Quantifying statistical regularities in the career achievements of scientists and professional athletes

Alexander M. Petersen
Department of Physics, Boston University
Thesis Advisor: H. Eugene Stanley

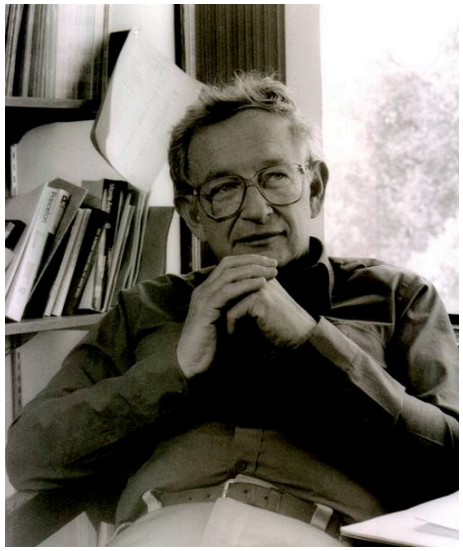
Final Oral Examination, March 8 2011

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

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

A. M. Petersen, H. E. Stanley, S. Succi. "Statistical regularities in the rank-citation profile of scientists". Under review.

Opening Questions



Using quantitative methods developed in statistical physics to address questions in *sociology*.....

- Are stellar careers an anomaly?
- Are there statistical regularities in *success*?
- Are there universal mechanisms that guide success?

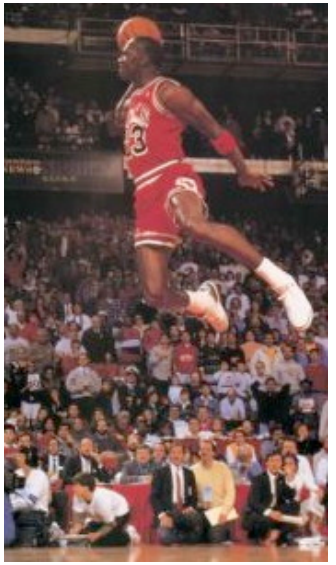
Outline

1. Question: How to quantify "success"?
2. Regularities in the career longevity and publication impact of scientists in academia
3. A quantitative model for career longevity that incorporates the "Matthew Effect"
4. Quantifying the rank-citation profile of individual scientists

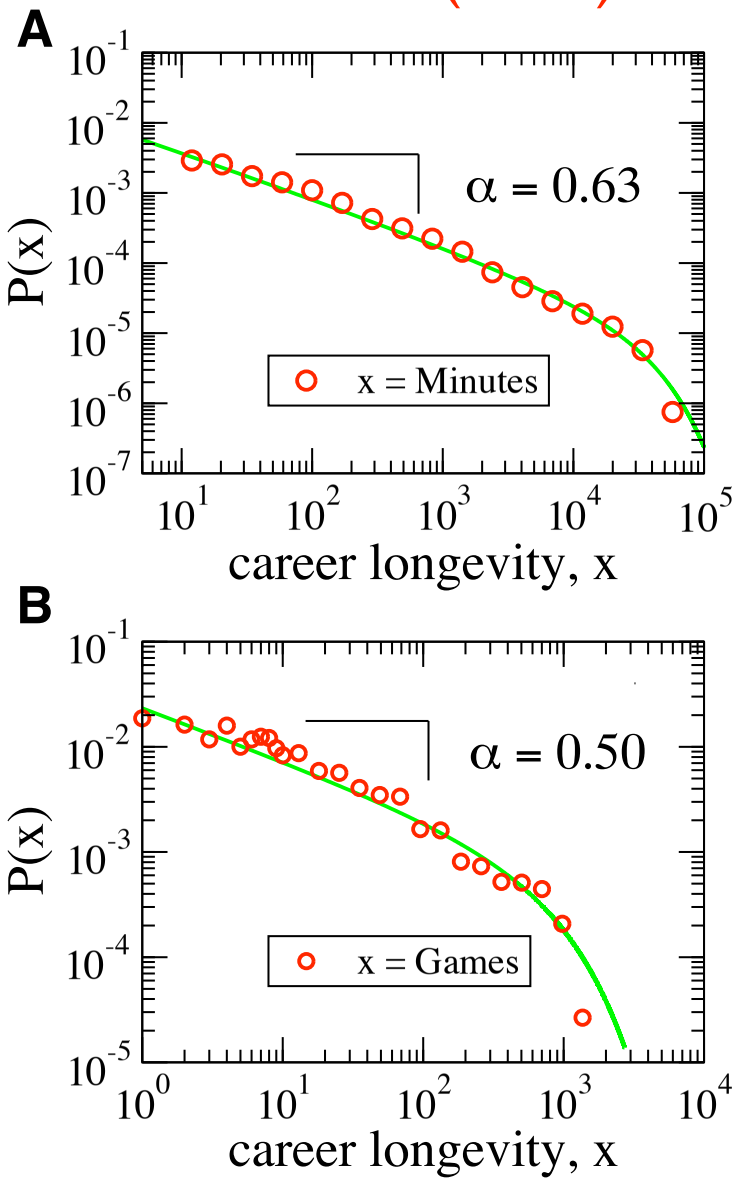
I

Quantifying career longevity & success in sports:

From Cap Anson and Robert Parish to Babe Ruth and Michael Jordan



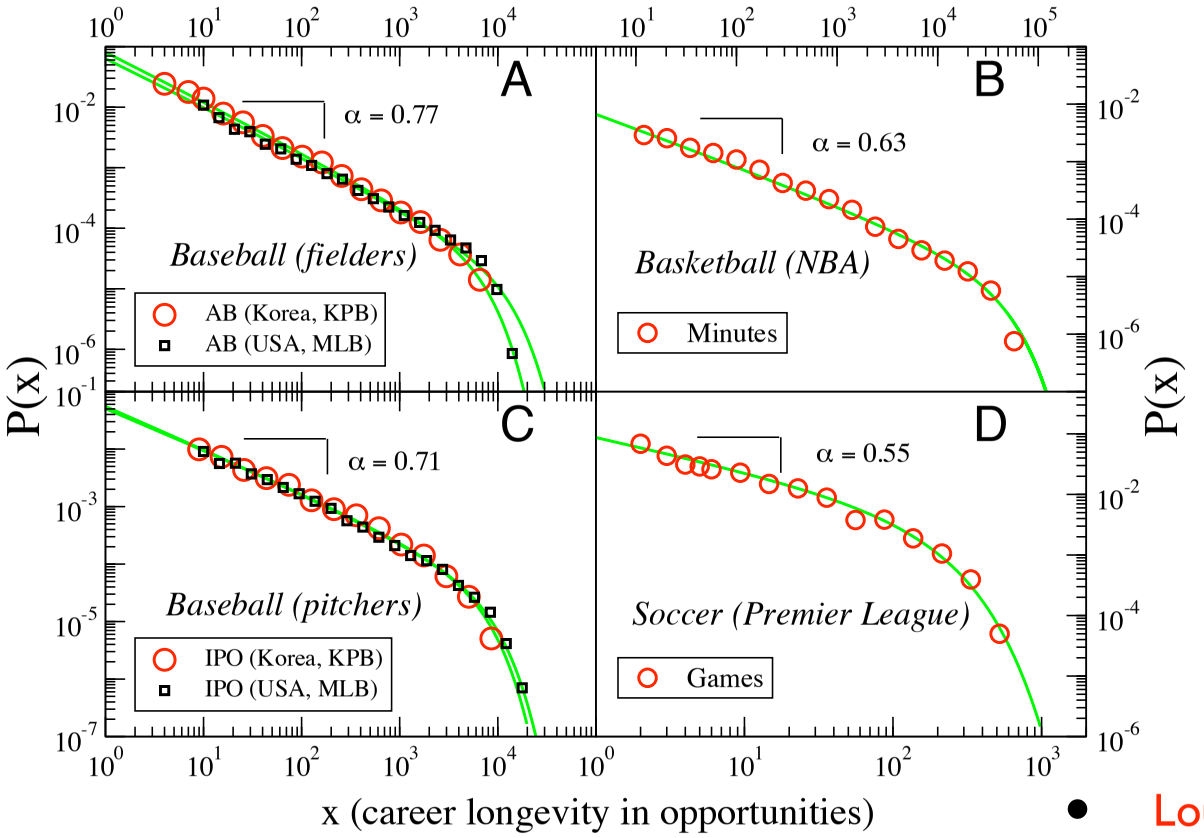
Empirical Results: Career longevity in professional basketball (NBA)



- Analyzed 2700+ completed careers over the 59-yr period 1946-2004
- $x \equiv$ *career longevity* (e.g. min. or games played)
- $P(x)$: probability density function (pdf) of career longevity x
- $P(x)$ is truncated power-law:
  - scaling exponent $\alpha \lesssim 1$
  - Exponential cutoff x_c: Finite-lifetime
  - Scale Free behavior: $P(x_1)/P(x_2) \cong (x_2/x_1)^{\alpha}$ *for* $x < x_c$

- 3% of players played between 1-12 minutes in their entire career! However, the average career length is approx. <x> = 6,500 min., Max(x) = 57,446 min. (Kareem A.-Jabbar)
- 2% of players played in only 1 game in their entire career! <x> = 273 games ~ 3 seasons, Max(x) = 1,611 games (R. Parish)

Career Longevity in 4 sports leagues



opportunities ~ time duration

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!

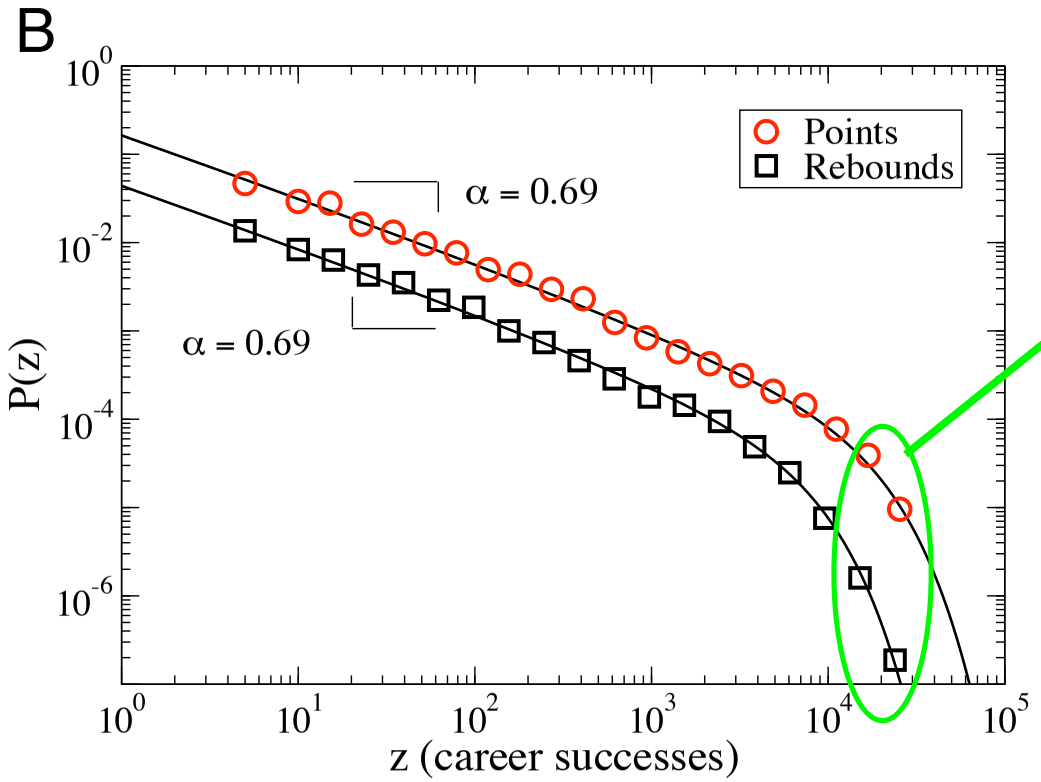
``Iron horses"

- 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

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

Implications of longevity on career success

American Basketball (NBA + ABA): 1946-2004

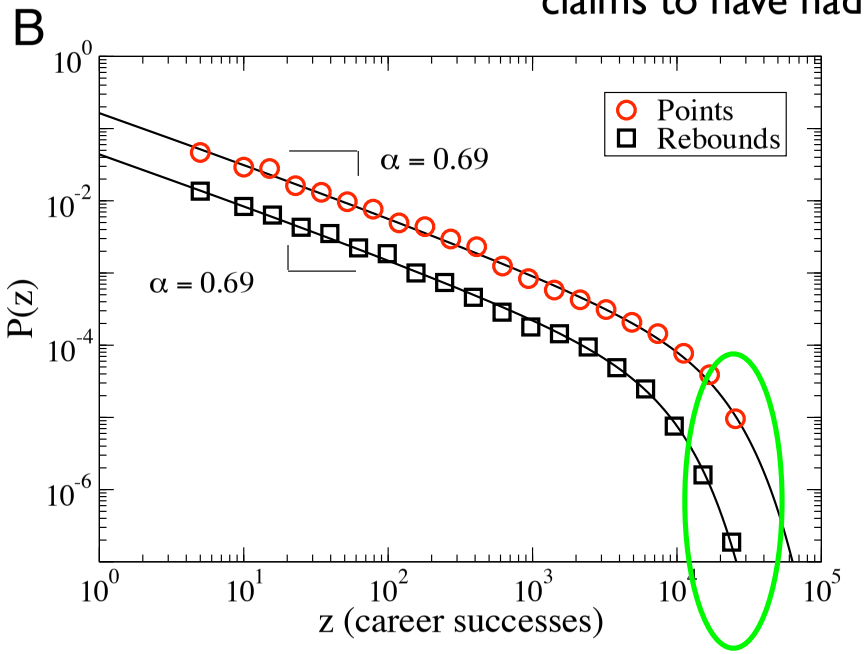


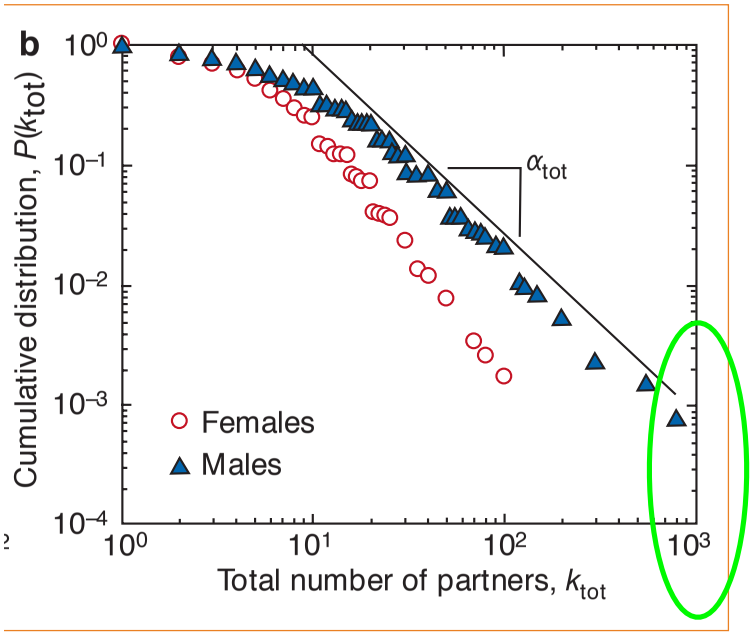
z = career success total ∝ career longevity

- Career longevity exponents α carry over naturally into career statistics

Right-skewed phenomena in the social sciences

Wilt
Chamberlain:
claims to have had over 20,000+ sexual partners....





F. Liljeros, et al.,"*The web of human sexual contacts,*" *Nature* **411**, 907 (2001)

"superstars" are not outliers, but are predicted and consistent with empirical heavy tailed distributions

11

Quantifying success and productivity in science

"Mathletes"

Publication careers of individual scientists within individual journals

Phys. Rev. Lett. 42, 673–676 (1979)

Scaling Theory of Localization: Absence of Quantum Diffusion in Two Dimensions

Abstract References Citing Articles (2,099) Page

Download: PDF (622 kB) Buy this article Export: BibTeX or EndNote (RIS)

E. Abrahams
Serin Physics Laboratory, Rutgers University, Piscataway, New Jersey 08854

P. W. Anderson*, D. C. Licciardello, and T. V. Ramakrishnan†
Joseph Henry Laboratories of Physics, Princeton University, Princeton, New Jersey 08540

PRL milestone Received 7 December 1978; published in the issue dated 5 March 1979

Arguments are presented that the T=0 conductance G of a disordered electronic system depends on its length scale L in a universal manner. Asymptotic forms are obtained for the scaling function $\beta(G)=d\ln G/d\ln L$, valid for both $G \ll G_c \simeq e^2/\hbar$ and $G \gg G_c$. In three dimensions, G_c is an unstable fixed point. In two dimensions, there is no true metallic behavior; the conductance crosses over smoothly from logarithmic or slower to exponential decrease with L.

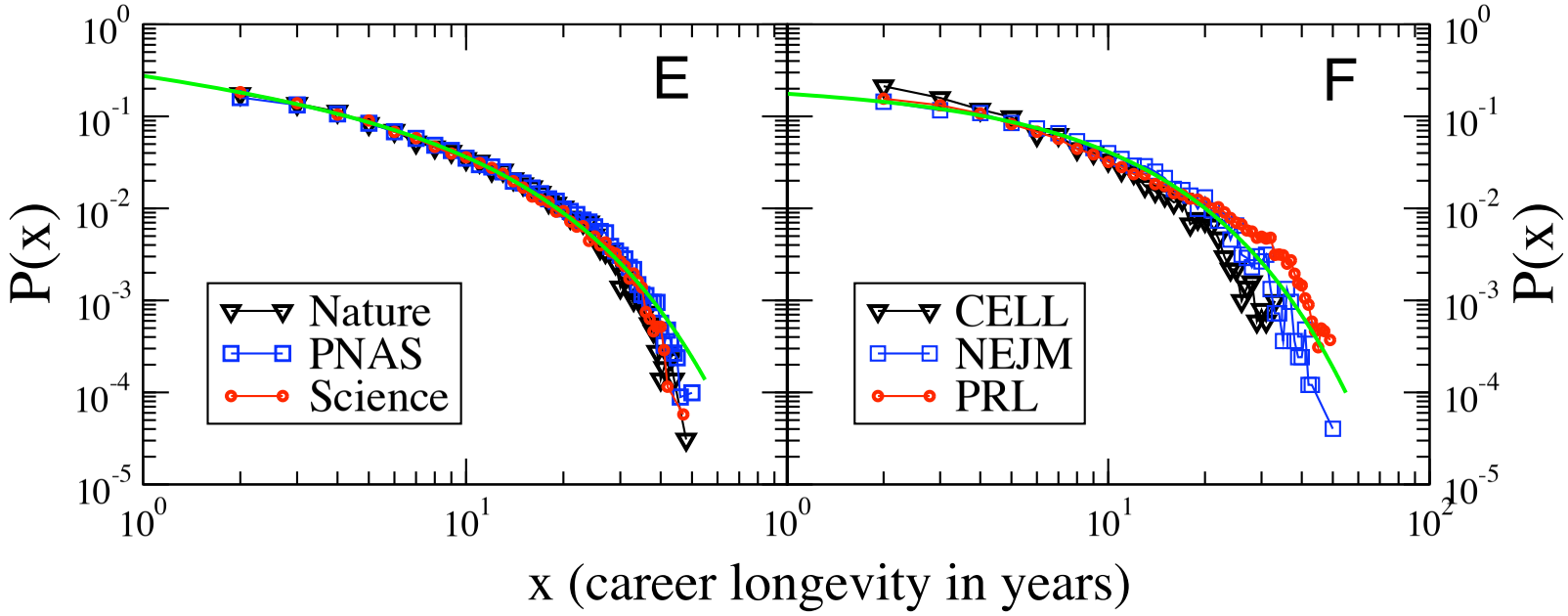
© 1979 The American Physical Society

For example, P.W. Anderson:
(n = 64 articles
published in *PRL* over this
51-year period)

TABLE I. Summary of data set size for each journal. Total number N of unique (but possibly degenerate) name identifications.

Journal	Years	Articles	Authors, N
CELL	1974–2008	53290	31918
NEJM	1958–2008	17088	66834
Nature	1958–2008	65709	130596
PNAS	1958–2008	84520	182761
PRL	1958–2008	85316	112660
Science	1958–2008	48169	109519

Career longevity in academia

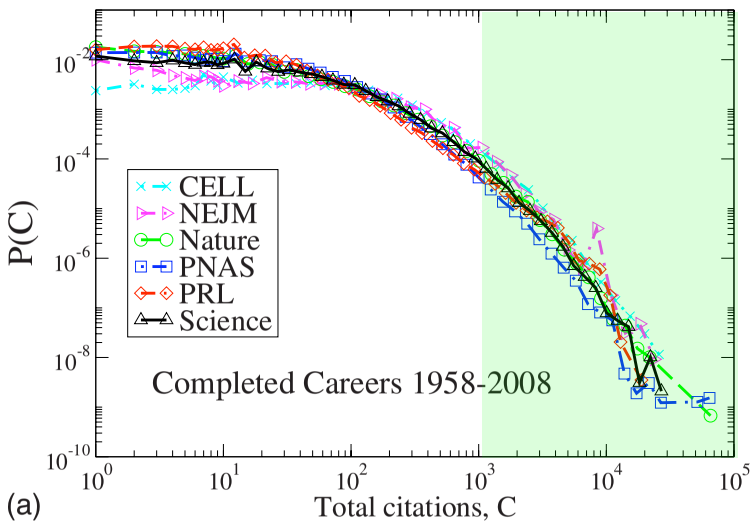


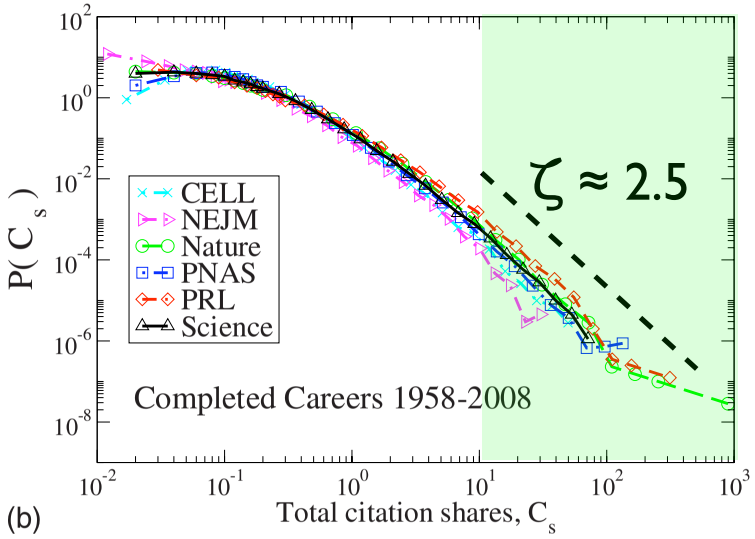
*A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley, "Quantitative and empirical demonstration of the Matthew effect in a study of career longevity." Proc. Natl. Acad. Sci. USA **108**, 18-23 (2011).*

- Each author i has n articles in a given journal j. As a proxy for career longevity in academia, we define the journal longevity x as the number of years separating his/her first and last publication in journal j :

$$x_{i,j} = y_{i,j}(f) - y_{i,j}(0) + 1$$

Journals as "arenas for competition"





Each author has n articles in a given journal j.
Each article i, published in year y, can be quantified by the number of citations c_i it has received at the time of data extraction.
(May, 2009)

Two possible ways to measure citations:

(i) Total citations: $$C = \sum_{i=1}^{n} c_i .$$

(ii) Total citations ``shares'': $$C_s = \sum_{i=1}^{n} \frac{1}{a_i} \frac{c_i(y)}{\langle c(y) \rangle} .$$

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

Top-20 "champions" of *Physical Review Letters*

Each author has n articles in a given journal j.
Each article i, published in year y, can be quantified by the number of citations c_i it has received at the time of data extraction.
(May, 2009)

Total citations ``shares'':

$$C_s = \sum_{i=1}^{n} \frac{1}{a_i} \frac{c_i(y)}{\langle c(y) \rangle}.$$

PRL		
Name	C_s	n
WEINBERG, S	313.3	49
ANDERSON, PW	137.4	64
WILCZEK, F	120.0	62
TERSOFF, J	105.1	76
HALDANE, FDM	102.3	38
YABLONOVITCH, E	87.5	21
PERDEW, JP	78.3	20
LEE, PA	74.6	76
PENDRY, JB	74.1	29
PARRINELLO, M	72.8	68
FISHER, ME	71.6	67
CIRAC, JI	66.7	97
HALPERIN, BI	66.7	50
RANDALL, L	63.4	14
BURKE, K	63.2	18
JOHN, S	62.8	20
GEORGI, H	61.9	26
CAR, R	59.8	51
GLASHOW, SL	59.6	37
CEPERLEY, DM	58.9	39

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

III

The "righ-get-richer" Matthew Effect:

"For to all those who have, more will be given, and they will have an abundance"

Gospel of St. Matthew 25: 29

A possible explanation: the Matthew Effect

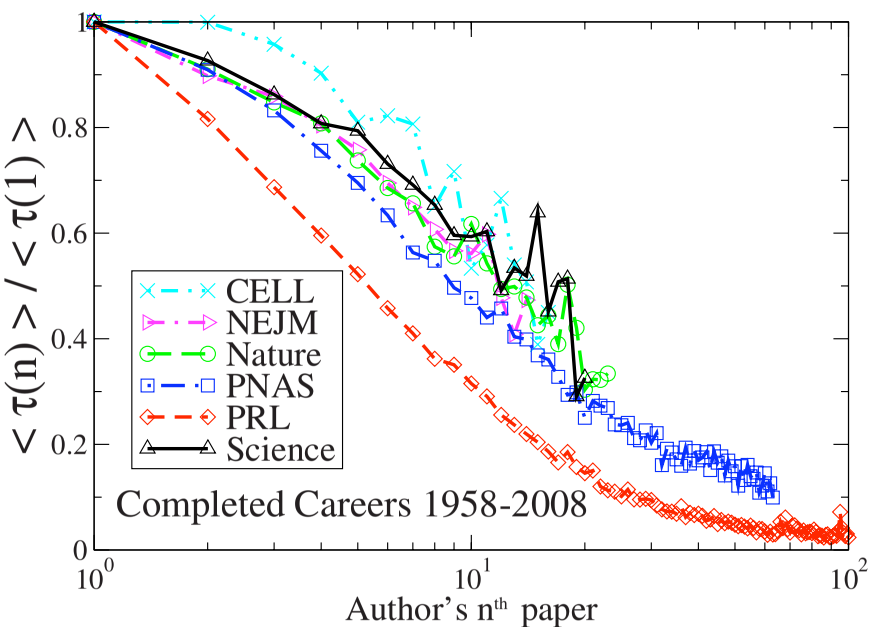


FIG. 7. (Color online) A decreasing waiting time $\tau(n)$ between publications in a given journal suggests that a longer publication career (larger n) facilitates future publications, as predicted by the Matthew effect. We plot $\langle\tau(n)\rangle/\langle\tau(1)\rangle$, the average waiting time $\langle\tau(n)\rangle$ between paper n and paper $n+1$, rescaled by the average waiting time between the first and second publication, $\langle\tau(1)\rangle$. The values of $\langle\tau(1)\rangle$ are 2.2 (*CELL*, *PRL*), 3.0 (*Nature*, *PNAS*, *Science*), and 3.5 (*NEJM*) years. *Physical Review Letters* exhibits a more

- For a given journal: the waiting time $\tau(n)$ is the number of years between an author's paper n and paper $n+1$
- A decreasing $\tau(n)$ indicates that it becomes "easier" to publish in a journal with each successive publication

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

A stochastic model for career longevity

- Ingredient I: Random forward progress

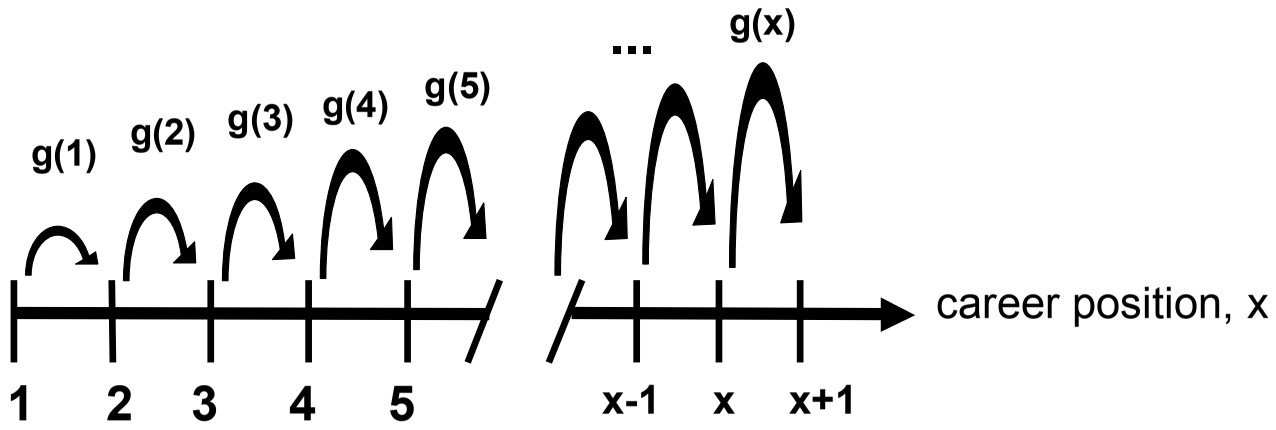
  Experience and reputation can provide positive feedback in sustaining a career (generic "rich-get-richer" effect)

- Ingredient II: Random termination time

  Career must survive through a horizon of hazards which eventually terminate the career

Ingredient 1: Random forward progress

- Forward progress is made according to the "progress rate" $g(x)$
- Matthew Effect: $g(x)$ increases with career position x



$P(x\,,\,t)$ = probability that career is at position x at time t

Master Equation approach →

Poisson Distribution

$$P(x,t) = \frac{e^{-\lambda t}(\lambda t)^{x-1}}{(x-1)!}$$

$$\lambda \equiv g(x)$$

Ingredient II: Random Termination Time

- Termination of career occurs for many reasons:
  career position at termination time T = career longevity

- Average *pdf* $P(x \mid T)$ over *pdf* $r(T)$ of termination (exit) times in

$$P(x) = \int_0^\infty P(x|T)P(T)$$

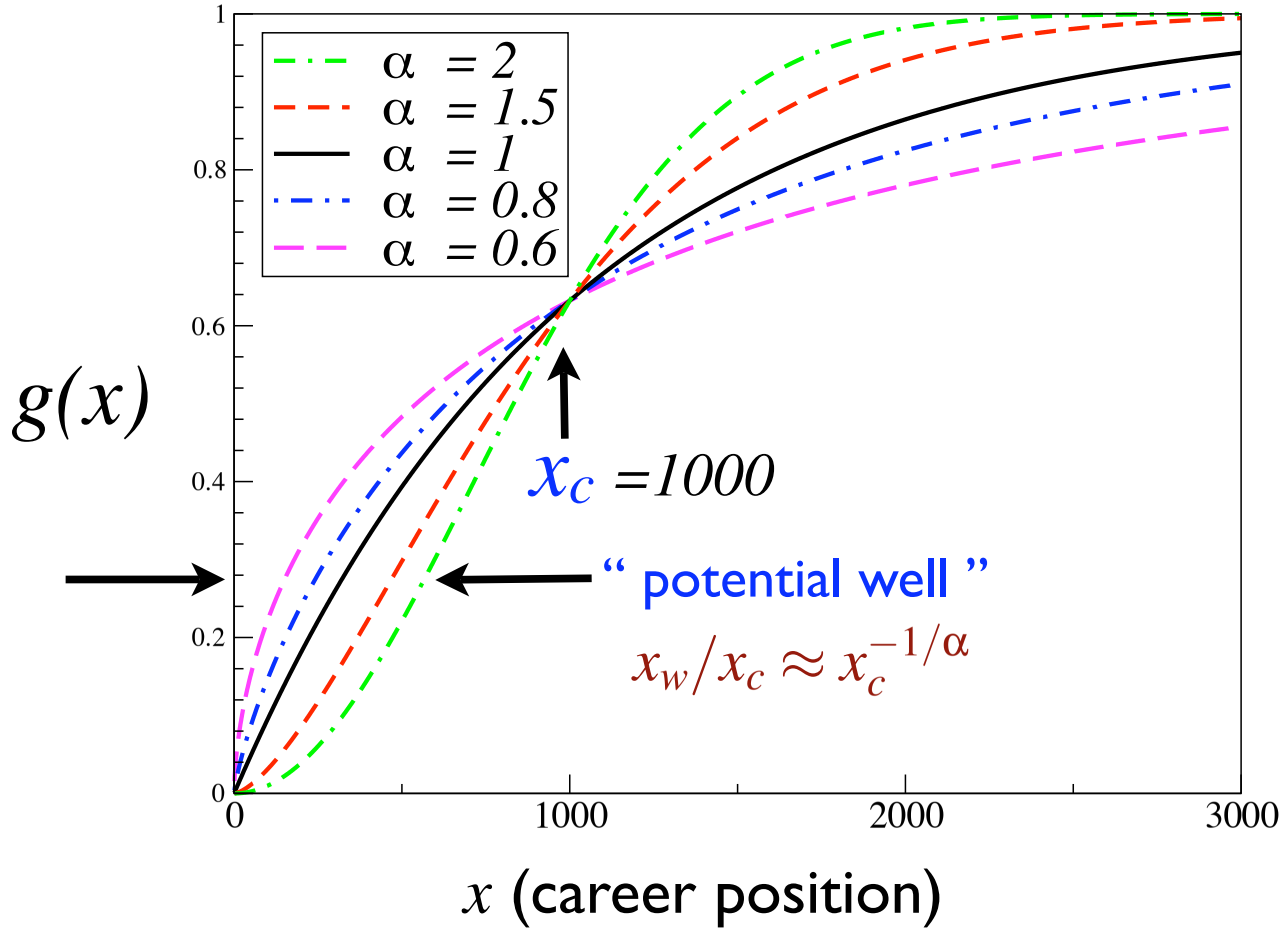
- Hazard rate $H(T)$: conditional probability that failure will occur at time $(T + \delta T)$ given that failure has not yet occurred at time T

$$H(T) = \frac{r(T)}{S(T)} = -\frac{\partial}{\partial T}\ln S(T) \qquad S(T) = 1 - \int_0^T r(t)dt$$

- So we choose an $r(T)$ that reflects constant hazards: $H(T) = 1/\mathrm{x_c}$
  which corresponds to :

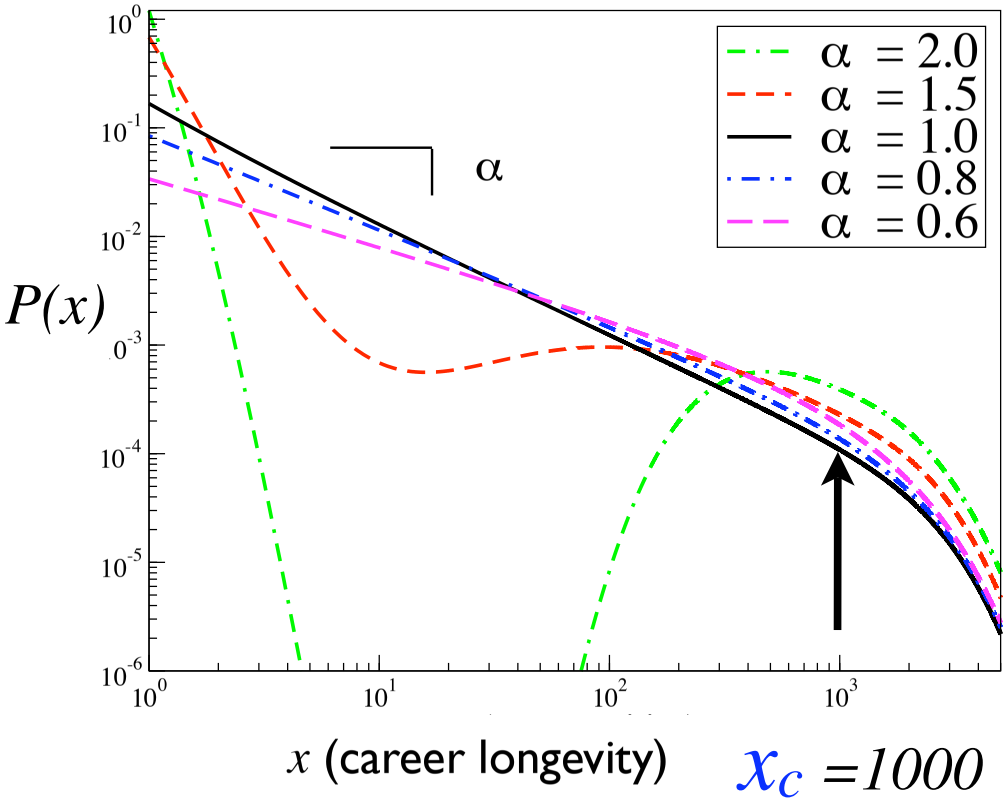
$$r(T) = \exp[-T/x_c]/x_c$$

Progress rate: $g(x) \equiv 1 - e^{-(x/x_c)^{\alpha}}$



- $x_c \equiv$ career position time-scale which separates veterans from newcomers.
- $\alpha \equiv$ quantifies the rate at which an individual climbs the "career ladder" : $g(x) \sim x^{\alpha}$ *for* $x \ll x_c$

Progress rate $g(x)$ → Career Longevity *pdf* $P(x)$



$$P(x) = \frac{g(x)^{x-1}}{x_c\left[\frac{1}{x_c} + g(x)\right]^x} \approx \frac{1}{g(x)x_c}\, e^{-\frac{x}{g(x)x_c}}$$

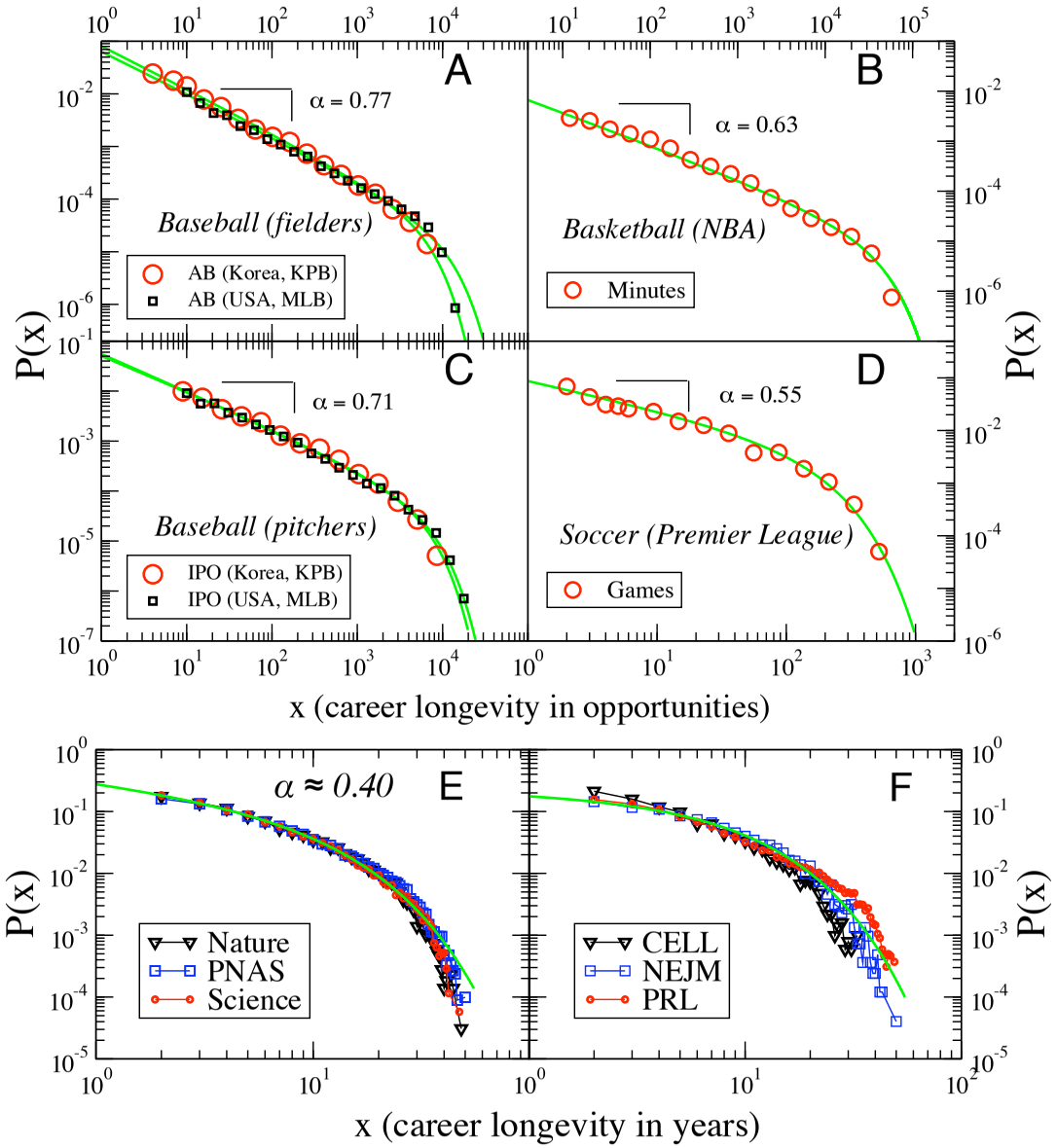
for convex $\alpha > 1$:

Bimodal

for concave $\alpha < 1$:

$$P(x) \propto \begin{cases} x^{-\alpha} & x < x_c \\ e^{-(x/x_c)} & x > x_c \end{cases}$$

- $\alpha \equiv$ power-law exponent for career longevity, which is intrinsically related to the rate at which individuals establish their reputation and secure future opportunity based on prior success.

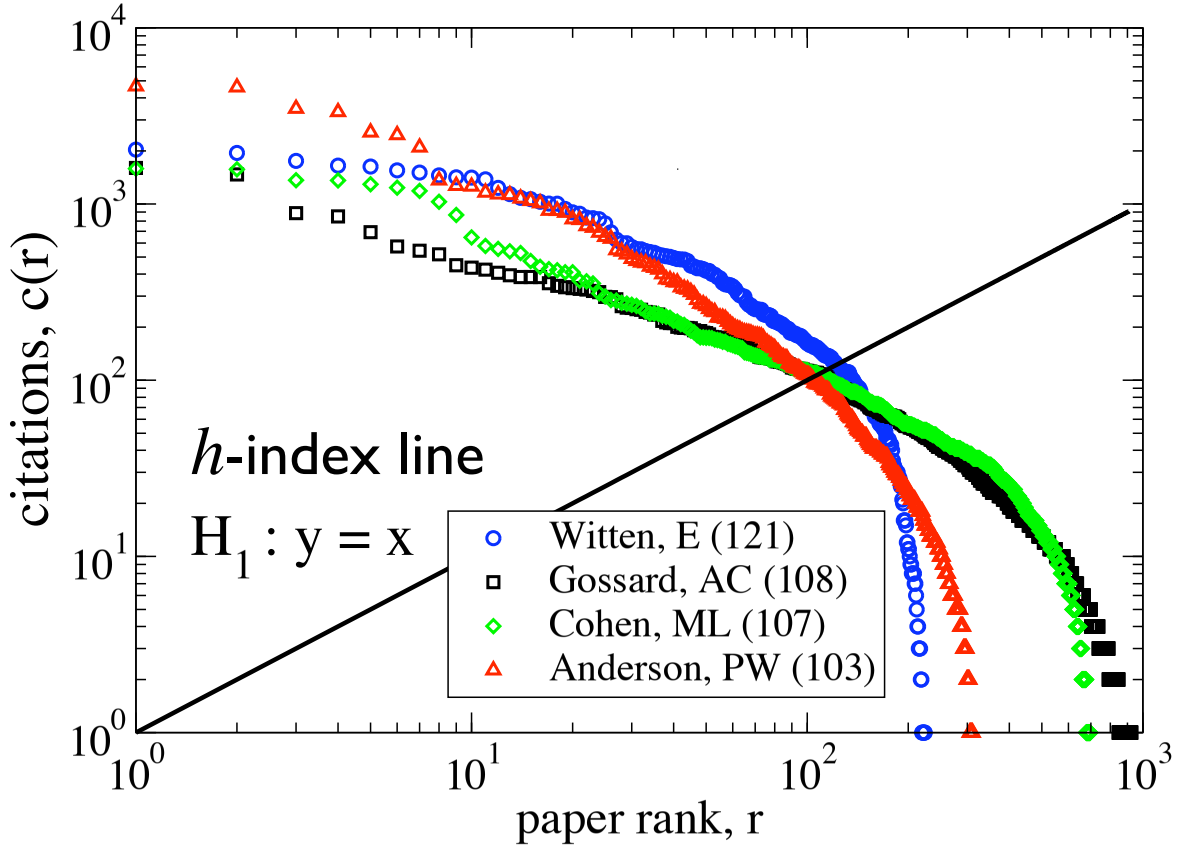


*A. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley, "Quantitative and empirical demonstration of the Matthew effect in a study of career longevity." Proc. Natl. Acad. Sci. USA **108**, 18-23 (2011).*

IV
How popular are your papers?

S. Redner, "How popular is your paper? An empirical study of the citation distribution."
Eur. Phys J. B (1998).

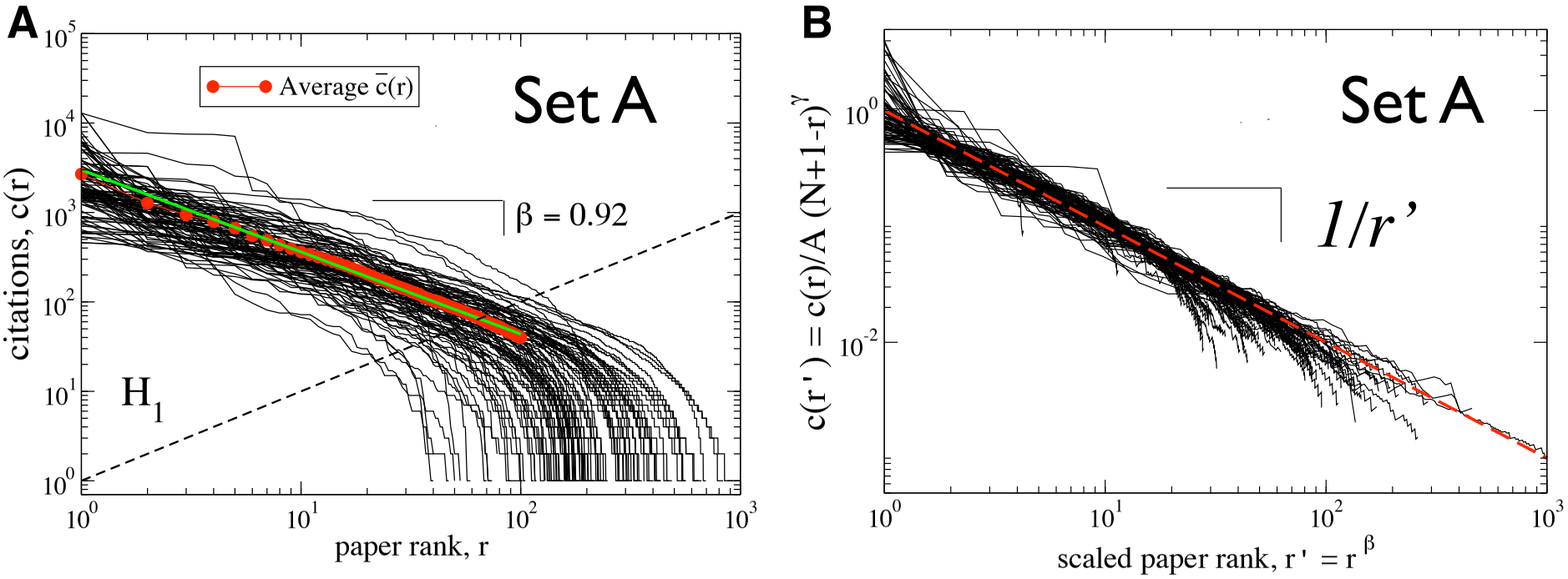
A closer look at scientific careers: the rank-citation profile $c_i(r)$



$c_i(r)$ is the rank-ordered (Zipf) citation distribution of the N papers published by individual i in his/her entire career

A. M. Petersen, H. E. Stanley, S. Succi. "Statistical regularities in the rank-citation profile of scientists." Under review.

A comparison of $c_i(r)$ the top-100 "champions" of PRL (Set A) with average h-index $\langle h \rangle = 61 \pm 21$



Discrete Generalized Beta Distribution(DGBD):

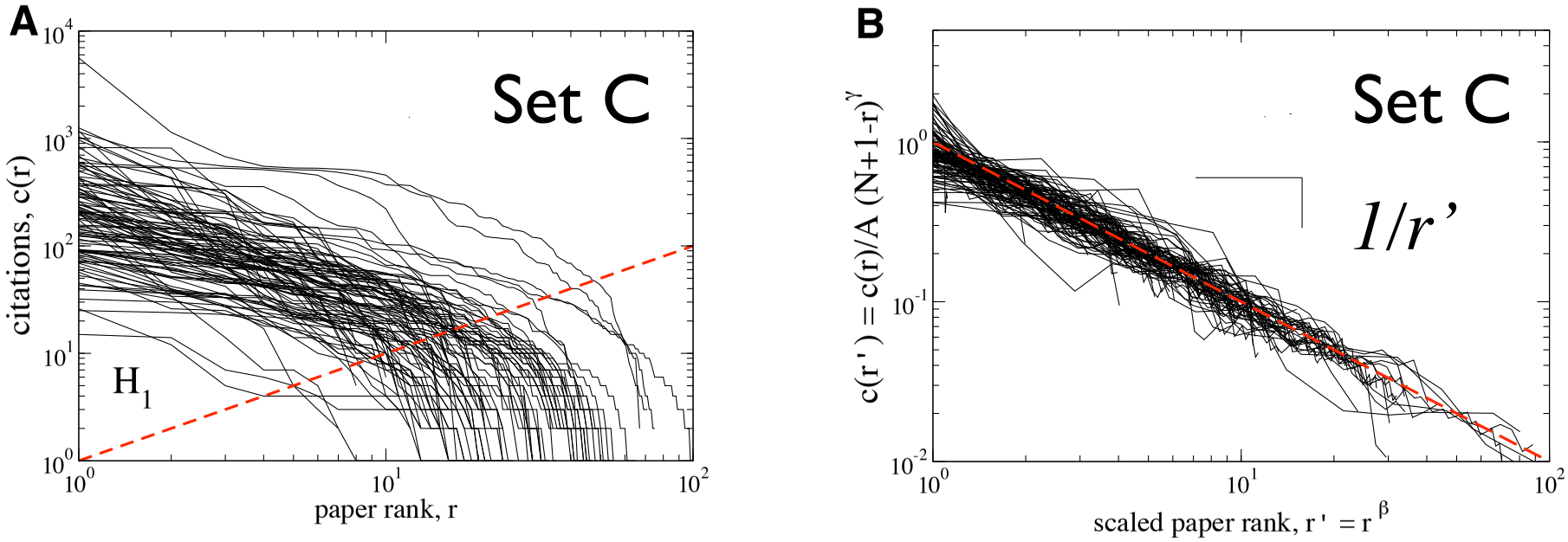
$$c(r) \equiv A r^{-\beta} (N + 1 - r)^{\gamma} .$$

Martinez-Mekler, et al. "Universality of rank-ordering distributions in the arts and sciences." PLoS ONE 4: e4791 (2009).

Average values of the DGBD model parameters:

$$\langle \beta \rangle = 0.83 \pm 0.23 \quad \text{and} \quad \langle \gamma \rangle = 0.67 \pm 0.19$$

Common functional form also describes even Assistant Professors with average h-index $\langle h \rangle = 15 \pm 7$



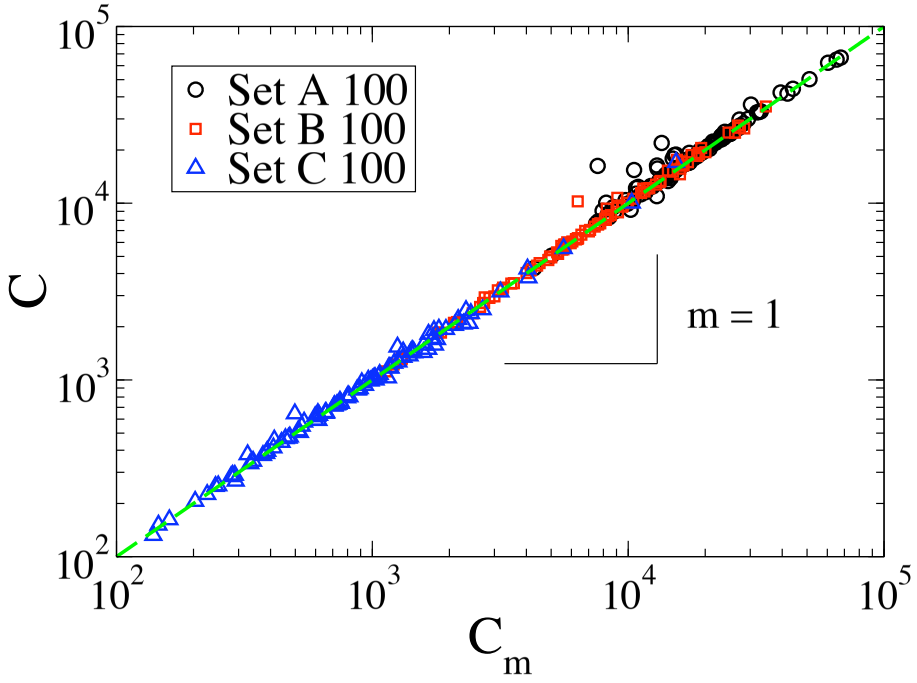
Set C: 100 Asst. Professors, 2 chosen from each of the top-50 U.S. physics departments

Average values of the DGBD model parameters:

$\langle \beta \rangle = 0.79 \pm 0.38$ and $\langle \gamma \rangle = 0.89 \pm 0.36$

Further validation of the DGBD model, comparing the *predicted* and *actual* total number of citations, C_i

$$C_i = \sum_{r=1}^{r=N} c_i(r)$$



$$C_{i,m} = \sum_{r=1}^{r=N} A r^{-\beta} (N + 1 - r)^{\gamma}$$

Scaling relation between C, h, and β

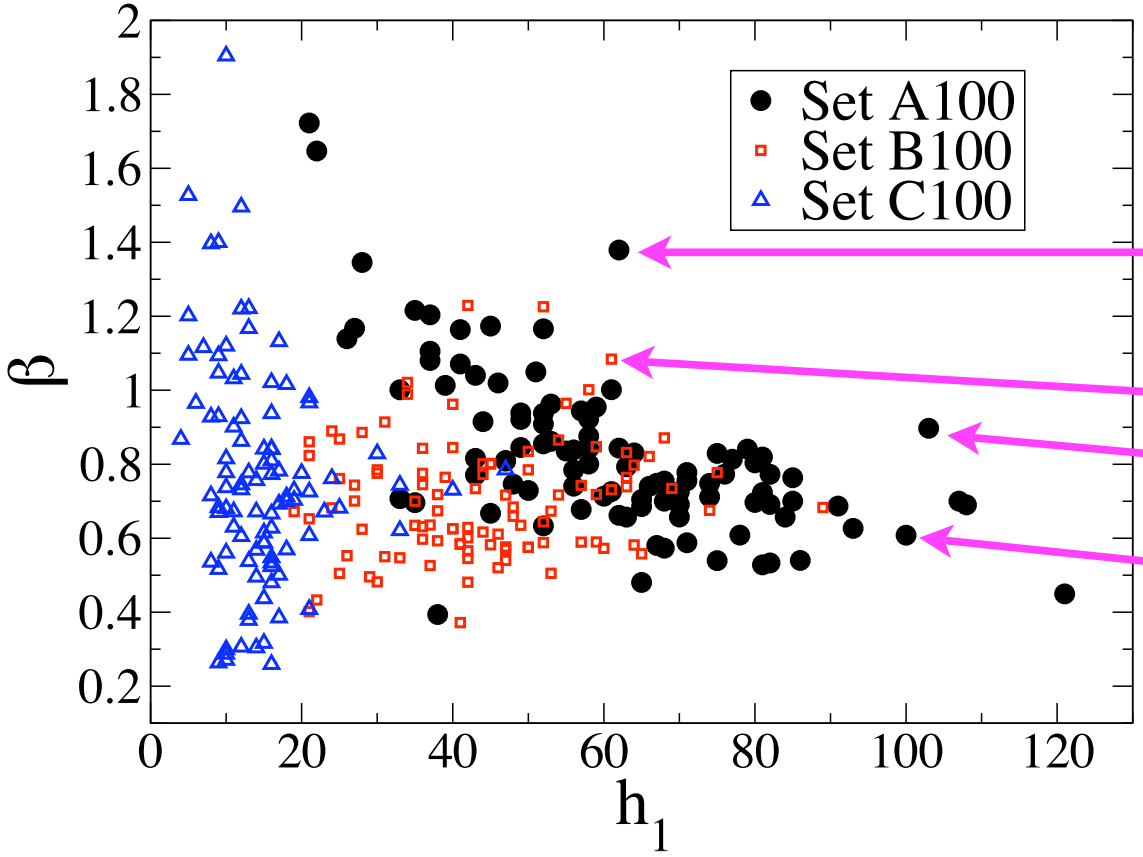
$$C_{\beta,h} \sim h^{1+\beta}$$

$$\downarrow \quad \beta \cong 1$$

* $C \approx 4h^2$

* S. Redner, "On the meaning of the h-index." J. Stat. Mech. 2010, L03005 (2010).

The β -vs- h parameter space



Matthias Ernzerhof (U. Montreal)
"Generalized gradient approximation made simple"
Perdew, JP; Burke, K; Ernzerhof, M
PRL 1996 (c = 16,314)

A.-L. Barabasi

P. W. Anderson

H. E. Stanley

For a given h, a large β value corresponds to a larger total citations,

$C_i \sim h^{1+\beta}$,

which is a proxy for career publication impact

Take home messages

- There is a beautiful statistical regularity that "bridges the gap" between the relatively short careers and the extremely long "stellar" careers.
    - Stellar careers are not an anomaly! They are predicted by pdf $P(x)$
    - The probability density function $P(x)$ corresponds to an exponentially truncated power-law with scaling exponent $\alpha \lesssim 1$
- The Matthew "rich-get-richer" effect can be used to explain the extremely right-skewed probability distributions that quantify both longevity and success.
    - evidence in the decreasing time duration $\tau(n)$ between publications and a model that predicts two classes of $P(x)$ depending on the choice of $g(x)$
- Quantifying the rank-citation profile $c_i(r)$ of individual scientists can provide a comprehensive evaluation of career impact and productivity.
  Moreover, it is surprising that all careers analyzed have common functional form, the Discrete Generalized Beta Distribution (DGBD)!
- There are many analogies between the superstars in science and the superstars in professional sports, possibly arising from the generic aspects of competition.

Thank You!

Also, a special thanks to my collaborators:
Woo-Sung Jung, Orion Penner, Gene Stanley, Sauro Succi, Fengzhong Wang, and Jae-Sook Yang
and to my Committee Members:
Plamen CH. Ivanov, Emanuel Katz, Anatoli Polkovnikov, William J. Skocpol, H. Eugene Stanley

I) A. M. Petersen, W.-S. Jung, H. E. Stanley, "On the distribution of career longevity and the evolution of home run prowess in professional baseball." Europhysics Letters ***83,*** *50010 (2008).*

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

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

IV) A. M. Petersen, O. Penner, H. E. Stanley, "Methods for detrending success metrics to account for inflationary and deflationary factors." Eur. Phys. J. B ***79****, 67 (2011). Pre-print title: Detrending career statistics in professional Baseball: accounting for the Steroids Era and beyond.*

V) A. M. Petersen, H. E. Stanley, S. Succi. "Statistical regularities in the rank-citation profile of scientists". Under review. (2011)

Least-square estimation of parameter values

TABLE S2: Data summary for the pdfs of career statistical metrics. The values α and x_c are determined for each career longevity pdf $P(x)$ and each career success pdf $P(z)$ via least-squares method using the functional form given by Eq. [5]. We calculate the Gamma pdf average $\langle x \rangle$, the standard deviation σ, and the extreme threshold value x^* at the $f = 0.019$ significance level using the corresponding values of α and x_c. The units for each metric are indicated in parenthesis alongside the league in the first column.

For publication distributions, the career longevity metric x is measured in years.

Professional League, (success metric)	Least-square values α	x_c	Gamma pdf values $\langle x \rangle$	σ	x^*	$\frac{x^*}{\langle x \rangle}$	$\frac{x^*}{\sigma}$
MLB, (H)	0.76 ± 0.02	1240 ± 150	300	610	2400	7.8	3.9
MLB, (RBI)	0.76 ± 0.02	570 ± 80	140	280	1100	7.8	3.9
NBA, (Pts)	0.69 ± 0.02	7840 ± 760	2400	4400	17000	7.0	3.9
NBA, (Reb)	0.69 ± 0.02	3500 ± 130	1100	2000	7600	6.9	3.9

Calculating milestone values based on player entry into the National Baseball Hall of Fame

$$x^* : \quad \int_{x^*}^{\infty} P(x)dx = f = 0.019$$

Professional League, (opportunities)	Least-square values α	x_c	Gamma pdf values $\langle x \rangle$	σ	x^*	$\frac{x^*}{\langle x \rangle}$	$\frac{x^*}{\sigma}$
KBB, (AB)	0.78 ± 0.02	2600 ± 320	580	1200	4700	8.2	3.9
MLB, (AB)	0.77 ± 0.02	5300 ± 870	1200	2500	9700	8.1	3.9
MLB, (IPO)	0.72 ± 0.02	3400 ± 240	950	1800	6900	7.3	3.9
KBB, (IPO)	0.69 ± 0.02	2800 ± 160	840	1500	5900	7.0	3.9
NBA, (Min)	0.64 ± 0.02	20600 ± 1900	7700	12600	48800	6.4	3.9
UK, (G)	0.56 ± 0.02	138 ± 14	61	92	360	5.8	3.9

Academic Journal, (career length in years)	Least-square values α	x_c
Nature	0.38 ± 0.03	9.1 ± 0.2
PNAS	0.30 ± 0.02	9.8 ± 0.2
Science	0.40 ± 0.02	8.7 ± 0.2
CELL	0.36 ± 0.05	6.9 ± 0.2
NEJM	0.10 ± 0.02	10.7 ± 0.2
PRL	0.31 ± 0.04	9.8 ± 0.3

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

Quantifying statistical regularities in the career achievements of scientists and professional athletes

Abstract:

For many professions, the quantitative analysis of individual careers is made difficult by the lack of comprehensive data and the difficulty in defining measures for productivity and longevity. However, comprehensive career data is recorded in professional sports and is perfectly tailored for studying human productivity. Similarly, the publication careers of scientists are also quantifiable using similar measures. Since both professions are subject to the common forces of competition, one motivating question in this talk is: "What are the statistical regularities in career achievement across an entire cohort of competitors?"

In this talk I will discuss the statistical regularities that describe the everyday topic of career achievement using comprehensive career data. In the first part of the talk, I will discuss the topic of career longevity, using as example the 60+ year history of the National Basketball Association and 2700+ complete careers over the period 1946-2004. Surprisingly, we find that a common career longevity distribution describes the careers of 20,000+ athletes from 4 sports leagues and 400,000+ scientists from 6 high-impact journals, where each journal serves as a generic arena for competition. In order to account for the regularities we observe across several professions, we develop an exactly solvable model for career longevity based on the Matthew "rich-get-richer" effect. Our model is in excellent agreement with empirical career longevity distributions for each profession analyzed. Our model follows from two general assumptions: (i) that there is random forward progress in the career, whereby it becomes easier to make progress the further along one is in his/her career, and (ii) that career termination follows from random hazards that are present throughout the career. The findings suggests that there is a common underlying mechanism which underlies career development in competitive professions. In the second part of the talk, I will discuss the publication careers of 300 individual scientists (ranging from very the very famous to current Assistant professors) and find remarkable statistical regularity in the functional form of the rank-citation distribution (analogous to the Zipf rank-frequency distribution) for each scientist studied.