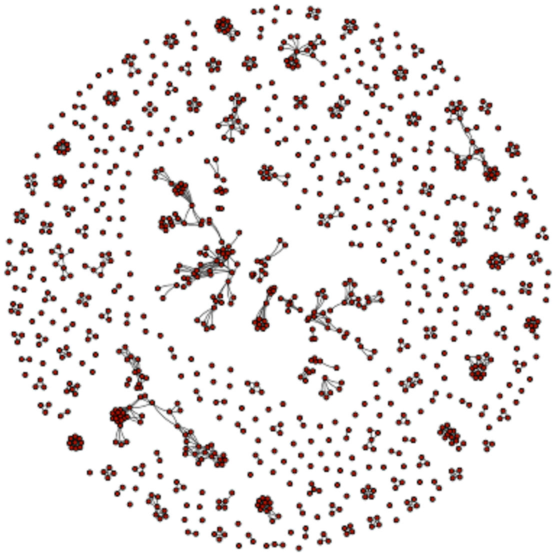
Quantifying growth trends in science careers with applications to bibliometric evaluation

ETH zürich Chair of Systems Design

COST workshop on
"Quantifying scientific impact: networks, measures, insights?"



INSTITUTE FOR ADVANCED STUDIES LUCCA

COST EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Alexander M. Petersen
IMT Institute for Advanced Studies, Lucca Italy

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

Interactions across scales

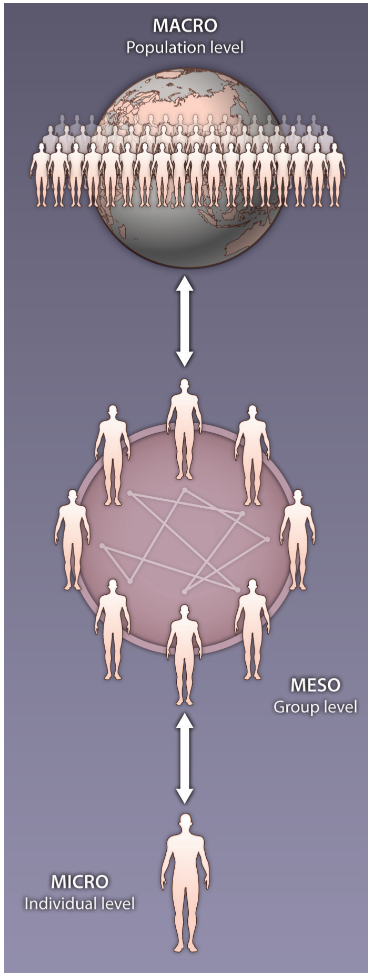
- Science
- Country
- Institution
- Lab / Team
- Individual
- Paper

Institutional factors

Organizational factors

Behavioral factors

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

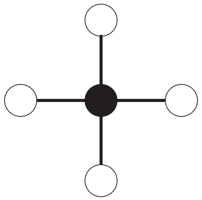


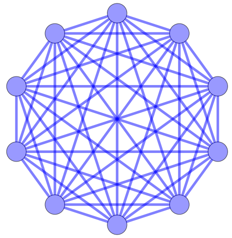
Motivating Questions

- How fast is science changing? How might paradigm shifts in science affect science careers?
- Are there quantifiable patterns of scientific success? Are they useful in the career evaluation process?
- Are the levels of competition in science efficient? Are there ways to improve the sustainability of science careers while at the same time maintaining a high level of competitive selection?
- How do metrics for individual achievement depend on collaboration and time-window factors? How to reduce the multiple-allocation of credit (by fractional citation counts?) without penalizing the incentives to collaborate?

Paradigm shifts

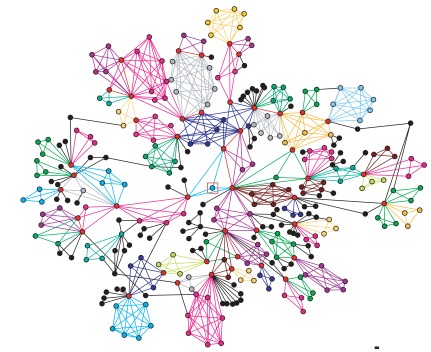
Limited complexity in small knowledge networks





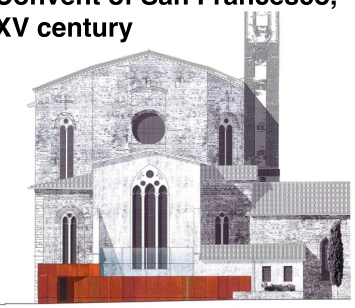
⇒

Emergent complexity in large knowledge networks

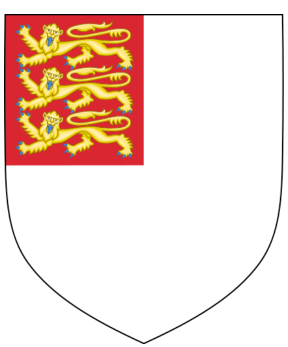


Early scholarly societies, e.g. national societies, scholastic monasteries, noble courts

Convent of San Francesco, XV century



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



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)

growth and increasing *organizational complexity*

⇒

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)

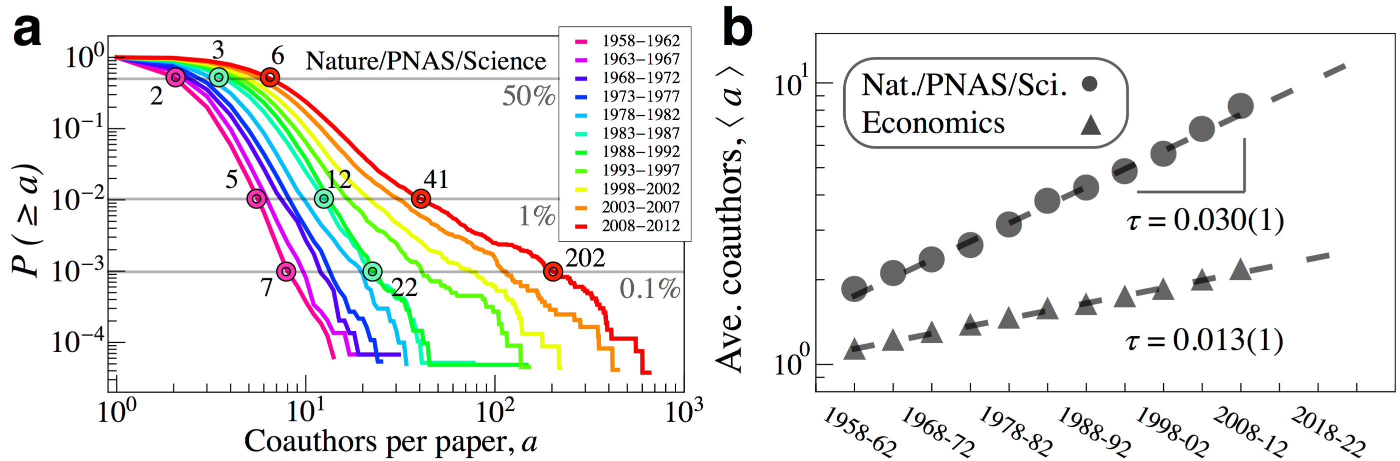
Academic staff
2,100

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)

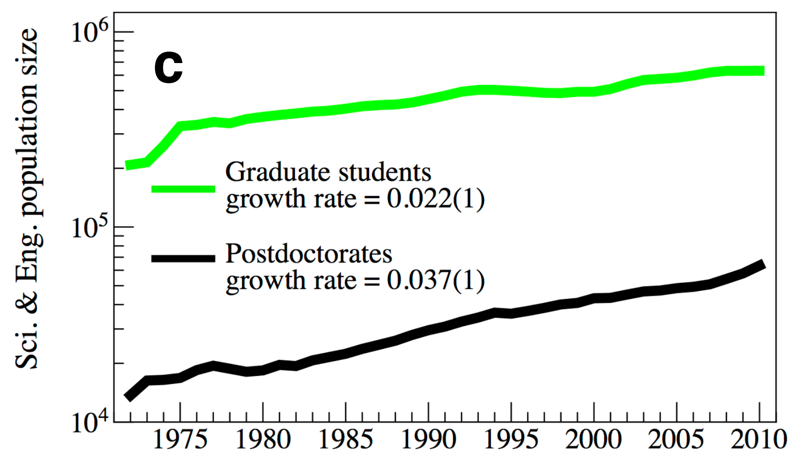
How might paradigm shifts in science affect science careers?

For example: Access to resources/opportunities is becoming increasingly dependent on an individual's 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



Micro (individual careers)

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

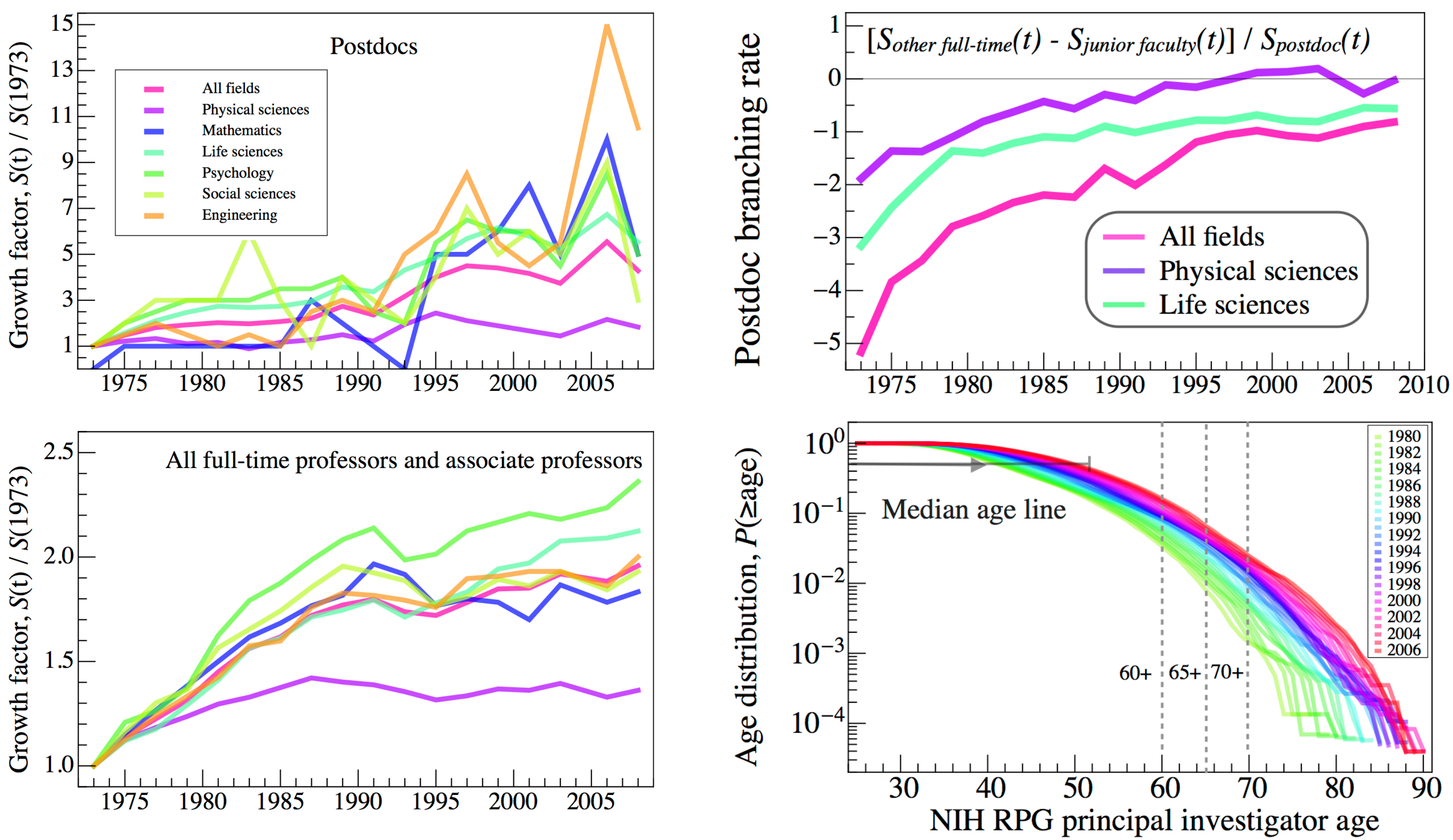
A quantitative perspective on ethics in large team science, *Sci. & Eng. Ethics (2014) A. M. Petersen, I. Pavlidis., I. Semendeferi.*

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

Increased competition in Future Academic Careers

– Bottle-neck in the tenure track model: redirection of PhDs into postdocs and non-tenure track personnel

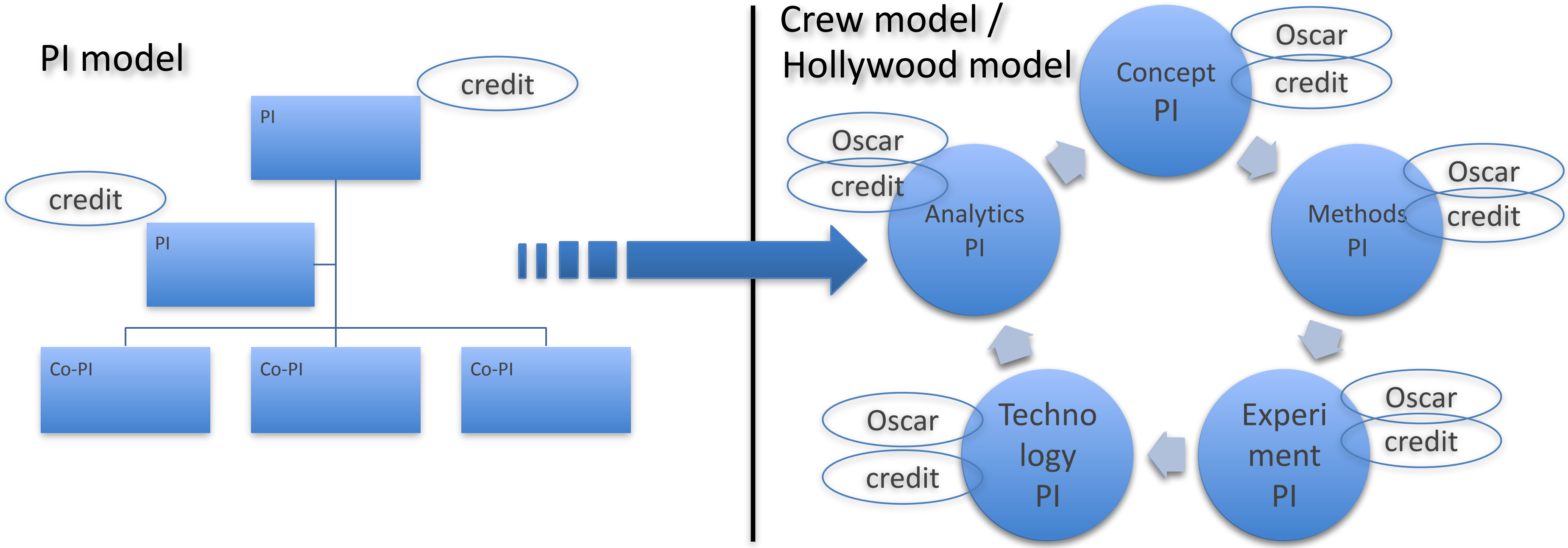
– Demographic shifts: aging, globalization and brain drain



Redesigning the credit system in science

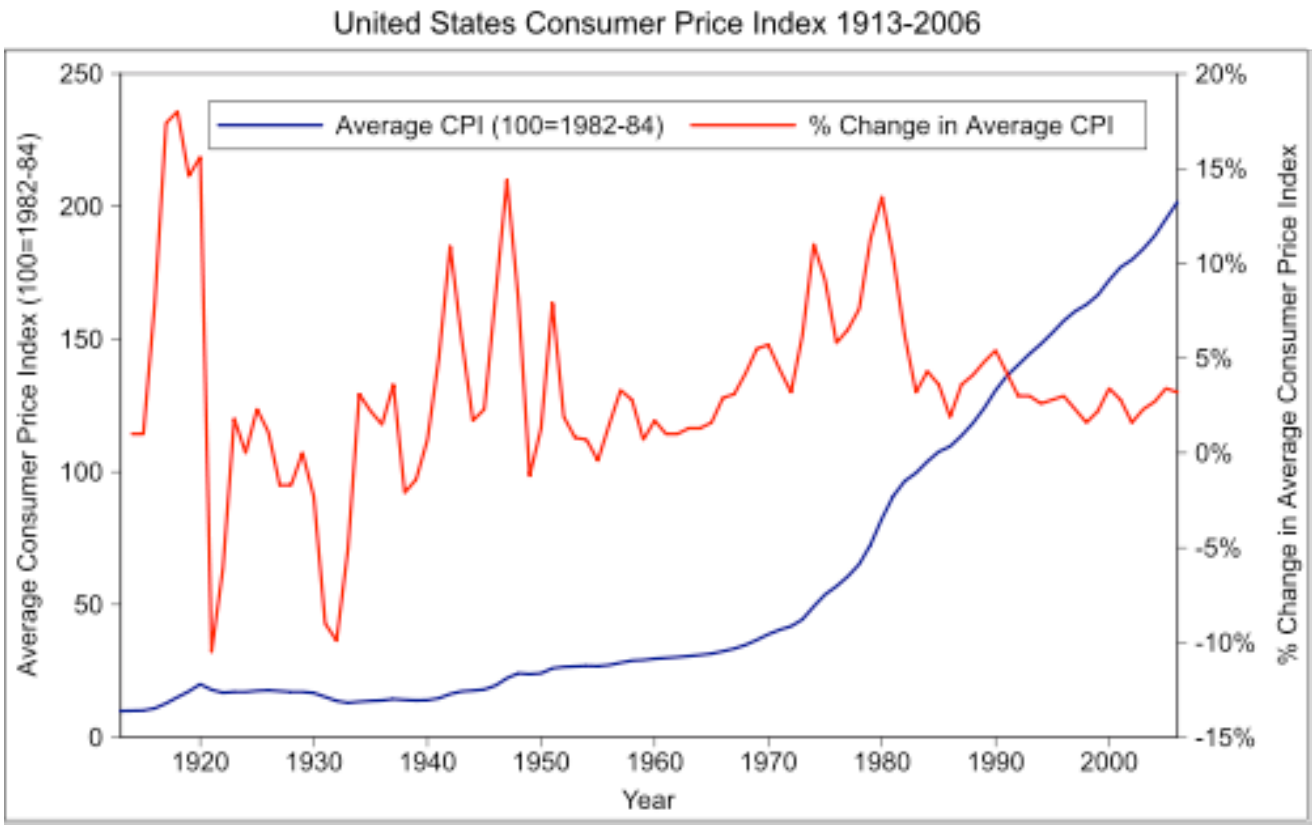
Adoption of career models from communities that embraced a team structure (e.g., filmmaking)

- 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.*

Citation deflator: accounting for the growth of scientific production

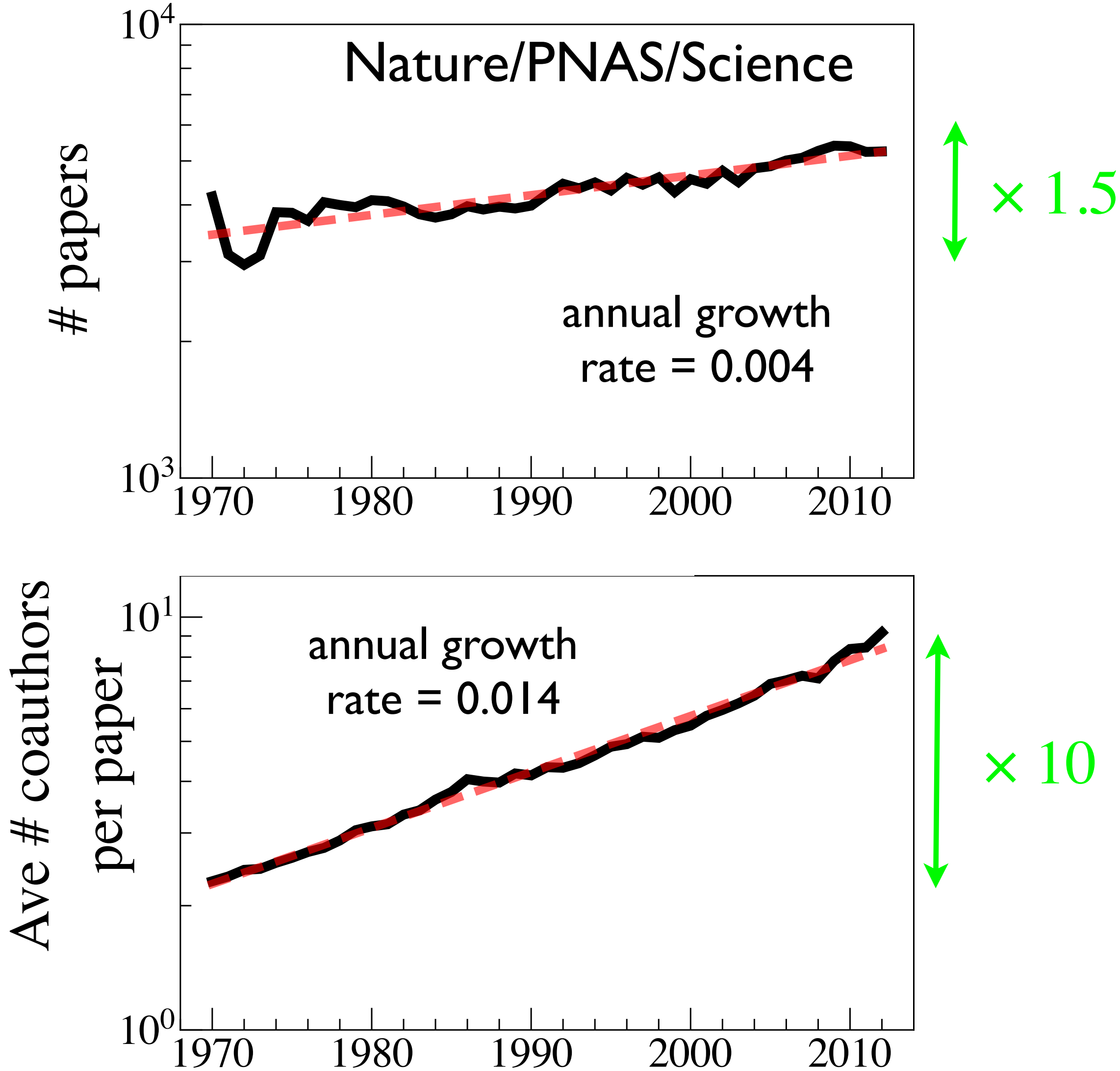


Reputation and impact in academic careers, *A. M. Petersen, S. Fortunato, R. K. Pan, K. Kaski, O. Penner, A. Rungi, M. Riccaboni, H. E. Stanley, F. Pammolli. Proc. Nat. Acad. Sci. USA 111, 15316-15321 (2014).*

Methods for detrending success metrics to account for inflationary and deflationary factors. *A. M. Petersen, O. Penner, H. E. Stanley. Eur. Phys. J. B 79, 67-78 (2011).*

Scientific output inflation

what is the relative impact/visibility of a publication today -vs- Y years ago?

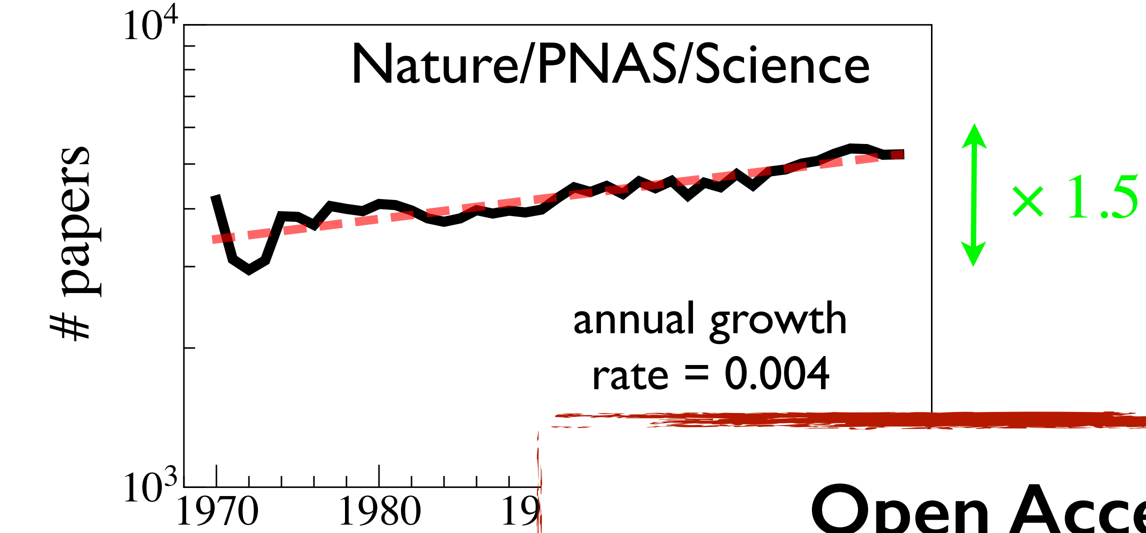


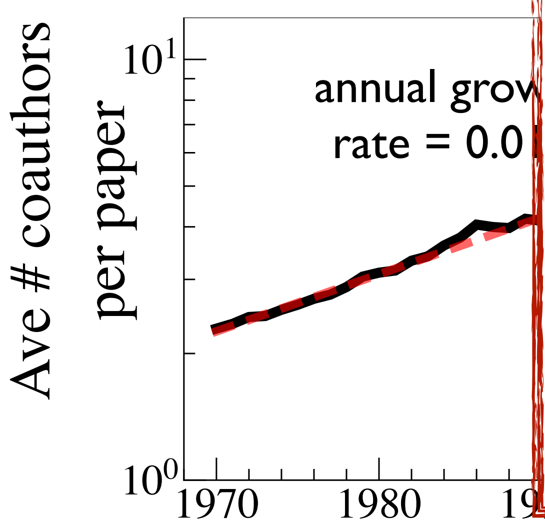
Scientific output increase due to technological factors, population growth, and "output inflation"

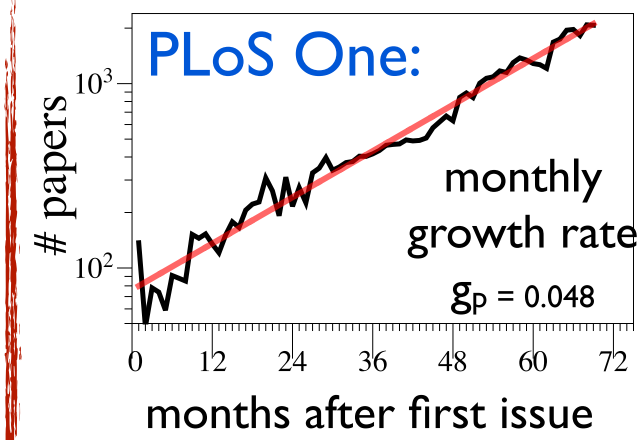
growth of team science

Scientific output inflation

what is the relative impact/visibility of a publication today -vs- Y years ago?







Open Access Journals

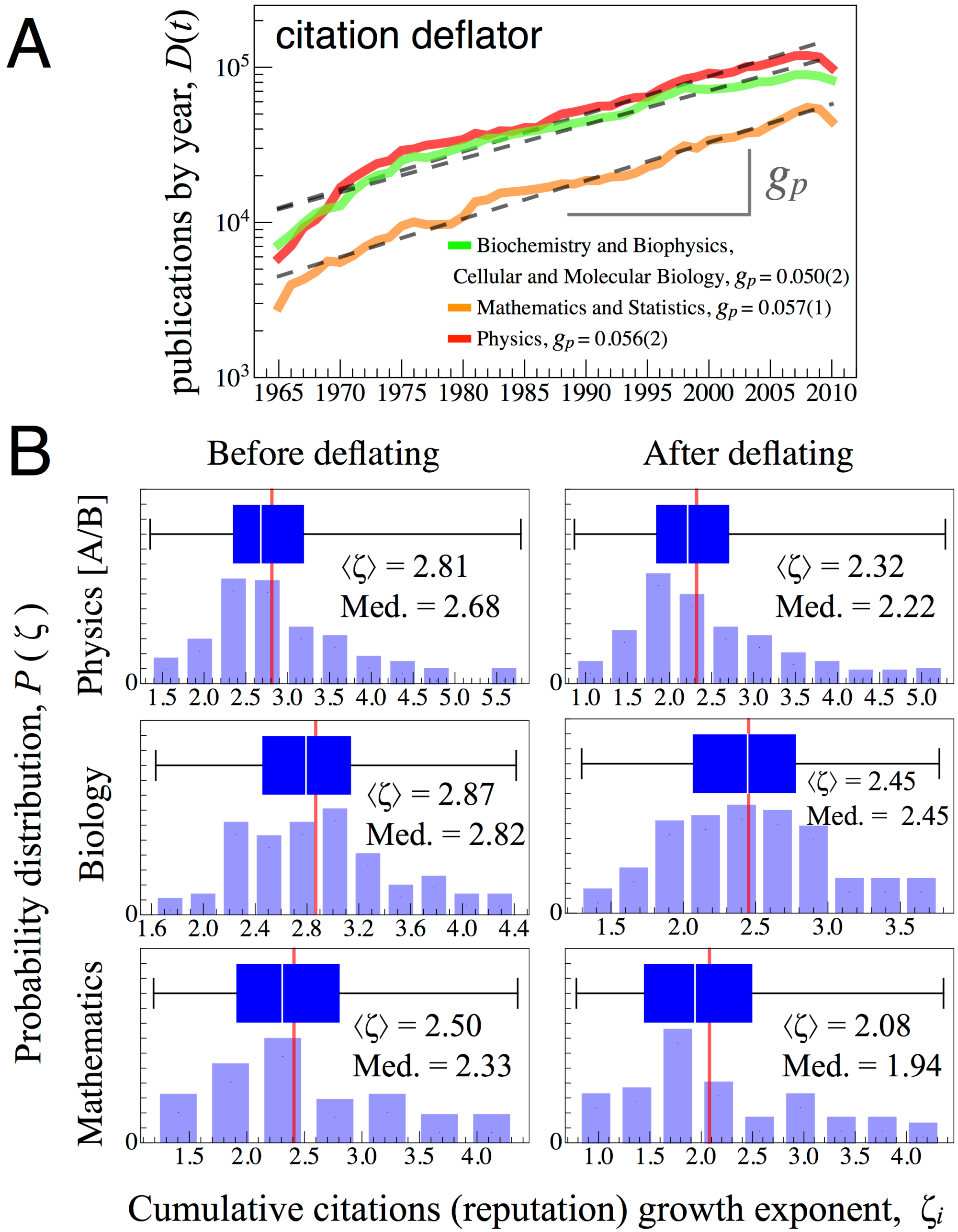
PLoS One:

~ 6,700 articles in 2010 and ~ 14,000 in 2011

⇒ × 2 growth in one year alone!

... who is reading/refereeing all these papers??

How much of career growth (ζ) can be explained by scientific inflation?



* the number of publications $D(t)$ within each discipline we analyzed is growing exponentially, roughly at a 5.5% per year (13-year doubling)

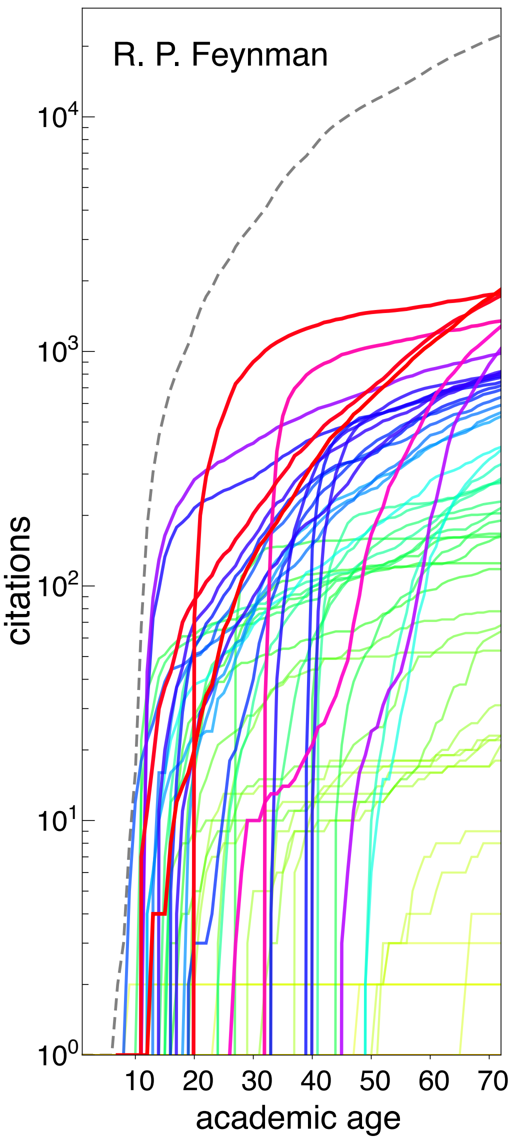
* Each new paper can cite another paper just once
  $\Rightarrow$ $D(t)$ a "deflator index"

$$\Delta c^D_{i,p}(t) \equiv \Delta c_{i,p}(t)/D(t)$$

$$\Delta C^D_i(t) \equiv \Delta C_i(t)/D(t)$$

ζ captures the significant reputation growth across the career, even when discounting for background inflation of scientific production

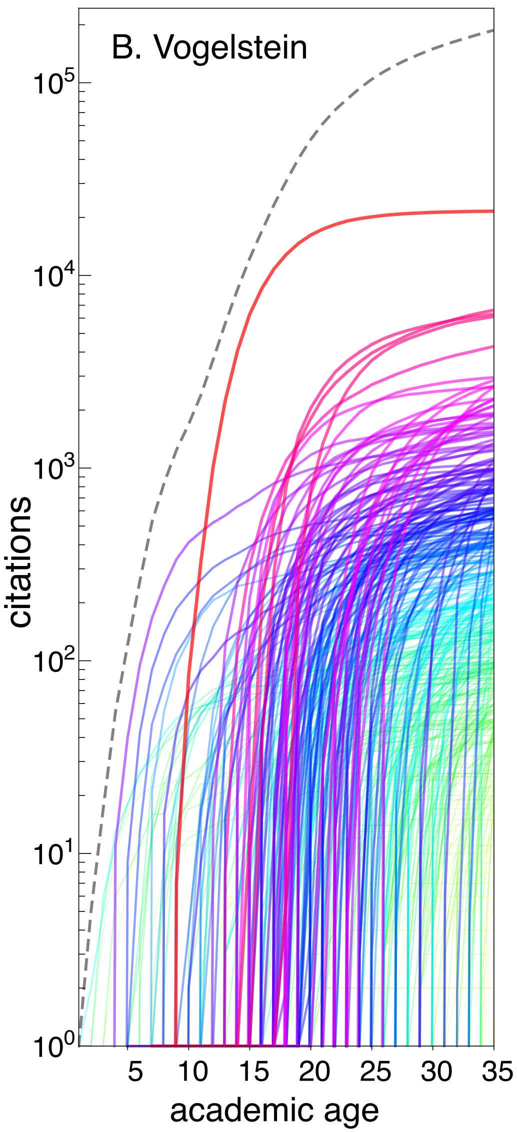
Patterns of growth in science careers



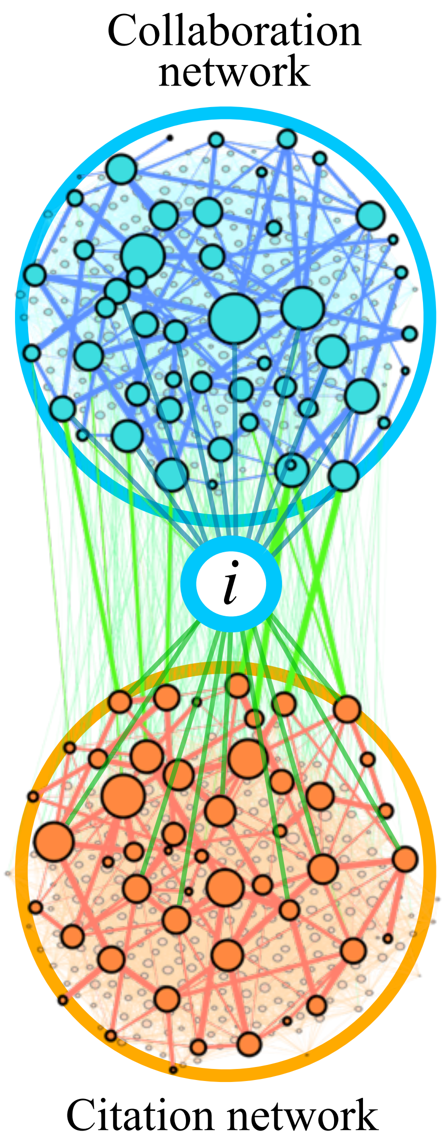
Statistical regularities in the rank-citation profile of scientists, *A. M. Petersen, H. E. Stanley, S. Succi. Scientific Reports 1, 181 (2011).*

The Z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile, *A. M. Petersen, S. Succi J. Informetrics 7, 823-832 (2013).*

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).*



Science careers embedded in a co-evolving network of networks



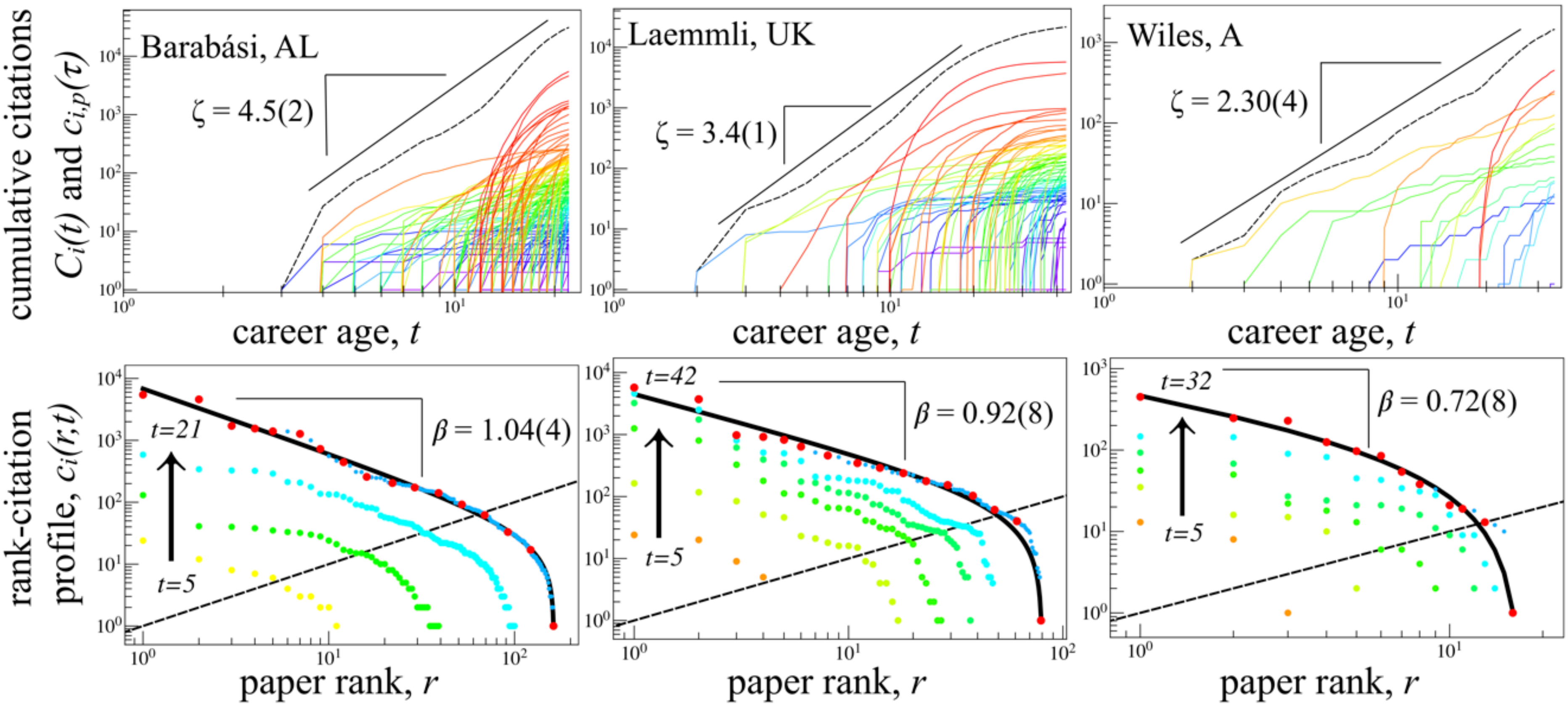
Complexity

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

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*



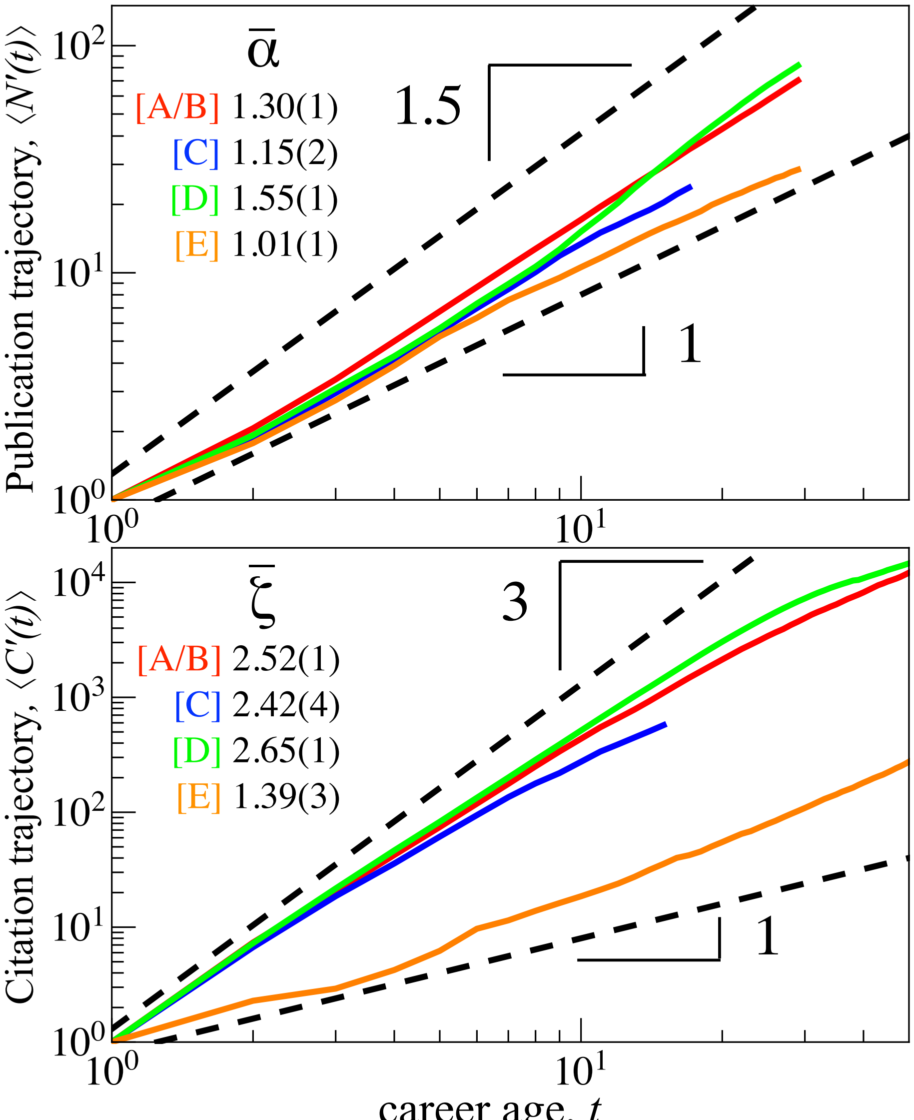
Statistical regularities in the rank-citation profile of scientists*, A. M. Petersen, H. E. Stanley, S. Succi. Scientific Reports 1, 181 (2011).*

$$c(r) \equiv A r^{-\beta}(N+1-r)^{\gamma}$$ discrete generalized Beta function (DGBD)

$$C_i \sim h_i^{1+\beta_i}$$ simple scaling relation between the *h*-index and *C*

The Z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile,
A. M. Petersen, S. Succi J. Informetrics 7, 823-832 (2013).

Patterns of “success”: publication and impact growth patterns of highly cited scientists



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

Set A: 100 most-cited physicists, average h-index, ⟨h⟩ = 61 ± 21

Set B: 100 additional highly-prolific physicists, ⟨h⟩ = 44 ± 15

Set C: 100 assistant professors from 50 US physics depts., ⟨h⟩ = 15 ± 7

Set D: 100 most-cited cell biologists, ⟨h⟩ = 98 ± 35

Set E: 50 highly-cited pure mathematicians, ⟨h⟩ = 20 ± 10

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

Potential pitfalls in the forecasting of careers?

On the Predictability of Future Impact in Science, *O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, S. Fortunato. Scientific Reports 3, 3052 (2013).*

The case for caution in predicting scientists' future impact, *O. Penner, A. M. Petersen, R. K. Pan, S. Fortunato, Physics Today 66, 8-9 (2013).*

Predicting scientific success

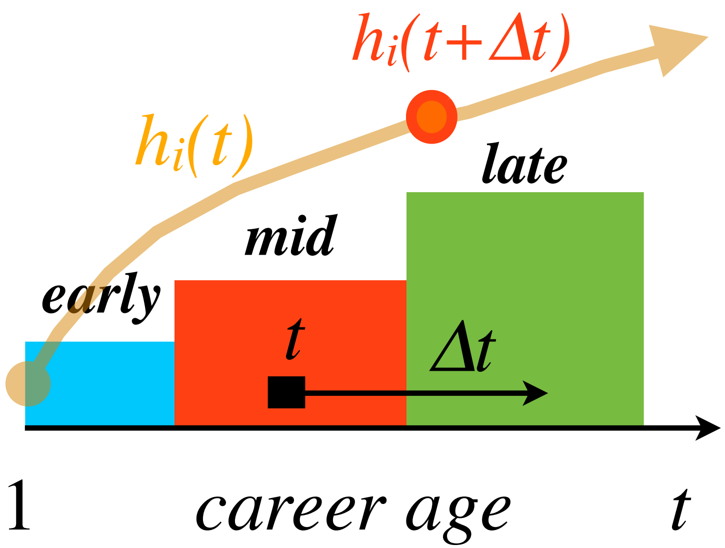
Daniel E. Acuna, **Stefano Allesina** and **Konrad P. Kording** present a formula to estimate the future h-index of life scientists.

13 SEPTEMBER 2012 | VOL 489 | NATURE | 201

Major Flaws!

1. Aggregating across different career-age cohorts
2. h-index is non-decreasing ⇒

   R^2 will be artificially large



METRICS

Predict your future h*-index*

These are approximate equations for predicting the h-index of neuroscientists in the future. They are probably reasonably precise for life scientists, but likely to be less meaningful for the other sciences. Try it for yourself online at go.nature.com/z4rroc.

- **Predicting next year** ($R^2=0.92$):

$$h_{+1}=0.76+0.37\sqrt{n}+0.97h-0.07y+0.02j+0.03q$$

- **Predicting 5 years into the future** ($R^2=0.67$):

$$h_{+5}=4+1.58\sqrt{n}+0.86h-0.35y+0.06j+0.2q$$

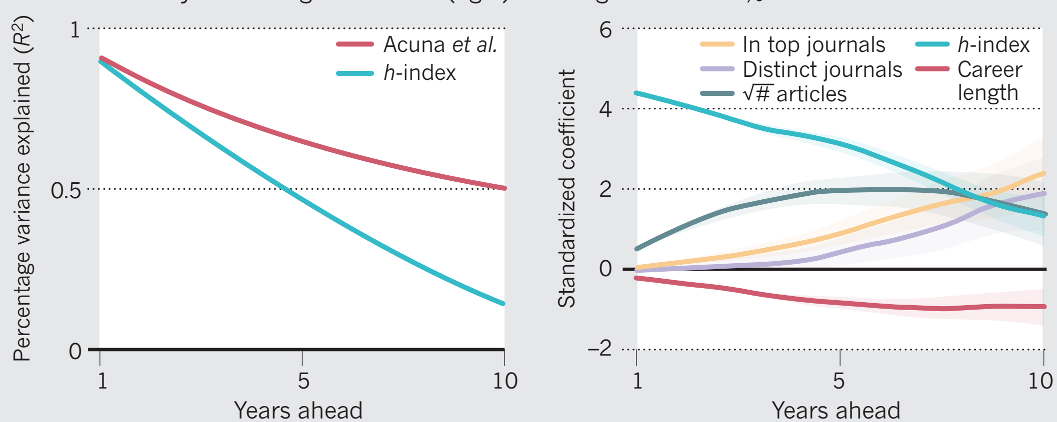
- **Predicting 10 years into the future** ($R^2=0.48$):

$$h_{+10}=8.73+1.33\sqrt{n}+0.48h-0.41y+0.52j+0.82q$$

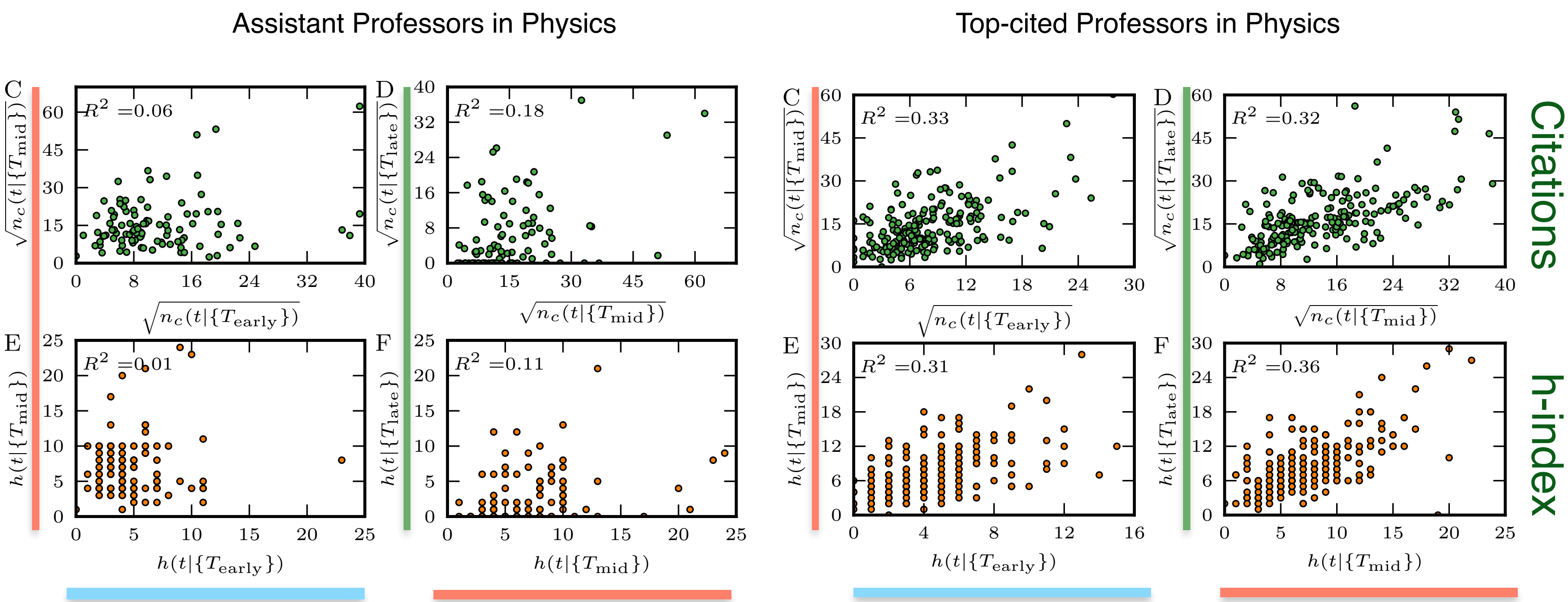
Key: n, number of articles written; h, current h-index; y, years since publishing first article; j, number of distinct journals published in; q, number of articles in *Nature*, *Science*, *Nature Neuroscience*, *Proceedings of the National Academy of Sciences* and *Neuron.*

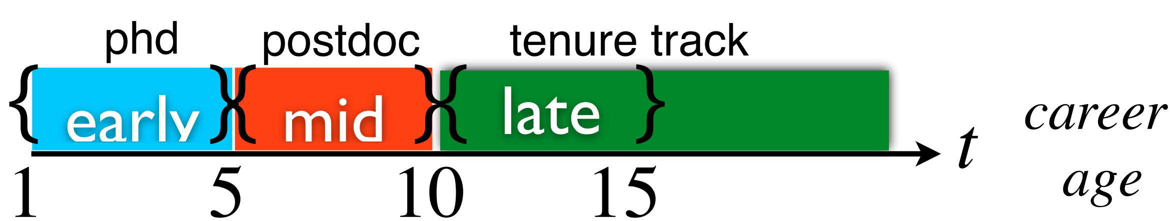
PATHS TO SUCCESS

The accuracy of future h-index prediction decreases over time, but the Acuna *et al*. formula predicts future h-index better than does current h-index alone (left). The contribution of each factor to the formula accuracy also changes over time (right). Shading indicates 95% confidence error bars.



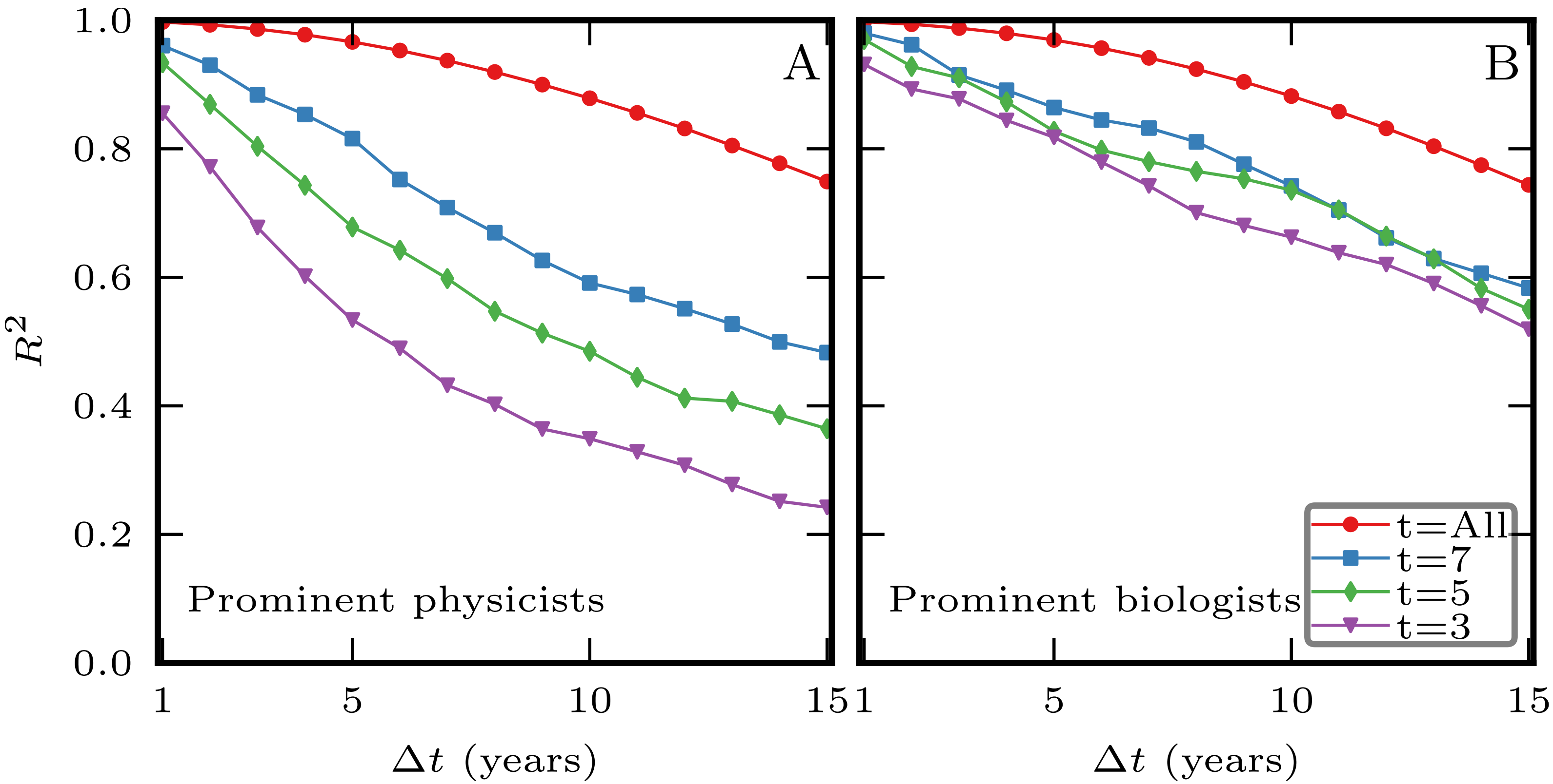
Difficulty in predicting scientists' future impact





n_c $(t \mid \{T_i\})$ = # of citations and h $(t \mid \{T_i\})$ = h-index computed at the end year t of each period, *ONLY* using papers produced in each period $\{T_i\}$. Comparing early, mid, late-career (non-overlapping) intervals shows that age and prestige affect the predictability!

The R^2 ("predictability") within younger age-cohorts is significantly less than the pooled (All)



200 Prolific authors of *Physical Review Letters* (PRL)

100 Prolific authors of *Cell*

Sources of uncertainty in predicting future impact

$$h(t) \xrightarrow{?} h(t+\Delta t)$$

use regression model for predicting *h(t+Δt)*

$$h(t+\Delta t) \text{ depends on}$$

$$h(t) = \text{H-index at career age t}$$
$$n_p(t) = \text{number of publications (co)authored}$$
$$j(t) = \text{number of distinct journals of publications}$$
$$q(t) = \text{number of papers in high impact journals}$$
$$t = \text{Career age of scientist}$$

- ***On the Predictability of Future Impact in Science,*** *O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, S. Fortunato. Scientific Reports 3, 3052 (2013).*
- ***The case for caution in predicting scientists' future impact,*** *O. Penner, A. M. Petersen, R. K. Pan, S. Fortunato, Physics Today 66, 8-9 (2013).*

Consider Non-Cumulative incremental measures

$$\Delta h(t, \Delta t) = h(t + \Delta t) - h(t)$$

$\Delta h(t, \Delta t)$ depends on

$$h(t) = \text{H-index at career age t}$$

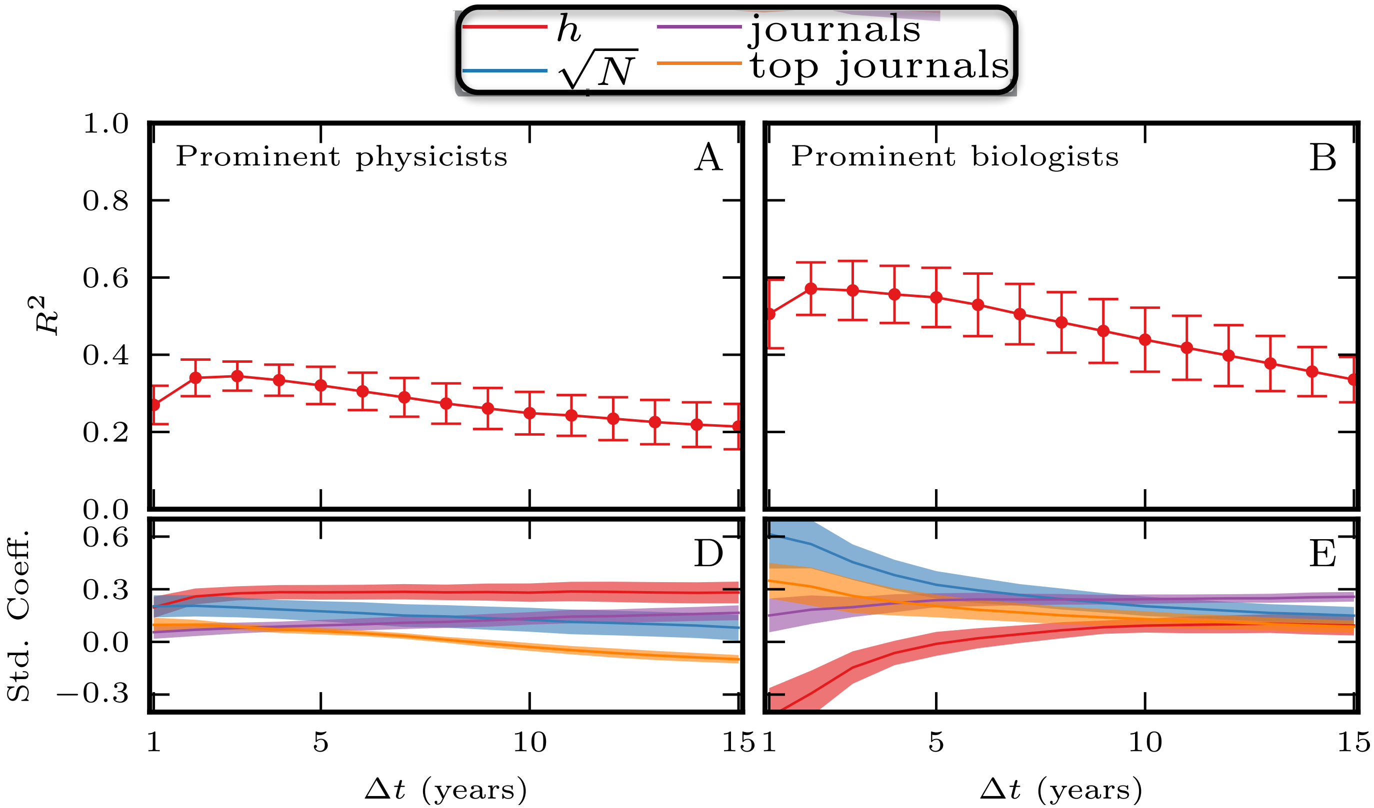
$$n_p(t) = \text{number of publications (co)authored}$$

$$j(t) = \text{number of distinct journals of publications}$$

$$q(t) = \text{number of papers in high impact journals}$$

… does not suffer from endogenous correlations

Modeling a non-cumulative measure: Δh(t+Δt,t)

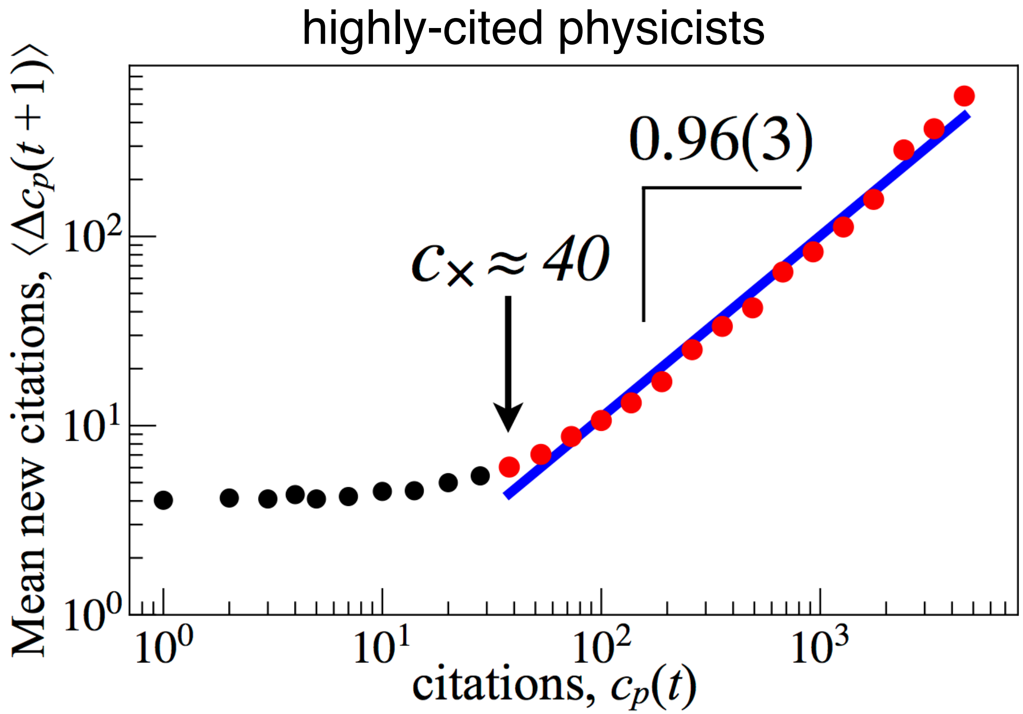


**** Reputation signaling:**

Interestingly, the predictive power due to publishing in top journals appears become less important further along the career

Lessons learned

1) Cumulative measures over-estimate prediction power
2) R^2 "predictive power" and regression parameters depend on career age
3) Early career scientists: "predictability" of *h(t)* is due to the non-decreasing incremental nature of *h(t)* *and not much more*
4) Important not to overfit models: separate age cohorts



Reputation effect citation model

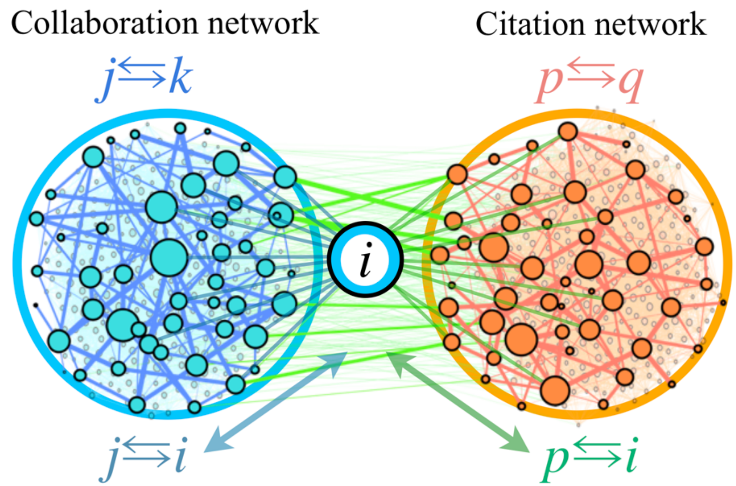
of new citations in year *t+1* = $\Delta c_{i,p}(t+1) \equiv \eta \times \Pi_p(t) \times A_p(\tau) \times R_i(t)$

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!

There are important yet quantifiable nuances to citation dynamics!!!

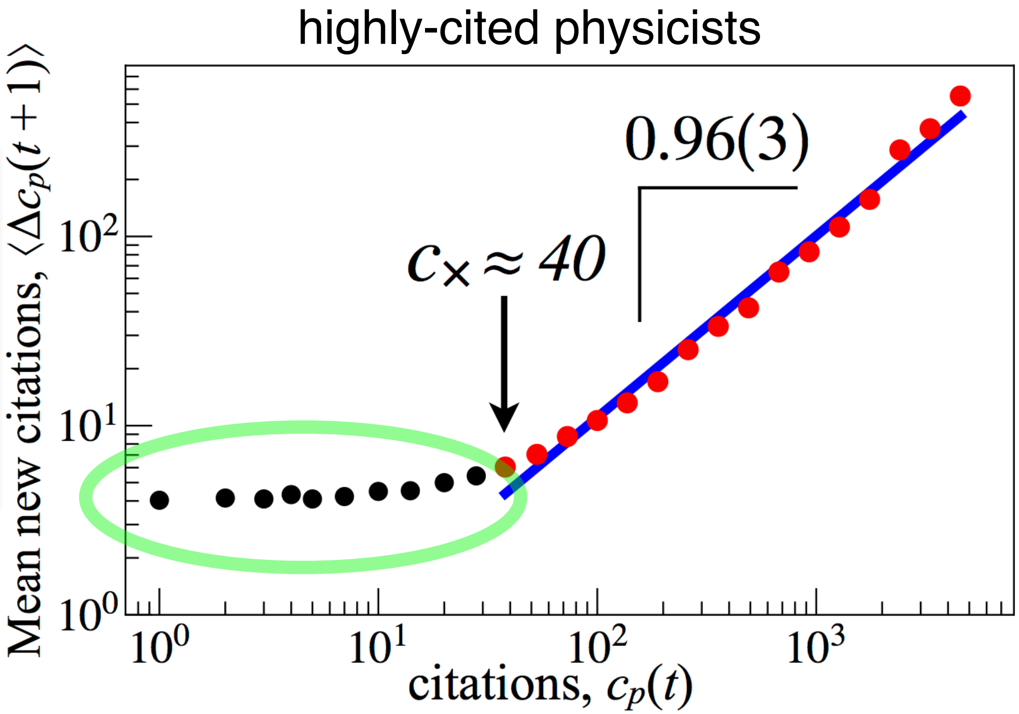
Measuring behavioral aspects: Reputation and Social Ties



Collaboration and citation networks provide channels for the flows of reputation signaling

We seek to quantify the impact of author reputation on the citation rate of his/her papers $(p \rightleftarrows i)$

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

Author-specific factors matter!

There are important yet quantifiable nuances to citation dynamics!!!

Author-specific features: $\pi_i, \overline{\tau}_i, \rho_i$

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.

		$c(t-1) < c_\times$			$c(t-1) \geq c_\times$			
	Name	π_i	$\overline{\tau}_i$	ρ_i	π_i	$\overline{\tau}_i$	ρ_i	$c_\times$
physics	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	40
	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	
biology	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	100
	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	
math	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	20
	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	

Take home message:

1) The reputation effect is strong for papers not yet highly cited

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

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

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

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

$\pi(c \geq c_\times) \approx 1$ (linear pref. attachment)

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.

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

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

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

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

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)

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

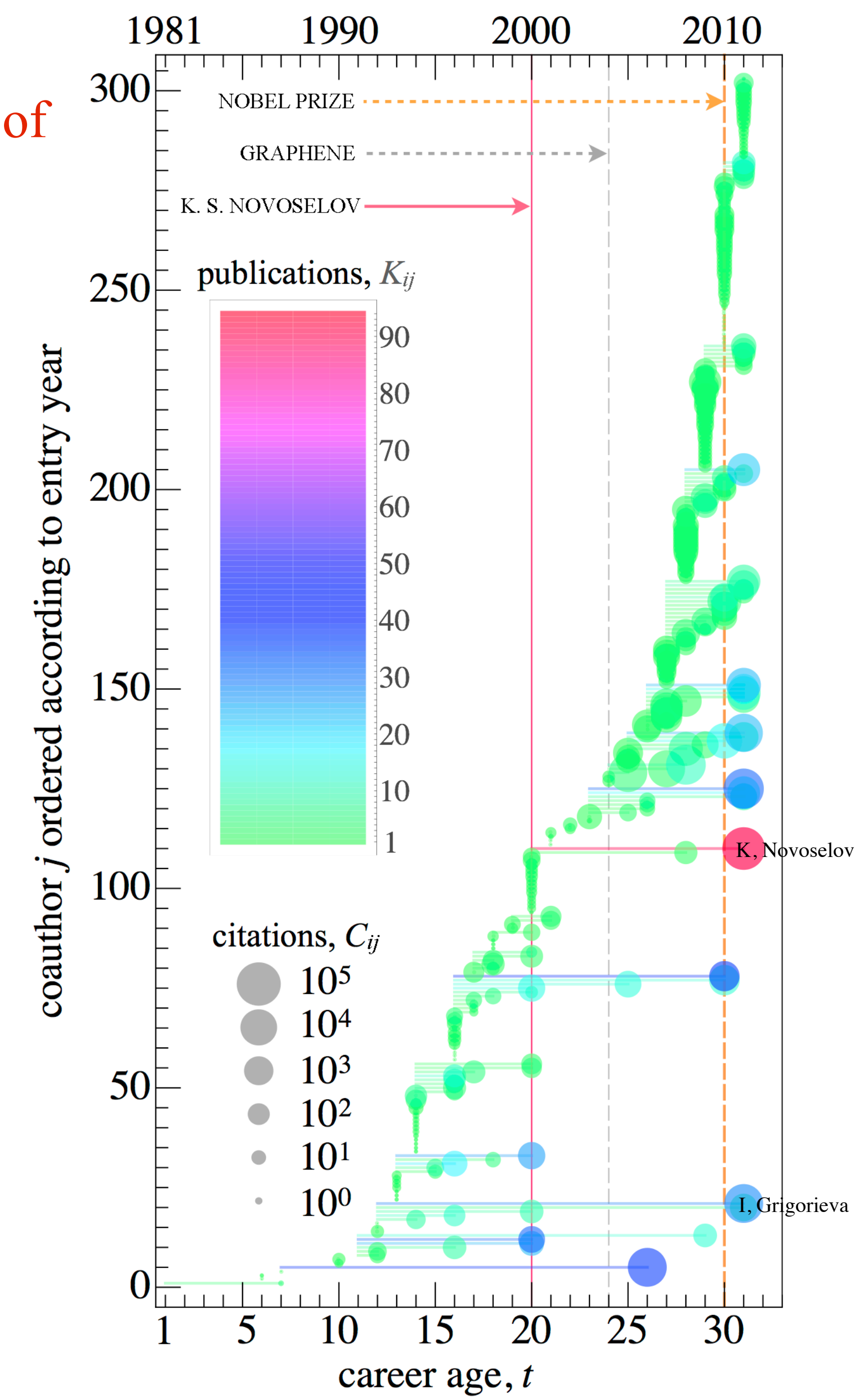
Sir Andre K. Geim
 # publications, N_i (2012) = 217
S_i = 303 coauthors
The average copublication duration $\langle L_i \rangle$ = 2.1 years, $\langle K_i \rangle$ = 3.7 pubs.

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

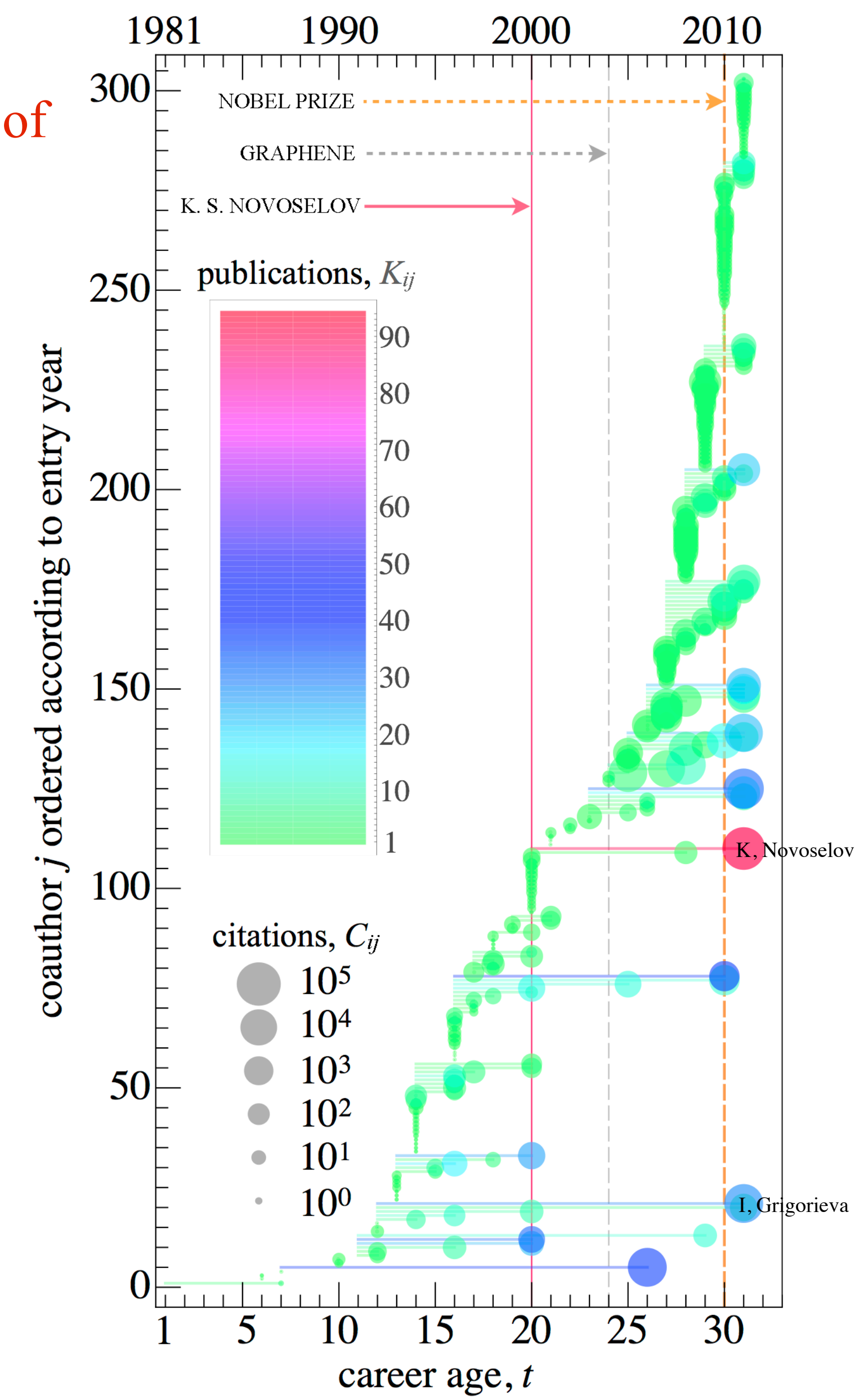
S_i = 303 coauthors

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

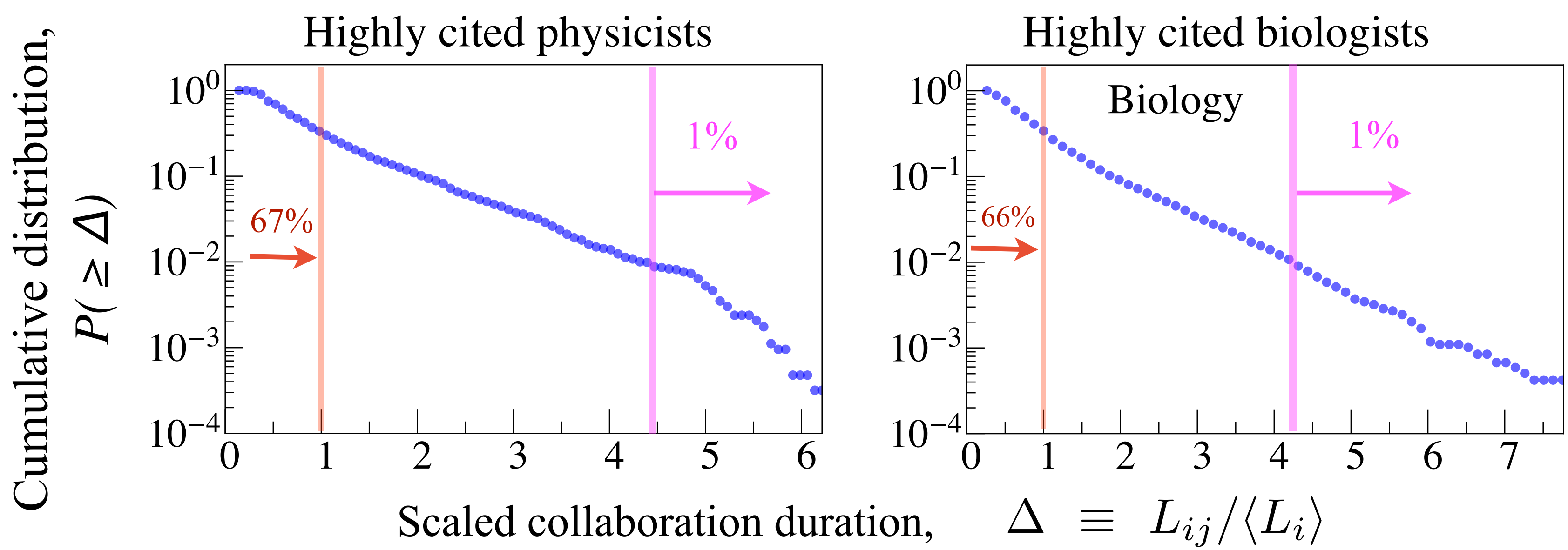
2) The effect of team heterogeneity on productivity is positive indicating the benefits of efficient team management via hierarchy / mentoring

3) Research life-partners — "a scientific marriage": The effect of strong 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).

Quantifying the impact of weak, strong, and super ties in scientific careers (2015) A. M. Petersen. Under Review



High collaboration turnover rate. Is this efficient?



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

- The "invisible college" is held together by weak ties
- Team formation/destruction costs are high; need to increase rates of meaningful and lasting collaboration
- Fractional counting could introduce a negative incentive to collaborate dragging on the innovative potential of science

How does publication and authorship inflation impact the citation credit economy?

Total credit C_y^T produced by all publications produced in year y using citation counts in year Y= y+Δy:

Partition credit equally into "shares"

$$C_{y,Y}^T = \sum_{p=1}^{N_y} a_p (c_{p,y,Y}/a_p)$$

$$= \sum_{p=1}^{N(y)} c_{p,y,Y} = N_y \langle c_{p,y,Y} \rangle$$

Reproduce (Multiply) credit for each author

$$C_{y,Y}^T = \sum_{p=1}^{N_y} a_p c_{p,y,Y}$$

using a crude approximation which also neglects correlations between team size and citations…..

$$\approx \langle a_p, y \rangle N_y \langle c_{p,y,Y} \rangle$$

- no penalty for unethical coauthorship behaviors such as "free-riding" or "tit-for-tat" partnering

- inflation in C_y^T the credit economy can have multiple sources (3 considered here)!

How might fractional counting affect career citation measures

Partition credit equally into “shares”:

$$\tilde{S}_i^j = \sum_{p=1}^{N_i^j} \frac{1}{a_p} \frac{c_{p,y,Y}}{\langle c_{y,Y} \rangle}$$

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

Reproduce (Multiply) credit for each author:

$$\tilde{C}_i^j = \sum_{p=1}^{N_i^j} \frac{c_{p,y,Y}}{\langle c_{y,Y} \rangle}$$

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).

i = author index
p = paper index
y = year paper p was published
Y = citation data download year (>y), also referred to as the census year
j = set of journals considered: Nature, PNAS, and Science research articles

Analyzed these journals over the years y =1958-2002 with Y=2009; roughly 200k papers, 40k career disambiguated profiles; median coauthor size across papers = 5, mean # papers across profiles = 2.5

Crucial difference:

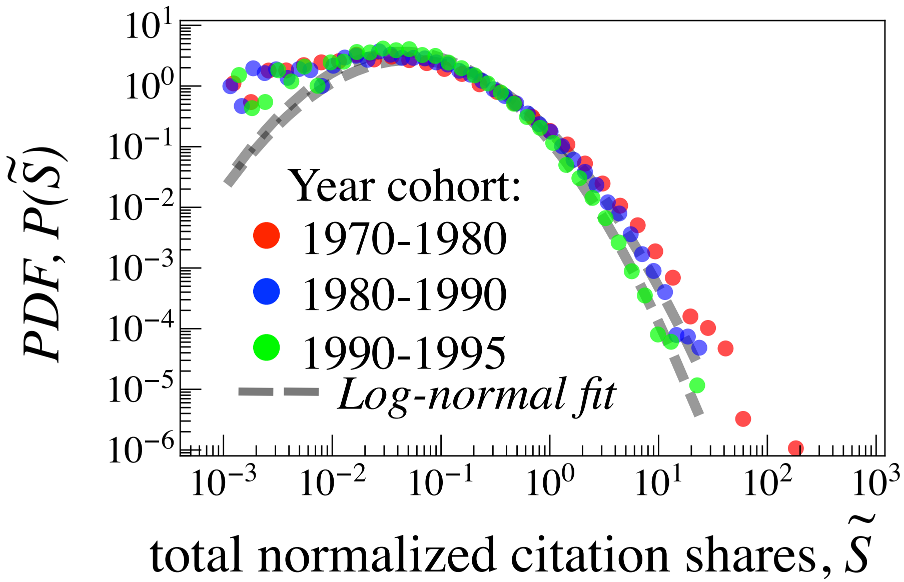
Total credit “issued” per paper = $\frac{c_{p,y,Y}}{\langle c_{y,Y} \rangle}$

Total credit “issued” per paper = $a_p \frac{c_{p,y,Y}}{\langle c_{y,Y} \rangle}$

Estimating the cumulative citation distribution across science careers

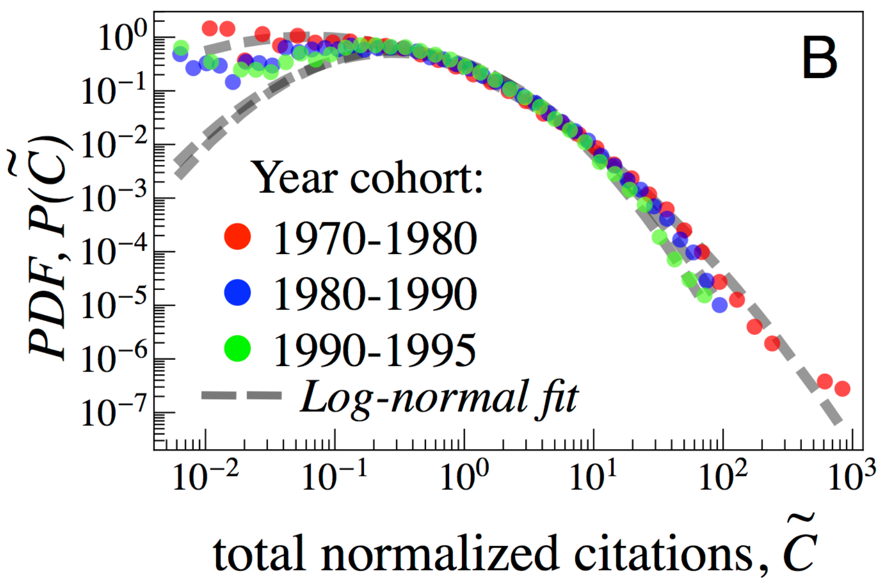
Fractional citations

$$\tilde{S}_i^j = \sum_{p=1}^{N_i^j} \frac{1}{a_p} \frac{c_{p,y,Y}}{\langle c_{y,Y} \rangle}$$



Multiplicative citations

$$\tilde{C}_i^j = \sum_{p=1}^{N_i^j} \frac{c_{p,y,Y}}{\langle c_{y,Y} \rangle}$$

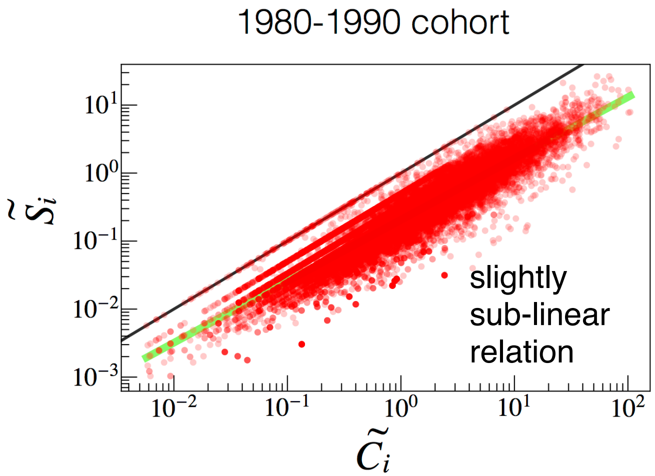


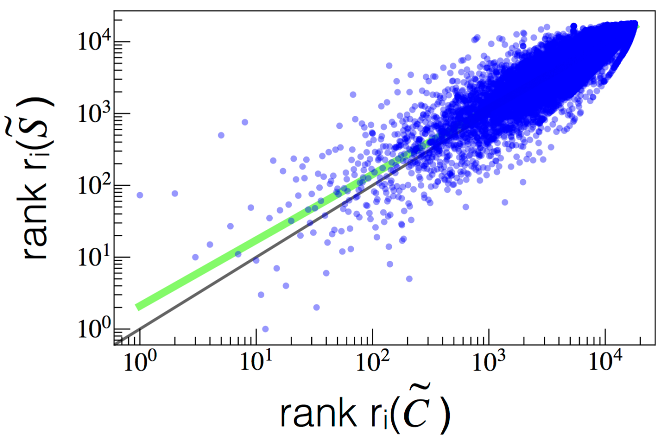
After controlling for censoring and cohort bias, Scientific careers exhibit a heavy-tailed "success" distribution that appears to be long-normally distributed for the bulk of the distribution.

Log-normal "size" distributions are indicative of Gibrat "proportional growth" processes. Moreover, the stability of the distribution for both measures indicates that the fractional citation method does not entirely disrupt the aggregate distribution of impact.

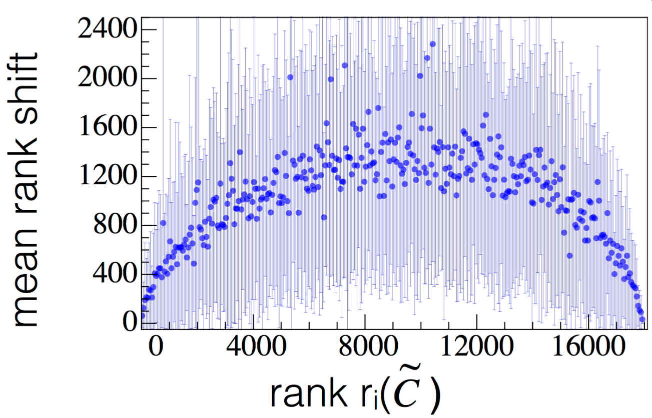
Journal set j	Cohort entry years	$G(\tilde{C})$	$f_{1\%}(\tilde{C})$	$G(\tilde{S})$	$f_{1\%}(\tilde{S})$
Nat./PNAS/Sci.	1970 – 1995	0.69	0.18	0.70	0.22
	1970 – 1980	0.74	0.22	0.74	0.27
	1980 – 1990	0.67	0.15	0.66	0.15
	1990 – 1995	0.63	0.12	0.62	0.13

TABLE I: Summary of the Gini index (G) and top-1% share ($f_{1\%}$) inequality measures calculated from the distributions of citation impact, using both normalized citations ($\tilde{C}$) and normalized citation shares ($\tilde{S}$) as the measure. The two G values are nearly the same, while $f_{1\%}(\tilde{S}) \gtrsim f_{1\%}(\tilde{C})$.





rank shift = $|\text{rank}_i(\tilde{C}) - \text{rank}_i(\tilde{S})|/\sqrt{2}$



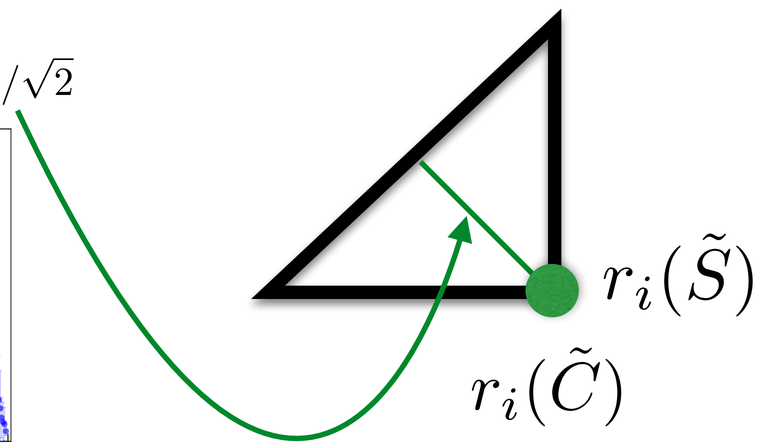
What is the potential impact of using fractional shares on the ranking of scientists?

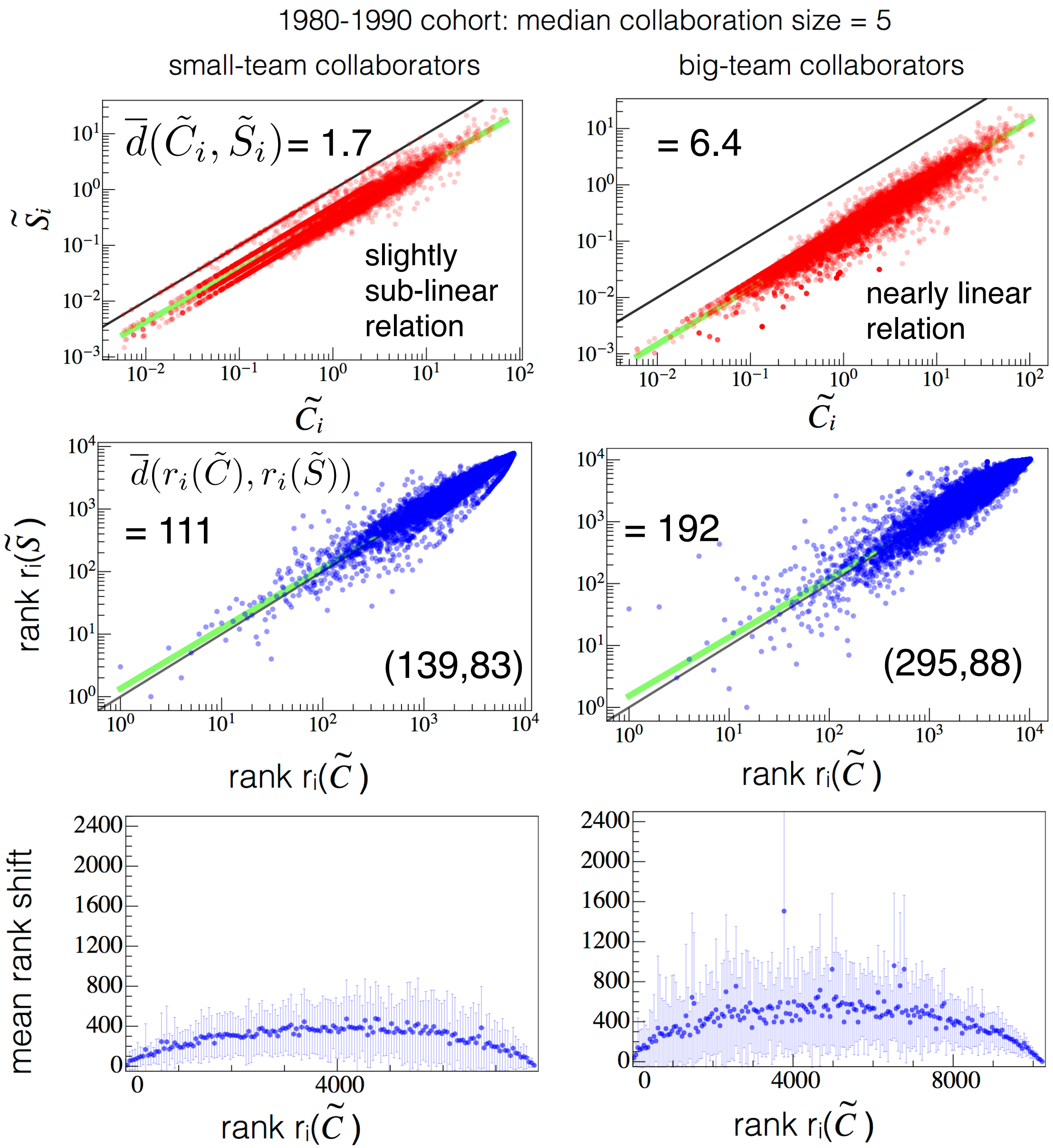
The new impact measure appears to be related by a quasi-linear relation

$$\tilde{S} \propto \tilde{C}^{\delta} \quad \text{with} \quad \delta \lesssim 1$$

However the noise in the subsequent ranking appears to be quite dependent on $\tilde{C}_i^j$

Leading to substantial rank reordering!





Is there team-size bias?

Each researcher profile is characterized by the M_i , the median # of coauthors calculated from their N_i publications (in j)

Separated profiles into two subsets, those with *$M_i \geq 5$ (big team) and $M_i < 5$ (small team)*

As one might suspect, there is larger noise in the ranking of big-team collaborators

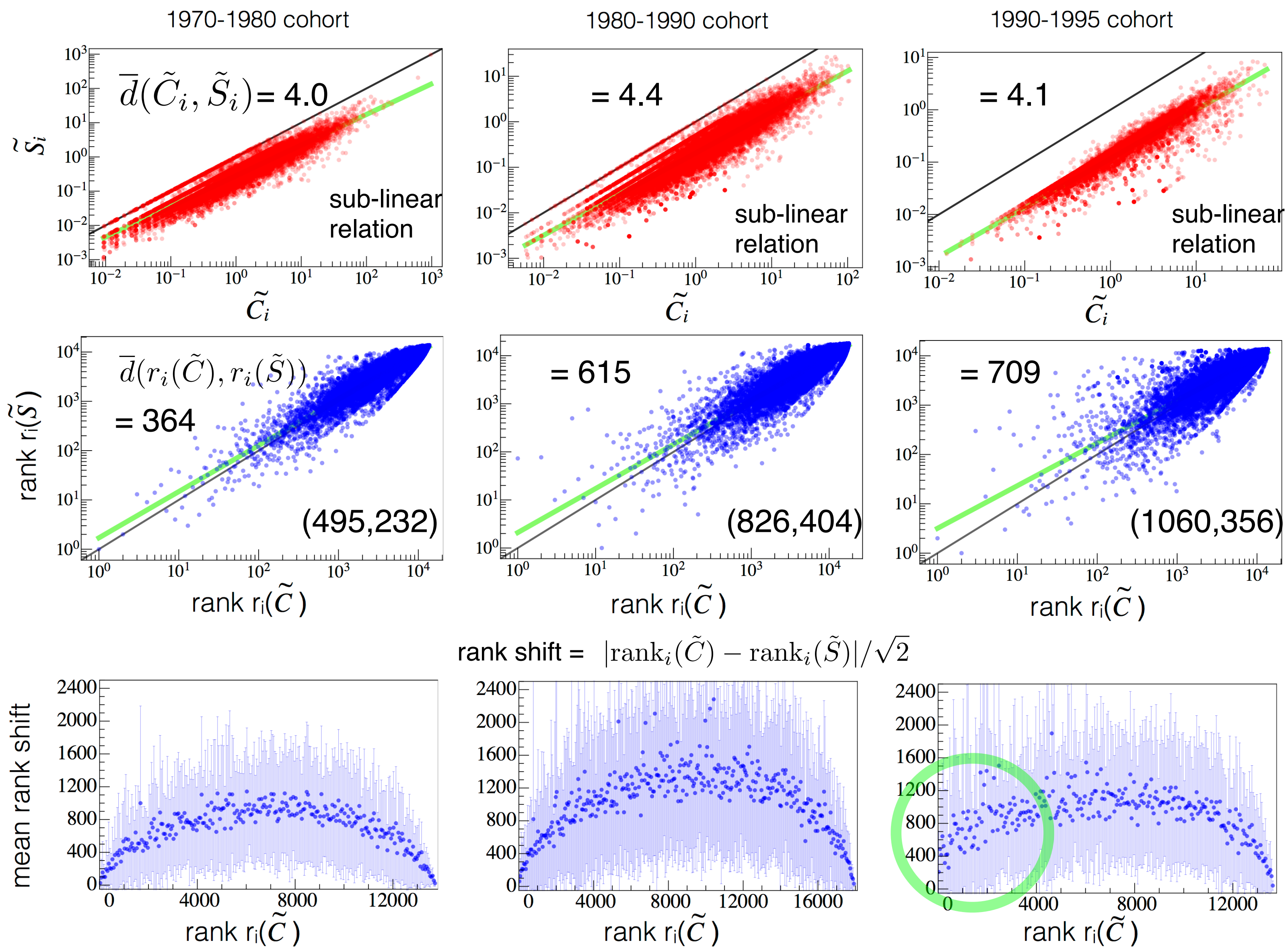
However, the mean rank-shift is significantly lower than when the two subsets were ranked together

In real academic ranking scenarios, consider rankings within variable team-size groups?….

Quantitative measure of rank instability:

Mean Kullback-Leibler relative entropy

$$\overline{d}(x_i, y_i) \equiv N_{authors}^{-1} \sum_{i=1}^{N_{authors}} \Big(x_i - y_i\Big) \ln \Big(x_i / y_i\Big)$$



Higher instability for the high-ranking profiles in the younger cohort

Institutions Markets Technologies
IMT
INSTITUTE FOR ADVANCED STUDIES LUCCA

Emergence of cumulative advantage in competitive arenas

Physical Review Letters
moving physics forward

Science
AAAS

nature

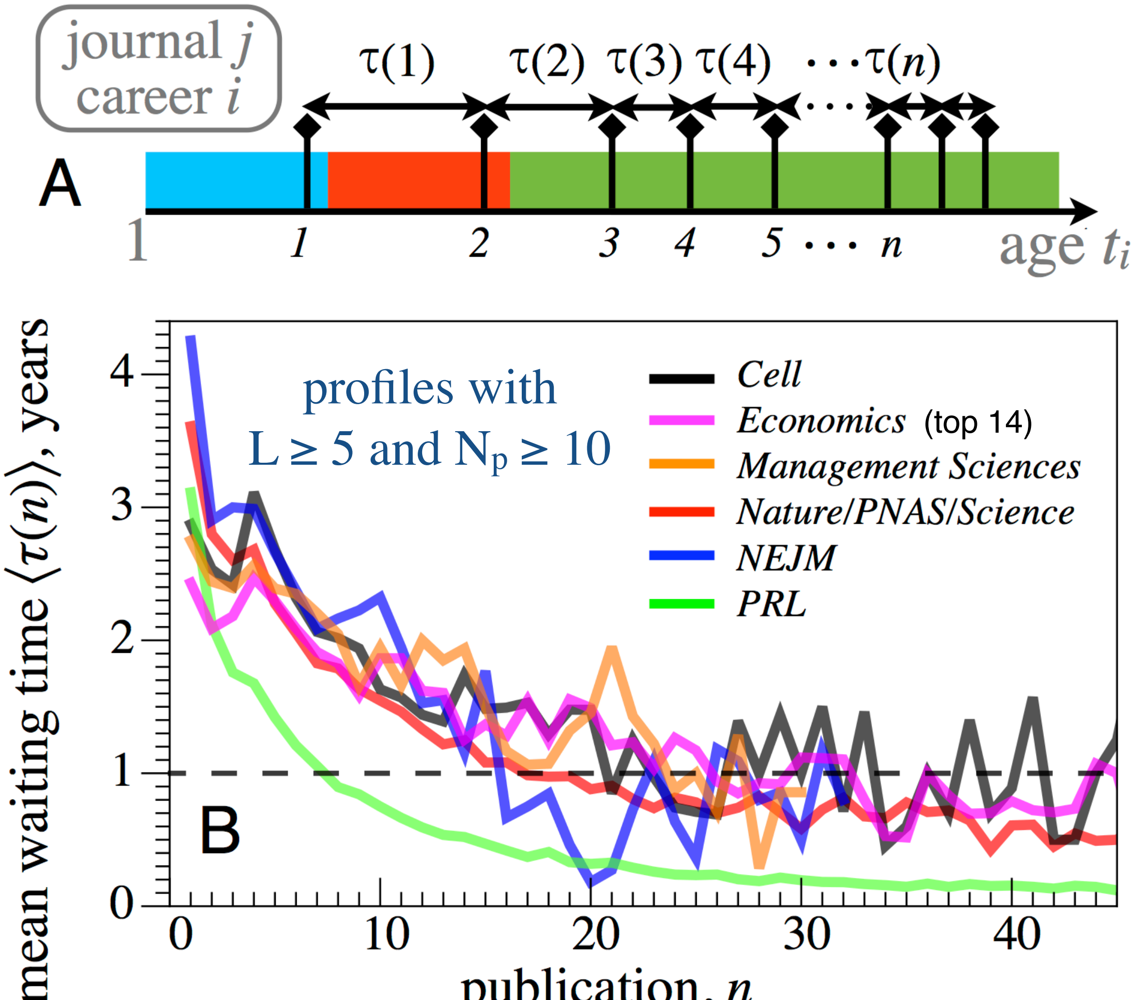
Cell

The NEW ENGLAND JOURNAL of MEDICINE

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

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

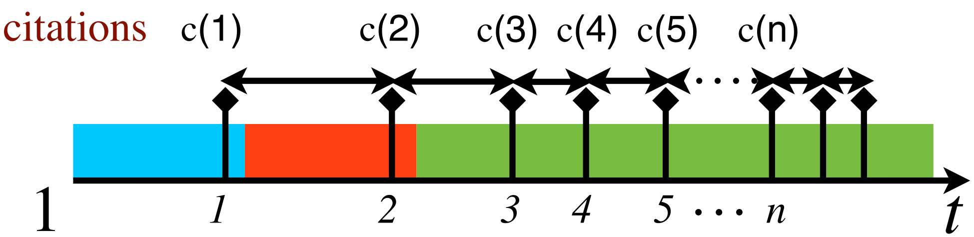
For each career *i* we track his/her longitudinal publication rate by aggregating over publications in a *specific set* of high-impact journals



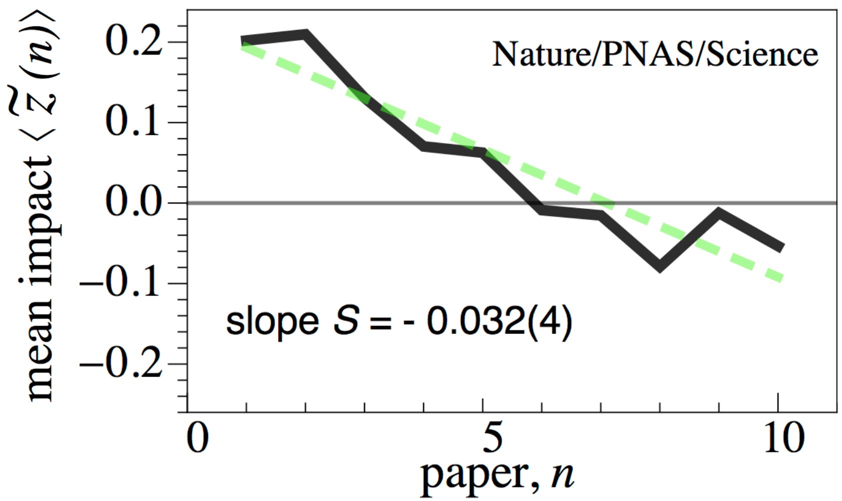
$\tau_i(n)$ is the waiting time between an author's n^{th} paper and $(n+1)^{th}$ paper?

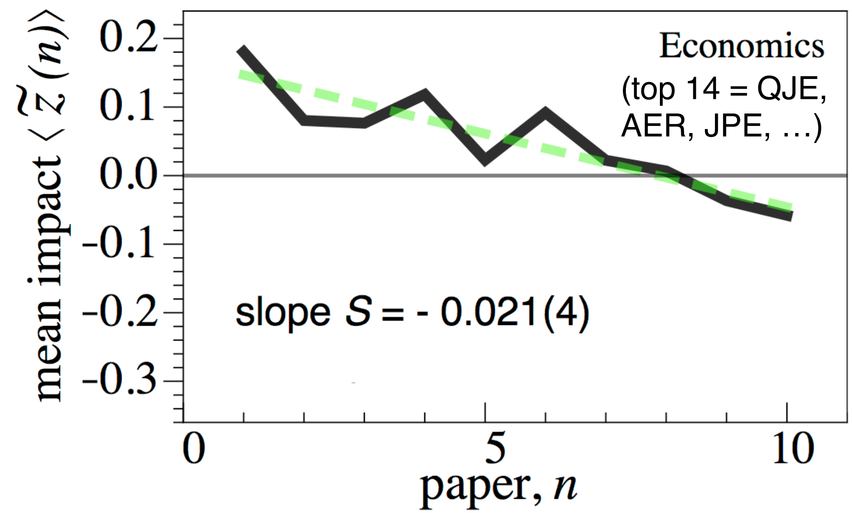
By the 10th paper, the waiting time between publications has decreased by ~ factor of 2 from $\tau_i(1)$!

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



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 cohort 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,

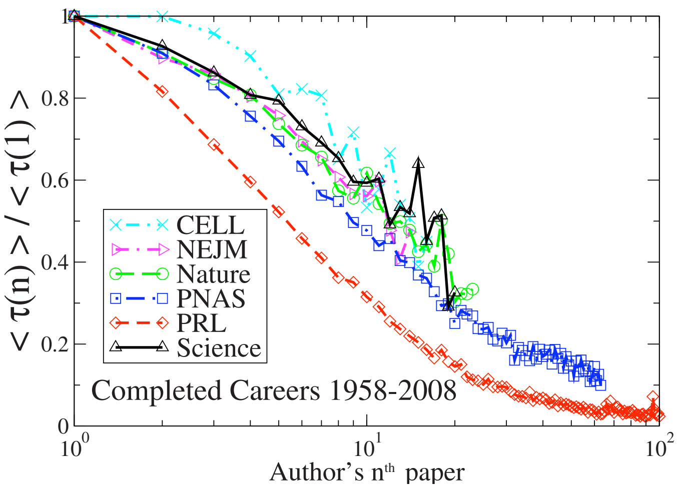
$$z_i(n) \equiv \frac{\ln c^j_{i,p,y}(n) - \langle \ln c^j_y \rangle}{\sigma[\ln c^j_y]} ,$$

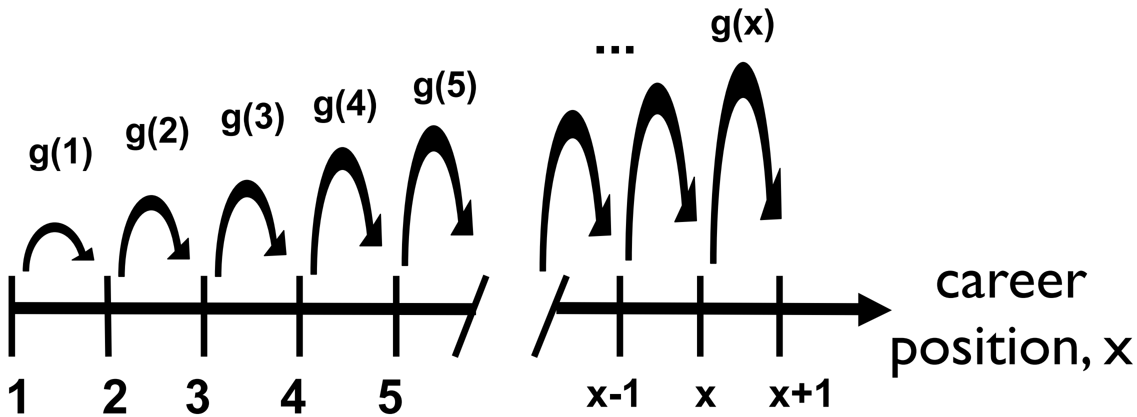
$$\tilde{z}_i(n) \equiv z_i(n) - \langle z_i \rangle$$

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.

Modeling the "Rich-get-richer" effect

- Forward progress follows a stochastic "progress rate" $g(x)$
- Cumulative advantage: $g(x)$ increases with career position x





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

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

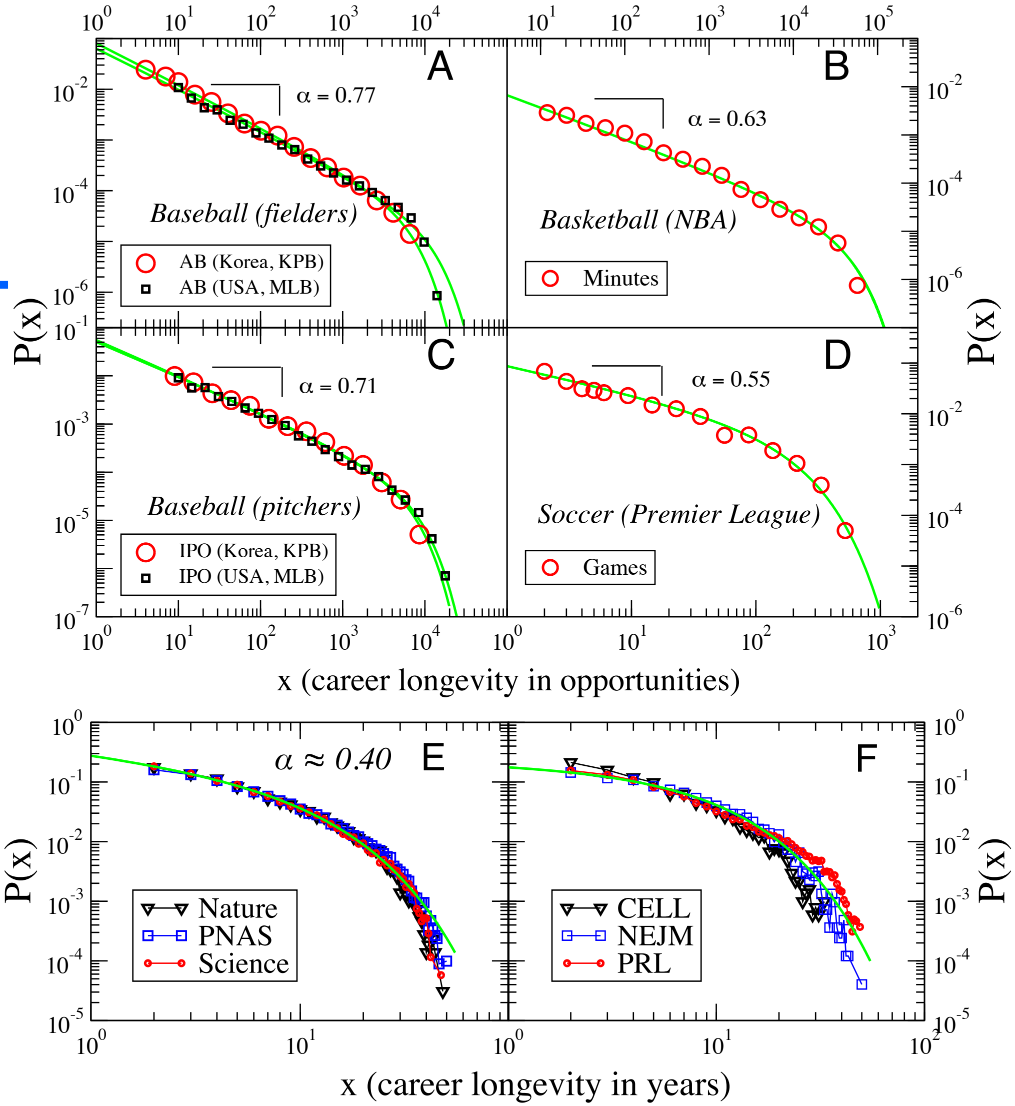
Methods for measuring the citations and productivity of scientists across time and discipline, *A. M. Petersen, F. Wang, H. E. Stanley. Phys. Rev. E 81, 036114 (2010).*

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 longev*ity,*
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”

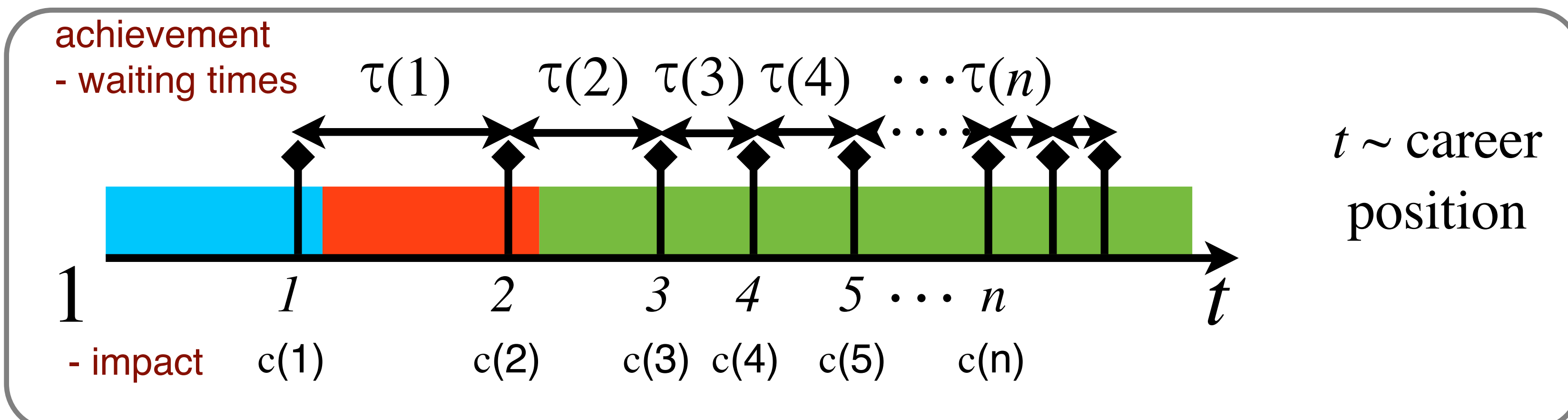
- 3% of all fielders finish their career with ONE at-bat!
- 3% of all pitchers finish their career with less than one inning pitched!

``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 lateral sclerosis (ALS), aka Lou Gehrig’s Disease

Sustainability of science careers

career i

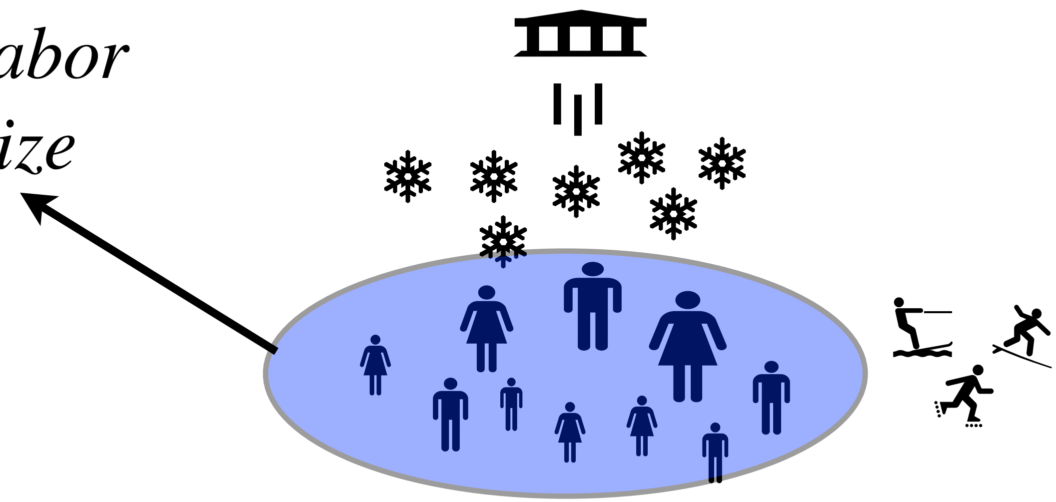


Appraisal of prior work: How important is cumulative advantage in a competitive system?

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

$$\text{capture rate} \propto w_i(t) \equiv \sum_{\Delta t=1}^{t-1} n_i(t-\Delta t)\underbrace{e^{-c\Delta t}}_{\text{exponential discount factor}}$$

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"

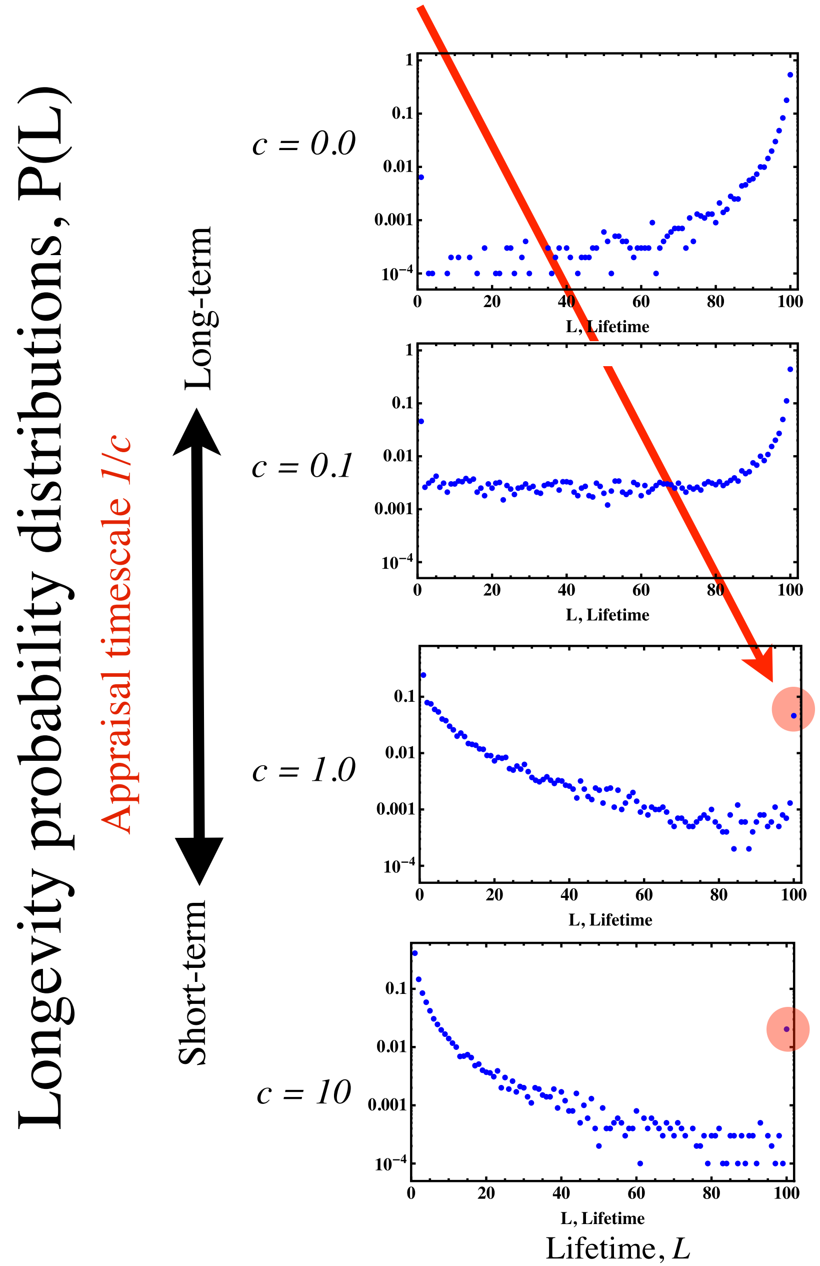
Our theoretical model suggests that

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.



Discounting time in the evaluation process: Insights from our appraisal model applied to real careers

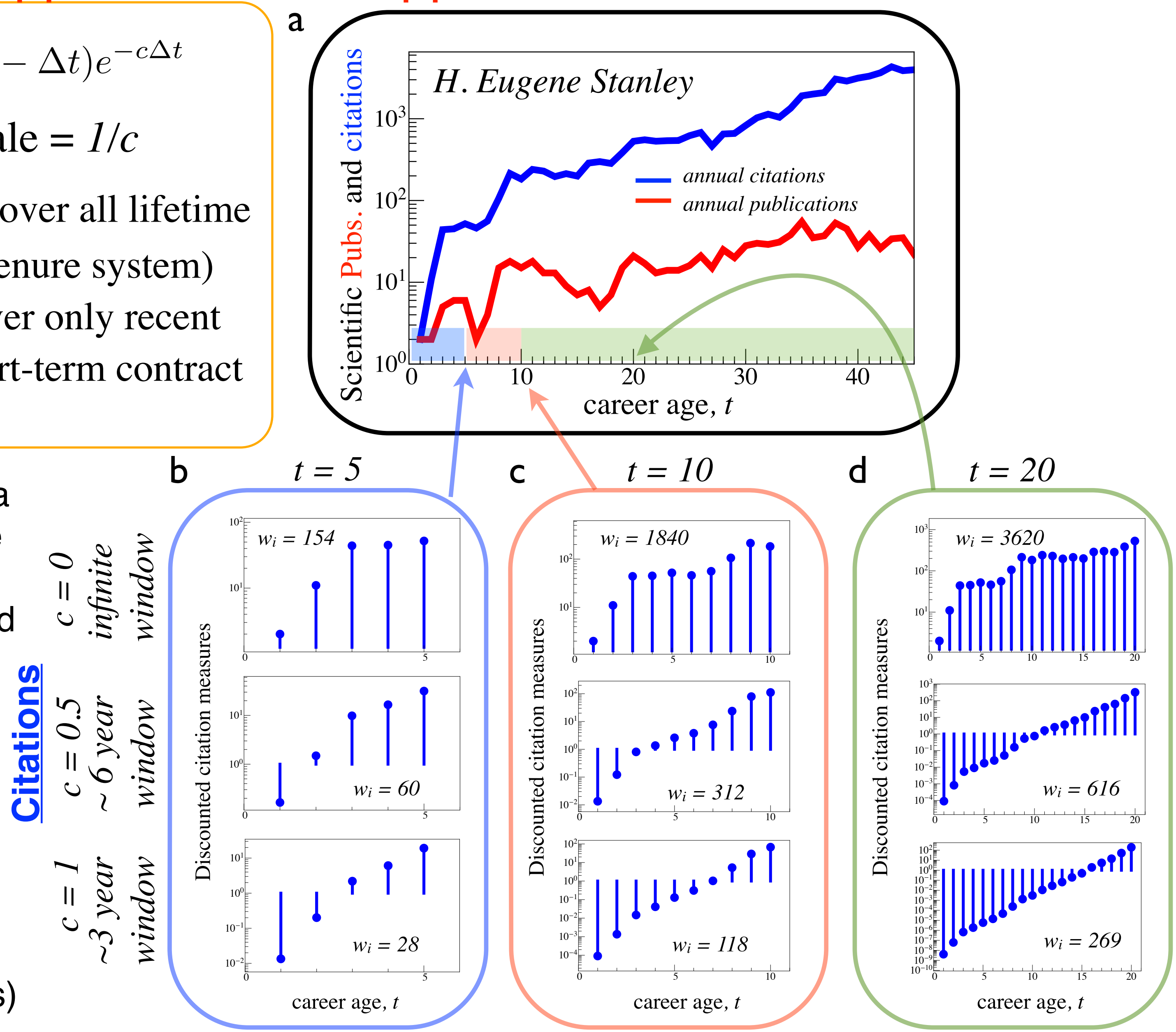
$$w_i(t) \equiv \sum_{\Delta t=1}^{t-1} n_i(t-\Delta t)e^{-c\Delta t}$$

appraisal timescale = *1/c*

c → 0 : appraisal over all lifetime achievements (~ tenure system)

c ≥1 : appraisal over only recent achievements (short-term contract system)

Question: Should a researcher's entire portfolio of prior work be considered in evaluation? In other competitive professions, more recent accomplishments are move valuable than more distant ones (e.g. professional sports)



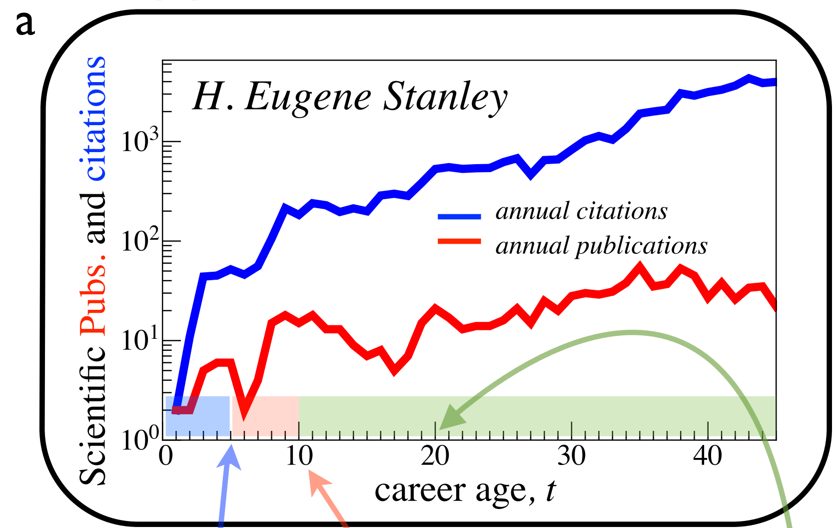
Discounting time in the evaluation process: Insights from our appraisal model applied to real careers

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

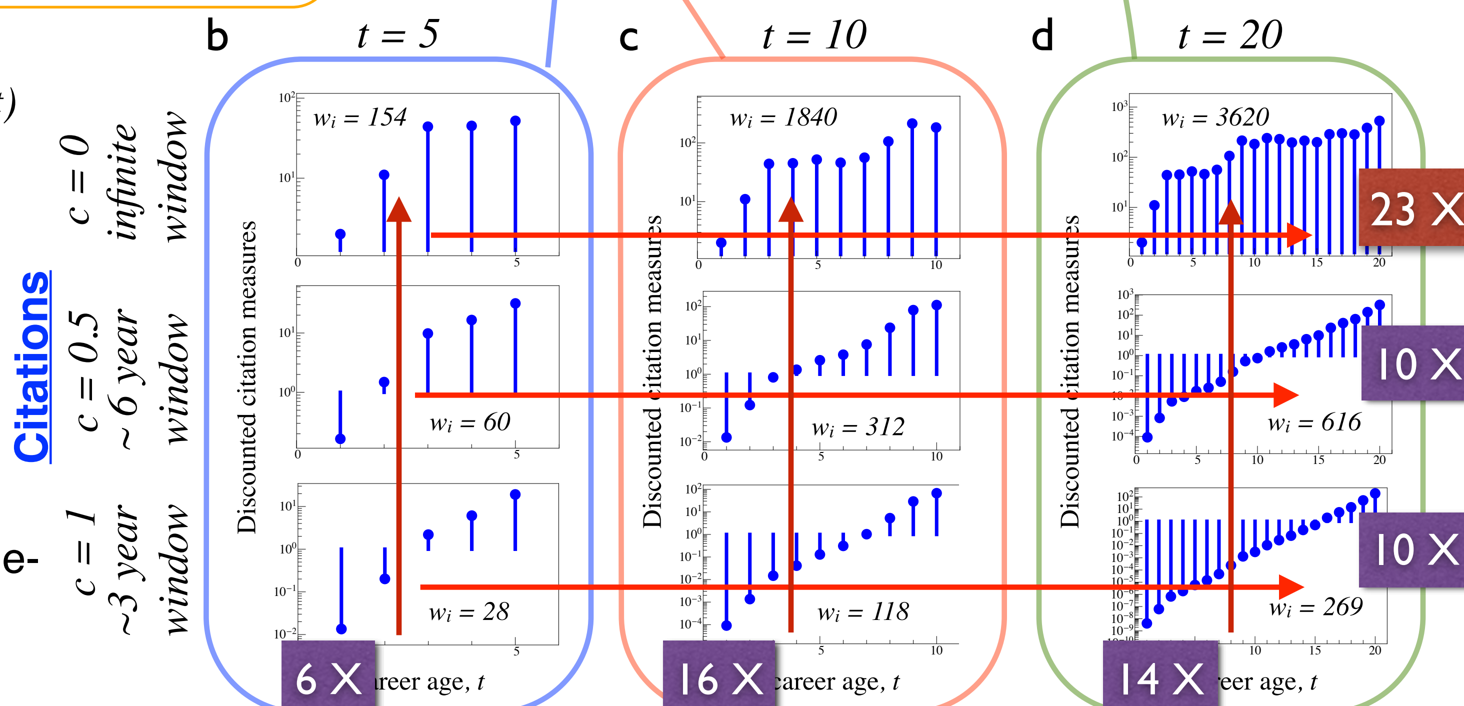
appraisal timescale = *1/c*

$c \rightarrow 0$: appraisal over all lifetime achievements (~ tenure system)

$c \geq 1$: appraisal over only recent achievements (short-term contract system)

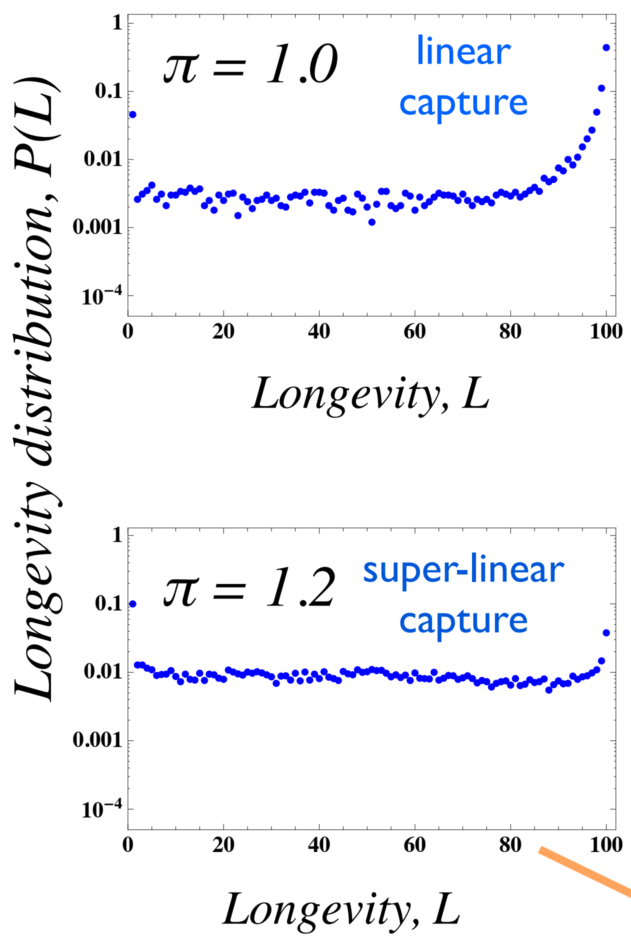


The increase in $w_i(t)$ (in the direction of the arrow) shows how different appraisal schemes can diminish the overwhelming cumulative advantage that can emerge in large time-window (small *c*) appraisal systems



Q: Is there an optimal appraisal-window size or contract length?

Model with *c = 0.1* ***(~ long term appraisal)***



counter-intuitive diminishing of the kingpin effect

non-linear preferential capture model

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

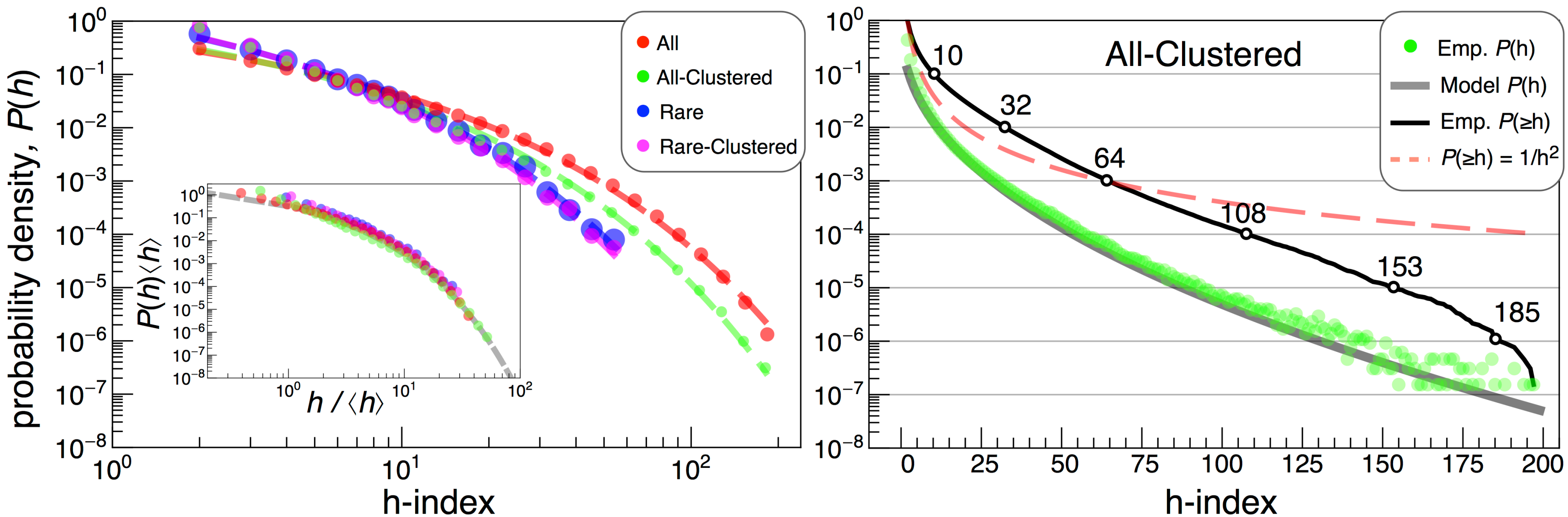
Hazard rate $H(L) = -d/dL\,[\ln P(L)]$: conditional probability that failure will occur at time $(L + \delta L)$ given that termination has not yet occurred at time L

$H(L) \approx 0$

*hazard rate is **almost** not dependent on career position!*

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)

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)

1. **How can we model the feedback of bibliometrics (IF) on scientists' (career, journal) decisions?**

Reputation, and other author-specific factors (age-cohort, collaboration style, etc.) matter. Even small differences can amplify over a career, resulting in a significant cumulative advantage.

Data-driven stochastic models that use empirical statistical patterns as benchmarks can be used to develop bibliometric indicators that (i) *properly* account for heterogeneity across careers and (ii) control for the growth (inflation) of science.

2. **Is fractional counting a solution to better capture the contribution of individuals?**

Indeed, fractional counting controls for paradigm shifts in the prevalence and role of teamwork on science careers and evaluation. However, the fractional counting method should not have the unintended consequence of dis-incentivizing collaboration.

Also, it should be known if the fractional counting introduces size-dependent bias — according to rank or collaboration style — by considering both the structural and dynamical aspects of collaboration.

- ***Together We Stand***, *I. Pavlidis, A. M. Petersen, I. Semendeferi. Nature Physics 10, 700-702 (2014).*
- ***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).*
- ***Exploiting citation networks for large-scale author name disambiguation***. *C. Schulz, A. Mazloumian, A. M. Petersen, O. Penner, D. Helbing. EPJ Data Science 3, 11 (2014).*
- ***A quantitative perspective on ethics in large team science***, *A. M. Petersen, I. Pavlidis, I. Semendeferi. Science & Engineering 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).*
- ***The Z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile,*** *A. M. Petersen, S. Succi J. Informetrics 7, 823-832 (2013).*
- ***On the Predictability of Future Impact in Science,*** *O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, S. Fortunato. Scientific Reports 3, 3052 (2013).*
- ***The case for caution in predicting scientists' future impact,*** *O. Penner, A. M. Petersen, R. K. Pan, S. Fortunato, Physics Today 66, 8-9 (2013).*
- ***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).*
- ***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 rank-citation profile of scientists***, *A. M. Petersen, H. E. Stanley, S. Succi. Scientific Reports 1, 181 (2011).*
- ***Methods for measuring the citations and productivity of scientists across time and discipline,*** *A. M. Petersen, F. Wang, H. E. Stanley. Physical Review E 81, 036114 (2010).*

Thank you!

A special thanks to my collaborators:

Santo Fortunato, Shlomo Havlin, Dirk Helbing, Woo-Sung Jung, Kimmo Kaski, Amin Mazloumian, Fabio Pammolli, Raj Pan, Ionnis Pavlidis, Orion Penner, Armando Rungi, Massimo Riccaboni, Christian Schulz, Ionna Semendeferi, Sauro Succi, Gene Stanley, Jae-Sook Yang

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

Title: Quantifying growth trends in science careers with application to bibliometric evaluation

Abstract: Research does not produce itself. Instead, there are idiosyncratic individuals involved, characterized by diverse backgrounds, interests, behaviors, strategies, and goals. As such, science is an extremely complex socio-economic system. I use data-driven computational methods to analyze and model the science of science, where the unit of analysis can vary across multiple scales, from publications, to individuals (careers), to teams, and large institutions such as countries. Against this multilevel backdrop, questions motivated from the theories of complex systems, management & organization science, labor economics, and research policy are often the starting point. Are there quantifiable patterns of scientific success? Are they useful in the career evaluation process? Are there ways to improve the sustainability of science careers while at the same time maintaining a high level of competitive selection? How do metrics for individual achievement depend on collaboration factors? How might paradigm shifts in science affect science careers?