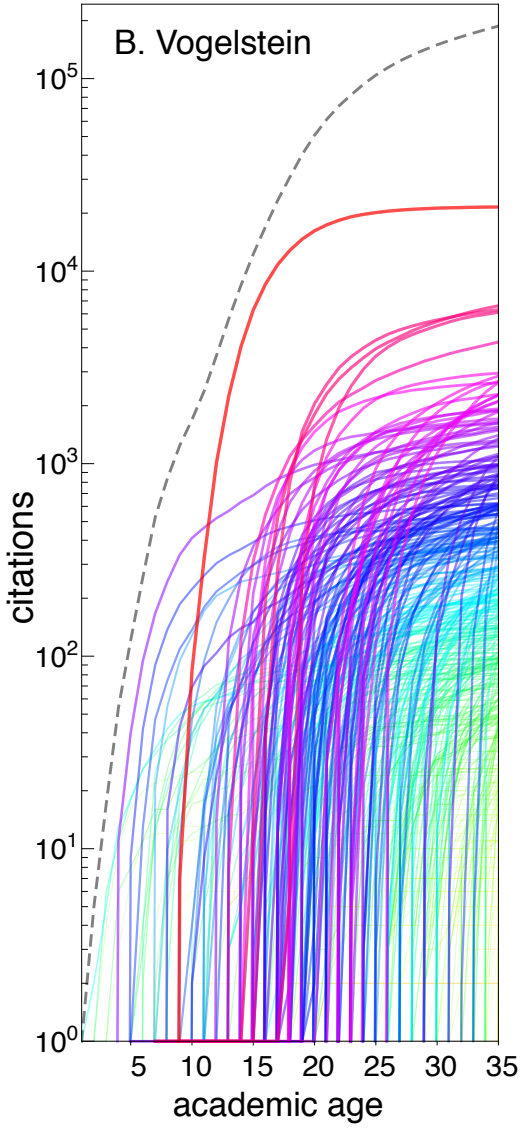
Together We Stand, *Nature Physics (2014) I. Pavlidis, A. M. Petersen, I. Semendeferi.*

Quantifying career growth in science

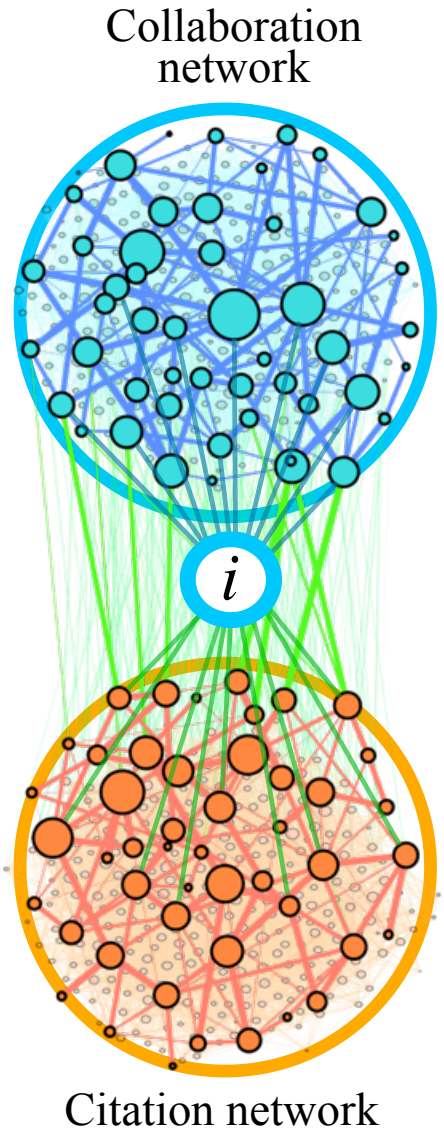


(A)
cumulative advantage
in the context of high-impact journals

(B)
microscopic reputation
mechanisms
operating at the level
of individual papers



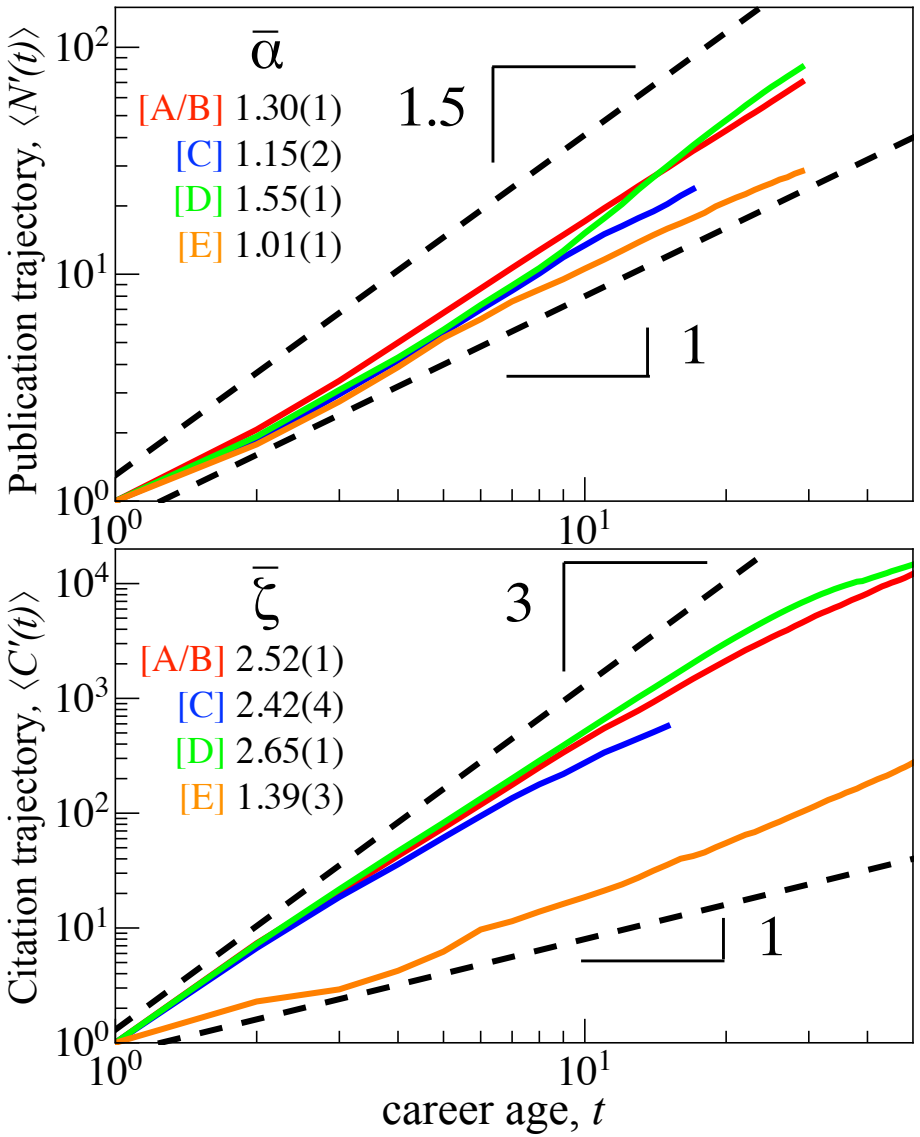
Science careers are embedded in a co-evolving network of networks



Complexity

- coevolutionary system:
  - knowledge
  - institutions
  - careers
- social processes:
  - behavioral aspects
  - economic incentives
  - cumulative advantage mechanisms
  - collaboration / competition

Quantitative evaluation in science is increasingly based on productivity and citations measures



The data: longitudinal Web of Science publication and citation data for 450 top scientists; 83,693 papers, 7,577,084 citations tracked over 387,103 years

Highly-cited scientists establish upper limits to longitudinal career growth

Set A: 100 most-cited physicists, average h-index, $\langle h \rangle = 61 \pm 21$

Set B: 100 additional highly-prolific physicists, $\langle h \rangle = 44 \pm 15$

Set C: 100 assistant professors from 50 US physics depts., $\langle h \rangle = 15 \pm 7$

Set D: 100 most-cited cell biologists, $\langle h \rangle = 98 \pm 35$

Set E: 50 highly-cited pure mathematicians, $\langle h \rangle = 20 \pm 10$

$\zeta > \alpha > 1$: knowledge, reputation, and collaboration spillovers contribute to sustainable growth across the academic career

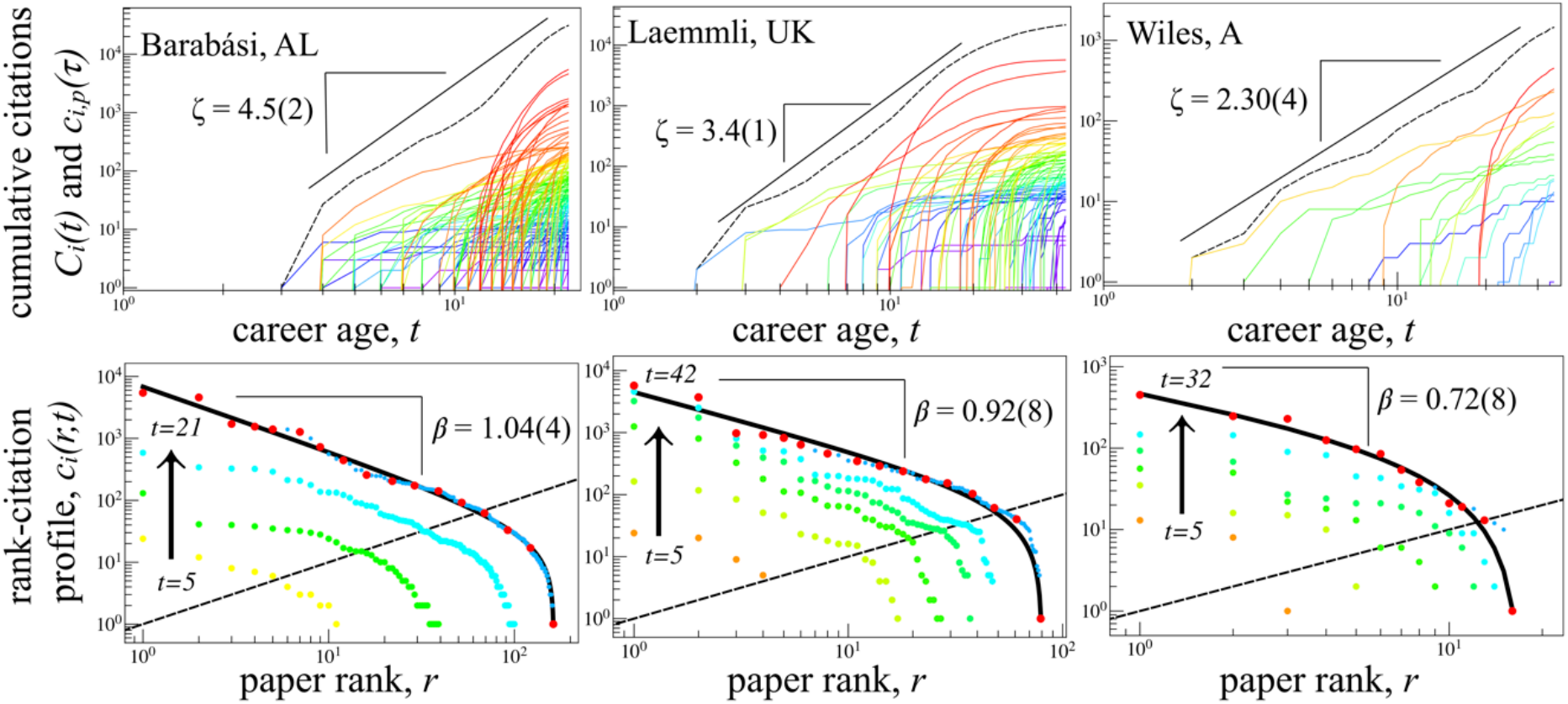
Benchmark patterns of microscopic career growth dynamics

$$c_p(\tau) = \sum_t \Delta c_p(t)$$

cumulative # of citations at paper age τ

$$C_i(t) = \sum_{r=1}^{N_i(t)} c_i(r,t) \sim t^{\zeta_i}$$

cumulative citations by career age *t*



The rank-citation profile illustrates the evolution of the publication-impact portfolio

$$c(r) \equiv A r^{-\beta}(N+1-r)^{\gamma}$$

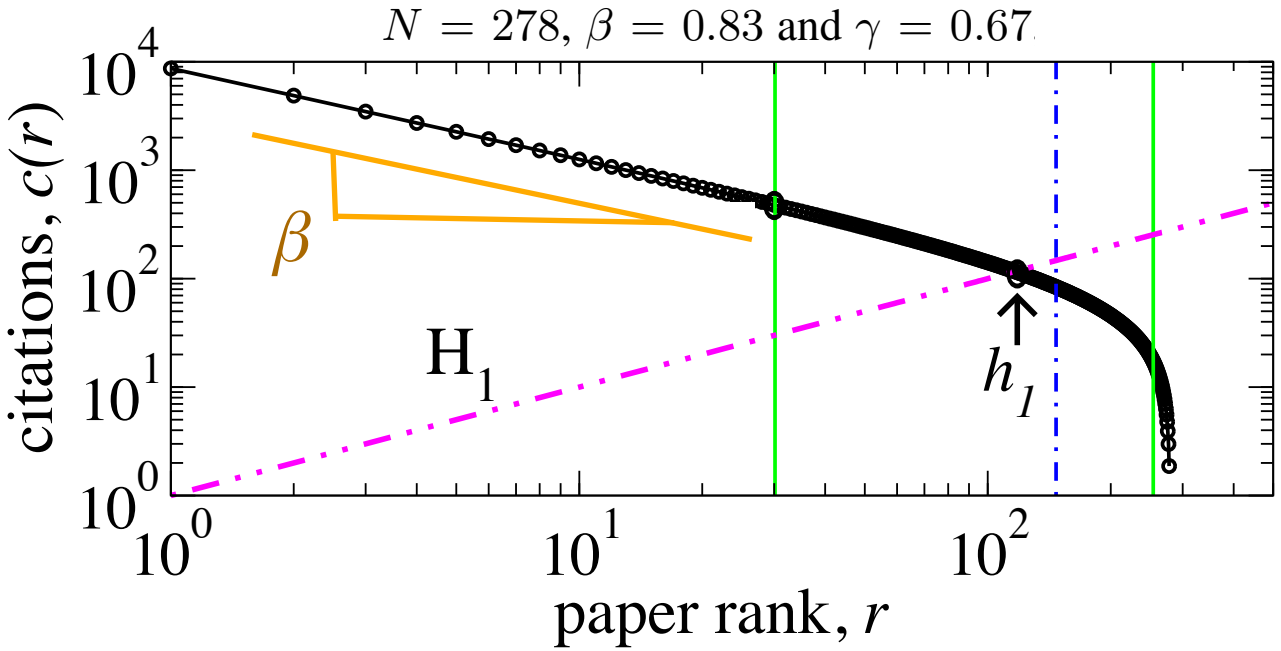
discrete generalized Beta function (DGBD)

$$C_i \sim h_i^{1+\beta_i}$$

simple scaling relation between the *h*-index and *C*

Statistical regularities in the rank-citation profile of scientists, A. M. Petersen, H. E. Stanley, S. Succi. Scientific Reports 1, 181 (2011).

Statistical regularity in the rank-citation distribution: the Discrete Generalized Beta Distribution (DGBD) model for $c_i(r)$



$$c(r) \equiv Ar^{-\beta}(N+1-r)^{\gamma} \ .$$

N_i = # of publications
β_i = scaling slope of top papers
γ_i = truncation scaling of less-cited papers
C_i = total citations from all papers

Scaling relation between C, h, and β

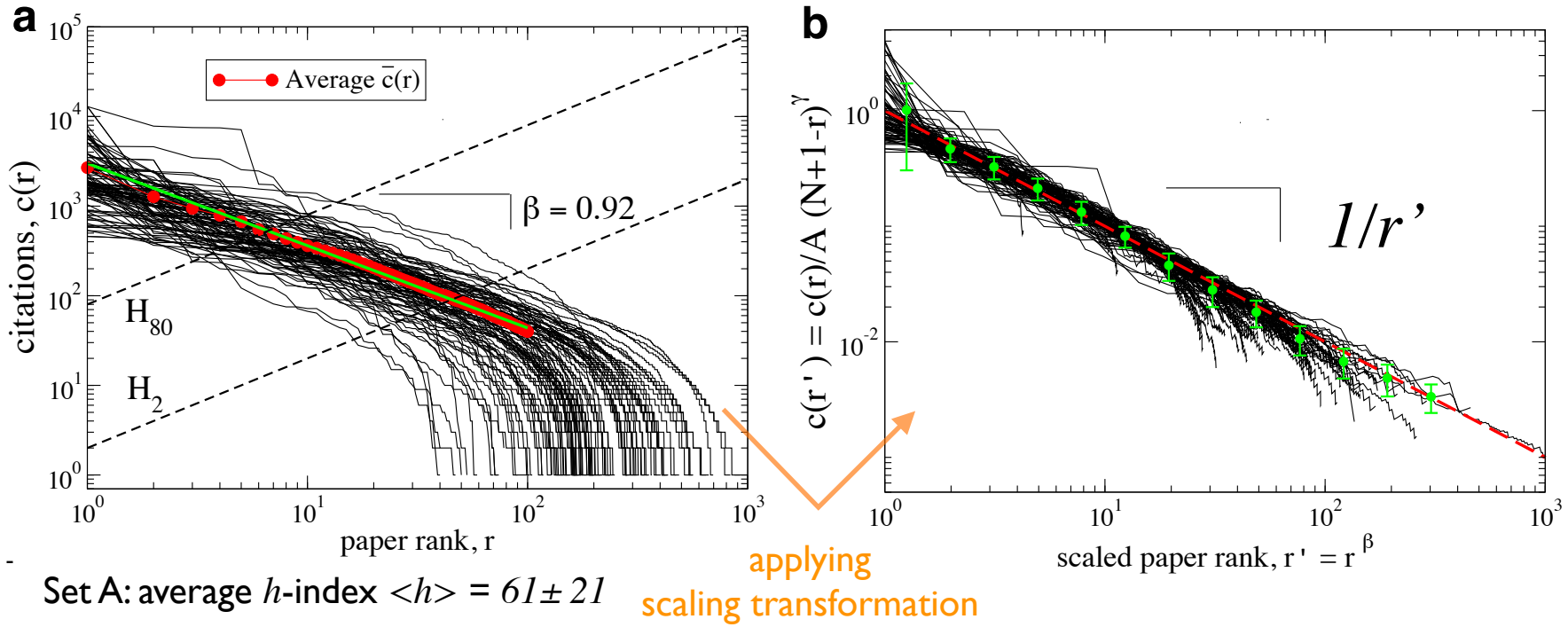
$$C \sim h^{1+\beta}$$

⇒ Hence,

knowing both the h-index and C is ≈ redundant

A. M. Petersen, H. E. Stanley, S. Succi. Statistical regularities in the rank-citation profile of scientists. Scientific Reports 1, 181 (2011).

A comparison of $c_i(r)$ for the top-100 "champions" of Physical Review Letters



Set A: average h-index $\langle h \rangle = 61 \pm 21$

Discrete Generalized Beta Distribution(DGBD):

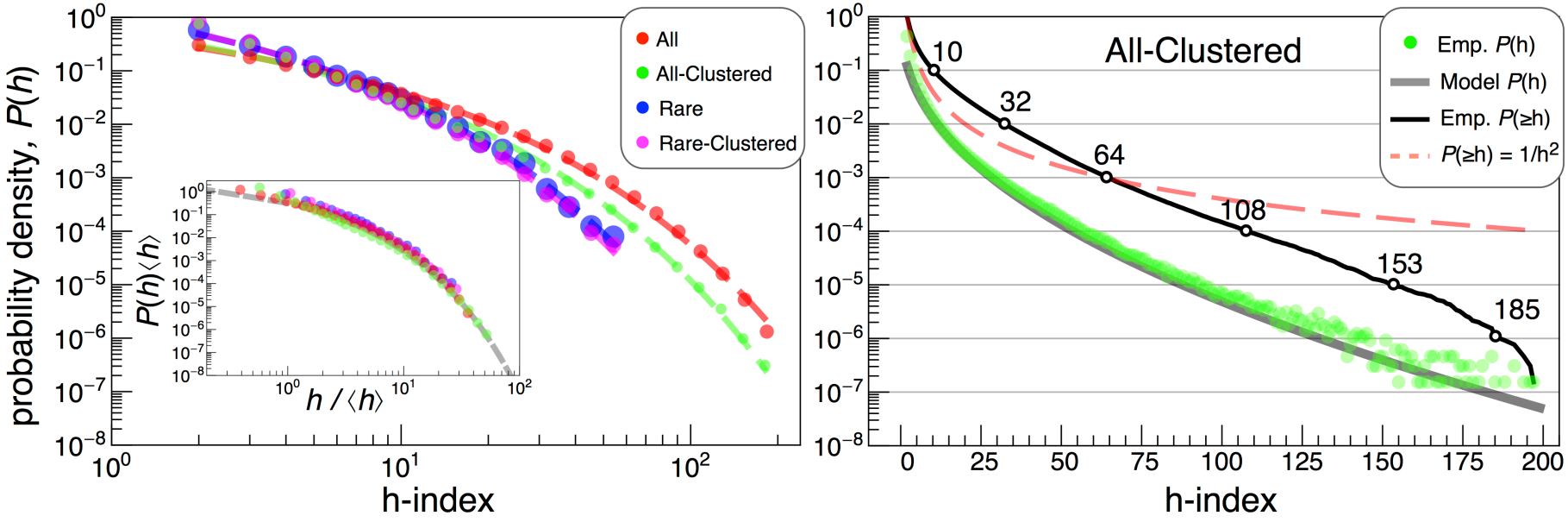
$$c(r) \equiv A r^{-\beta} (N + 1 - r)^{\gamma} .$$

Average values of the DGBD model parameters:

$\langle \beta \rangle = 0.83 \pm 0.23$ and $\langle \gamma \rangle = 0.67 \pm 0.19$

Inequality in science careers

The h-index distribution derived from the full Web of Science citation index: 6,498,286 research profiles



Exploiting citation networks for large-scale author name disambiguation.
C. Schulz, A. Mazloumian, A. M. Petersen, O. Penner, D. Helbing.
EPJ Data Science (2014)

Data-driven investigation of cumulative advantage processes in competitive arenas

Physical Review Letters
moving physics forward

Science
AAAS

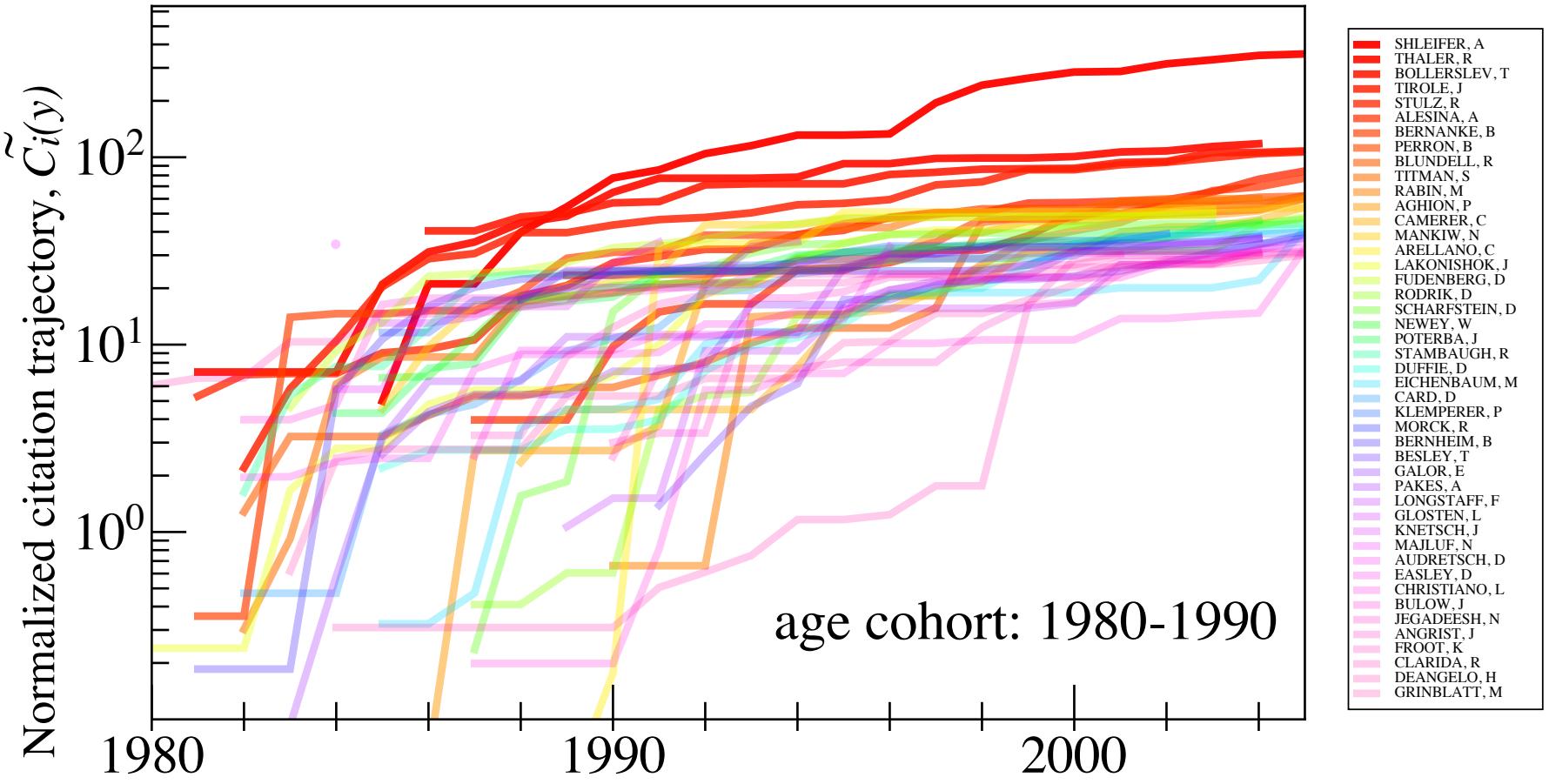
nature

Cell

The NEW ENGLAND JOURNAL of MEDICINE

PNAS
Proceedings of the National Academy of Sciences of the United States of America

Champions of Economics



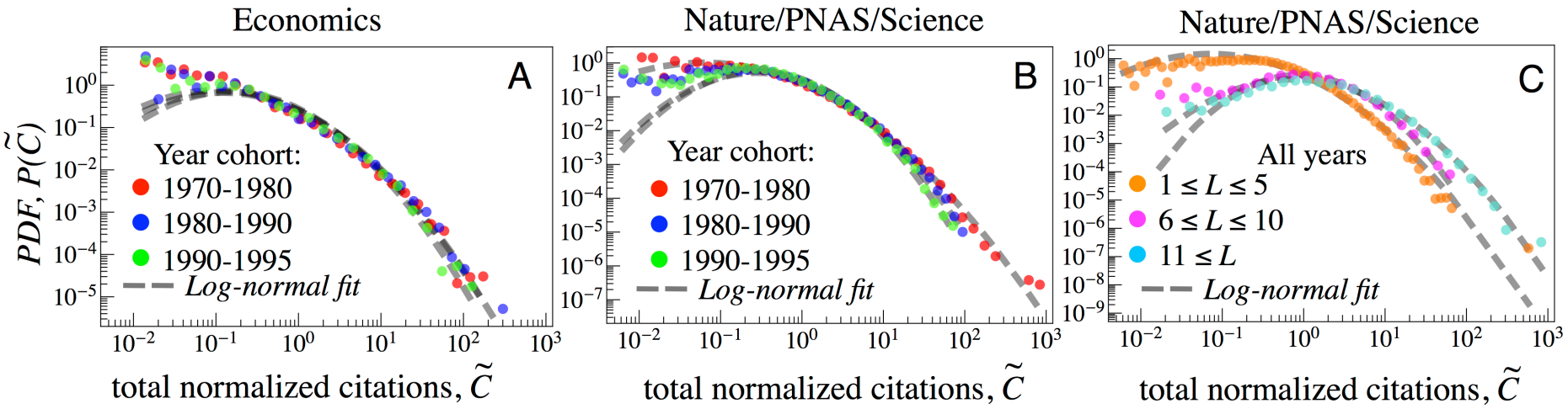
Normalized citation impact

$$\tilde{c}_{i,p}(y) = c^{j}_{i,p}(y)/\langle c^{j}(y)\rangle$$

Aggregate impact (color value)

$$\tilde{C}_i(y) = \sum_{p=1}^{N^j(y)} \tilde{c}_{i,p}(y) \ ,$$

Levels of inequality in science careers



Detrending citation counts to account for cohort bias

$$\tilde{c}^{\,j}_{i,p}(y) = c^{\,j}_{i,p,Y}(y) \;/\; \langle c^{j}_{Y}(y) \rangle,$$

Measuring cumulative citation impact within high-impact journals

$$\tilde{C}^{j}_{i}(y) = \sum_{p=1}^{N^{j}_{p}(y)} \tilde{c}^{\,j}_{i,p}(y)$$

Inequality and cumulative advantage in science careers: a case study of high-impact journals.
A. M Petersen, O. Penner.
EPJ Data Science (2014).

Scientific careers demonstrate a wide range of long-normally distributed citation impact, even after controlling for censoring bias, cohort bias, and controlling for career longevity.

Log-normal "size" distributions suggests that $\tilde{C}^{j}_{i}$ follows a Gibrat proportional growth process

Gini index and top-1% share of total citations in high-impact journals

Journal set j	Cohort entry years	$G(\tilde{C})$	$f_{1\%}(\tilde{C})$	$G(N_p)$	$f_{1\%}(N_p)$
Economics	1970 – 1995	0.80	0.23	0.54	0.09
	1970 – 1980	0.83	0.26	0.56	0.10
	1980 – 1990	0.79	0.21	0.55	0.09
	1990 – 1995	0.74	0.19	0.47	0.07
Nat./PNAS/Sci.	1970 – 1995	0.69	0.18	0.46	0.10
	1970 – 1980	0.74	0.22	0.53	0.12
	1980 – 1990	0.67	0.15	0.45	0.08
	1990 – 1995	0.63	0.12	0.35	0.06

Decreasing levels of inequality over time

Summary of the Gini index (G) and top-1% share ($f_{1\%}$) inequality measures calculated from the distributions of citation impact ($\tilde{C}$) and productivity (N_p) for the cohorts of scientists whose first publication occurred in the indicated time intervals.

Interestingly, this story seems to be opposite of what has been observed in a recent analysis of US research institute funding, which indicates a slow but steady increase in the *G* across U.S. universities over the last 20 years, with current estimates of the Gini inequality index for university expenditure around $G \approx 0.8$ (Xie, Science, 2014).

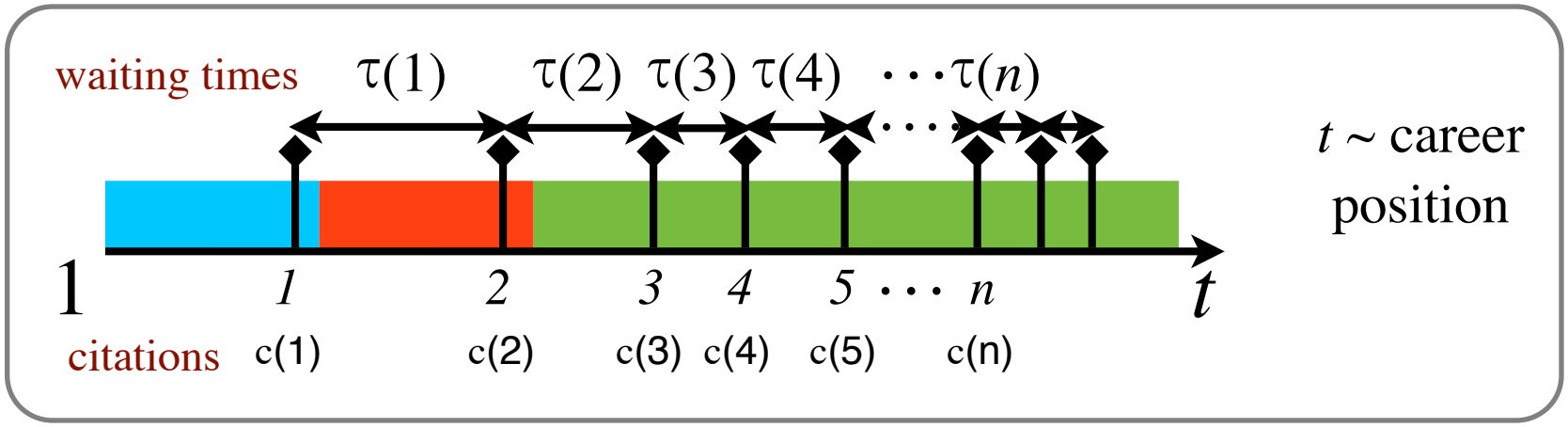
For comparison, the 2010 U.S. income Gini coefficient was $G = 0.4$, and the top 1% share of individual income (USA) has increased from roughly 10% to 20% over the last half century.

Citation inequality levels are high, but over time, science appears to becoming more equitable! (**Possibly a collaboration effect)

A) cumulative advantage: a case-study of high-impact journals

Macro-level of career trajectories

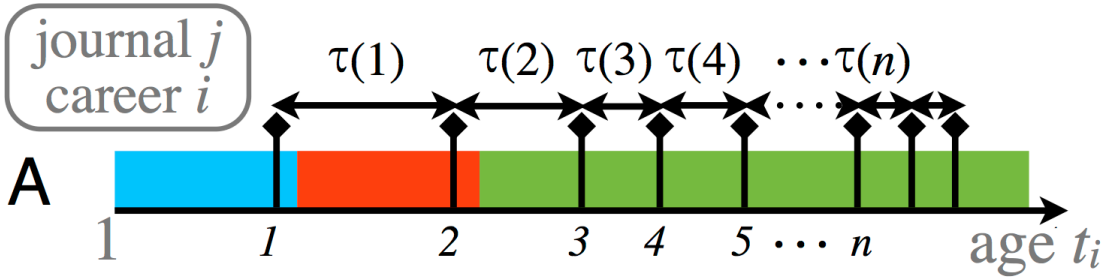
career i

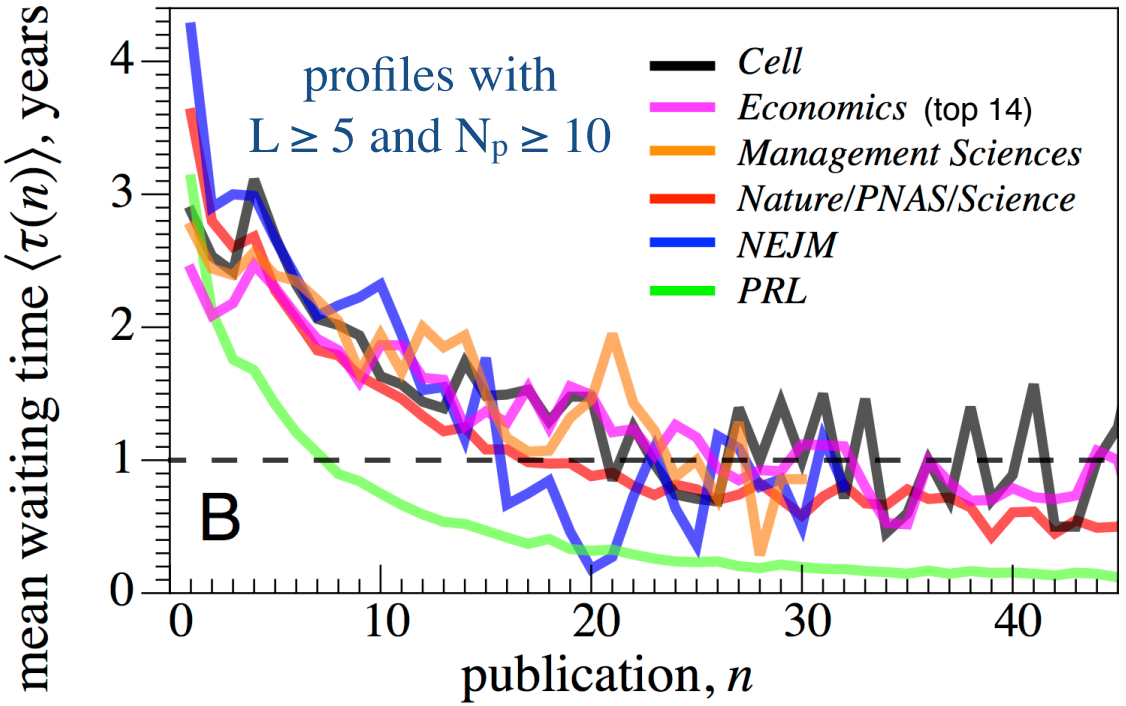


What generic processes might contribute to sustained growth across the career?

How long does a researcher typically wait before his/her next publication in a prestigious journal?

For each career *i* we track his/her longitudinal publication rate by aggregating over publications in a *specific set* of high-impact journals



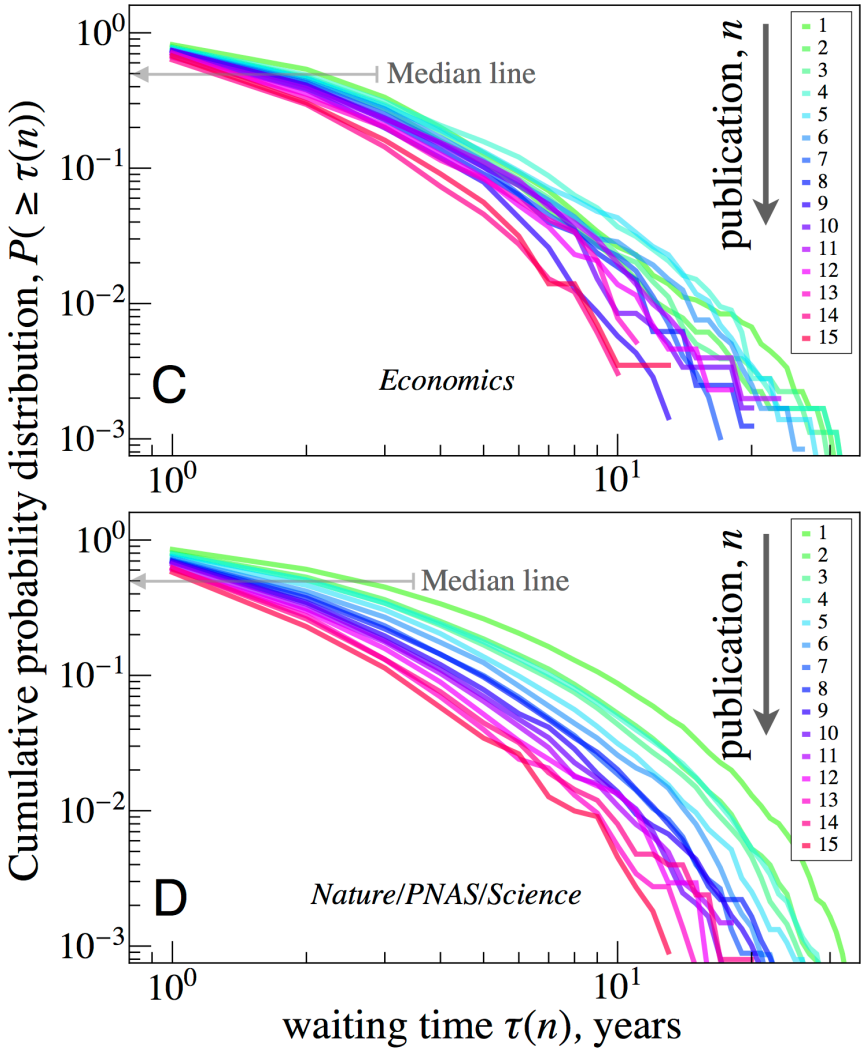


$\tau_i(n)$ is the waiting time between an author's n^{th} paper and $(n+1)^{th}$ paper?

By the 10th paper, the waiting time between publications has decreased by ~ factor of 2 from $\tau_i(1)$!

How long does a researcher typically wait before his/her next publication in a prestigious journal?

For each career i we track his/her longitudinal publication rate by aggregating over publications in a *specific set* of high-impact journals



$\tau_i(n)$ is the waiting time between an author's n^{th} paper and $(n+1)^{th}$ paper?

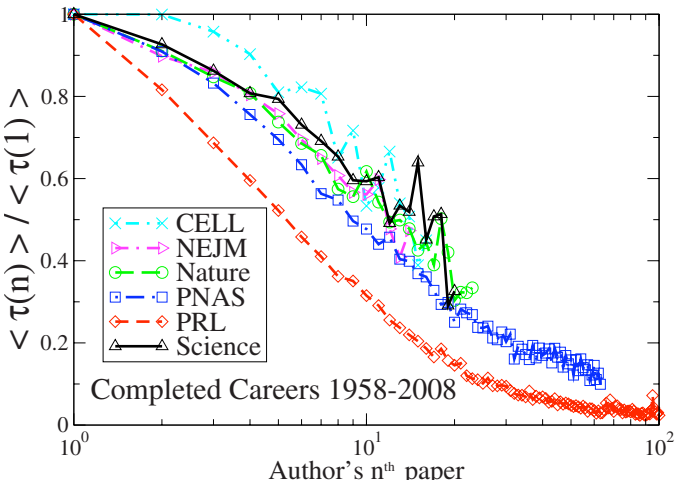
The trend in the average $\langle \tau(n) \rangle$ is not just a "tail effect" , but is apparent in a shift in the entire distribution of waiting times

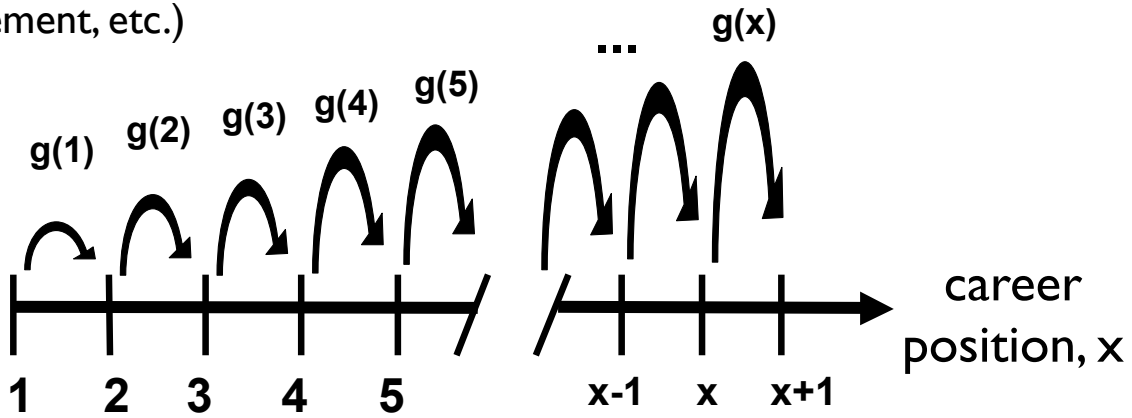
These empirical findings of decreasing waiting time are consistent with a "Matthew Effect" rich-get-richer model

Cumulative advantage model: Two main ingredients

1) Forward progress follows a stochastic "progress rate" $g(x)$. Cumulative advantage corresponds to $g(x)$ increasing with career position x

2) Random termination of the career due to hazards (e.g. decreased work performance, economic down, economic downturn, health, retirement, etc.)





$$g(x) = 1 / \langle \tau(x) \rangle$$

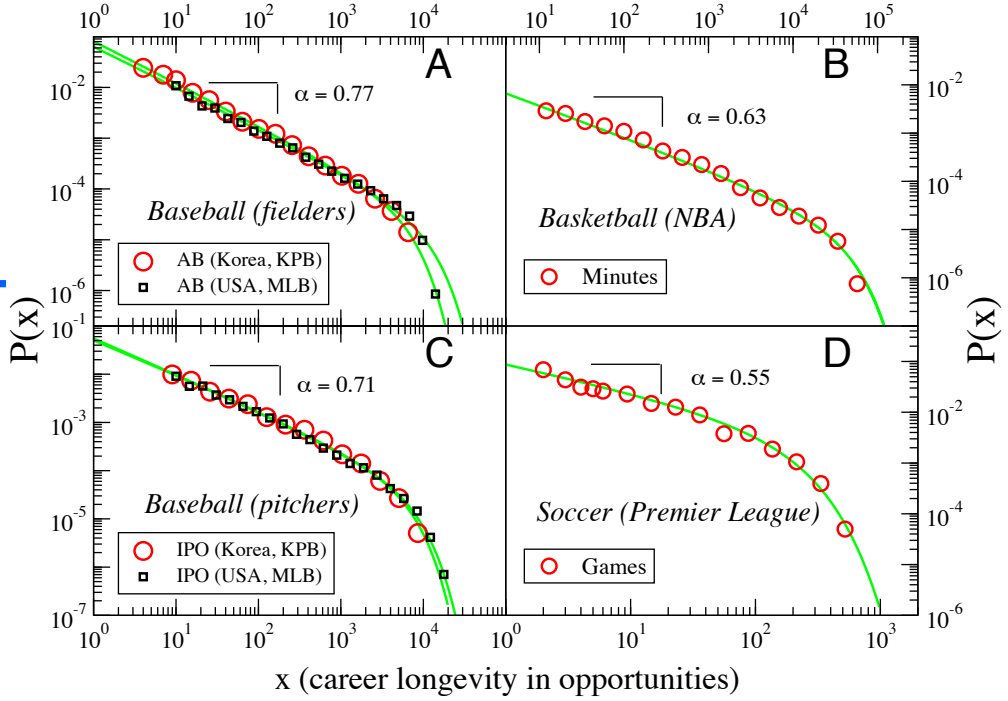
The progress probability g is the inverse of the mean waiting time τ

Methods for measuring the citations and productivity of scientists across time and discipline, A. M. Petersen, F. Wang, H. E. Stanley. Phys. Rev. E 81, 036114 (2010).

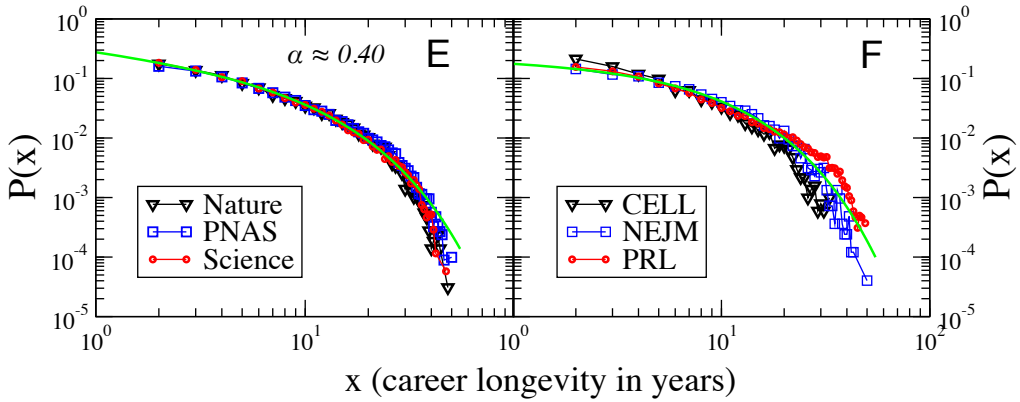
Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley. Proc. Natl. Acad. Sci. USA 108, 18-23 (2011).

Statistical regularities in the career longevity distribution

Pro Sports



Academia



opportunities ~ time duration

Quantitative and empirical demonstration of the Matthew effect in a study of career longevity, A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley. Proc. Natl. Acad. Sci. USA 108, 18-23 (2011).

Major League Baseball

- 130+ years of player statistics, ~ 15,000 careers

``One-hit wonders"

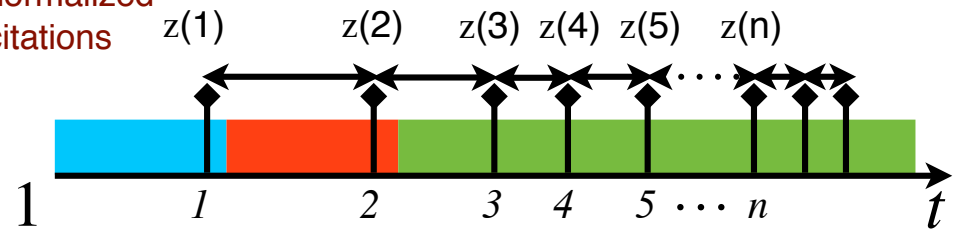
- 3% of all fielders finish their career with ONE at-bat!
- 3% of all pitchers finish their career with less than one inning pitched!

``Iron horses"

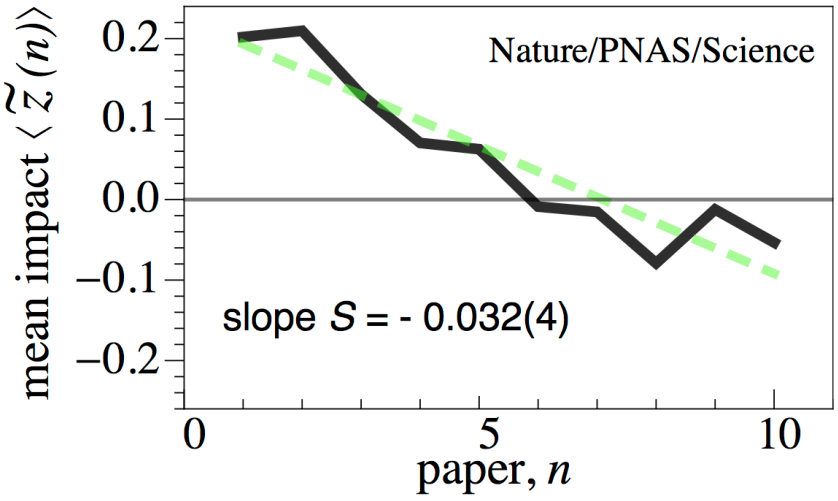
- Lou Gehrig (the Iron Horse): NY Yankees (1923-1939)
- Played in 2,130 consecutive games in 15 seasons! 8001 career at-bats!
- Career & life stunted by the fatal neuromuscular disease, amyotrophic later sclerosis (ALS), aka Lou Gehrig's Disease

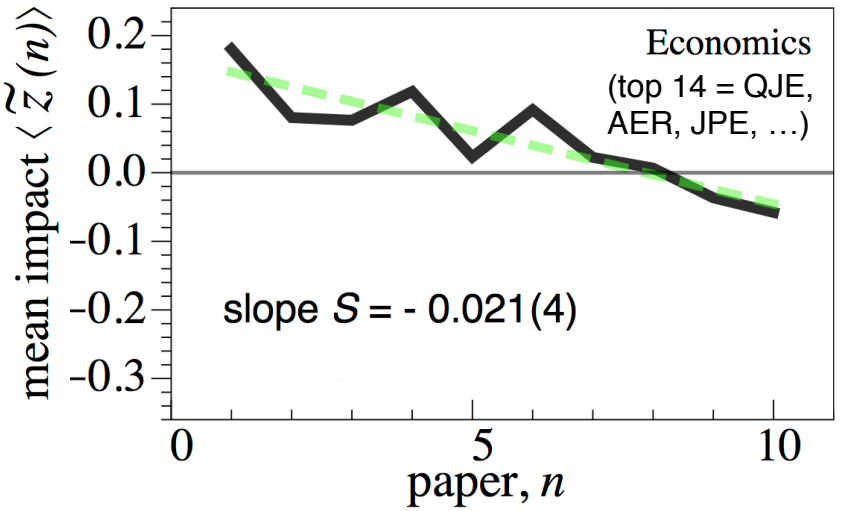
Are researcher's later publications more or less cited than their previous publications?

normalized citations



Inequality and cumulative advantage in science careers: a case study of high-impact journals.
A. M Petersen, O. Penner.
EPJ Data Science (2014).





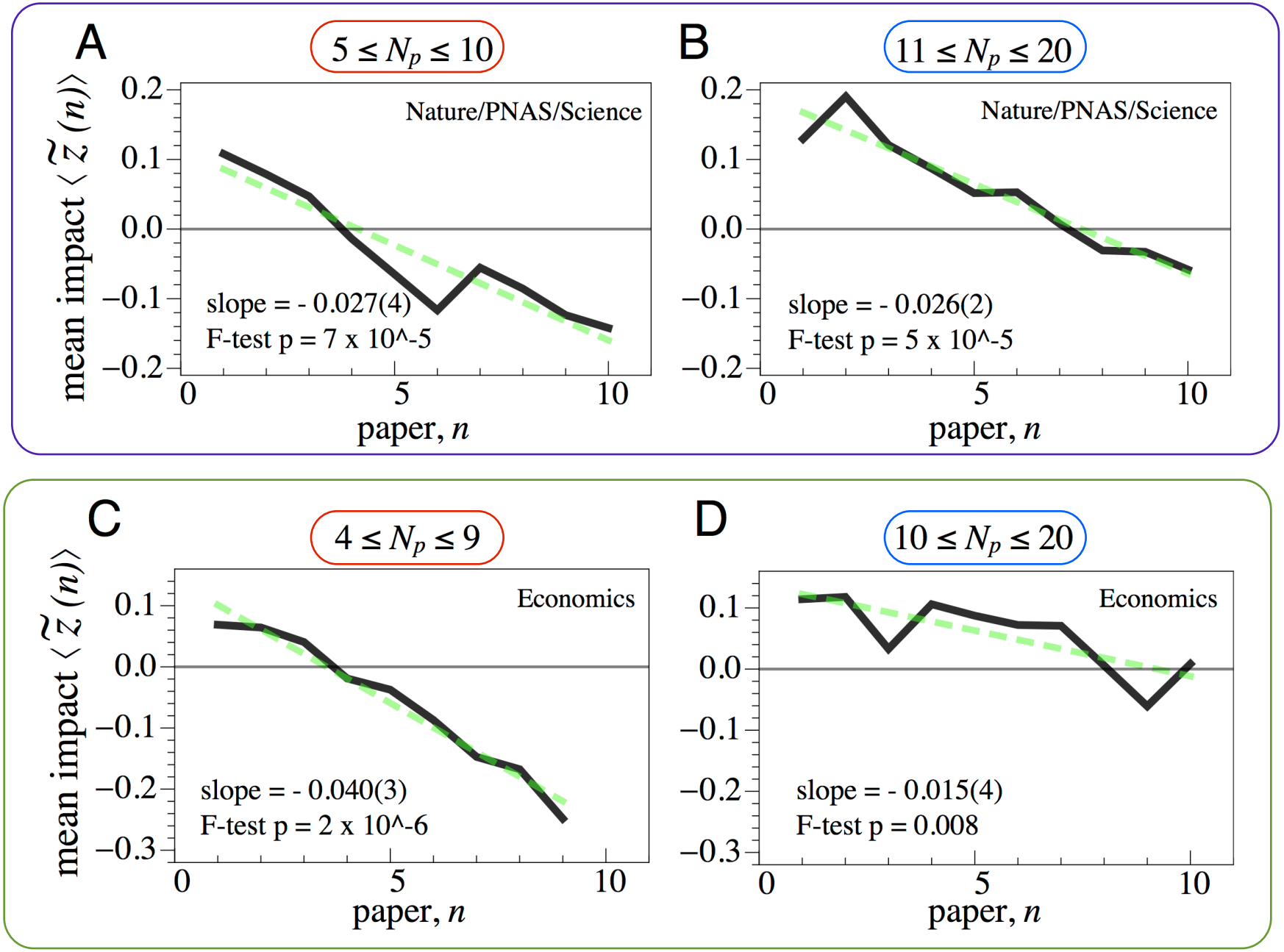
How to account for temporal bias? To investigate the longitudinal variation in the citation impact, we map the citation count $c^j_{i,p,y}$ of the n^{th} publication of researcher i, published in journal set j to a z-score,

$$z_i(n) \equiv \frac{\ln c^j_{i,p,y}(n) - \langle \ln c^j_y \rangle}{\sigma[\ln c^j_y]} \,,$$

$$\tilde{z}_i(n) \equiv z_i(n) - \langle z_i \rangle$$

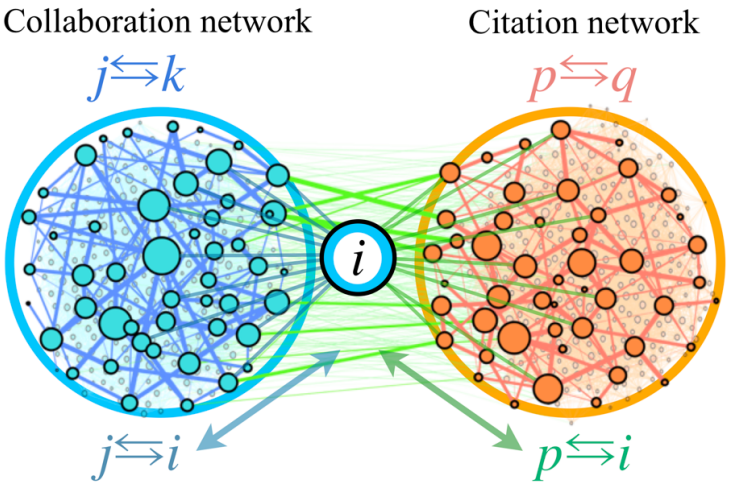
This decreasing impact pattern highlights the difficulty of repeatedly producing research findings in the highest citation-impact echelon, as well as the role played by finite career and knowledge life-cycles.

Indication of confirmation bias in science career evaluation?



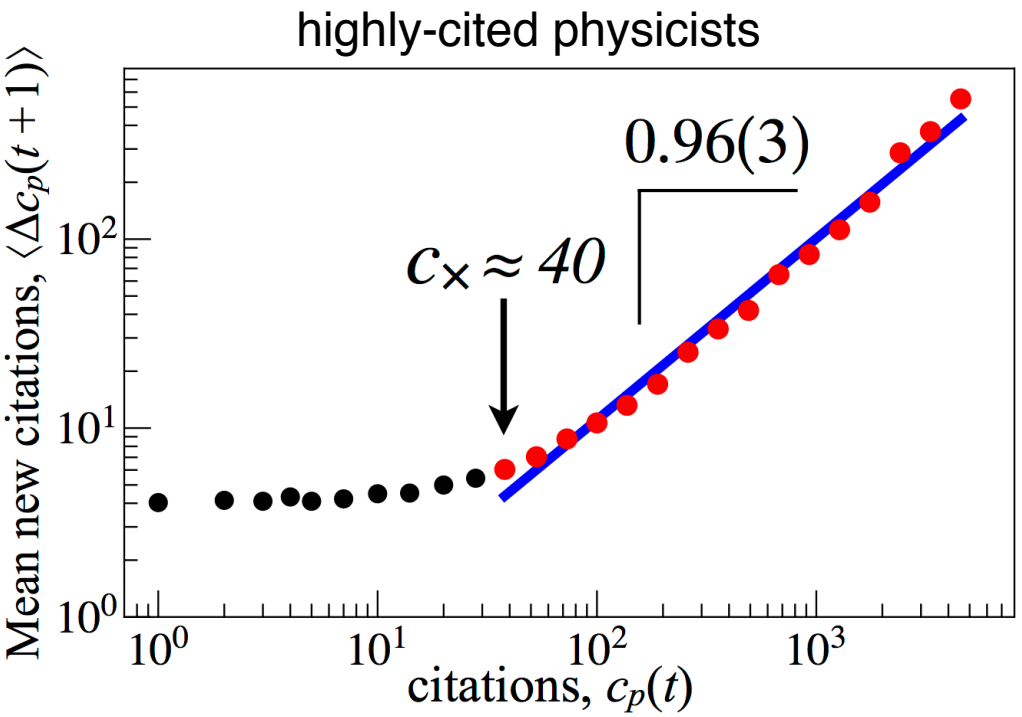
B) Reputation flows in the collaboration-citation network

Micro-level of citation trajectories



Collaboration and citation networks provide channels for reputation signaling

What is the impact of author reputation on a paper's citation rate $(i \longrightarrow p)$?

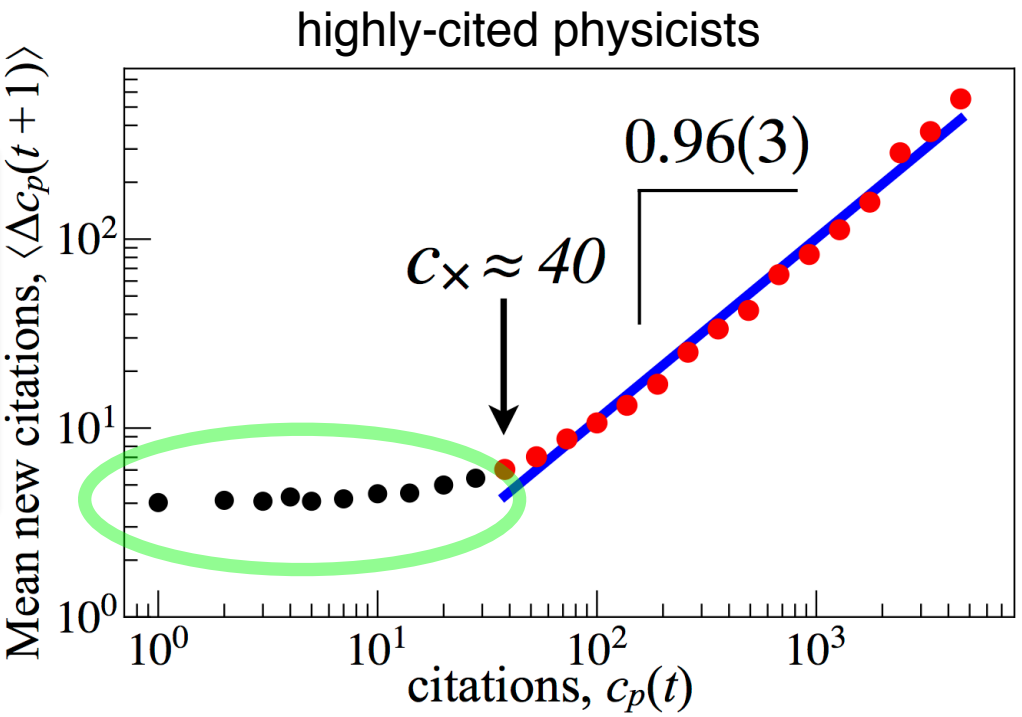


Reputation effect citation model

of new citations in year *t+1* = $\Delta c_{i,p}(t+1) \equiv \eta \times \Pi_p(t) \times A_p(\tau) \times R_i(t)$

1. preferential attachment $\Pi_p(t) \equiv [c_p(t)]^\pi$
2. citation life-cycles $A_p(\tau) \equiv \exp[-\tau_p/\overline{\tau}]$
3. author reputation effect $R_i(t) \equiv [C_i(t)]^\rho$

An excess citation rate above what you would expect from linear preferential attachment alone



Reputation $C_i(t)$ is estimated by the total citations of the most highly cited coauthor (here assumed to be i)

Reputation effect citation model

of new citations in year $t+1$ = $\Delta c_{i,p}(t+1) \equiv \eta \times \Pi_p(t) \times A_p(\tau) \times R_i(t)$

1. preferential attachment $\Pi_p(t) \equiv [c_p(t)]^\pi$
2. citation life-cycles $A_p(\tau) \equiv \exp[-\tau_p/\overline{\tau}]$
3. author reputation effect $R_i(t) \equiv [C_i(t)]^\rho$

Author-specific factors matter,
corresponding to important quantifiable nuances underlying citation dynamics!!!

Author-specific features: π_i, $\overline{\tau}_i$, ρ_i

TABLE I: Best-fit parameters for individual careers and the average values within disciplinary datasets. The three features of the citation model are parameterized by π, the paper citation effect, $\overline{\tau}$, the life-cycle effect, and ρ, the reputation effect.

math | biology | physics

	$c(t-1) < c_\times$			$c(t-1) \geq c_\times$		
Name	π_i	$\overline{\tau}_i$	ρ_i	π_i	$\overline{\tau}_i$	ρ_i
GOSSARD, AC	0.34 ± 0.027	4.92 ± 0.261	0.25 ± 0.008	0.80 ± 0.048	4.73 ± 0.184	0.09 ± 0.024
BARABÁSI, AL	0.42 ± 0.036	3.00 ± 0.155	0.29 ± 0.010	1.06 ± 0.016	3.65 ± 0.111	0.01 ± 0.011
Ave. $\pm$ Std. Dev. [A]	0.43 ± 0.14	5.67 ± 2.52	0.22 ± 0.06	0.96 ± 0.19	8.93 ± 4.09	-0.07 ± 0.11
BALTIMORE, D	0.32 ± 0.018	4.64 ± 0.148	0.28 ± 0.006	0.62 ± 0.047	5.92 ± 0.250	0.15 ± 0.026
LAEMMLI, UK	0.54 ± 0.036	5.09 ± 0.297	0.21 ± 0.014	1.09 ± 0.025	6.40 ± 0.255	-0.12 ± 0.019
Ave. $\pm$ Std. Dev. [D]	0.40 ± 0.14	6.64 ± 6.24	0.26 ± 0.05	0.99 ± 0.22	9.55 ± 26.30	-0.06 ± 0.14
SERRE, JP	0.33 ± 0.095	15.90 ± 3.724	0.14 ± 0.026	0.66 ± 0.065	20.50 ± 3.862	-0.03 ± 0.039
WILES, A	0.56 ± 0.208	5.23 ± 1.187	0.24 ± 0.052	0.70 ± 0.059	9.04 ± 0.633	0.10 ± 0.042
Ave. $\pm$ Std. Dev. [E]	0.27 ± 0.17	30.60 ± 56.80	0.14 ± 0.07	0.54 ± 0.25	21.40 ± 54.30	0.01 ± 0.11

$c_\times$
40
100
20

* A fixed-effects model yields consistent results

Take home message:

1) The reputation effect is stronger for newer publications ($c < c_\times$)

$$\rho(c < c_\times) > \rho(c \geq c_\times)$$

2) The citation rate of highly-cited papers is largely independent of the author reputation

$$\pi(c < c_\times) < \pi(c \geq c_\times)$$

$$\rho(c \geq c_\times) \approx 0$$

$$\pi(c \geq c_\times) \approx 1 \quad \text{(linear pref. attachment)}$$

How strong is the citation boosts attributable to author reputation?

TABLE I: Best-fit parameters for individual careers and the average values within disciplinary datasets. The three features of the citation model are parameterized by π, the paper citation effect, $\overline{\tau}$, the life-cycle effect, and ρ, the reputation effect.

	$c(t-1) < c_\times$			$c(t-1) \geq c_\times$		
Name	π_i	$\overline{\tau}_i$	ρ_i	π_i	$\overline{\tau}_i$	ρ_i
GOSSARD, AC	0.34 ± 0.027	4.92 ± 0.261	0.25 ± 0.008	0.80 ± 0.048	4.73 ± 0.184	0.09 ± 0.024
BARABÁSI, AL	0.42 ± 0.036	3.00 ± 0.155	0.29 ± 0.010	1.06 ± 0.016	3.65 ± 0.111	0.01 ± 0.011
Ave. $\pm$ Std. Dev. [A]	0.43 ± 0.14	5.67 ± 2.52	0.22 ± 0.06	0.96 ± 0.19	8.93 ± 4.09	-0.07 ± 0.11
BALTIMORE, D	0.32 ± 0.018	4.64 ± 0.148	0.28 ± 0.006	0.62 ± 0.047	5.92 ± 0.250	0.15 ± 0.026
LAEMMLI, UK	0.54 ± 0.036	5.09 ± 0.297	0.21 ± 0.014	1.09 ± 0.025	6.40 ± 0.255	-0.12 ± 0.019
Ave. $\pm$ Std. Dev. [D]	0.40 ± 0.14	6.64 ± 6.24	0.26 ± 0.05	0.99 ± 0.22	9.55 ± 26.30	-0.06 ± 0.14
SERRE, JP	0.33 ± 0.095	15.90 ± 3.724	0.14 ± 0.026	0.66 ± 0.065	20.50 ± 3.862	-0.03 ± 0.039
WILES, A	0.56 ± 0.208	5.23 ± 1.187	0.24 ± 0.052	0.70 ± 0.059	9.04 ± 0.633	0.10 ± 0.042
Ave. $\pm$ Std. Dev. [E]	0.27 ± 0.17	30.60 ± 56.80	0.14 ± 0.07	0.54 ± 0.25	21.40 ± 54.30	0.01 ± 0.11

math | biology | physics

$c_\times$
40
100
20

The reputation premium: A 66% increase in the citation rate for every 10-fold increase in reputation, C_i

Incentive for Quality > Quantity!
Since ~ 10-15% of an author's C_i comes from his/her highest-cited paper

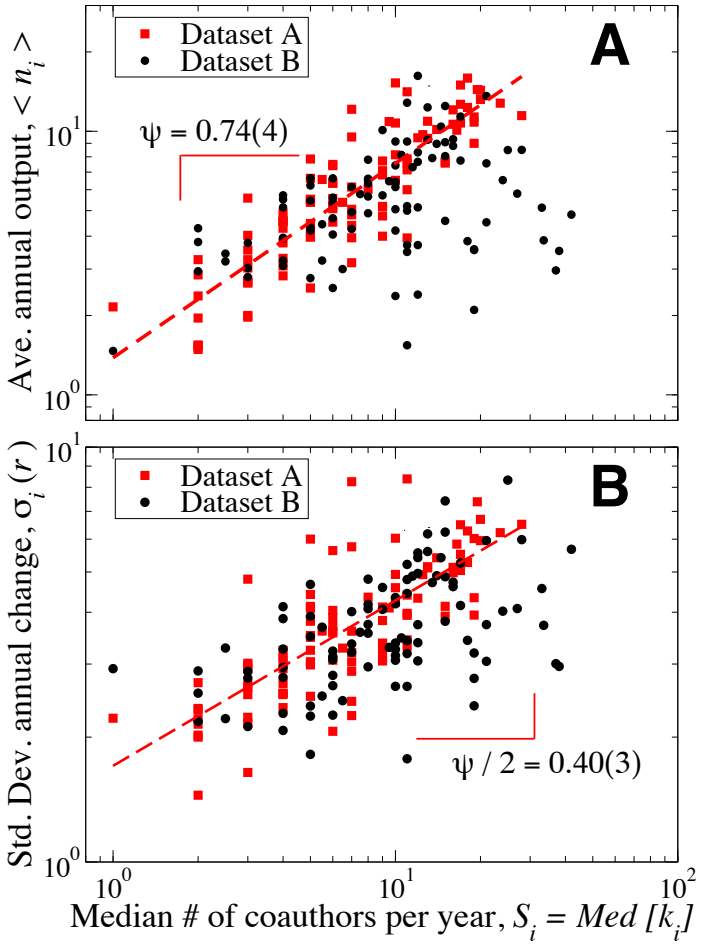
consider 2 scientists, one with 10× as many total citations as the other,
$C_1(t) = 10\, C_2(t)$,
then for 2 new papers,
Ceterus paribus:

$$\frac{\Delta c_{1,p}(t+1)}{\Delta c_{2,p}(t+1)} = 10^{\rho} = 1.66$$

Reputation and Impact in Academic Careers,
A. M. Petersen, S. Fortunato, R. K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. E. Stanley, F. Pammolli, Proc. Nat. Acad. Sci. (2014)

Team Assembly

The challenge of optimizing between redundancy and specialty, incumbents and newcomers



Collaboration radius and team efficiency

Dataset A: Top physicists
Dataset B: random set of prolific physicists

Towards a micro-level production function:

$$\langle n_i \rangle \sim S_i^{\psi}$$

average number of publications per year

S_i is median number of coauthors per year

Output change ("growth fluctuation"),

$$r_i(t) \equiv n_i(t) - n_i(t - \Delta t)$$

std. deviation of publication change

$$\sigma_i(r) \sim S_i^{\psi/2}.$$

output volatility

team efficiency parameter ψ

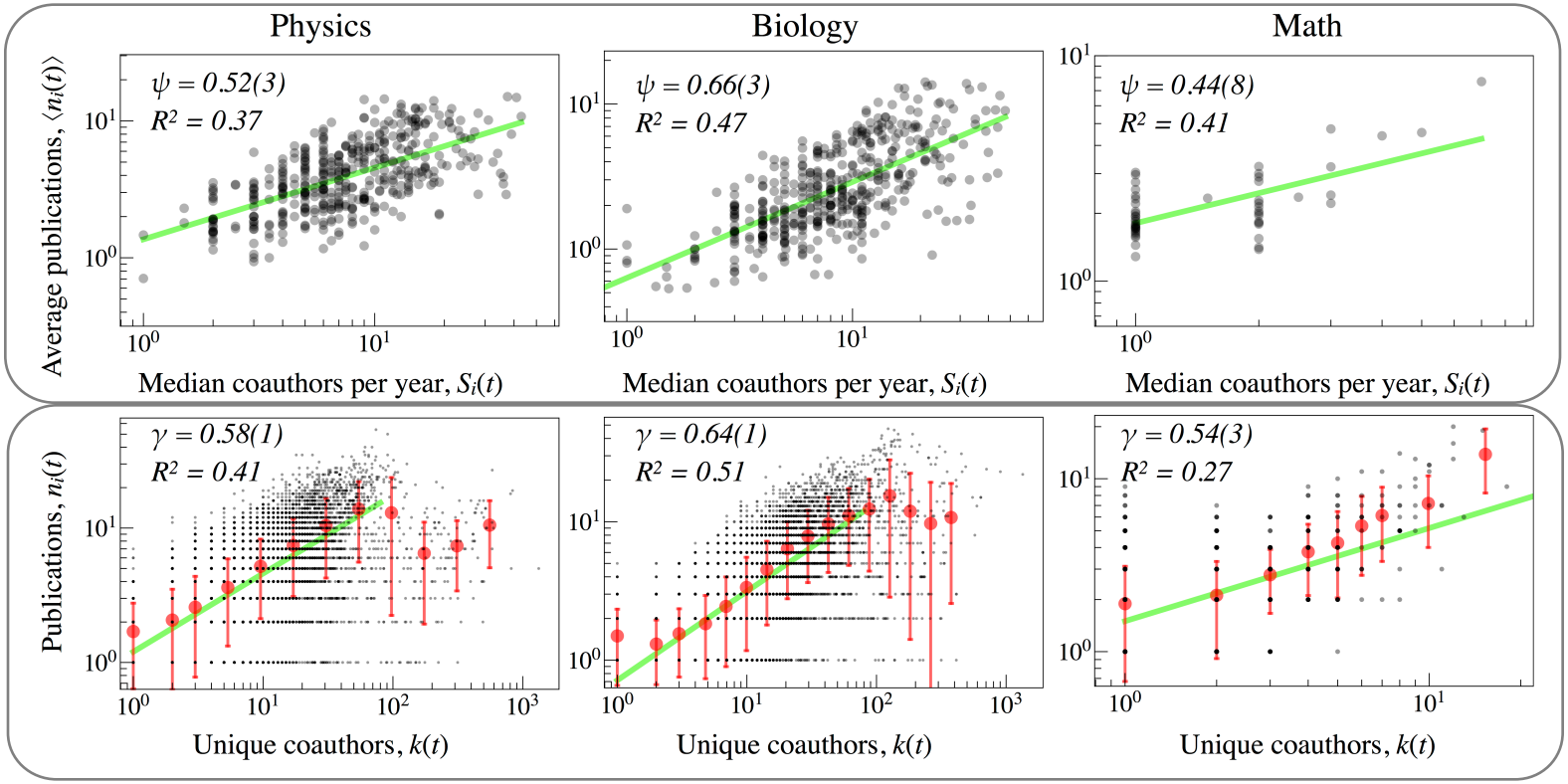
Not surprisingly, there is a decreasing marginal returns with increasing collaboration radius, likely attributable to team management inefficiencies, however inefficiencies aggregate sub-linearally, $\psi < 1$

Persistence and Uncertainty in the Academic Career,
A. M. Petersen, M. Riccaboni, H. E. Stanley, F. Pammolli.
Proc. Natl. Acad. Sci. USA 109, 5213-5218 (2012).

Team (in)efficiency

Q: How does annual productivity depend on the number of "labor inputs" ?

Q: Are there disciplinary variations ?



We measure the input-output relation using two aggregation methods, which both yield sub-linear scaling relations with efficiency parameters $\psi \approx \gamma$ and $\psi, \gamma < 1$

Interestingly, for scientists not in the top cohort we observe smaller ψ and γ values, suggesting that team management skills are an important factor related to success

$\gamma_{\text{Top100 Phyics}} = 0.68(1) > \gamma_{\text{Prolific Physics}} = 0.52(1), \gamma_{\text{AsstProfessor Physics}} = 0.51(2)$

Ego collaboration network: quantifying *dynamic* & *heterogenous* patterns of collaboration within scientific careers

Sir Andre K. Geim

 # publications, N_i (2012) = 217

S_i = 303 coauthors

The average copublication duration $\langle L_i \rangle$ = 2.1 years, $\langle K_i \rangle$ = 3.7 pubs.

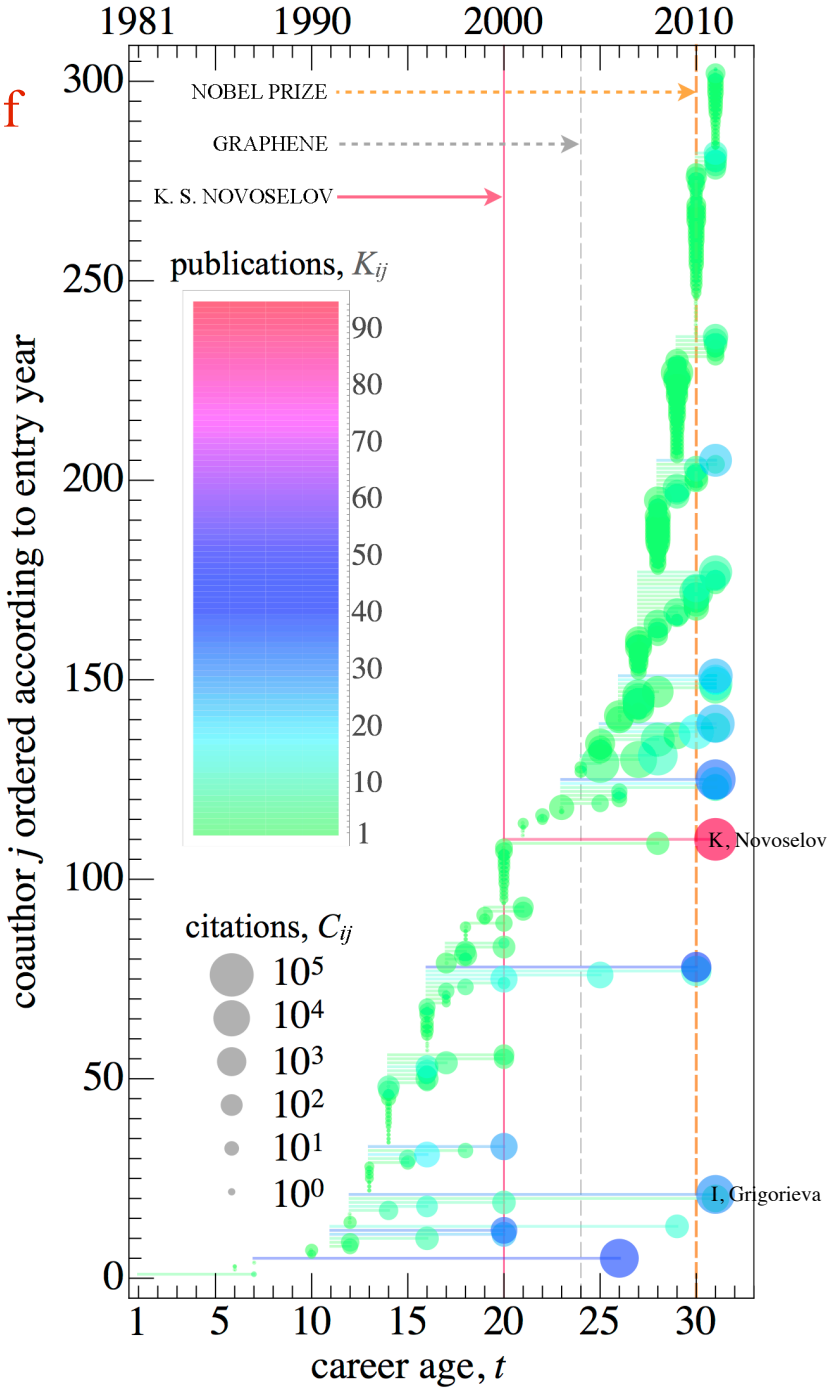
I) Measuring the duration L_{ij} of the tie (time b/w 1st and last copublication)

II) Measuring the intensity K_{ij} of the tie (# of copublications)

III) Measuring the value C_{ij} of the tie (citation impact)

How important are academic "Life partners"?

- Division/Diversity of labor
- Risk/Reward sharing
- Ethics of credit distribution & free-riding



Ego collaboration network: quantifying *dynamic* & *heterogenous* patterns of collaboration within scientific careers

Sir Andre K. Geim

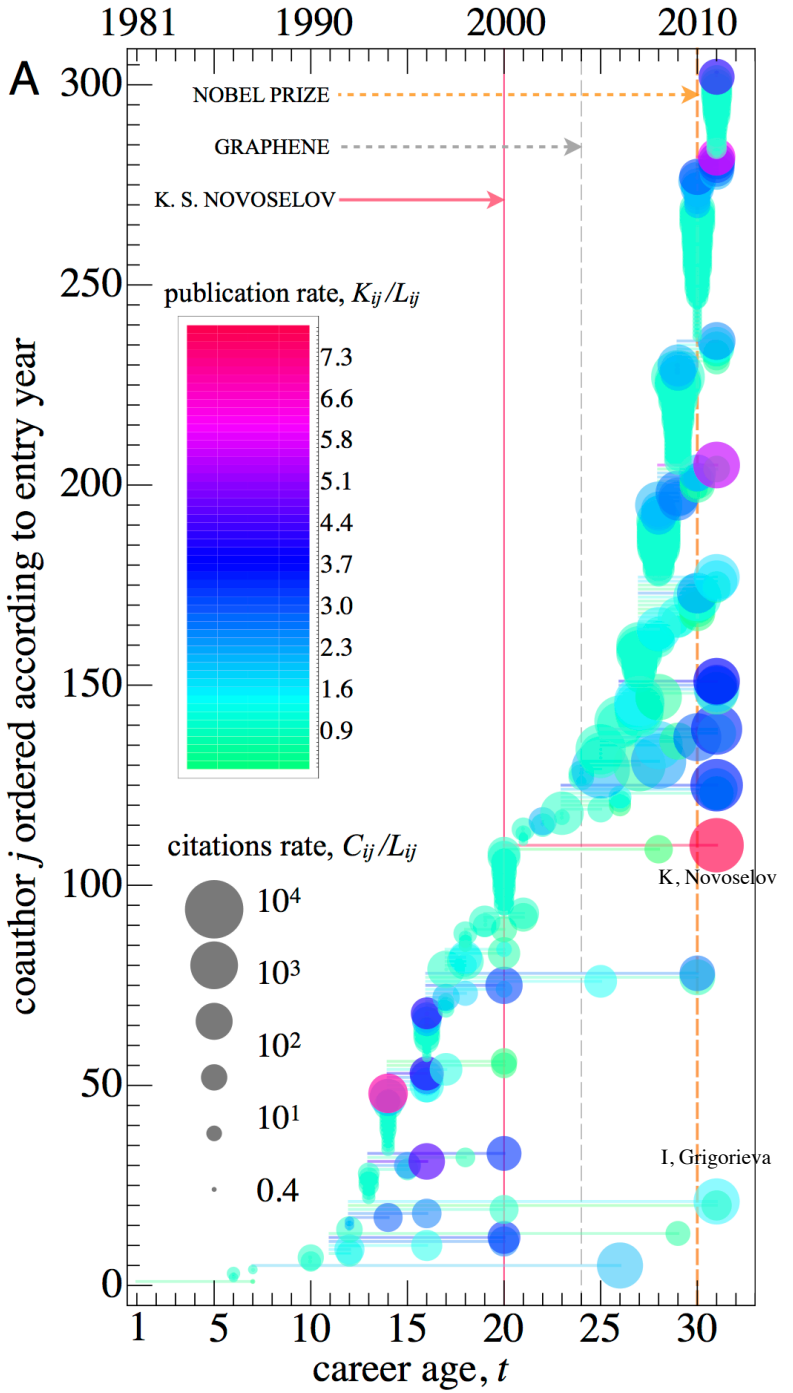
 # publications, N_i (2012) = 217

S_i = 303 coauthors

1) high churning of new entrants (new ideas, new methods, new resources) correlates with higher productivity; however, it represents inefficiencies on the team-formation process and the career trajectory

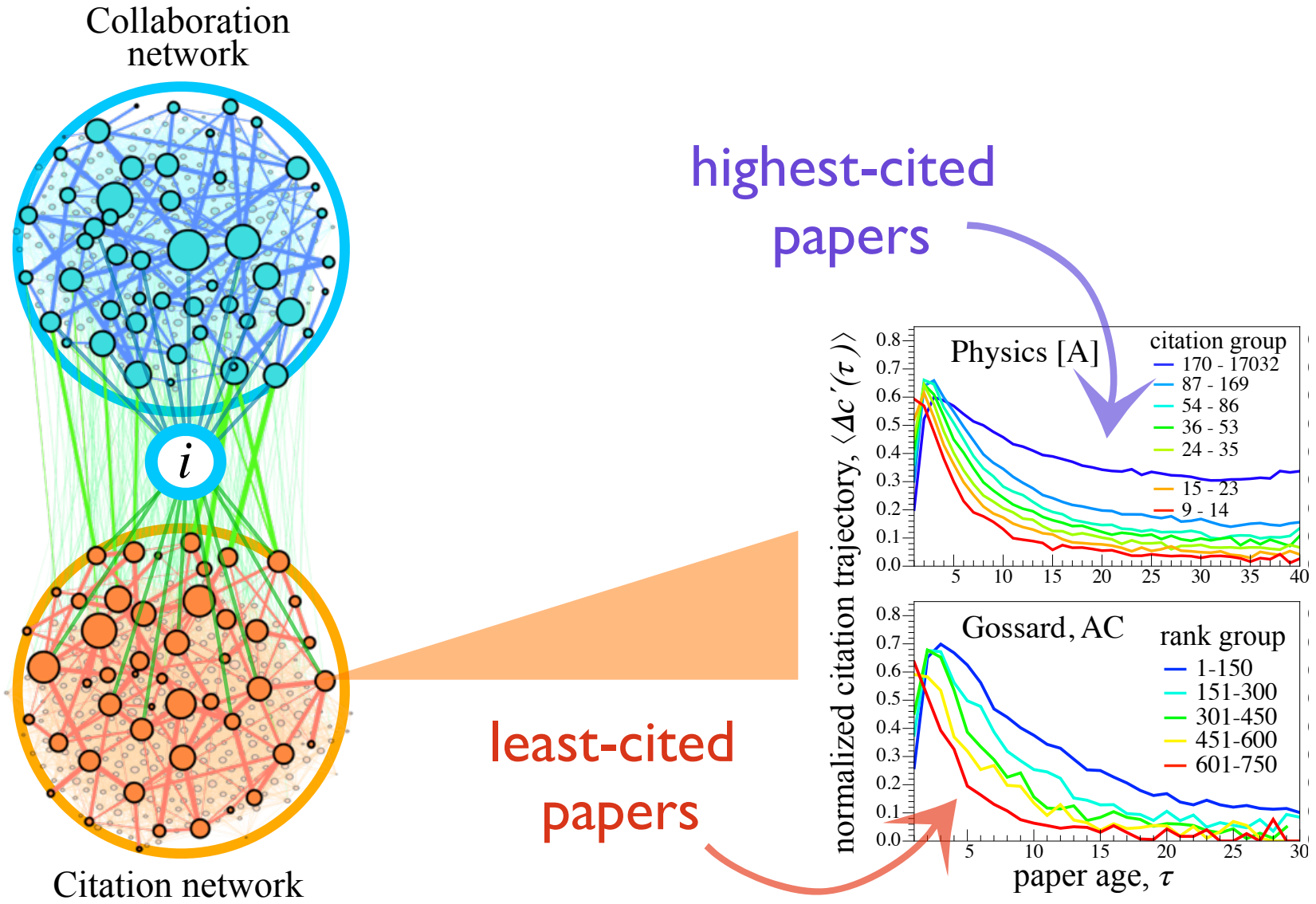
2) The effect of team heterogeneity on productivity is positive indicating the benefits of efficient team management via hierarchy / mentoring

3) Research life-partners — "a scientific marriage": The effect of super ties on productivity is positive indicating the benefits of matching complementary capabilities and beneficial roles. Also points to the profit-sharing of a tit-for-tat publication strategy (free-riding).

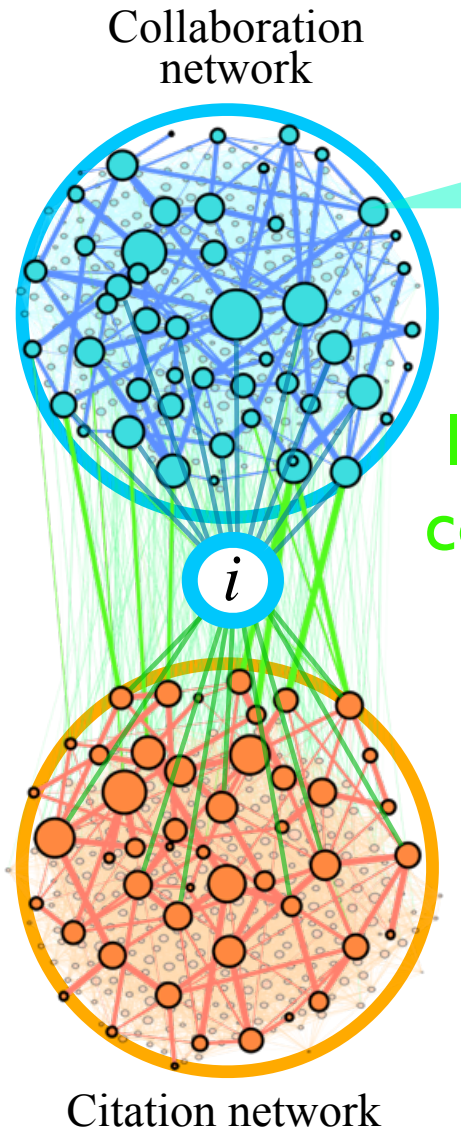


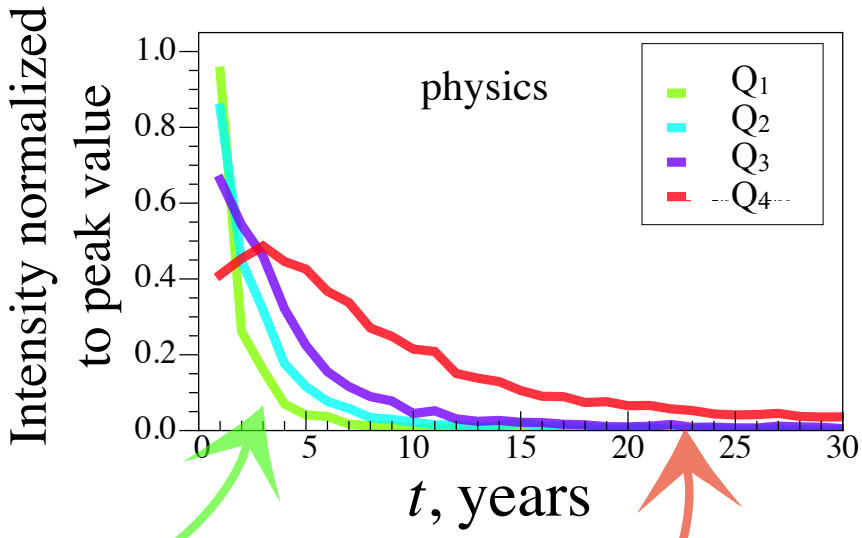
A. M. Petersen **Quantifying the impact of weak, strong, and super ties in scientific careers** Proc. Nat. Acad. Sci. (2015)

Dynamic network characterized by life-cycles



Dynamic network characterized by life-cycles





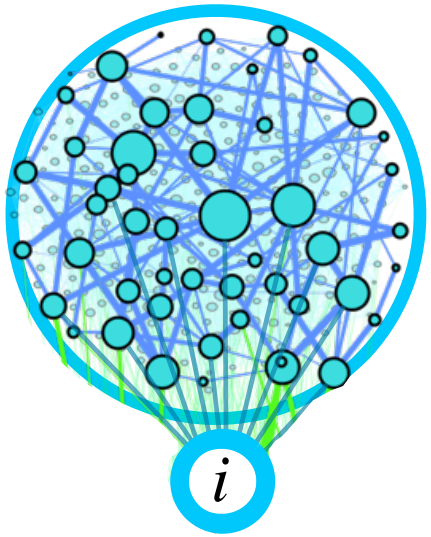
low intensity collaborations

high intensity collaborations

Variable collaboration life-cycles reveal tie-formation dynamics characterized by a complex dichotomy of burstiness and persistence.

An ego-centric perspective of collaboration in science reveals diverse collaboration strategies

Collaboration network



central author *i*

Interactions mediated by social "forces":

- Collaboration (attractive)
- Competition (repulsive)
- Knowledge (an "exchange particle")

Binary-star strategy:

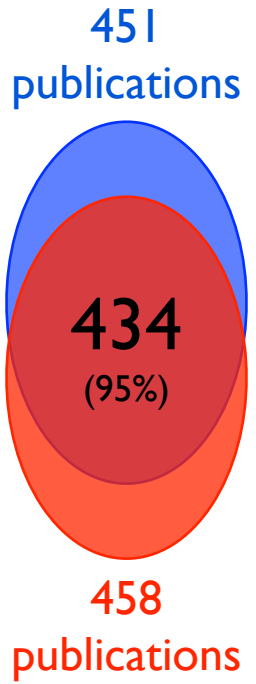
* **Michael Stuart Brown**
* **Joseph L. Goldstein**

Recipients of the 1985 Nobel Prize in Physiology or Medicine for describing the regulation of cholesterol metabolism.

Solo-artist strategy:

* **Marilyn Kozak** (also cell biologist)

$N = 70$, $N_{solo} = 59$



Measures of collaboration intensity

K_{ij} **Individual level**: How strong/weak is the collaboration tie?

$a_{i,p}$, $\overline{a}_{i,t}$ **Team level:** How big is the team?

$G^{K}_{t,i}$ **Group level**: How concentrated are the collaborator tie strengths?

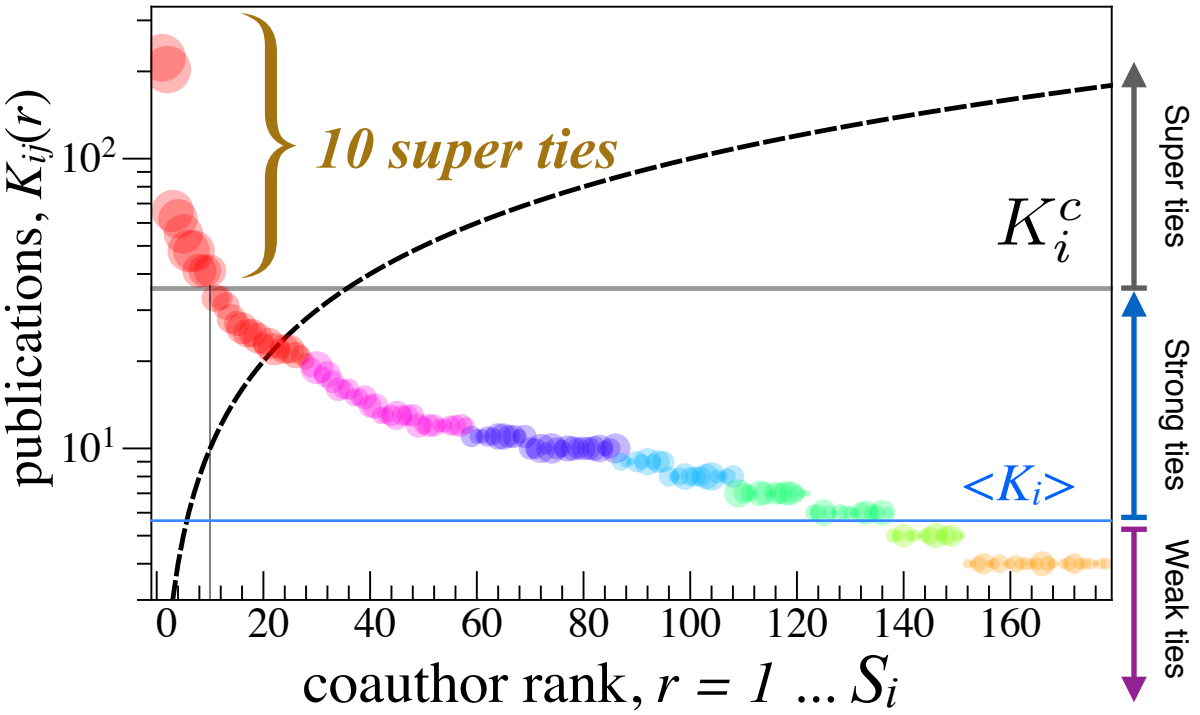
Weak ties, strong ties, and super ties

H. Eugene Stanley

$N_i\,(2010) = 909$ publications

$S_i = 541$ coauthors

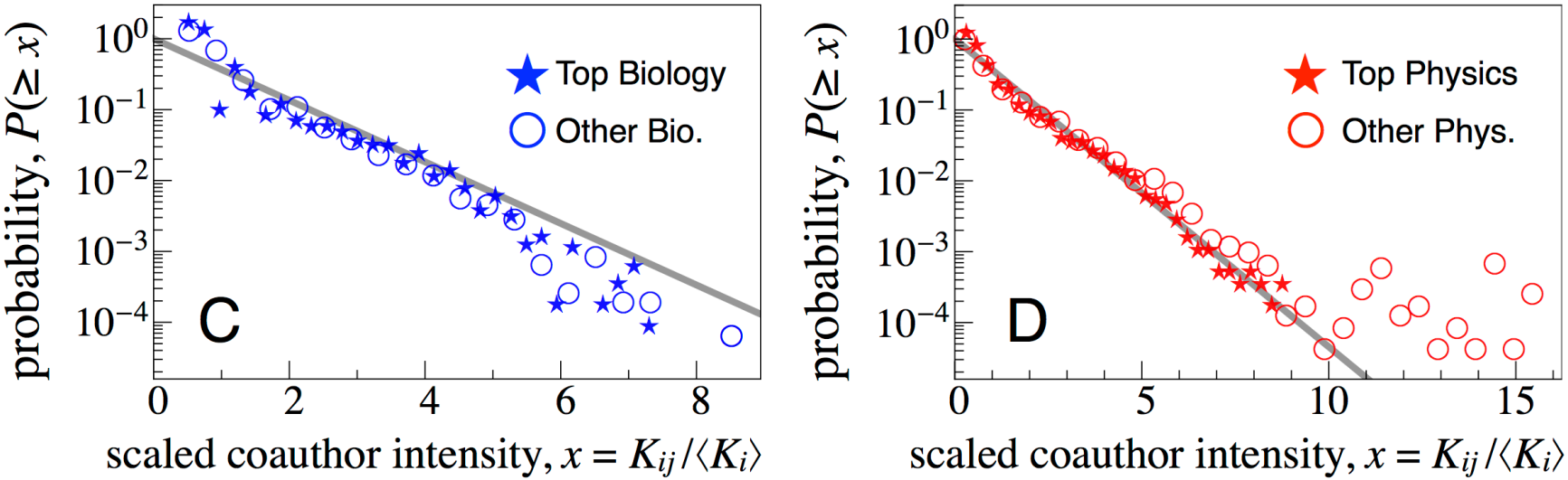
$\langle K_i \rangle = 5.7$ papers



rank		K_{ij}	%
1	HAVLIN, S	223	.25
2	BULDYREV, SV	203	.22
3	AMARAL, LAN	66	.07
4	SCIORTINO, F	62	.06
5	IVANOV, PC	55	.05
6	GOLDBERGER, AL	48	
7	PENG, CK	48	
8	GOPIKRISHNAN, P	41	.04
9	PLEROU, V	41	
10	STARR, FW	41	.03
11	DOKHOLYAN, NV	33	
12	PAUL, G	33	
13	BUNDE, A	31	
14	GIOVAMBATTISTA, N	28	
15	MAKSE, HA	27	
16	CONIGLIO, A	26	.02
17	URBANC, B	25	
18	CRUZ, L	25	
19	SCALA, A	24	
20	LARRALDE, H	23	
21	MANTEGNA, RN	23	
22	POOLE, PH	22	

Q: How to define collaboration super-tie "outliers" ~ i.e. research life partners

Is there a characteristic collaboration intensity scale?



In order to aggregate across careers with varying coauthorship patterns, we use the normalized variable $x = K_{ij}/\langle K_i \rangle$

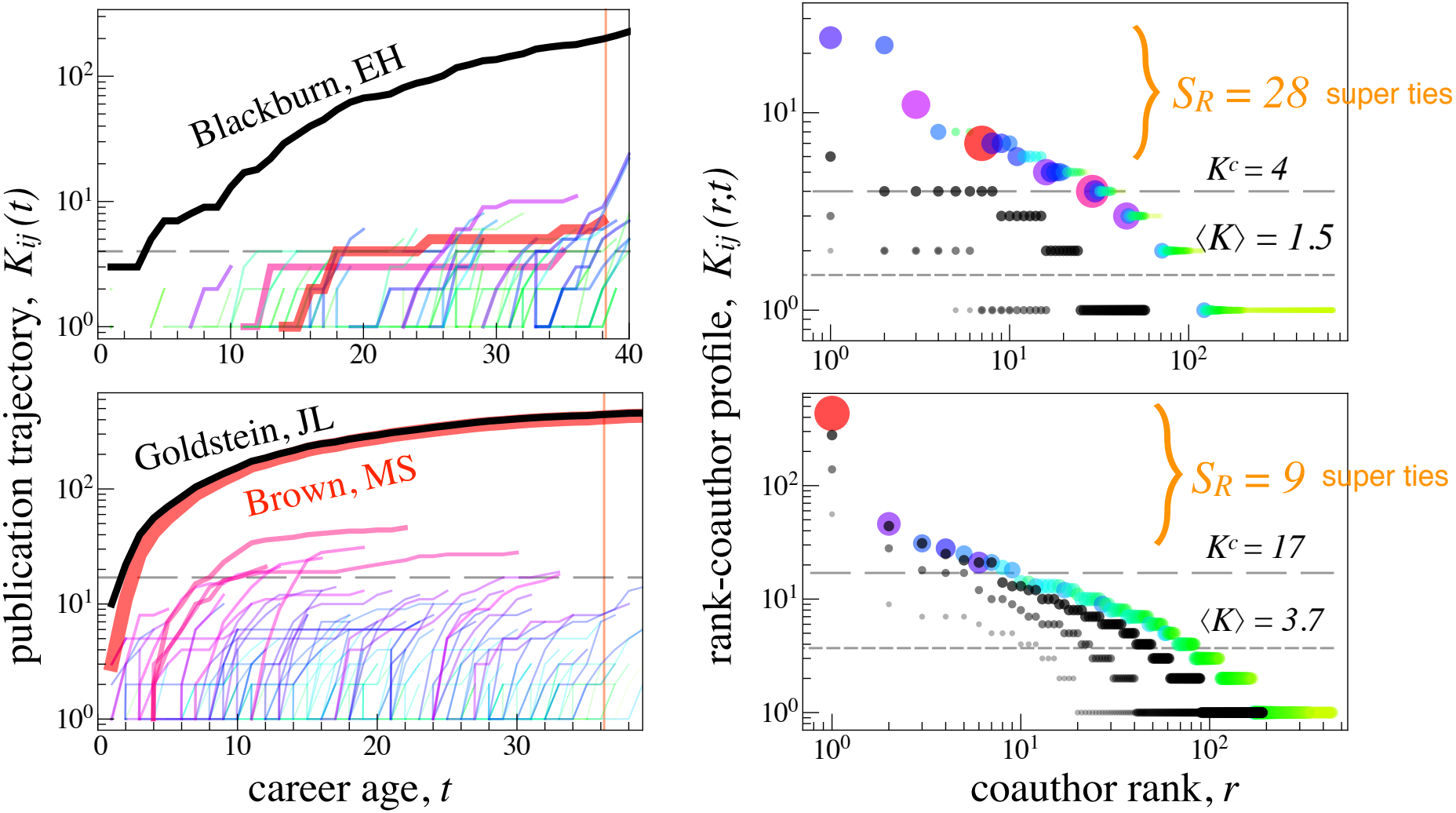
$P(\geq x)$ is well-described by an exponential distribution, for which there is a closed-form solution to the extreme value equation:

$$1/S_i = \sum_{K_{ij} > K_i^c}^{\infty} P(K_{ij}) = \exp(-\kappa_i K_i^c)$$

which has the simple solution

"super tie" threshold $K_i^c = (\langle K_i \rangle - 1)\ \mathrm{Ln}(S_i)$

The (collaboration) path to the Nobel

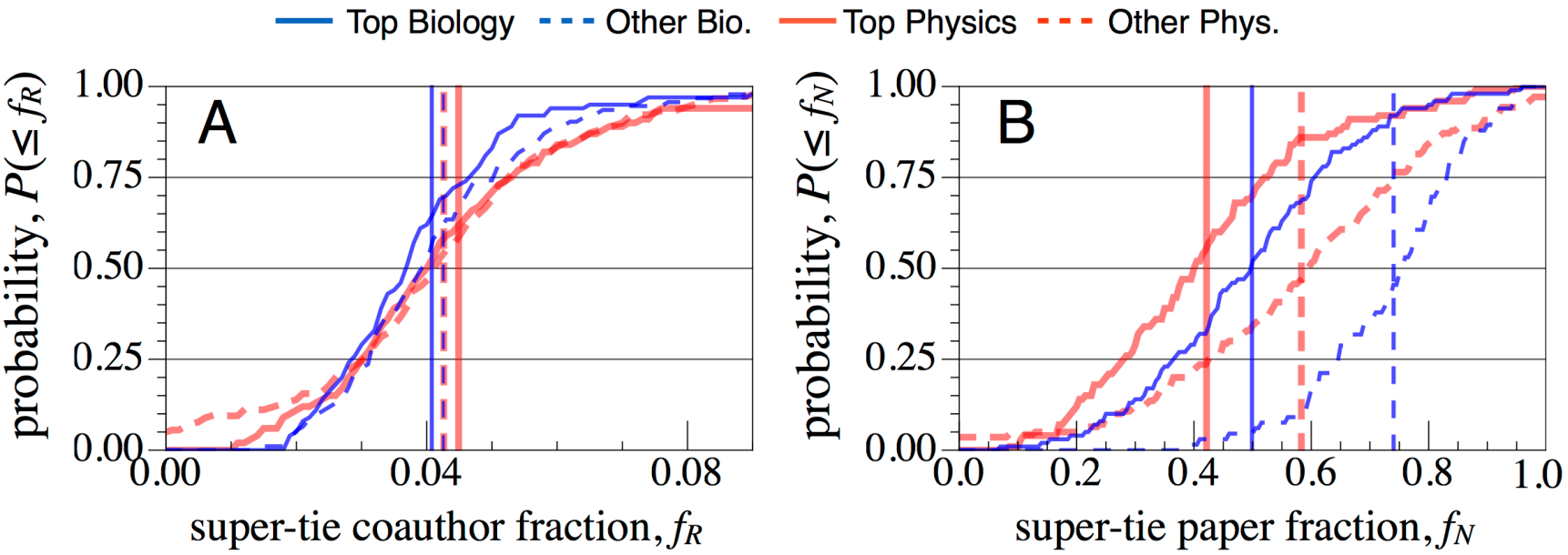


It is quite clear that none of these researchers were anywhere close to being solo-artists

For example, we found that 9% of the biologists and 20% of the physicists shared 50% or more of their papers with their top collaborator.

Q: What is the career impact of the relatively common and indispensable super ties?

Superstars are typically not lone stars - Super ties are rather common



f_R : fraction of coauthors that are super ties: on average 1 in 25

f_N : fraction of publications that include a super tie: on average 50%

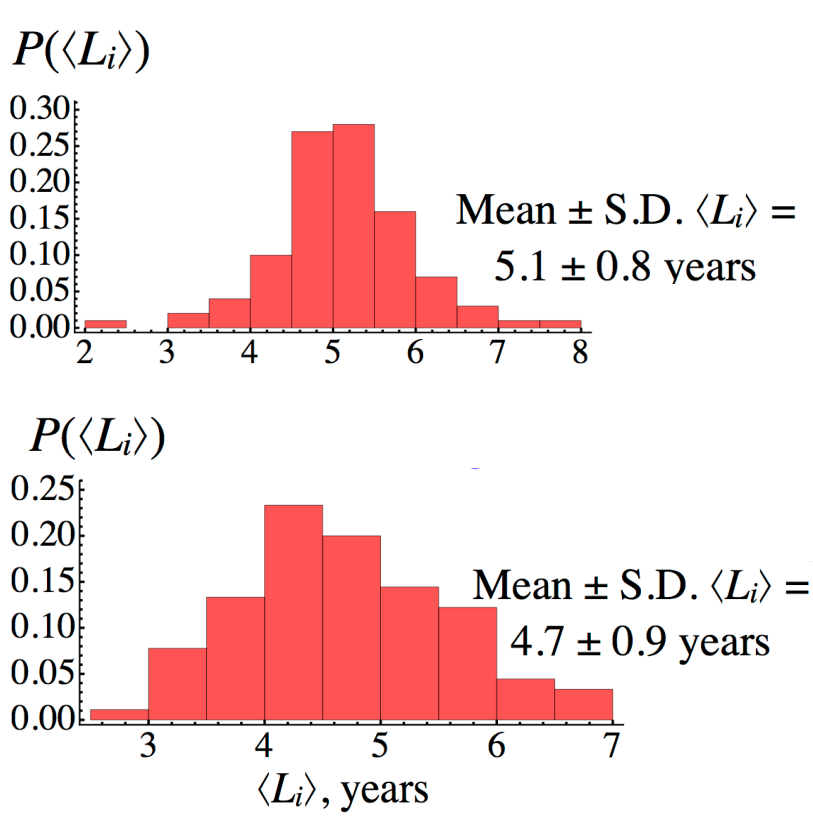
Considerable variation across publication profiles, however the datasets are well-matched with regard to f_R.

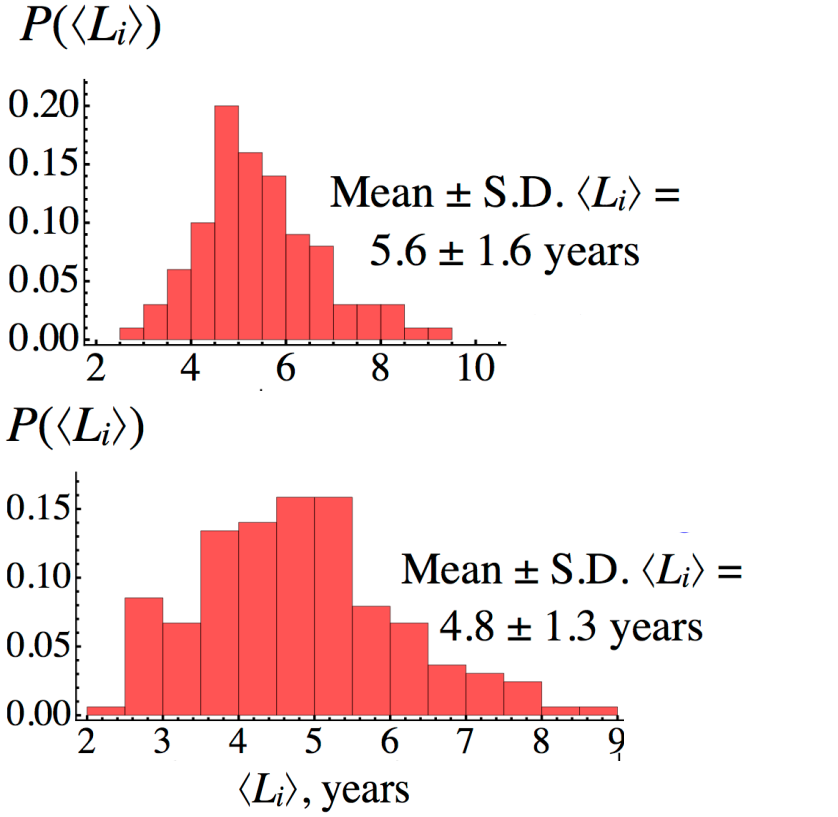
Measures of collaboration duration

$\langle L_i \rangle$ **Individual career level**: What is the characteristic collaboration length?

$\overline{L}_{i,t}$ **Team level**: What is the team's experience together?

Characteristic collaboration duration $\langle L_i \rangle$

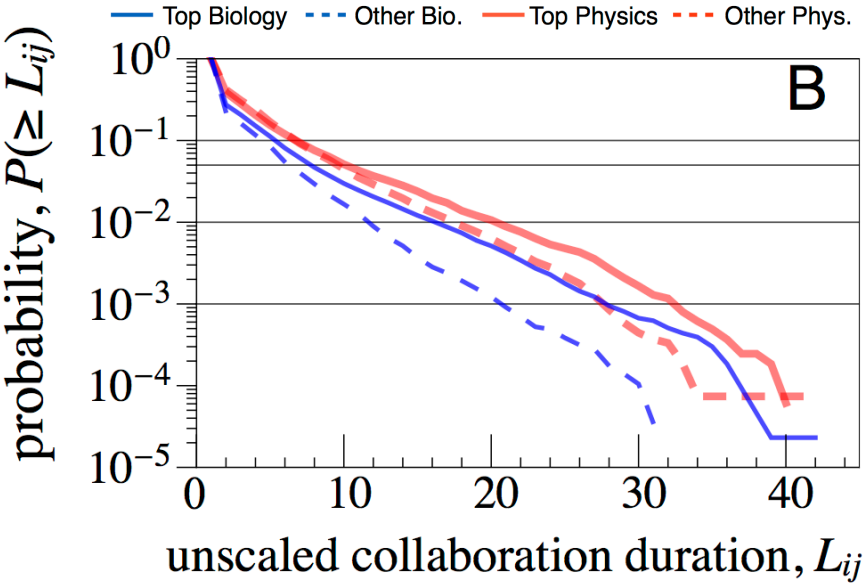


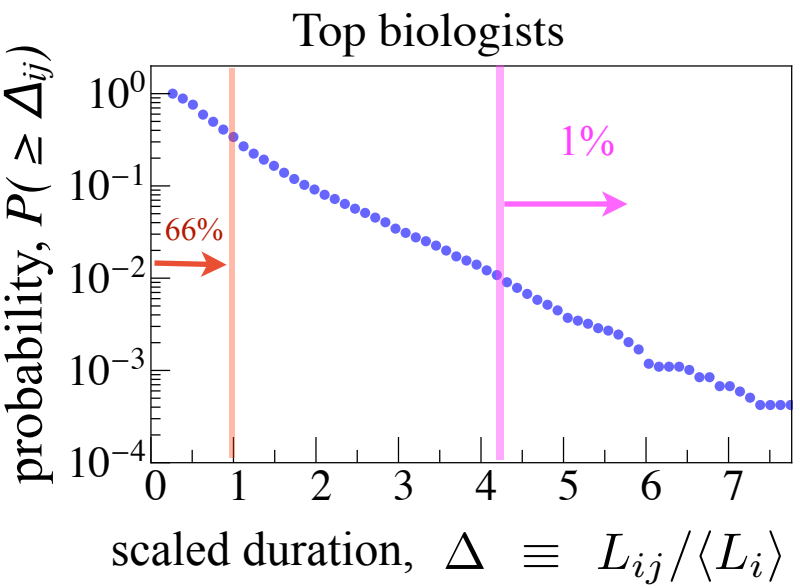


$\langle L_i \rangle$ ranges from 4 to 6 years, consistent with the typical duration of an early career phase position (e.g. graduate school, postdoctoral, assistant professor).

**These averages were calculated after excluding the collaborations with L_{ij} = 1, which account for a remarkable *70 to 80 percent of all collaborations*! Including the L_{ij} = 1 values, the $\langle L_i \rangle$ instead are in the range of 2 to 3 years.

High collaboration turnover rate: a source of inefficiency?



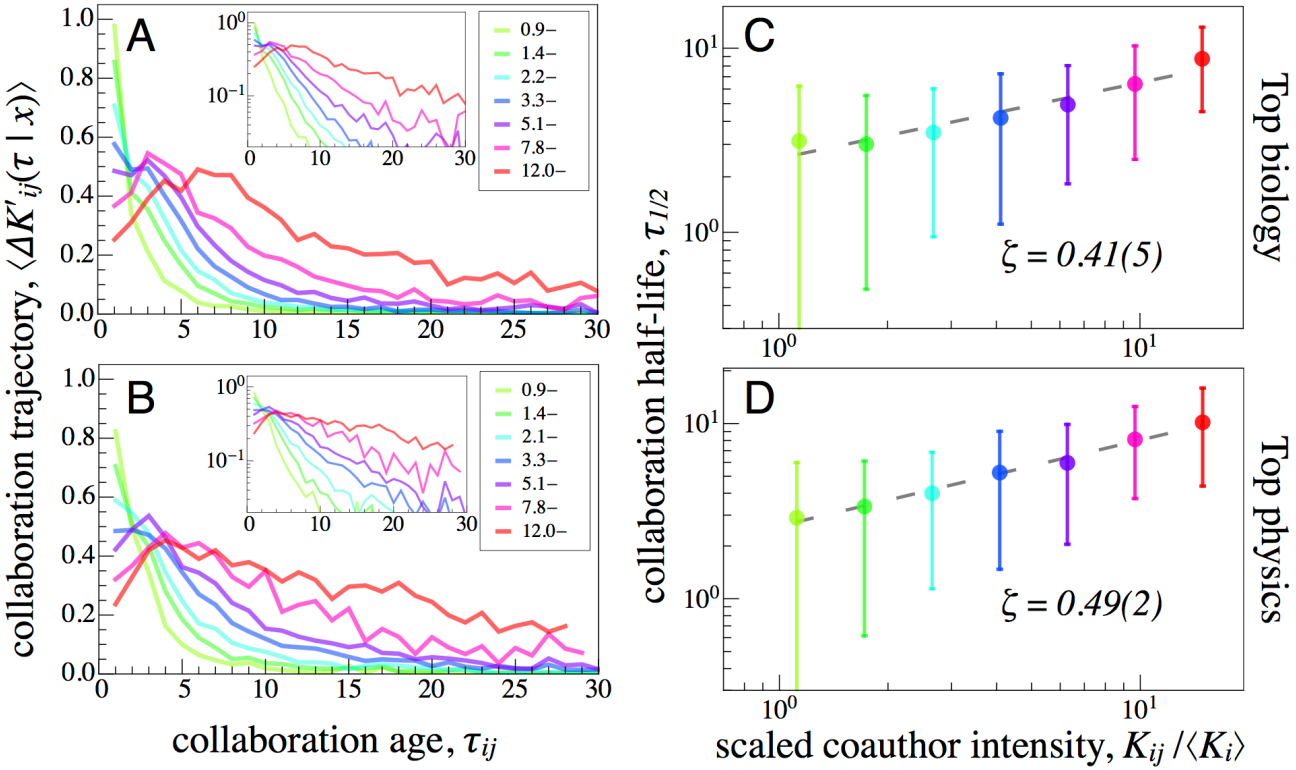


Spurious ties: ~2/3 collaborations have $L_{ij} < \langle L \rangle \sim 5$ years

Lifelong ties: only ~1% last longer than $\sim 4\langle L \rangle \sim 20$ years

1. An egocentric perspective (as apposed to a cross-sectional perspective) reveals that the "invisible college" is held together by weak transient ties
2. Team assembly: collaboration formation/destruction costs are high, both for PI and for transient scientists
3. Credit distribution: Fractional publication/citation counting could reduce incentives to collaborate, thereby reducing the innovative potential of science

Dichotomy of burstiness and persistence in collaboration ties



Growth and decay of collaboration ties. A/B: (Mean) collaboration rate $\Delta K'_{ij}(\tau)$ (# of publications normalized to peak value), measured τ_{ij} years after the initiation of the collaboration. C/D: The ζ value quantifies the scaling of $\langle \tau_{1/2} \rangle$ as a function of the normalized tie strength $x_{ij} \equiv K_{ij}/\langle K_i \rangle$.

- The sub linear ($\zeta < 1$) values indicate that longer collaborations are relatively more productive: increasing marginal returns with increasing collaboration duration ($\tau_{1/2}$).
- These results signal the productivity benefits of long-term collaborations characterized by formalized roles, mutual trust, experience, and group learning that together facilitate efficient interactions.

The 'Apostle' effect

Is there any advantage associated with heavily investing in a select group of research partners?

The advantage of extremely strong social ties characterized by trust, conviction, commitment, experience, skill complementarity, effective collocation and a larger formal and informal social network, moral support, reputation spillovers, risk-profit sharing …

The apostle effect I: annual productivity

Unit of analysis = career period t_i

Dependent variable = $\frac{n_{i,t}}{\langle n_i \rangle}$ = annual productivity normalized to the average over the study period $t \in [6,29]$

All quantities are defined over Δt-year non-overlapping periods. Overall results are robust for Δt =1,3.

$\overline{a}_{p,t,i}$ The average number of authors per publication, a proxy for team management costs and also the technological level of the research

$\overline{L}_{t,i}$ The average collaboration length using the $L_{ij}(t\text{-}1)$ values for only the coauthors with $\Delta K_{ij}(t) > 0$, measuring team experience.

$G^K_{t,i}$ The gini index calculated using the $K_{ij}(t)$ values in the previous period, measuring the tie-strength concentration, ranging from 0 (all team members contributed equally) to 1 (extreme inequality in the team participation)

$$\rho_{t,i} \equiv \frac{\sum_{j|R=1} \Delta K_{ij}(t)}{\sum_{j|R=0} \Delta K_{ij}(t)}$$

The (productivity) apostle effect: The ratio of collaboration inputs from super ties to non-super ties measuring the relative intensity of super tie collaborators within the period.

fixed-effects model

$$\frac{n_{i,t}}{\langle n_i \rangle} = \beta_{i,0} + \beta_{\overline{a}} \ln \overline{a}_{p,i,t} + \beta_{\overline{L}} \overline{L}_{i,t} + \beta_G G^K_{i,t} + \beta_\rho \rho_{i,t} + \beta_t t_i + \epsilon_{i,t}$$

The apostle effect I: annual productivity

Apostle effect I: productivity model ($n_{i,t}$)

Dataset	A	$\ln \overline{a}_t$	$\overline{L}_t$	G_t^K	ρ_t	t	$N_{obs.}$	Adj. R^2
All	466	0.002 ± 0.029	$\mathbf{-0.054 \pm 0.008}$	$\mathbf{1.788 \pm 0.134}$	$\mathbf{0.110 \pm 0.013}$	$\mathbf{0.029 \pm 0.002}$	8483	0.19
(Std. coeff.)		0.002 ± 0.033	$\mathbf{-0.140 \pm 0.021}$	$\mathbf{0.320 \pm 0.024}$	$\mathbf{0.140 \pm 0.016}$	$\mathbf{0.049 \pm 0.004}$		
p-value		0.943	0.000	0.000	0.000	0.000		

1) The extra labor appears to balance out the coordination costs ($\beta_{\overline{a}} \approx 0$)

2) The role of team stagnation (inflexibility : fixation of redundancies/inefficiencies) is negative ($\beta_{\overline{L}} < 0$) indicating that high churning of new entrants (new ideas, new methods, new resources) is positively related with higher productivity

3) The effect of team heterogeneity on productivity is positive ($\beta_G > 0$): indicating the benefits of efficient team management via hierarchy / mentoring

4) The apostle effect ($\beta_\rho > 0$): significant productivity boost due to higher relative contribution by super ties, indicating the benefits of matching complementary capabilities, experience, prosocial rewards of working together, etc.

5) Aging effects ($\beta_t > 0$) : indicating increasing productivity with career stage (due to increasing access to resources: labor, $$, knowledge stock, reputation and other cumulative advantage sources)

The apostle effect II: long-term citation impact

Unit of analysis = publication quality (proxied by long-term citations)
Only papers at least 6 years old (Y-$t_p \geq 6$ years) were analyzed.

Dependent variable = $z_{i,p,Y}$ = the citation impact $c_{i,p,Y}(y)$ of publication *p* normalized to baseline citation levels defined by other papers published in the same year *y*. Y is the census year for the citation counts.

$$z_{i,p,y} \equiv \frac{\ln c_{i,p,Y}(y) - \langle \ln c_Y^m(y) \rangle}{\sigma[\ln c_Y^m(y)]}$$

This measure is approximately normally distributed within pub. age (y) cohorts, and is appropriate for comparing citation impact across time. *m* represents the set of publications from the high-impact journals Nature, PNAS, and Science, aggregated for each year, providing baseline measures which control for the time-dependence of citation counts (Petersen & Penner EPJ Data Science 2014)

The apostle effect II: long-term citation impact

Unit of analysis = publication quality (proxied by long-term citations)
Only papers at least 6 years old (Y-$t_p \geq 6$ years) were analyzed.

Dependent variable = $z_{i,p,Y}$ = the citation impact $c_{i,p,Y}(y)$ of publication *p* normalized to baseline citation levels defined by other papers published in the same year *y*. Y is the census year for the citation counts.

$$z_{i,p,y} \equiv \frac{\ln c_{i,p,Y}(y) - \langle \ln c_Y^m(y) \rangle}{\sigma[\ln c_Y^m(y)]}$$

This measure is approximately normally distributed within pub. age (y) cohorts, and is appropriate for comparing citation impact across time. *m* represents the set of publications from the high-impact journals Nature, PNAS, and Science, aggregated for each year, providing baseline measures which control for the time-dependence of citation counts (Petersen & Penner EPJ Data Science 2014)

$a_{i,p}$ number of coauthors ≈ proxy of coordination costs and tech. level and number of paper-level "apostles"

$N_i(t_p)$ number of papers up to year t_p ≈ prestige measure

$R_{i,p}$ The (citation) apostle effect: a super-tie indicator variable. 1 if at least one of the coauthors is a super tie, and 0 otherwise.

$S_i(t_p)$ number of distinct coauthors up to year t_p ≈ collaboration radius measuring access to new/old team members

t_p publication year measured relative to career age, accounting for aging and cumulative advantage effects, learning and prestige

Fixed-effects model

$$z_{i,p} = \beta_{i,0} + \beta_a \ln a_{i,p} + \beta_R R_{i,p} + \beta_t t_{i,p} + \beta_N \ln N_i(t_p) + \beta_S \ln S_i(t_p) + \epsilon_{i,p}$$

The apostle effect II: long-term citation impact

Apostle effect II: citation model ($z_{i,p}$)

Dataset	A	$\ln a_p$	R_p	t_p	$\ln N_i(t_p)$	$\ln S_i(t_p)$	$N_{obs.}$	Adj. R^2
All	377	$\mathbf{0.263 \pm 0.024}$	$\mathbf{0.202 \pm 0.023}$	$\mathbf{-0.061 \pm 0.004}$	0.062 ± 0.066	0.065 ± 0.072	68589	0.27
(Std. coeff.)		$\mathbf{0.135 \pm 0.012}$	$\mathbf{0.129 \pm 0.015}$	$\mathbf{-0.039 \pm 0.003}$	0.044 ± 0.046	0.050 ± 0.055		
p-value		0.000	0.000	0.000	0.347	0.367		

1) Positive impact of having more disciples ($\beta_a > 0$) : indicating the benefits of having more contributors (towards big endeavor) and also more messengers (of the results)

2) Positive citation boost due to super ties ($\beta_R > 0$) : possibly reflecting the promoting power of super ties via self-citation and reputation growth, but also skill complementarity

3) Aging effects ($\beta_t < 0$) : indicates a decreasing citation impact with increasing career age. A decreasing trend in researchers' normalized impact is possibly due to finite career (staying motivated) and knowledge life-cycles (staying on the crest of the knowledge front) and possibly reflects the role of confirmation bias in the career growth process (Petersen & Penner EPJ Data Science 2014)

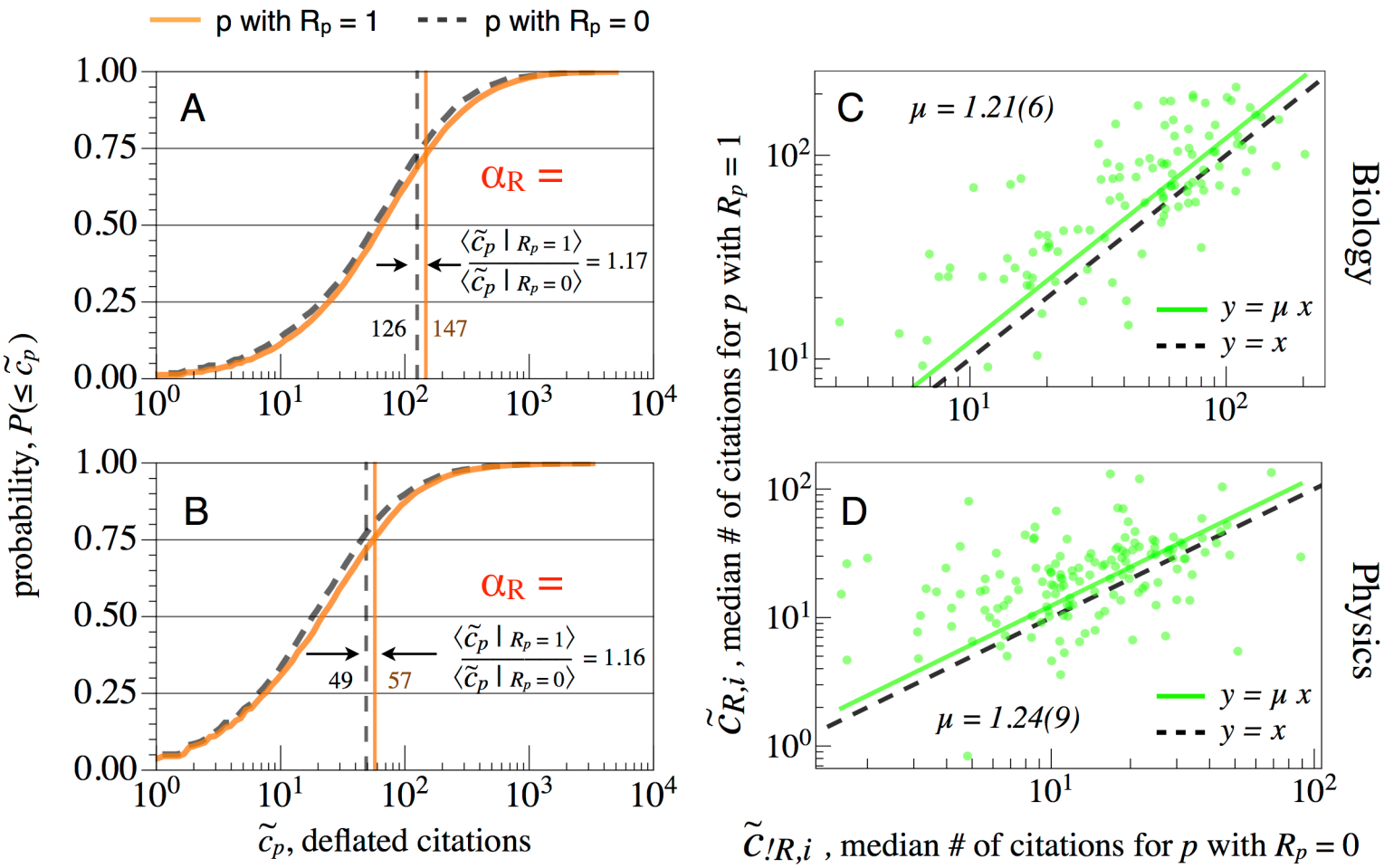
4) Neither prestige nor collaborator radius show a significant effect ($\beta_S, \beta_N = 0$)

So what is the added value measured in terms of citations?

We analyzed all the publications from 1990-2000, splitting them into two groups depending on $R_p = 0,1$

A) Aggregate level

B) Career level



These two methods provide consistent estimates of the citation boost at the publication level, $\mu \approx \alpha_R$, corresponding to a 16%-24% citation increase attributable to super ties:

$$P(\tilde{c}|R_p = 1) \approx P(\alpha_R \tilde{c}|R_p = 0)$$

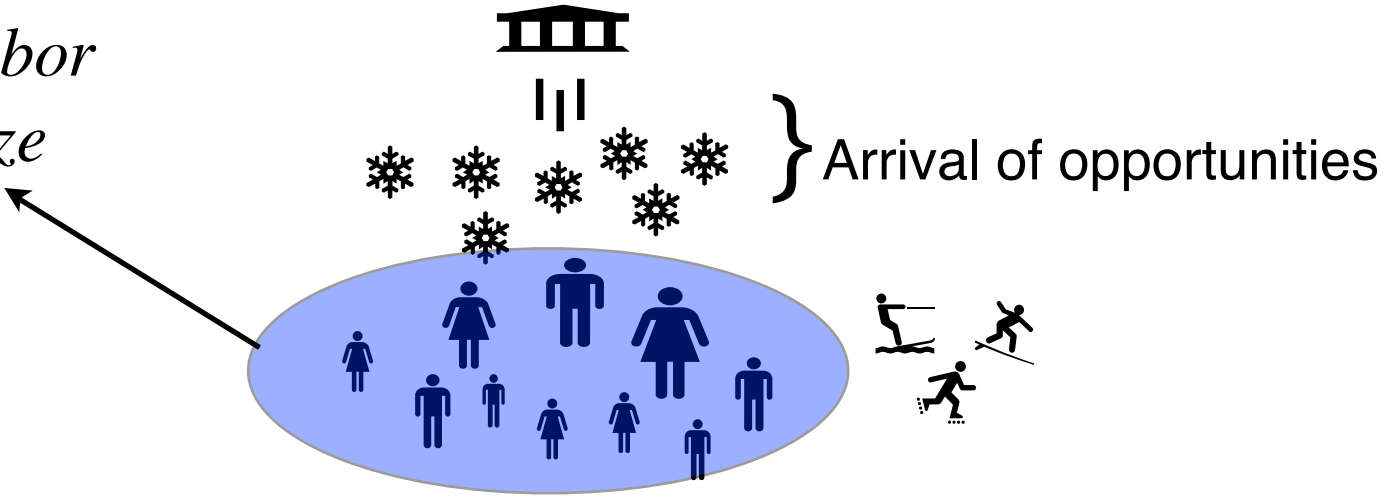
Competition and contract length

How does competition affect career sustainability?

An agent-based competition model with cumulative achievement appraisal (evaluation)

Achievement measured by $n_i(t)$, the number of opportunities (ex. publications) captured in time period t

I = *finite labor force size*



Persistence and Uncertainty in the Academic Career,
A. M. Petersen, M. Riccaboni, H. E. Stanley, F. Pammolli.
Proc. Natl. Acad. Sci. USA 109, 5213-5218 (2012).

Appraising prior achievement

Achievement measured by $n_i(t)$, the number of opportunities captured in time period t

The cohort of I agents compete for a fixed number of opportunities in each period over a lifespan of $t = 1 \ldots T$ periods.

In each period, the capture rate of a given individual i is calculated by an appraisal of the achievement history

capture rate $\propto$ $$w_i(t) \equiv \sum_{\Delta t=1}^{t-1} n_i(t-\Delta t) e^{-c\Delta t}$$

exponential discount factor

Appraisal timescale $1/c$

$c \to 0$: appraisal over all lifetime achievements (~ tenure system)
$c > 1$: appraisal over only recent achievements (short-term contract system)

Crowding out by "kingpins"

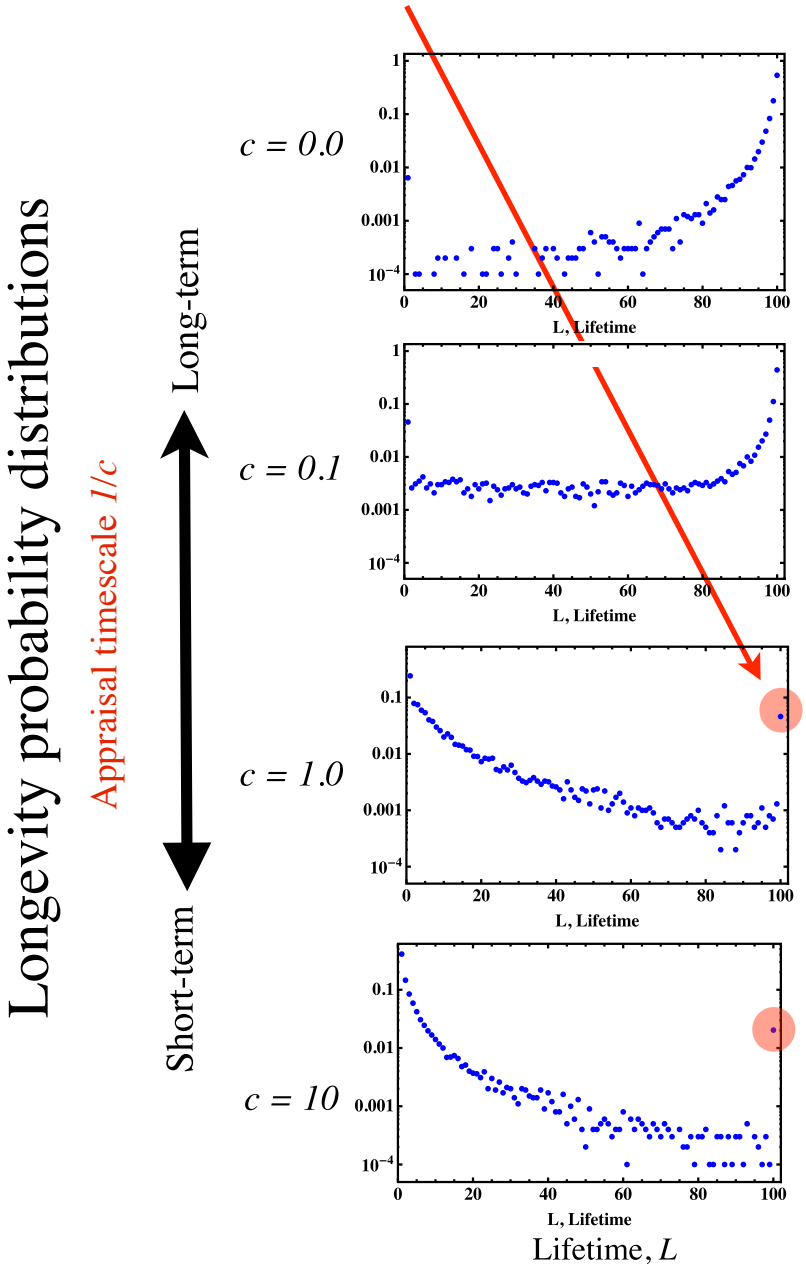
Our theoretical model suggests that

short-term appraisal systems:

* can amplify the effects of competition and uncertainty making careers more vulnerable to early termination, not necessarily due to lack of individual talent and persistence, but because of random negative production shocks.

* effectively discount the cumulative achievements of the individual.

* may reduce the incentives for a young scientist to invest in human and social capital accumulation.



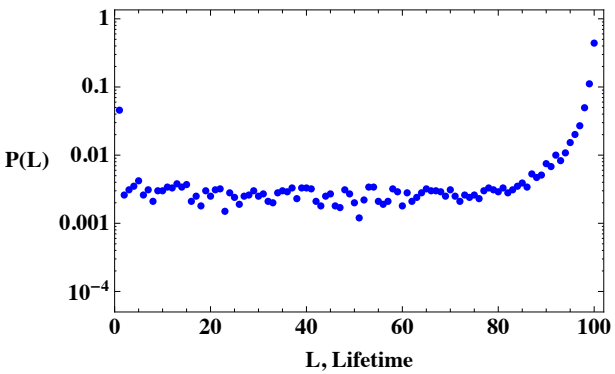
Q: Is there an optimal appraisal (contract) length?

c = 0.1 (~ long term appraisal)

Longevity, L

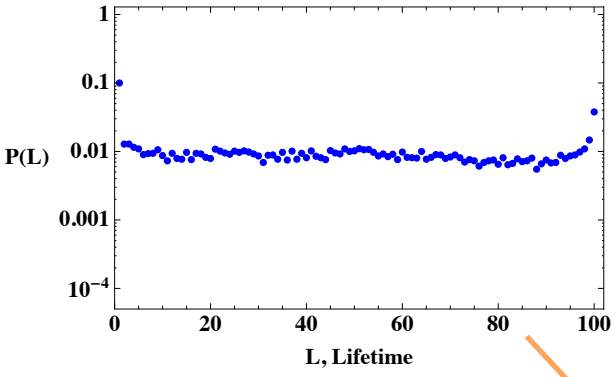
linear capture

$\pi = 1.0$



$\pi = 1.2$

super-linear capture



non-linear preferential capture model

$$\mathcal{P}_i(t) = \frac{w_i(t)^\pi}{\sum_{i=1}^{I} w_i(t)^\pi} \ .$$

Hazard rate $H(L)=-d/dL\,[\ln P(L)]$: conditional probability that failure will occur at time $(L + \delta L)$ given that termination has not yet occurred at time L

$H(L) \approx 0$

hazard rate is not dependent on career position!

Closing thoughts:

Science is a challenging but fascinating system to analyze using data-driven methods: lots of open data, various levels of aggregation, and paradigm shifts offer many natural experiments. It also goes without saying that the science of science is highly relevant: formally — Science of Science policy, and informally — "the hunter becomes the hunted" — informing young scientists on the paths to success in science

Quantitative insights into social/institutional processes underlying science: competition, reputation, team formation, productivity spillovers, career uncertainty, social & economic impact, peer review, etc.

Reputation: Nuances in the interpretation of citation rates. On the one hand, we show there is a significant boost in the early phase of the citation lifecycle due to reputation. Nevertheless, it is not strong enough to explain the phenomena of extremely highly-cited publications, nor should a single publication represent the accomplishments of a career. Reputation offers a strategic incentive to work with established researchers.

Super ties (often career partners): A scientist will encounter many potential collaborators throughout the career. As such, the choice to start or terminate a collaboration can also be an important strategic consideration with long- term implications. The formation of super ties can benefit a career, which can gain a competitive advantage from collective experience, skill complementarity, effective collocation (larger formal and informal social network), moral support, reputation spillovers, cost-risk-profit sharing, etc. We measure the significance of these super ties using an ego-centric perspective which quantifies the added-value of super ties on productivity and citations.

Policy recommendations: One particularly relevant scenario is in career award and tenure evaluations, where it is a common practice to consider "independence from one's thesis advisor" as a selection criteria. We show that in order to assess a researcher's independence, evaluation committees should also take into consideration the level of publication overlap between a researcher and his/her strongest collaborator(s) and the citation impact attributable to working with highly cited scientists due to the reputation effect. Yet at the same time, the beneficial role of super ties should also be acknowledged and supported. For example, funding programs might consider career awards that are specifically multipolar, aimed at life partners (possibly real ones). Policies on credit sharing should make sure to avoid penalizing the incentives to collaborate.

Institutions Markets Technologies
IMT
INSTITUTE FOR ADVANCED STUDIES LUCCA

Thank you!

A special thanks to my collaborators:
Santo Fortunato, Woo-Sung Jung, Kimmo Kaski, Fabio Pammolli, Raj Pan, Ioannis Pavlidis, Orion Penner, Armando Rungi, Massimo Riccaboni, Ionna Semendeferi, Sauro Succi, Gene Stanley, Jae-Sook Yang

Papers available at: *http://physics.bu.edu/~amp17/*

- ***Quantitative and empirical demonstration of the Matthew effect in a study of career longevity***, *A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley. Proc. Natl. Acad. Sci. USA 108, 18-23 (2011).*
- ***Persistence and Uncertainty in the Academic Career***, *A. M. Petersen, M. Riccaboni, H. E. Stanley, F. Pammolli. Proc. Natl. Acad. Sci. USA 109, 5213-5218 (2012).*
- ***Statistical regularities in the rank-citation profile of scientists***, *A. M. Petersen, H. E. Stanley, S. Succi. Scientific Reports 1, 181 (2011).*
- ***Reputation and impact in academic careers***, *A. M. Petersen, S. Fortunato, R. K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. E. Stanley, F. Pammolli. Proc. Nat. Acad. Sci. USA 111, 15316-15321 (2014)*
- ***Together We Stand***, *I. Pavlidis, A. M. Petersen, I. Semendeferi. Nature Physics 10, 700-702 (2014).*
- ***A quantitative perspective on ethics in large team science***, *A. M. Petersen, I. Pavlidis, I. Semendeferi. Sci. & Eng. Ethics 20, 923-945 (2014)*
- ***Inequality and cumulative advantage in science careers: a case study of high-impact journals***. *A. M Petersen, O. Penner. EPJ Data Science 3, 24 (2014).*
- ***Quantifying the impact of weak, strong, and super ties in scientific careers***. *A. M. Petersen. Proc. Nat. Acad. Sci. USA (2015)*

Title: The computational social science of academic career growth

Abstract: Quantitative measures are becoming increasingly prevalent at all scales of scientific evaluation, from countries, to universities, departments, laboratories, and individuals. In the first part of this talk I will share our recent work on the reputation effect in science based upon an analysis of comprehensive career data for several hundred leading scientists from biology, mathematics, and physics. Reputation is an important social construct in science, which enables informed quality assessments of both publications and careers of scientists in the absence of complete systemic information. However, the relation between reputation and career growth of an individual remains poorly understood, despite recent proliferation of quantitative research evaluation methods. I will discuss an original framework for measuring how a publication's citation rate depends on the reputation of its central author. We find that a new publication may gain a significant early advantage corresponding to roughly a 66% increase in the citation rate for each tenfold increase in author reputation. I will conclude with recent evidence on how cumulative advantage underlies trends in waiting times and citation patterns of individual researchers within high-impact "arenas". In the second part of the talk I will discuss new results on the role of tie strength in egocentric collaboration networks. This study is motivated by the fact that a scientist will encounter many potential collaborators throughout the career. As such, the choice to start or terminate a collaboration can be an important strategic consideration with long- term implications. While previous studies have focused primarily on aggregate cross-sectional patterns of collaboration, here we analyze the 'egocentric' patterns of collaboration along individual careers, focusing on tie-formation dynamics characterized by a complex dichotomy of burstiness and persistence. We develop a framework for quantifying collaborative tie strength, revealing a new class of 'super tie', the analog of a research life partner. Accounting for author-specific features, we measure a significant positive impact of super ties on a researcher's productivity and citations – the 'apostle effect' – representing the advantage of extremely tight social ties characterized by trust, conviction, and commitment.