Dynamical graph-based impact metrics

Filippo Radicchi

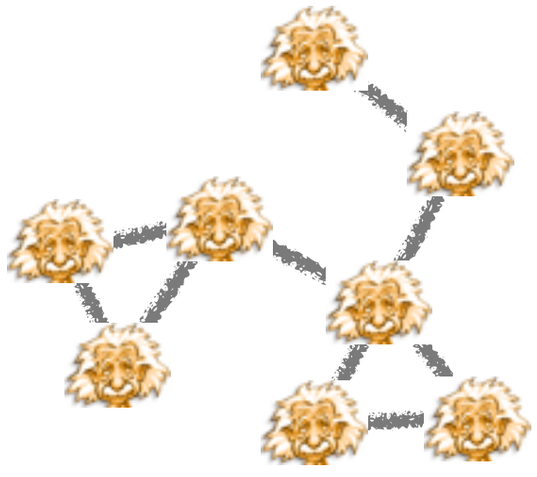
filiradi@indiana.edu filrad.homelinux.org

SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY
Bloomington

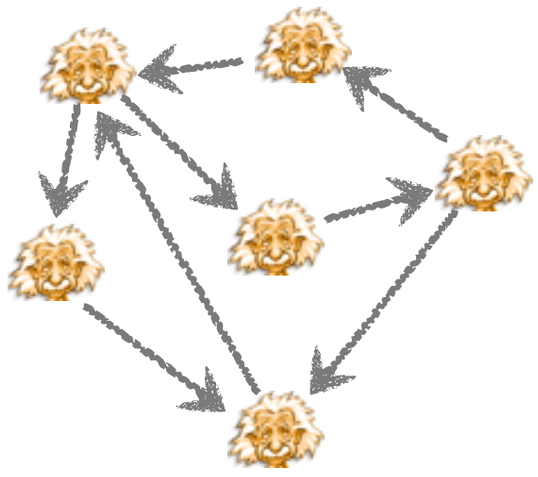
Bibliographic data

only in 2006: 10^4 journals, 10^6 papers, 10^7 references

Scientific motivations

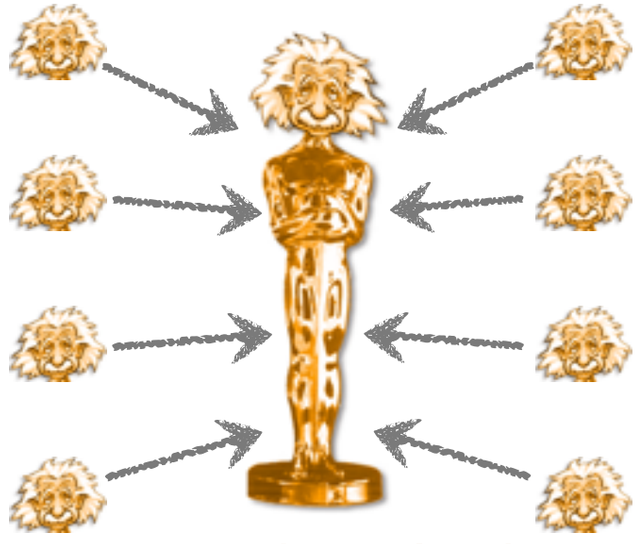


collaboration networks



citation networks

Practical motivations



research evaluation

A practical example

the Italian National Scientific Qualification

1) Number of papers: $I(N_p, A_A) = \frac{10\,N_p}{A_A}$

2) Number of citations: $I(N_C, A_A) = \frac{N_C}{A_A}$

3) Contemporary h-index: $S(i,t_i,t) = \frac{4}{(t-t_i+1)} C(i,t_i,t)$

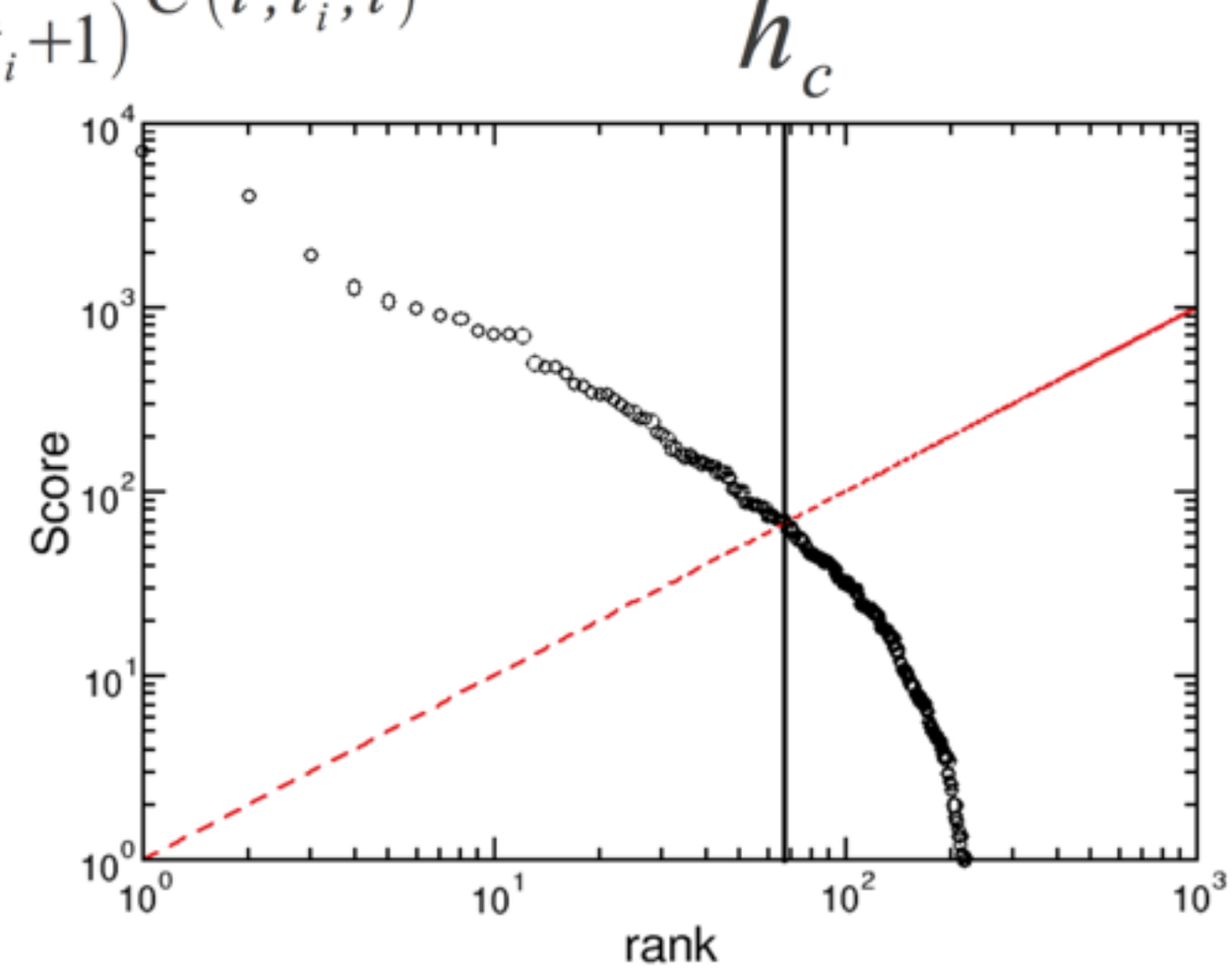
Sidiropoulos A et al. Scientometrics 72, 253 (2007)

N_p total number of publications

A_A academic age

N_C total number of citations

$C(i,t_i,t)$ citations accumulated up to year t by paper i published in year t



source: ”Abilitazione Scientifica Nazionale – La normalizzazione degli indicatori per l'eta' accademica”, ANVUR, Jul. 2012

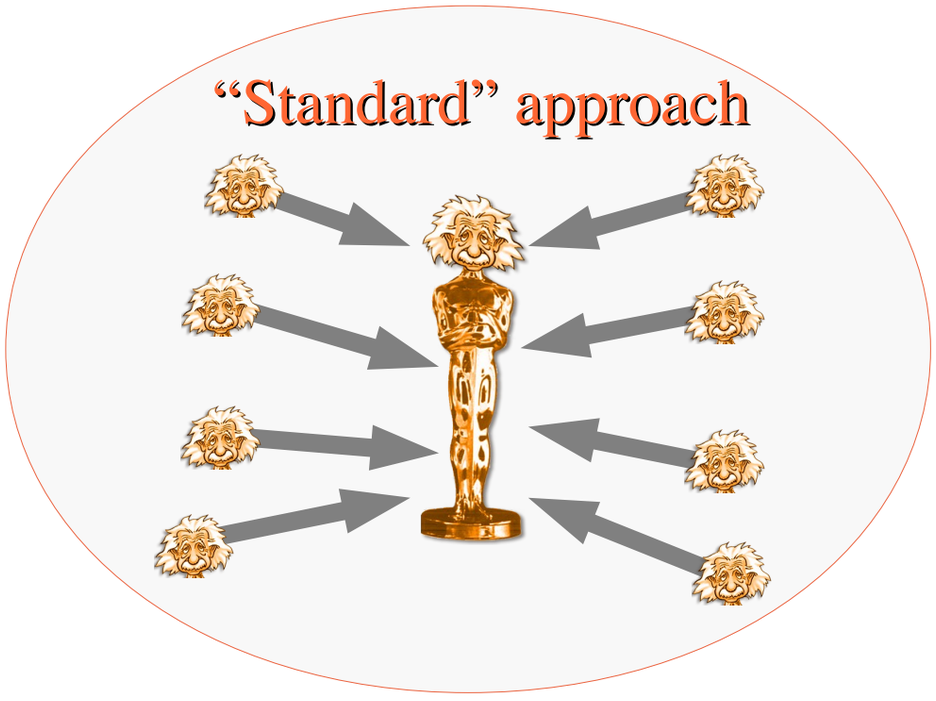
A practical example

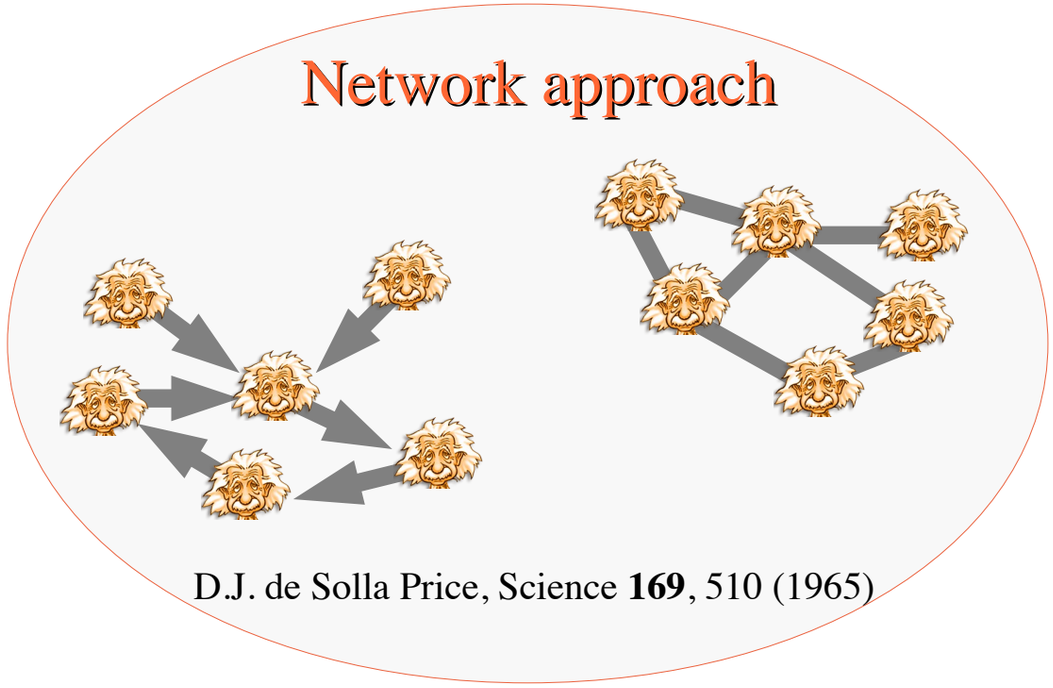
the Italian National Scientific Qualification

		associate professor			full professor		
		norm. pub.s	norm. citations	h-c index	norm. pub.s	norm. citations	h-c index
Mathematics	01/A1	5	1.74	2	4	1.37	2
	01/A2	8	1.65	2	9	3.23	3
	01/A3	10	4.34	4	14	8	5
Physics	02/A1	59.5	104.08	18	78	105.03	22
	02/B2	**37.5**	**40.08**	**11**	**47.5**	**75.94**	**14**
Biology	05/A2	14	24.45	8.5	20	37.47	10
	05/C1	21.5	15.77	8	26	18.63	9
Chemistry	03/A1	26	29.47	9	41	53.81	12
	03/B1	31	47.05	11	49.5	62.38	13

source: ”Mediane per candidati all’abilitazione scientifica nazionale a professore associato/ordinario”, ANVUR, Jul. 2012

The network structure of citation data is often neglected in research evaluation





D.J. de Solla Price, Science **169**, 510 (1965)

papers	Citation counts	CiteRank
journals	Impact factor	Eigenfactor
scientists	h-index, g-index, ...	?

Graph-based ranking of scientists

Physical Review Series I (**PRI**), Physical Review (**PR**), Physical Review Letters (**PRL**), Physical Review A (**PRA**), Physical Review B (**PRB**), Physical Review C (**PRC**), Physical Review D (**PRD**), Physical Review E (**PRE**), Reviews of Modern Physics (**RMP**) between **1893** and **2006**

PHYSICAL REVIEW B VOLUME 23, NUMBER 10 15 MAY 1981

Self-interaction correction to density-functional approximations for many-electron systems

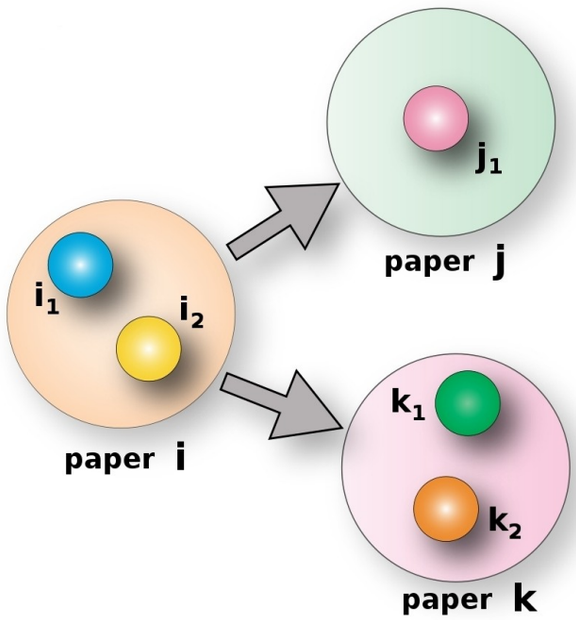
J. P. Perdew
Department of Physics and Quantum Theory Group, Tulane University, New Orleans, Louisiana 70118

Alex Zunger
Solar Energy Research Institute, Golden, Colorado 80401
and Department of Physics, University of Colorado, Boulder, Colorado 80302

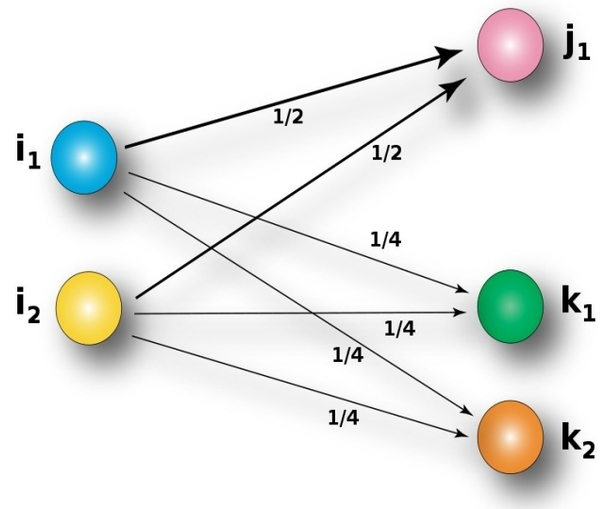
(Received 31 October 1980)

[1]E. Fermi and E. Amaldi, Accad. Ital. Rome 6, 119 (1934).
[2]J. C. Slater and J. H. Wood, Int. J. Quantum Chem. 4, 3 (1971).
[3]N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Holt, Rinehart and Winston, New York, 1976).
[4]A. B. Kunz, Phys. Rev. B 12, 5890 (1975).
[5]J. C. Slater, *The Self-Consistent Field for Molecules and Solids* (McGraw-Hill, New York, 1974).
[6]P. Hohenberg and W. Kohn, Phys. Rev. 136, B864 (1964).
[7]W. Kohn and L. J. Sham, Phys. Rev. 140, A1133 (1965).
[8]U. von Barth and L. Hedin, J. Phys. C 5, 1629 (1972). Also A. K. Rajagopal and J. Callaway, Phys. Rev. B 7, 1912 (1973).
[9]O. Gunnarsson, B. I. Lundqvist, and J. W. Wilkins, Phys. Rev. B 10, 1319 (1974).
[10]O. Gunnarsson, J. Harris, and R. O. Jones, J. Chem. Phys. 67, 3970 (1977).
[11]B. I. Dunlap, J. W. Connolly, and J. R. Sabin, J. Chem. Phys. 71, 4993 (1979).
[12]V. L. Moruzzi, J. F. Janak, and A. R. Williams, *Calculated Electronic Properties of Metals* (Pergamon, New York, 1978).

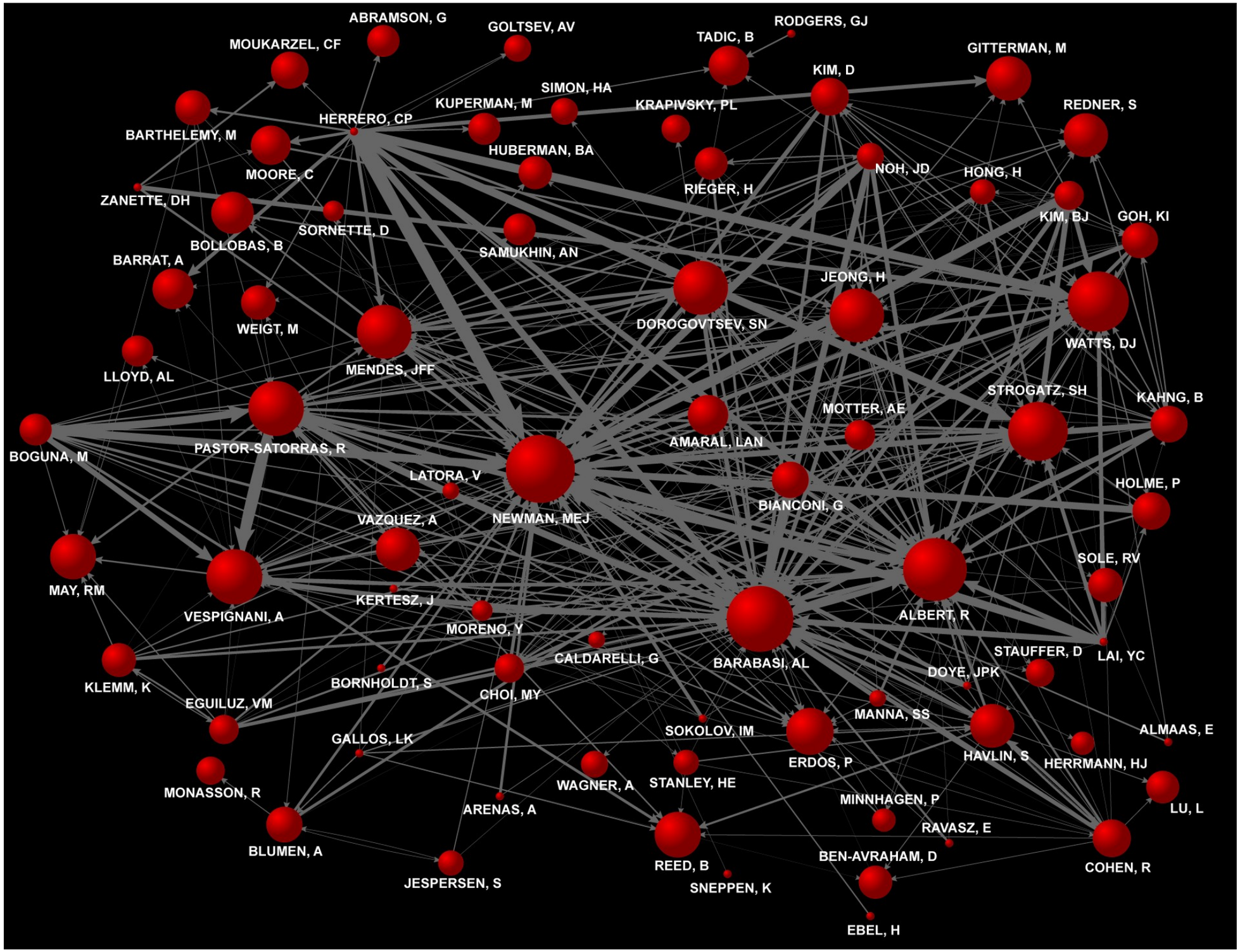
Paper Citation Network



Weighted Author Citation Network



Weighted author citation network

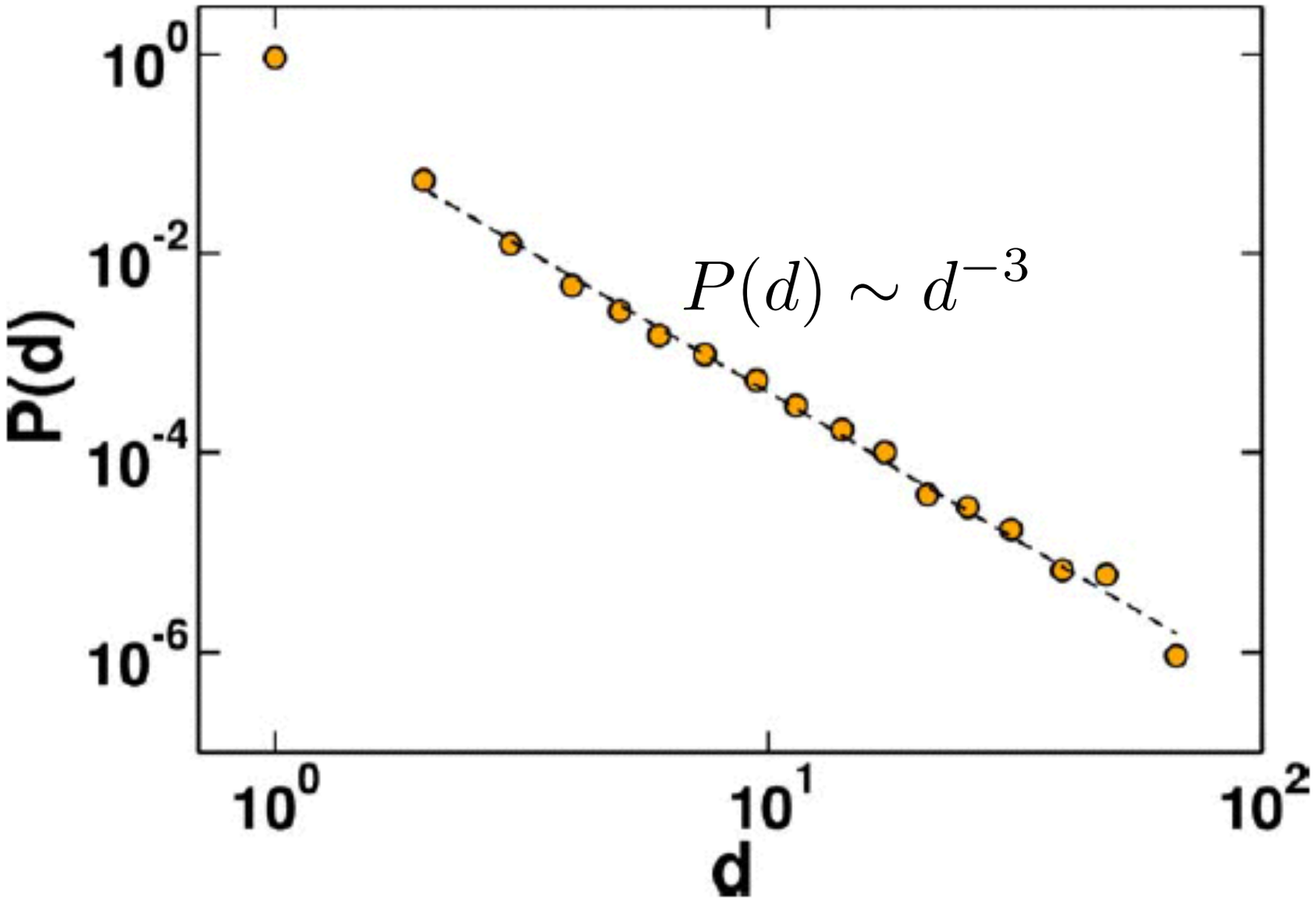


key-words: ”**complex network**” , ”**scale-free network**”, ”**small-world network**”, etc..

We didn't perform any disambiguation

Last name, First name → Last name, Initials

Radicchi, Filippo → Radicchi, F



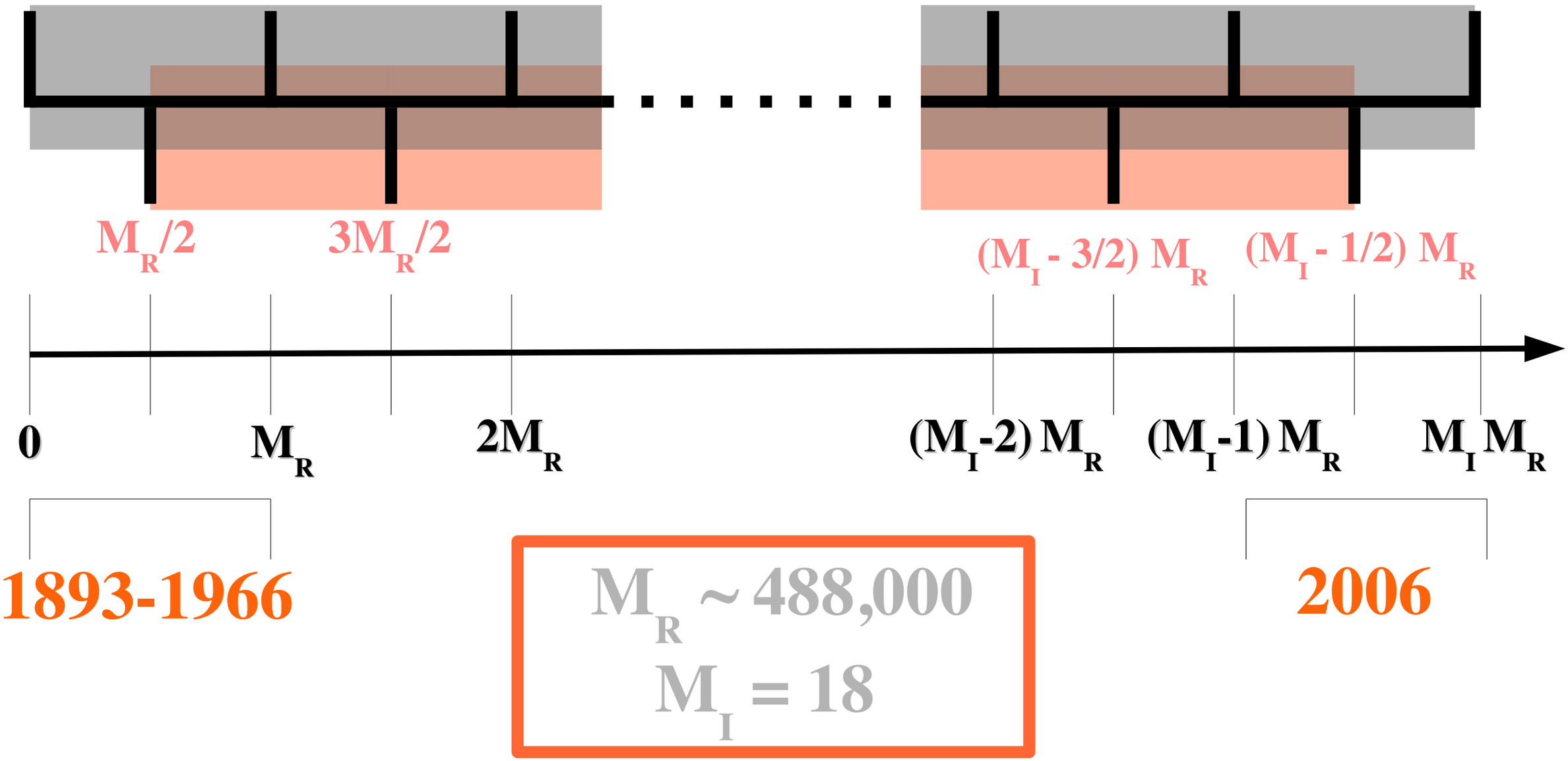
d = # "distinct" scholars with the same abbreviation

Dynamical representation

Divide 8,783,994 total references into homogeneous intervals

M_I = # of intervals

M_R = # of references in each interval



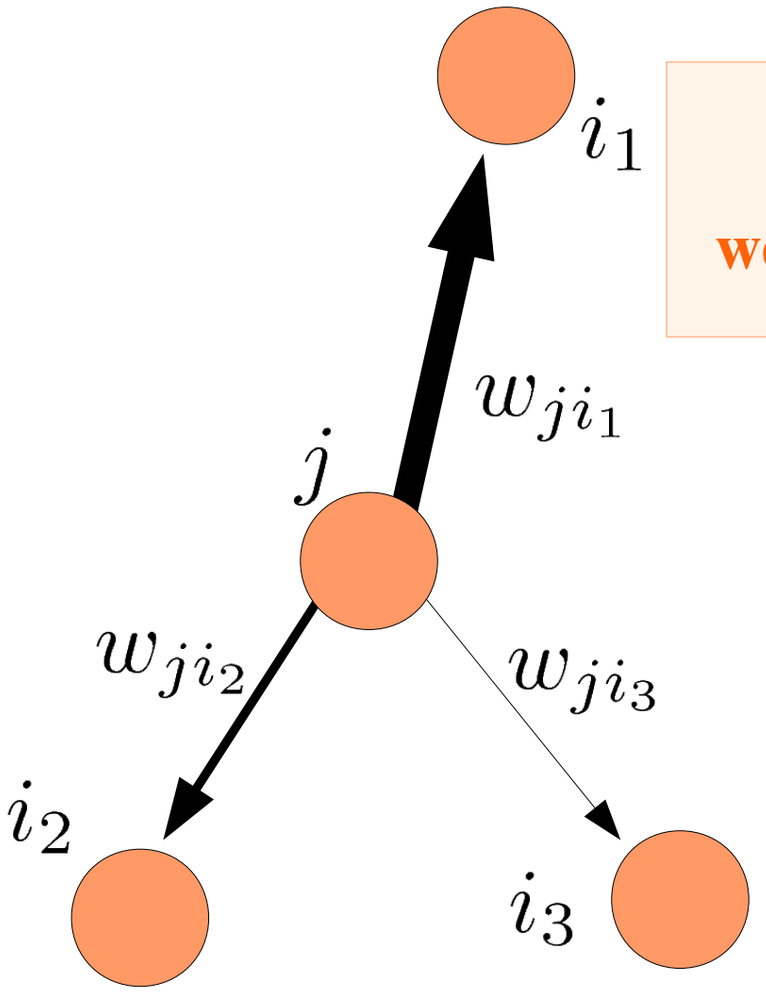
$M_R \sim 488{,}000$

$M_I = 18$

Science Author Rank Algorithm

Diffusion equation

$$P_i = (1-q)\sum_j \frac{P_j}{s_j^{out}} w_{ji} + qz_i + (1-q)\, z_i \sum_j P_j\, \delta\left(s_j^{out}\right)$$



w_{ji}

weight of the arc from j to i

$$s_j^{out} = \sum_i w_{ji}$$

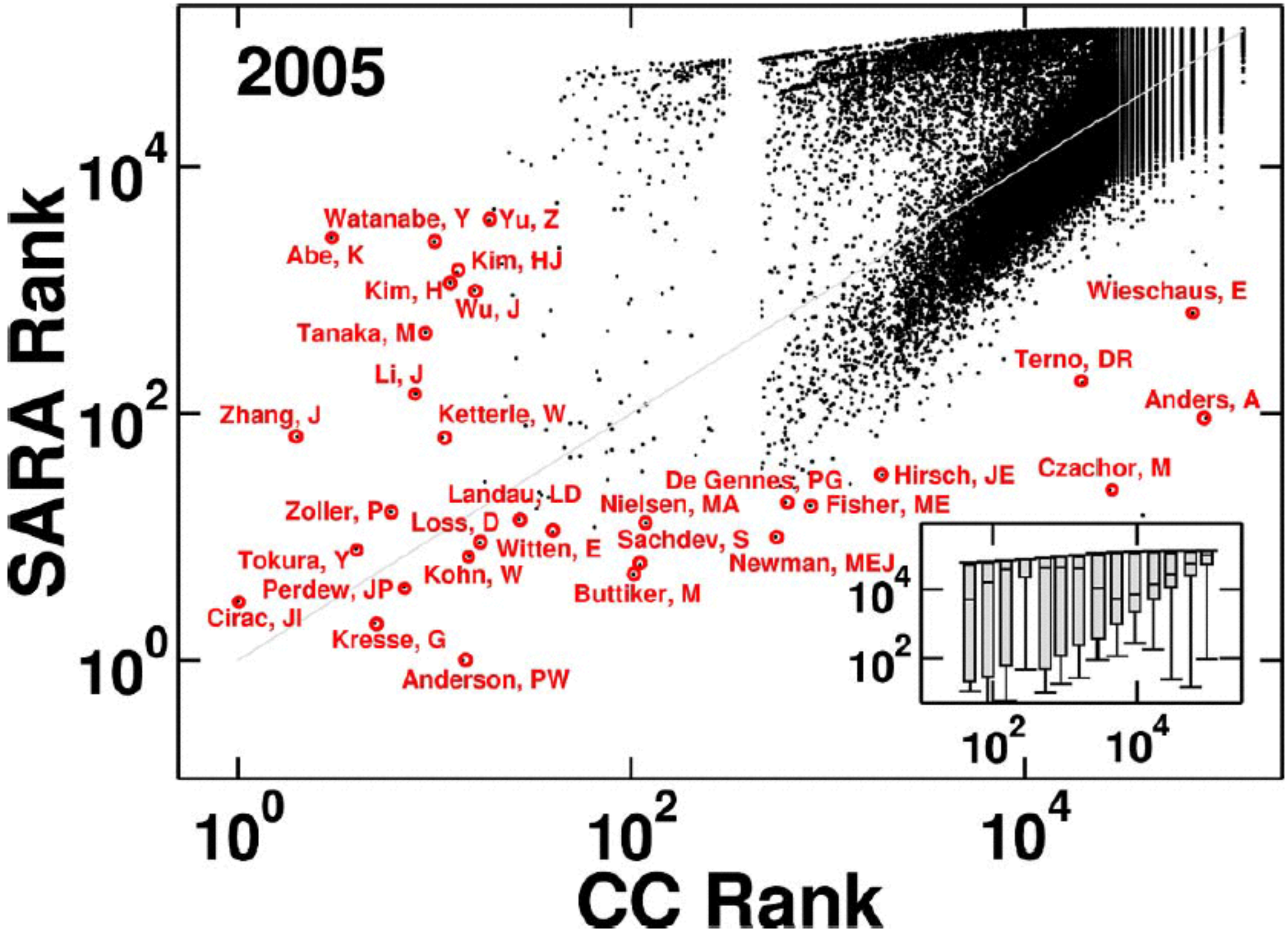
out-strength of the node j

$$z_i = \frac{\sum_p \delta_{p,i}\ 1/n_p}{\sum_j \sum_p \delta_{p,j}\ 1/n_p}$$

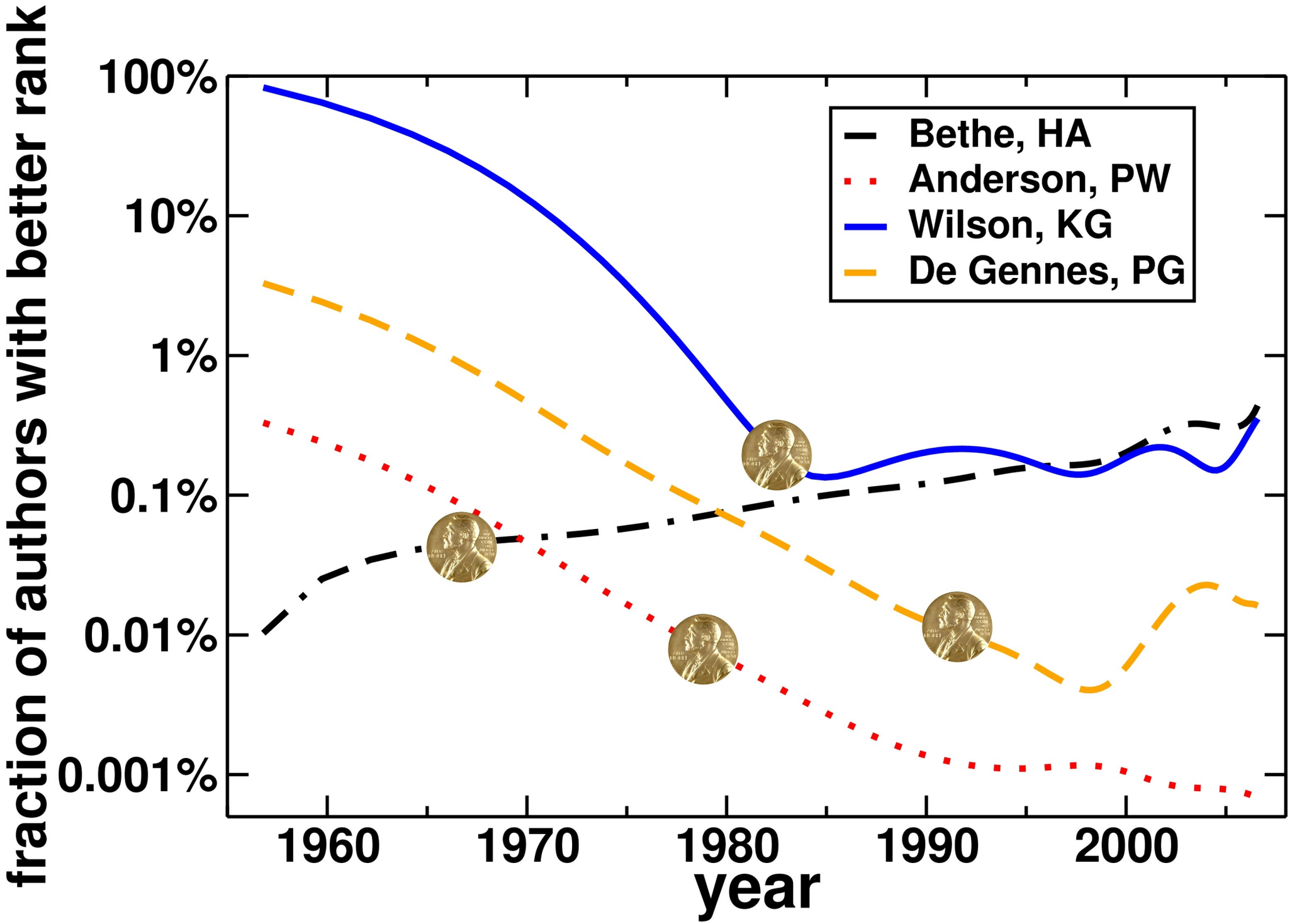
each paper carries a ”scientif c credit”, equally divided among its authors

SARA scores depend on the choice of the redistribution probability q

SARA vs. Citation Count



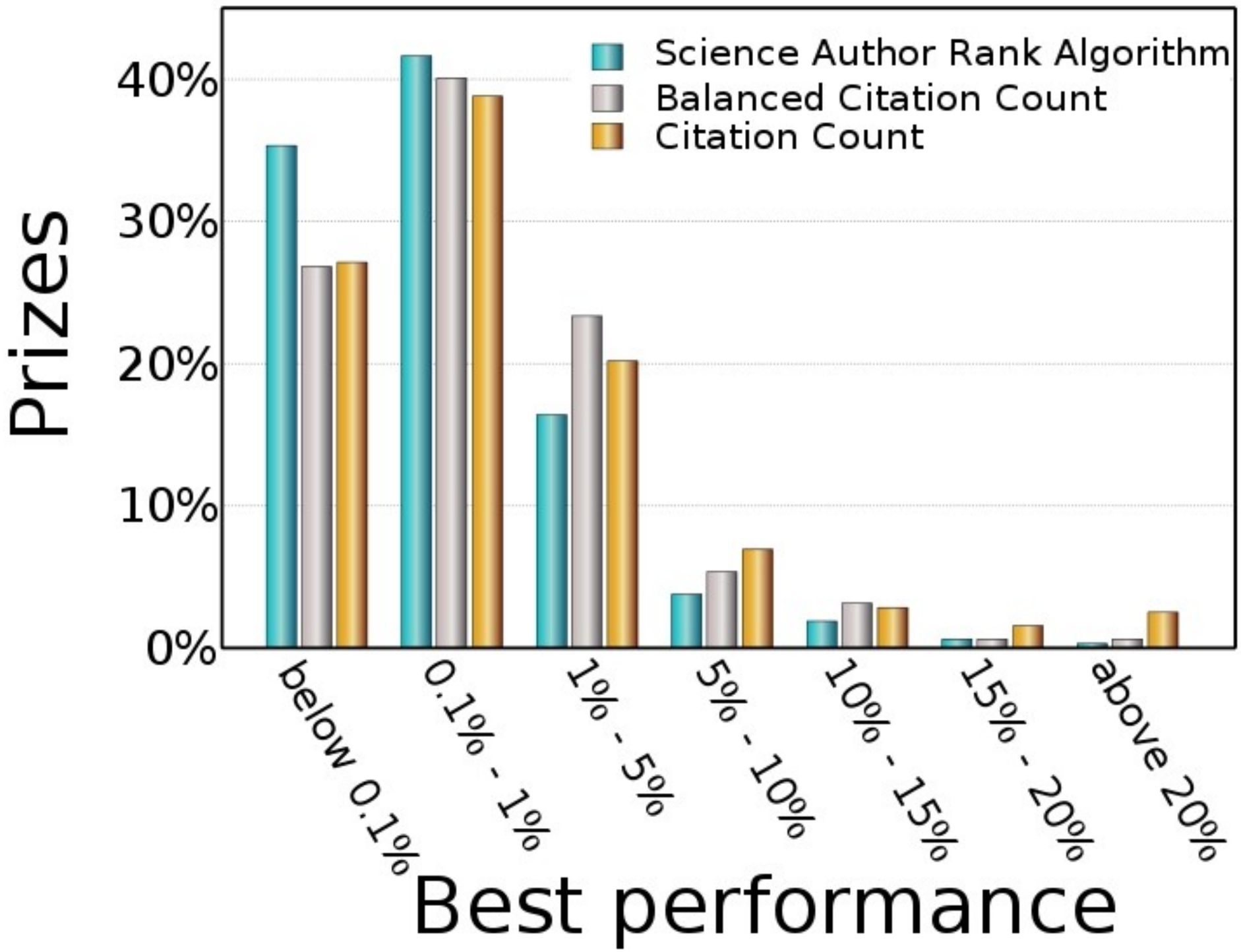
Science Author Rank Algorithm



$$R_i = 1/N \sum_{j \neq i} \theta \left(P_j - P_i \right)$$

Comparison with different metrics

Benchmarking SARA



Considered prizes: Nobel prize, Wolf prize, Boltzmann medal, Dirac medal and Planck medal

Best physicists according to SARA

1973

Rank	Author	NP	WP	BM	DM	PM
1	GELL-MANN, M	1969	-	-	-	-
2	WEINBERG, S	1979	-	-	-	-
3	SCHWINGER, J	1965	-	-	-	-
4	FEYNMAN, RP	1965	-	-	-	-
5	LEE, TD	1957	-	-	-	-
6	ANDERSON, PW	1977	-	-	-	-
7	BJORKEN, JD	-	-	-	2004	-
8	YANG, CN	1957	-	-	-	-
9	SLATER, JC	-	-	-	-	-
10	ADLER, SL	-	-	-	1998	-
11	GLAUBER, RJ	2005	-	-	-	-
12	CHEW, GF	-	-	-	-	-
13	WIGNER, EP	1963	-	-	-	1961
14	LOVELACE, C	-	-	-	-	-
15	SATCHLER, GR	-	-	-	-	-
16	MOTT, NF	1977	-	-	1985	-
17	FISHER, ME	-	1980	1983	-	-
18	MANDELSTAM, S	-	-	-	1991	-
19	BETHE, HA	1967	-	-	-	1955
20	PHILLIPS, JC	-	-	-	-	-

2004

Rank	Author	NP	WP	BM	DM	PM
1	ANDERSON, PW	1977	-	-	-	-
2	WITTEN, E	-	-	-	1985	-
3	TOKURA, Y	-	-	-	-	-
4	PERDEW, JP	-	-	-	-	-
5	KOHN, W	-	-	-	-	-
6	KRESSE, G	-	-	-	-	-
7	BÜTTIKER, M	-	-	-	-	-
8	WEINBERG, S	1979	-	-	-	-
9	CIRAC, JI	-	-	-	-	-
10	ZUNGER, A	-	-	-	-	-
11	BARABÁSI, AL	-	-	-	-	-
12	LEE, PA	-	-	-	2005	-
13	VANDERBILT, D	-	-	-	-	-
14	SACHDEV, S	-	-	-	-	-
15	NEWMAN, MEJ	-	-	-	-	-
16	AFFLECK, I	-	-	-	-	-
17	MACDONALD, AH	-	-	-	-	-
18	HIRSCH, JE	-	-	-	-	-
19	ZOLLER, P	-	-	-	2006	2005
20	PARISI, G	-	-	1992	1999	-

NP= Nobel prize, WP= Wolf prize, BM= Boltzmann medal, DM= Dirac medal, and PM= Planck medal

login

Phys Author Rank Algorithm

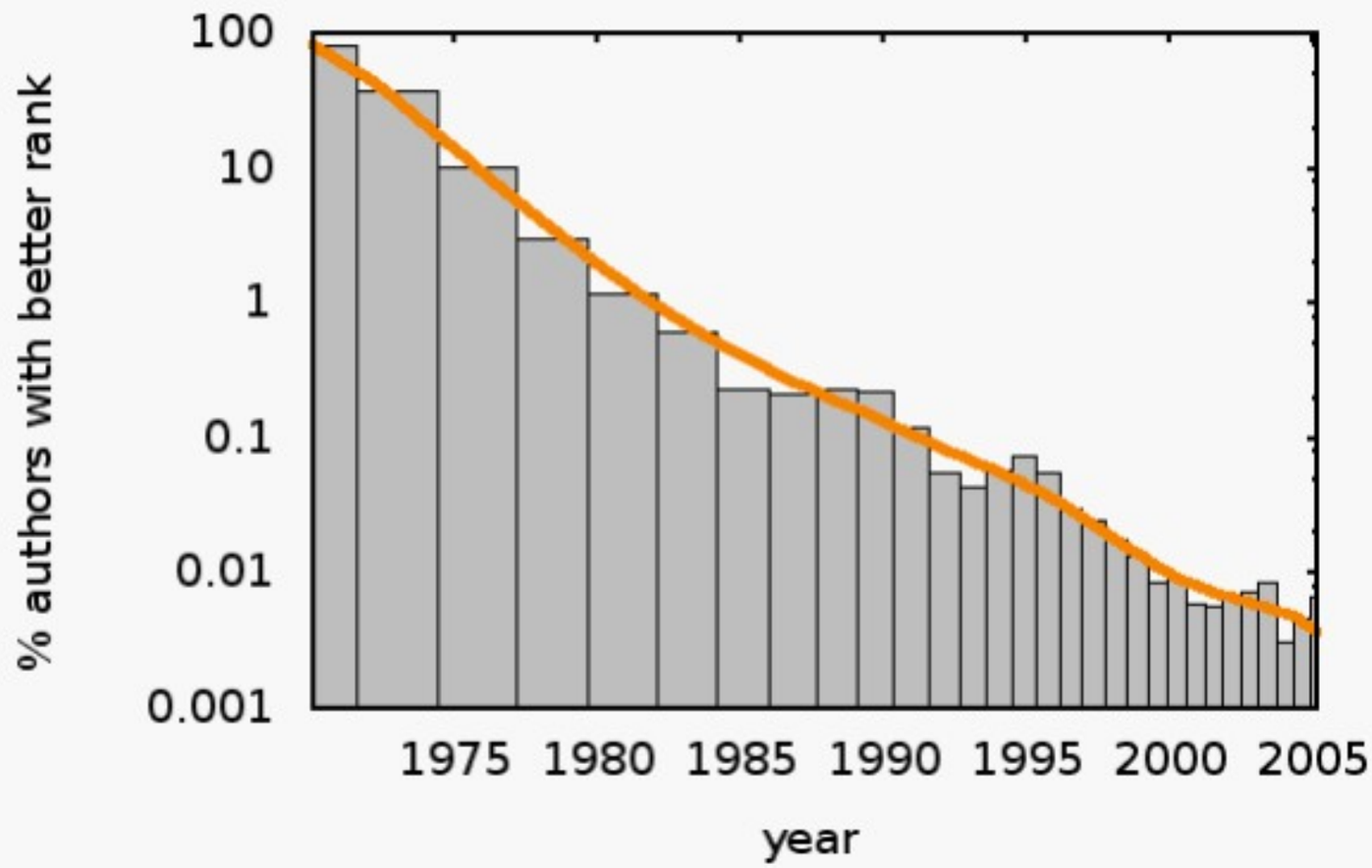
Author Search perdew jp

Search/Rank

Insert the lastname eventually followed by the initials of the author as in "Bethe HA"

Rank's evolution

Rank analysis for **PERDEW, JP**



Best performance **0.0031%**

Last performance **0.0069%**

physauthorsrank.org

Trying to address the workshop questions

What are the prospects of machine learning in rankings?
Who decides about training data?

Good (maybe too small) training data are lists of prominent scientists, such as awardees of prestigious prizes.

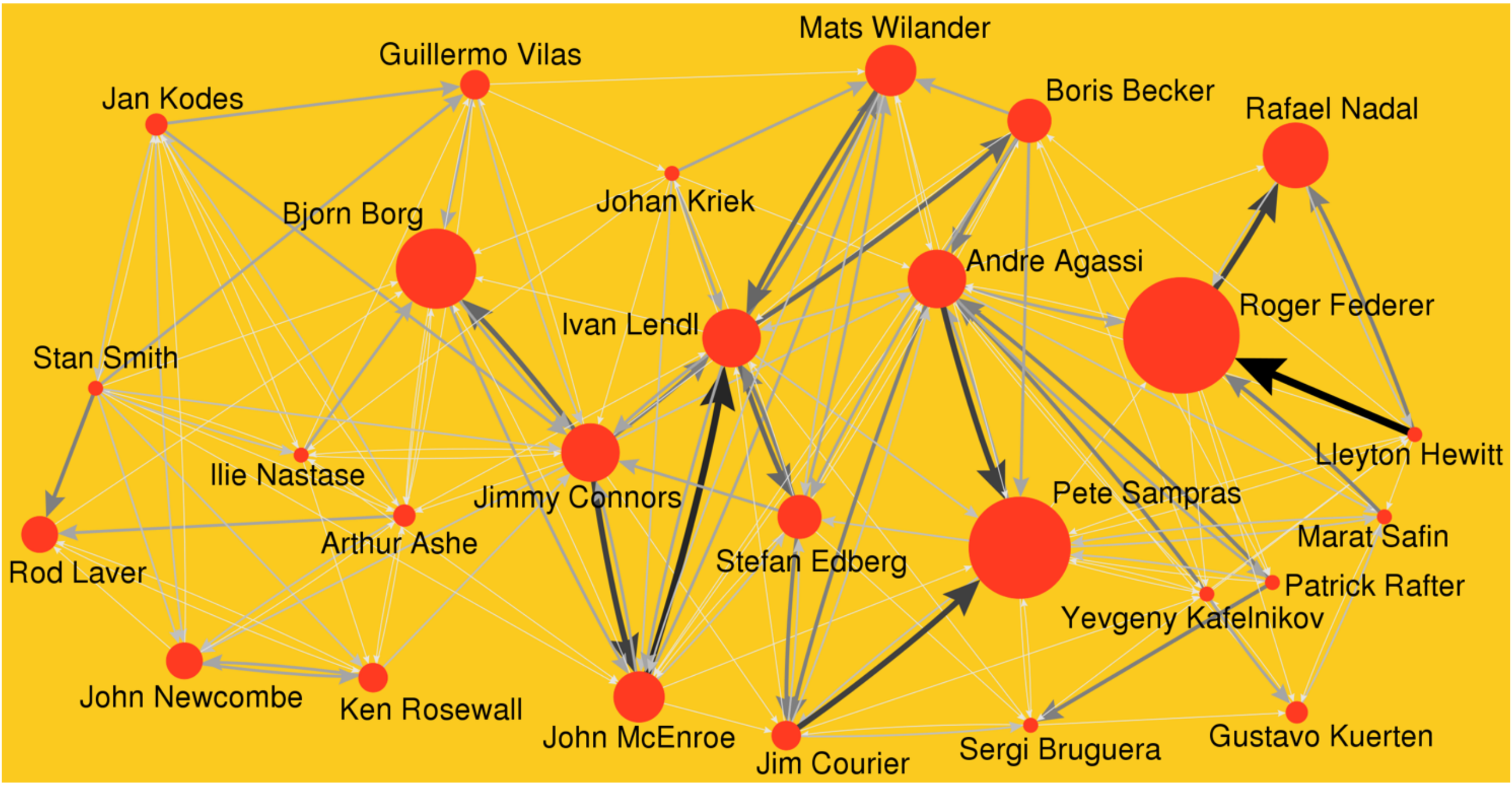
How can the time dimension be included in network-based ranking?

Natural time is not the unique and optimal option

How can we improve name disambiguation methods?

Disambiguation seems necessary to treat a few pathological cases. Possible improvements are: user managed profiles (e.g., ResearcherID, GS citations); machine learning algorithms (cleaner and larger training sets are required).

Tennis Prestige Score



http://tennisprestige.soic.indiana.edu

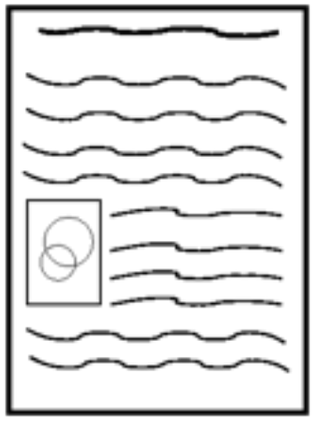
Are we ready to use bibliographic data for research evaluation?

discipline	scientist	h-index
Physics	Edward Witten	110
	Marvin Cohen	94
	Philip W. Anderson	91
	Manuel Cardona	86
	Franck Wilczek	68
Chemistry	George Whitesides	135
	Elias J. Corey	132
	Martin Karplus	129
	Alan Hegeer	114
	Kurt Wuthrich	113
Computer science	Hector Garcia-Molina	70
	Deborah Estrin	68
	Ian Foster	67
	Scott Shenker Jeffrey D. Ullman Don Towsley	65

Ball P. Nature **448**, 737 (2007)

Different scientific disciplines

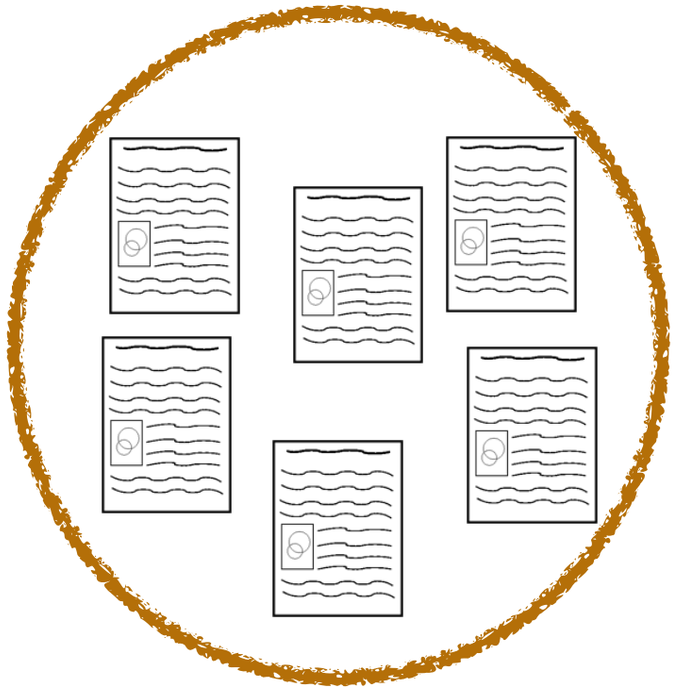
Papers



pub. year

citations

Journals



Journal Citation Reports (JCR) subject-categories

Journals are classified in 172 categories

from acoustics to zoology

ISI Web of KNOWLEDGE

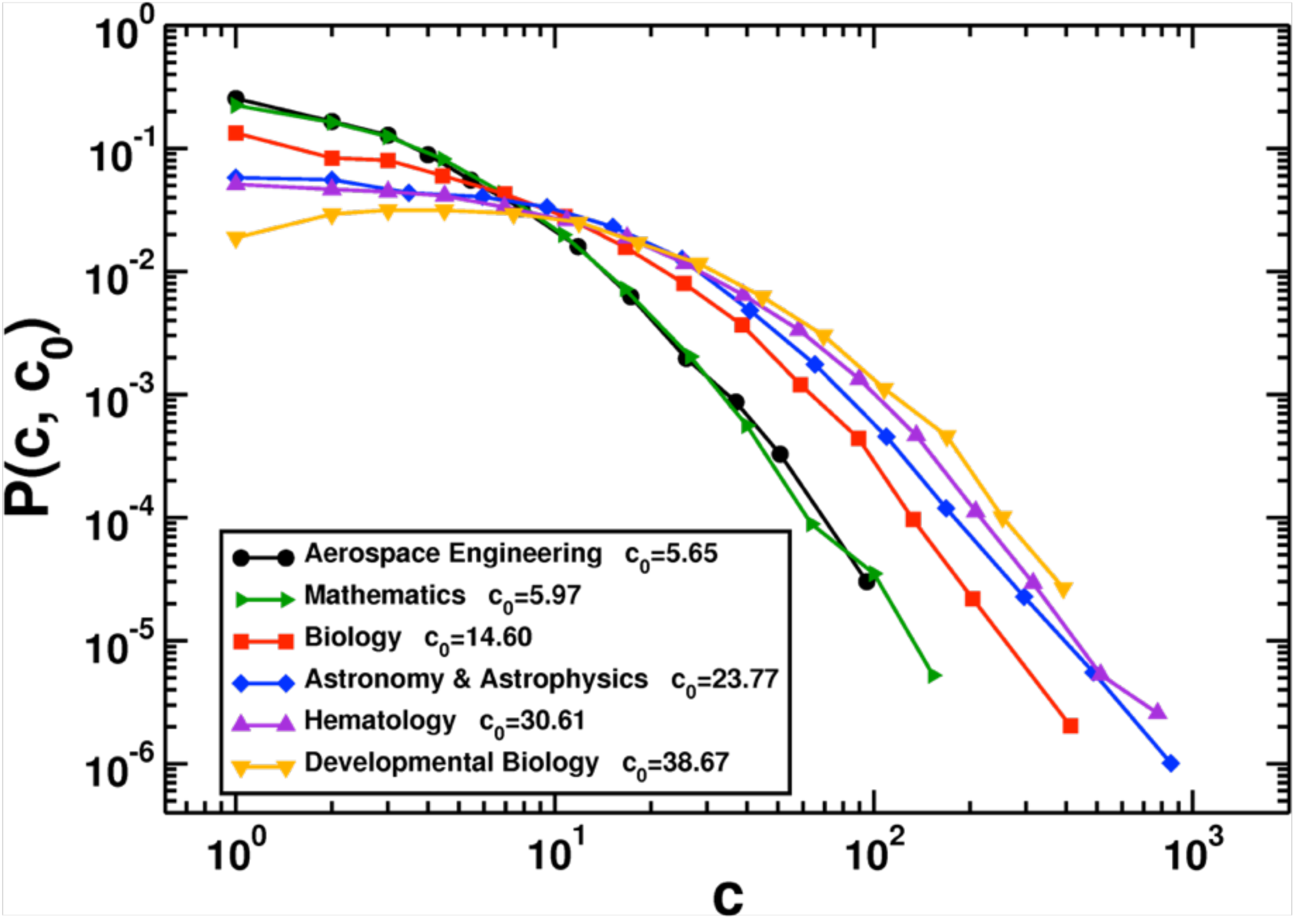
http://science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D

Citation patterns in different scientific disciplines

Subject category	Year	N_p	c_0	c_{max}
Agricultural economics and policy	1999	266	6.88	42
Allergy	1999	1,530	17.39	271
Anesthesiology	1999	3,472	13.25	282
Astronomy and astrophysics	1999	7,399	23.77	1,028
Biology	1999	3,400	14.6	413
Computer science, cybernetics	1999	704	8.49	100
Developmental biology	1999	2,982	38.67	520
Engineering, aerospace	1999	1,070	5.65	95
Hematology	1990	4,423	41.05	1,424
Hematology	1999	6,920	30.61	966
Hematology	2004	8,695	15.66	1,014
Mathematics	1999	8,440	5.97	191
Microbiology	1999	9,761	21.54	803
Neuroimaging	1990	444	25.26	518
Neuroimaging	1999	1,073	23.16	463
Neuroimaging	2004	1,395	12.68	132
Physics, nuclear	1990	3,670	13.75	387
Physics, nuclear	1999	3,965	10.92	434
Physics, nuclear	2004	4,164	6.94	218
Tropical medicine	1999	1,038	12.35	126

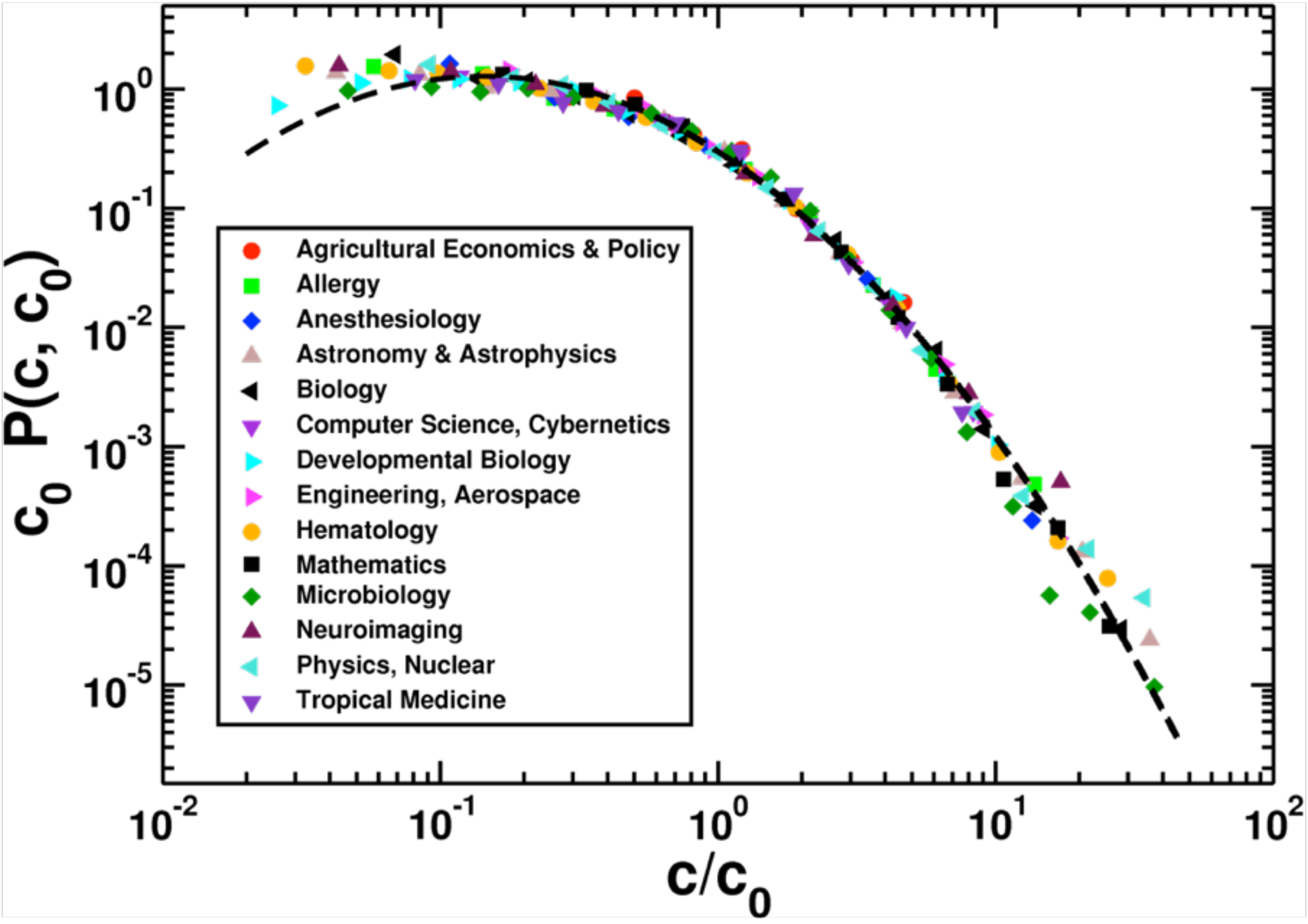
citation data collected from ISI Web of KNOWLEDGE in March 2008

Citation patterns in different scientific disciplines



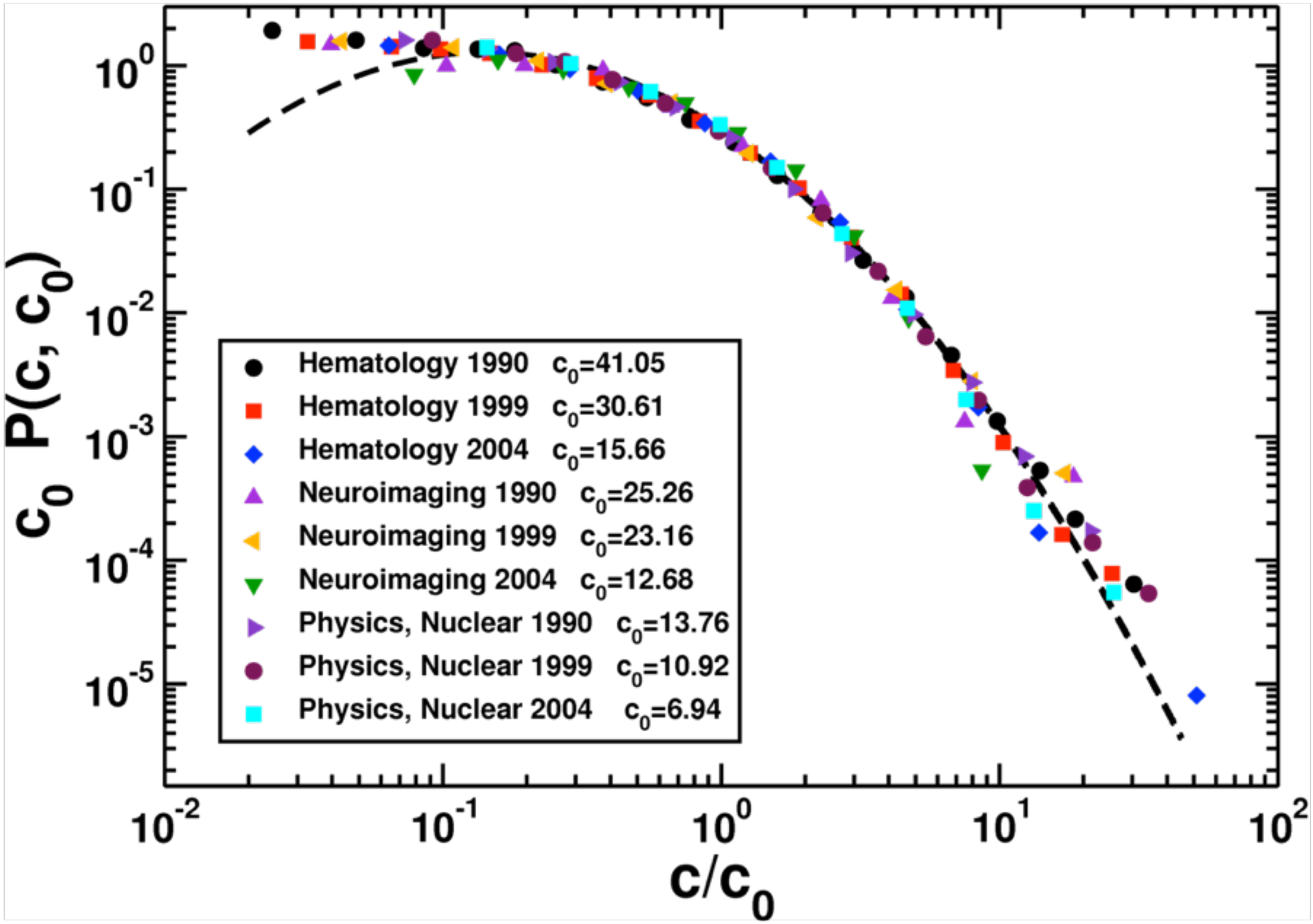
Radicchi F, et al. Proc Natl Acad Sci USA 105, 17268–17272 (2008)

Universality of citation distributions



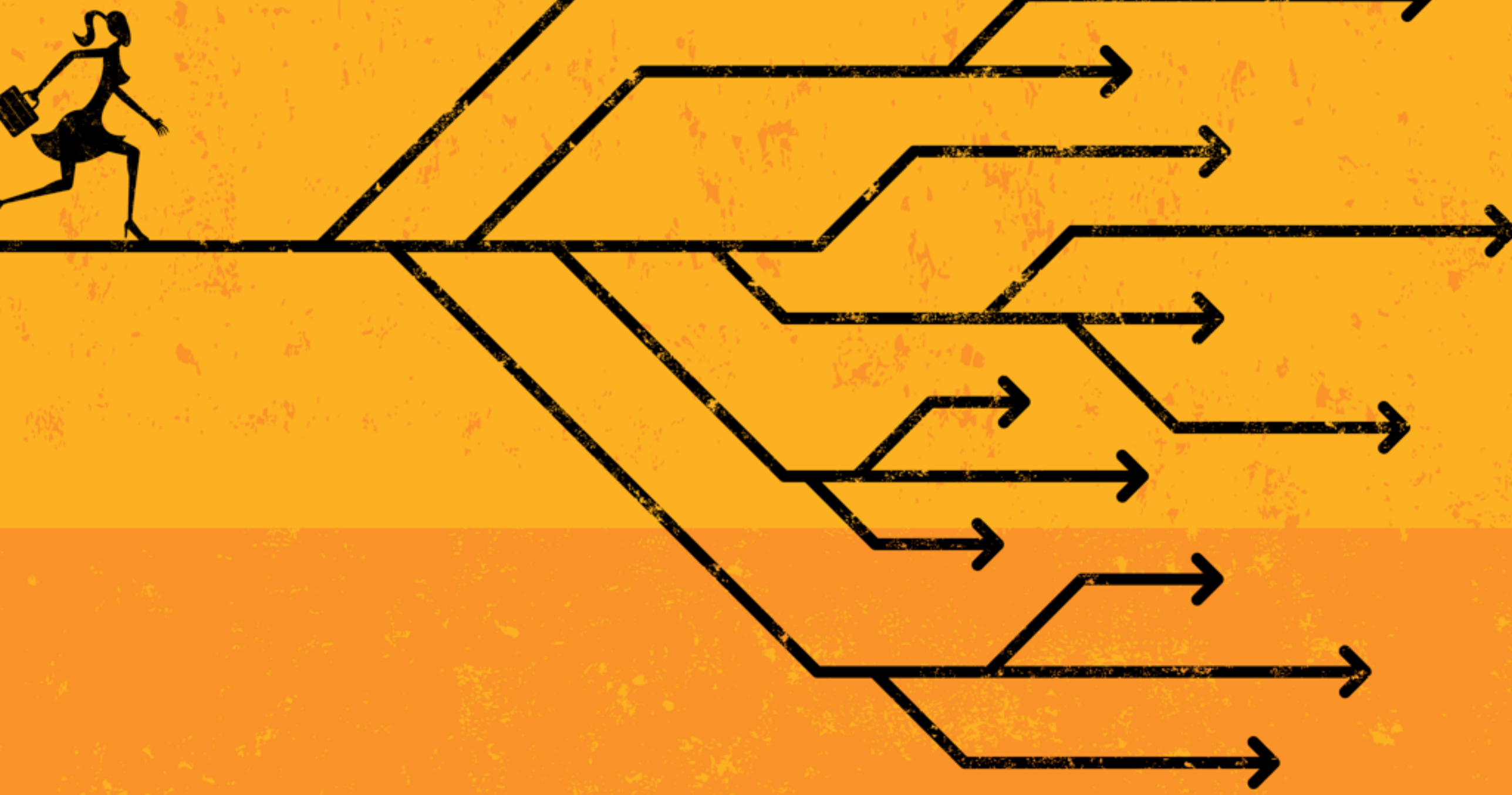
Radicchi F, et al. Proc Natl Acad Sci USA 105, 17268–17272 (2008)

Universality of citation distributions

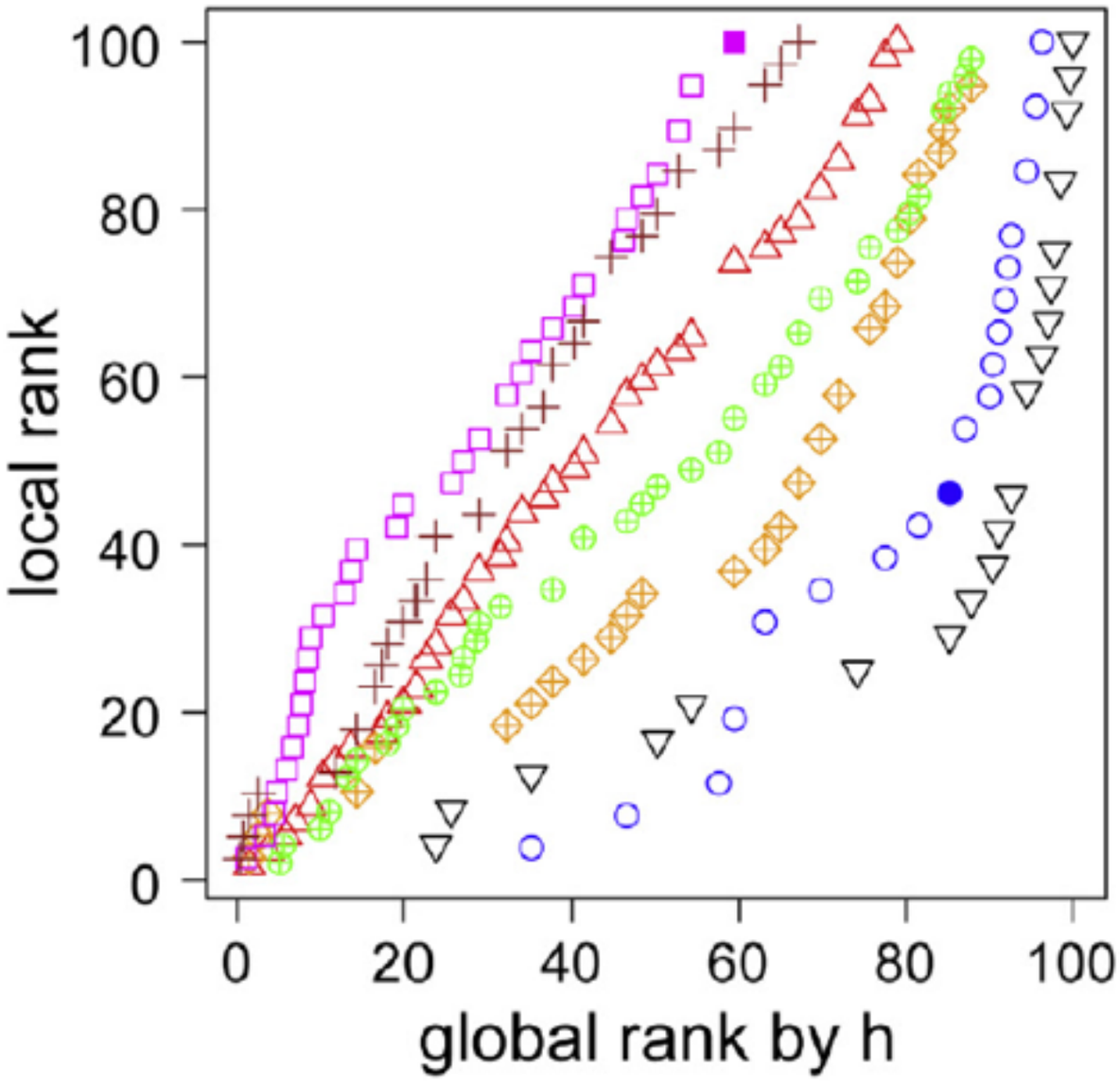


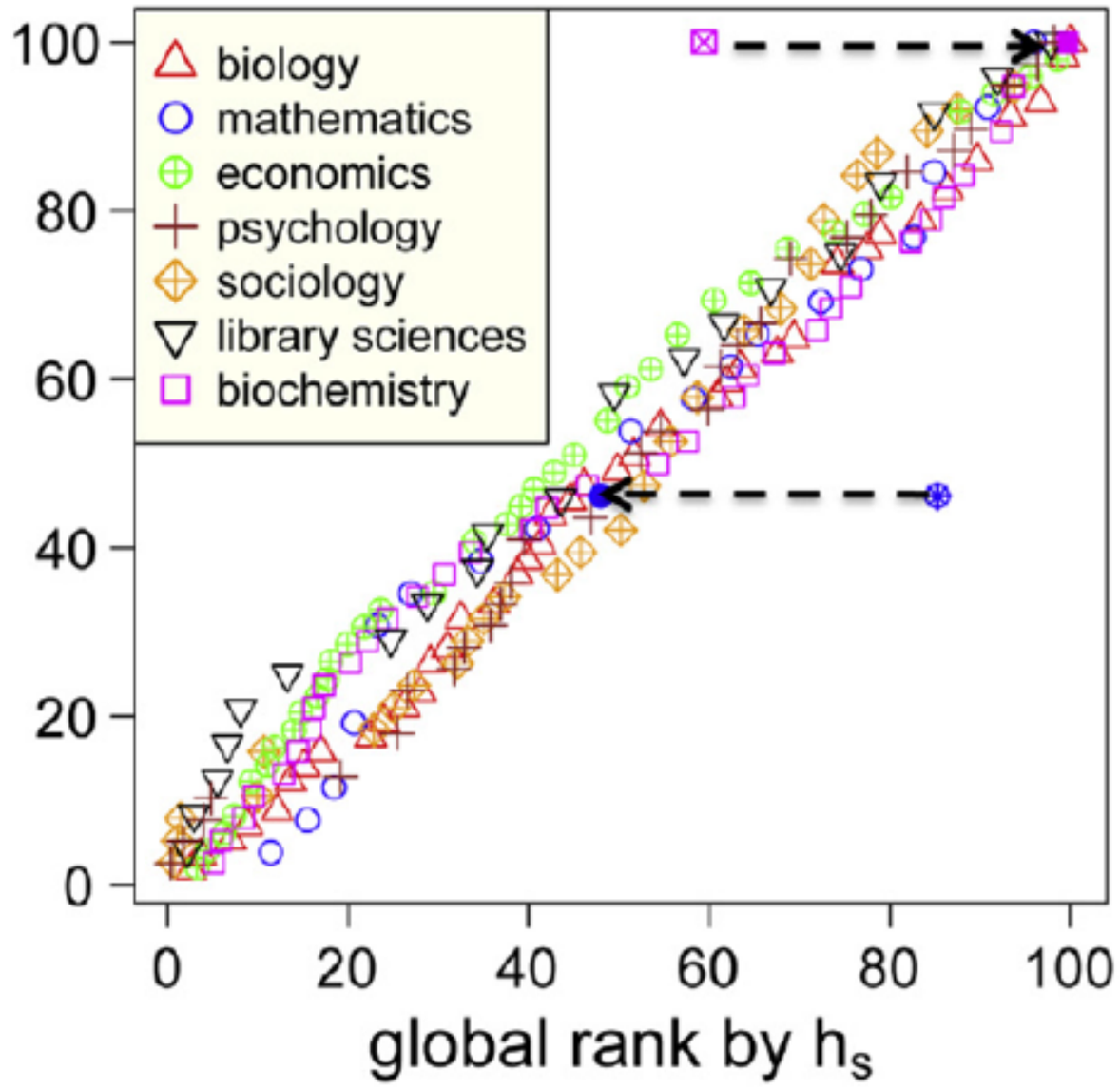
Radicchi F, et al. Proc Natl Acad Sci USA 105, 17268–17272 (2008)

What about scholars?



Universality of scholarly impact metrics





Kaur J, Radicchi F, Menczer F, J Informetr 7, 924 (2013)

Problems in the evaluation of scholars

age : should it be quantified in terms of number of papers or number of years of activity?

discipline : can we really classify people in specific categories?

discipline and age : are there common patterns in the development of a scientific career?

Direct comparisons among scientists are complicated because it is very hard to generate homogenous categories composed of a sufficiently large number of scholars.

Our proposal

terms of comparisons specifically tailored for each scientist

real publication record

$$\{y\} = \{y_1, y_2, \ldots, y_{N_r}\} \qquad \{d\} = \{d_1, d_2, \ldots, d_{N_r}\}$$

years of publication — subject-categories

$$\{c\} = \{c_1, c_2, \ldots, c_{N_r}\}$$

citations accumulated

synthetic publication record

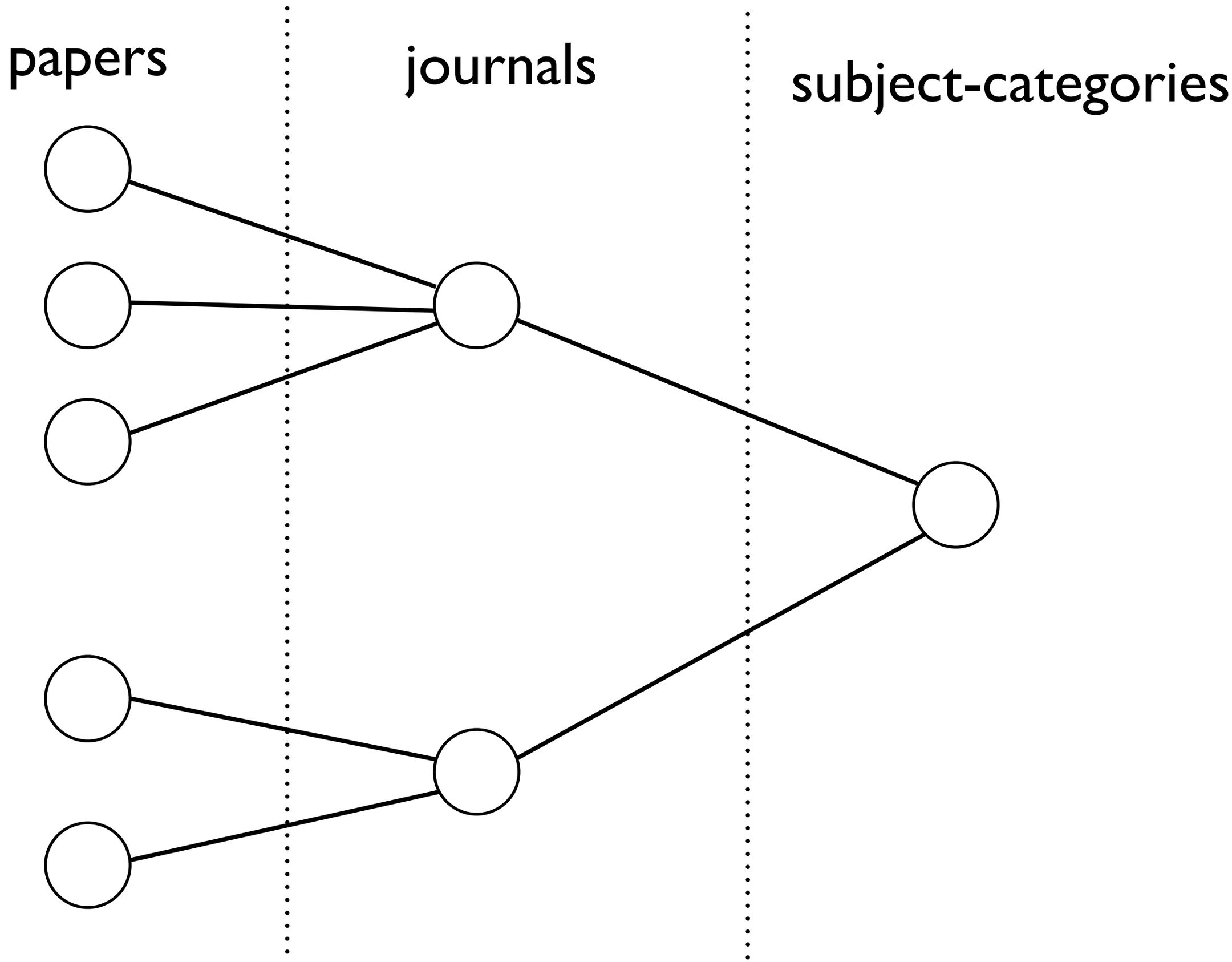
$$\{y\} = \{y_1, y_2, \ldots, y_{N_r}\} \qquad \{d\} = \{d_1, d_2, \ldots, d_{N_r}\}$$

years of publication — subject-categories

$$\{\tilde{c}\} = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_{N_r}\}$$

citations accumulated are randomly extracted from the set of papers with same age and subject-category

Resampling strategy



q-score

real publication record

$$\{y\} = \{y_1, y_2, \ldots, y_{N_r}\}$$

$$\{d\} = \{d_1, d_2, \ldots, d_{N_r}\}$$

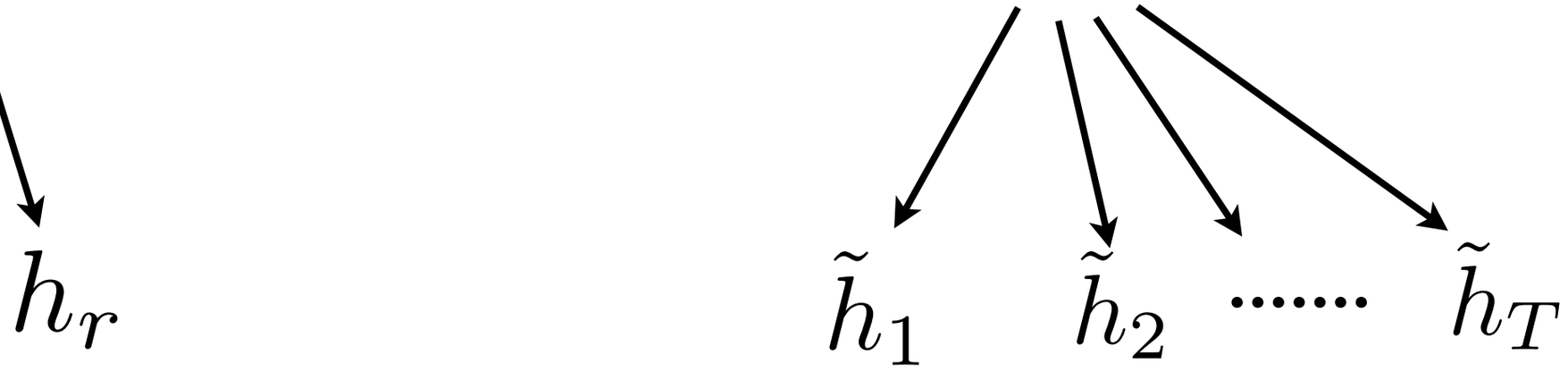
$$\{c\} = \{c_1, c_2, \ldots, c_{N_r}\}$$

synthetic publication record

$$\{y\} = \{y_1, y_2, \ldots, y_{N_r}\}$$

$$\{d\} = \{d_1, d_2, \ldots, d_{N_r}\}$$

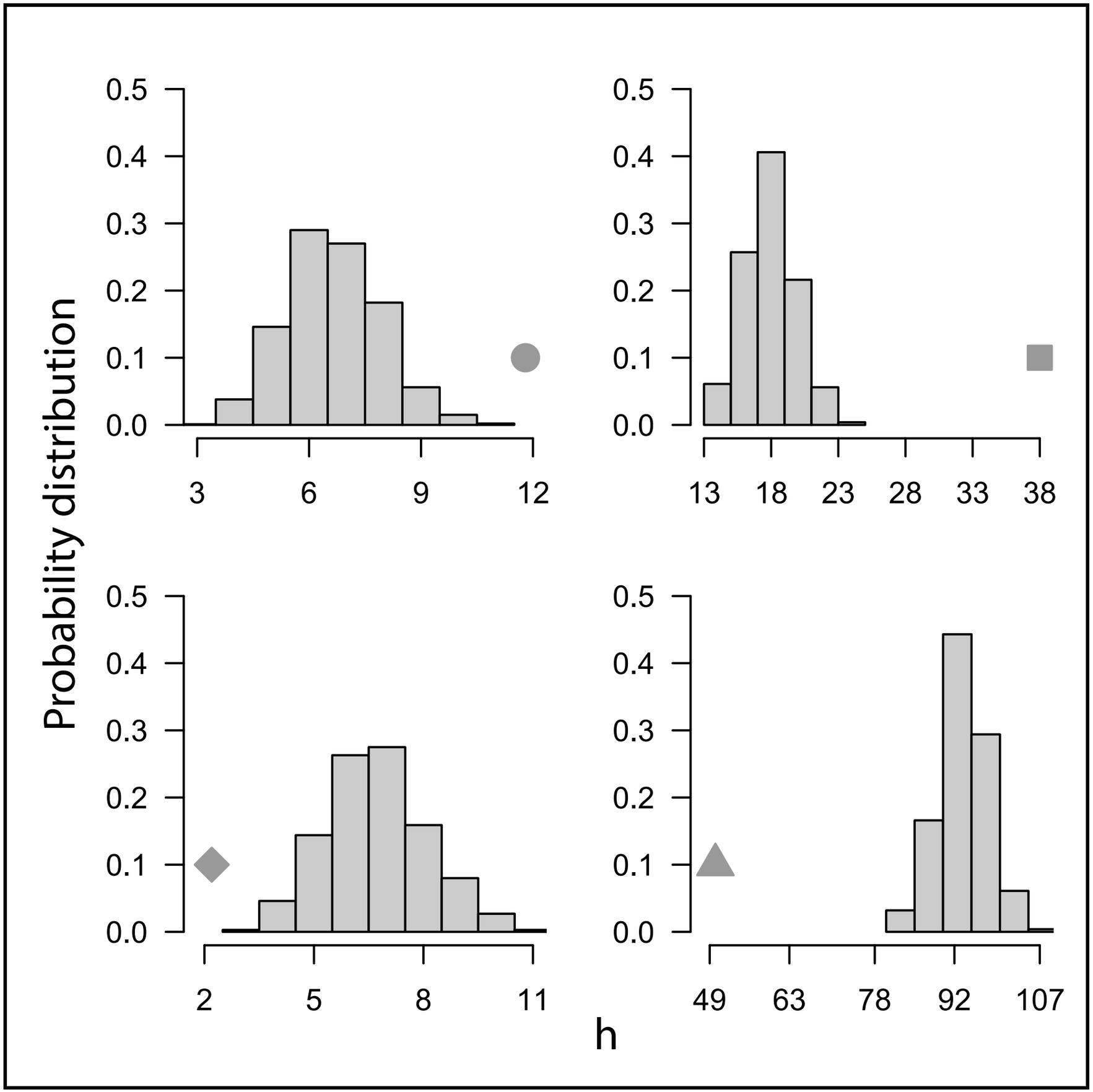
$$\{\tilde{c}\} = \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_{N_r}\}$$



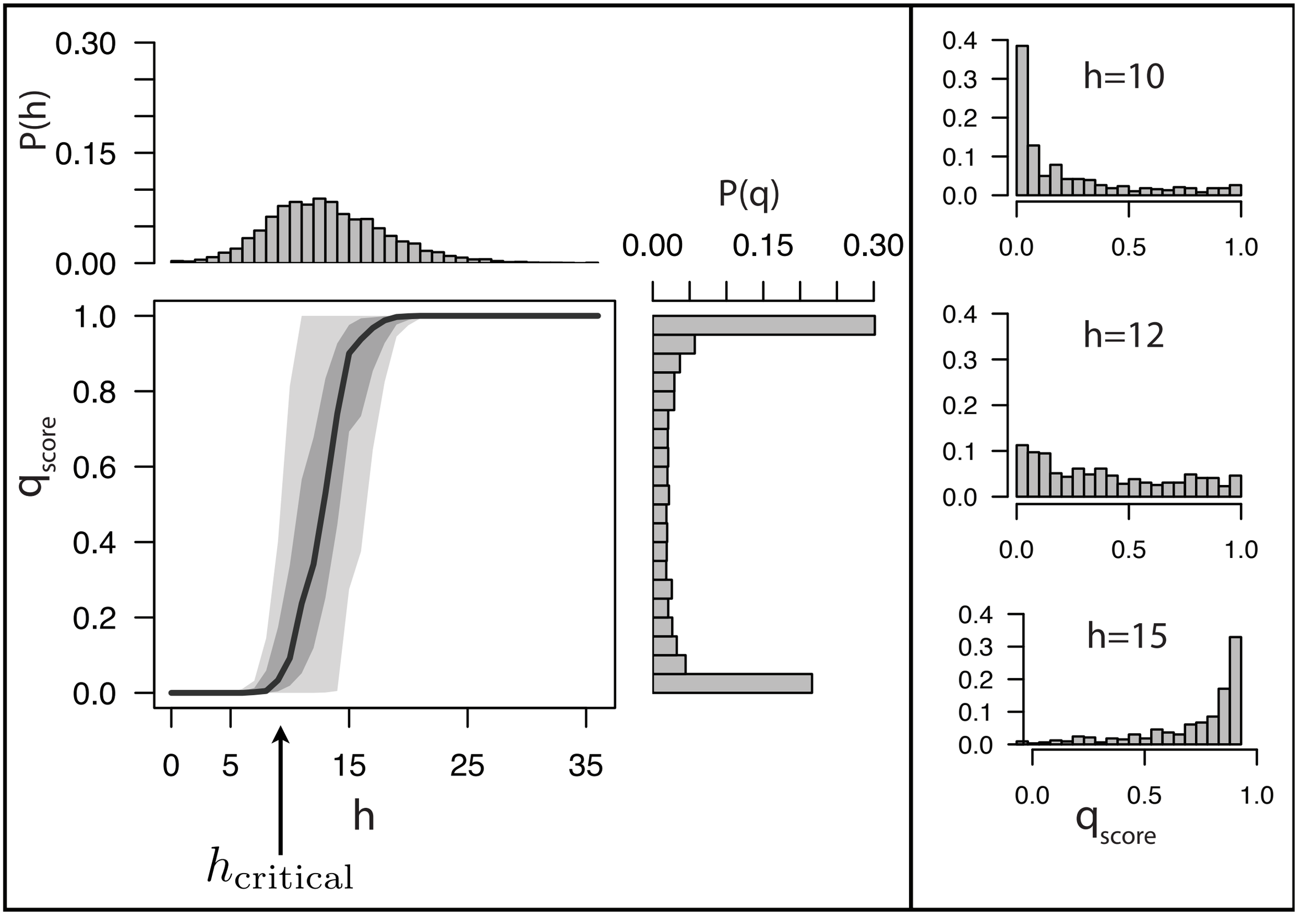
$$\text{q-score} = \frac{1}{T} \sum_{i=1}^{T} \theta\left(h_r - \tilde{h}_i\right)$$

$$\theta(x) = \begin{cases} 1 & , \text{if } x \geq 0 \\ 0 & , \text{oth.} \end{cases}$$

q-score

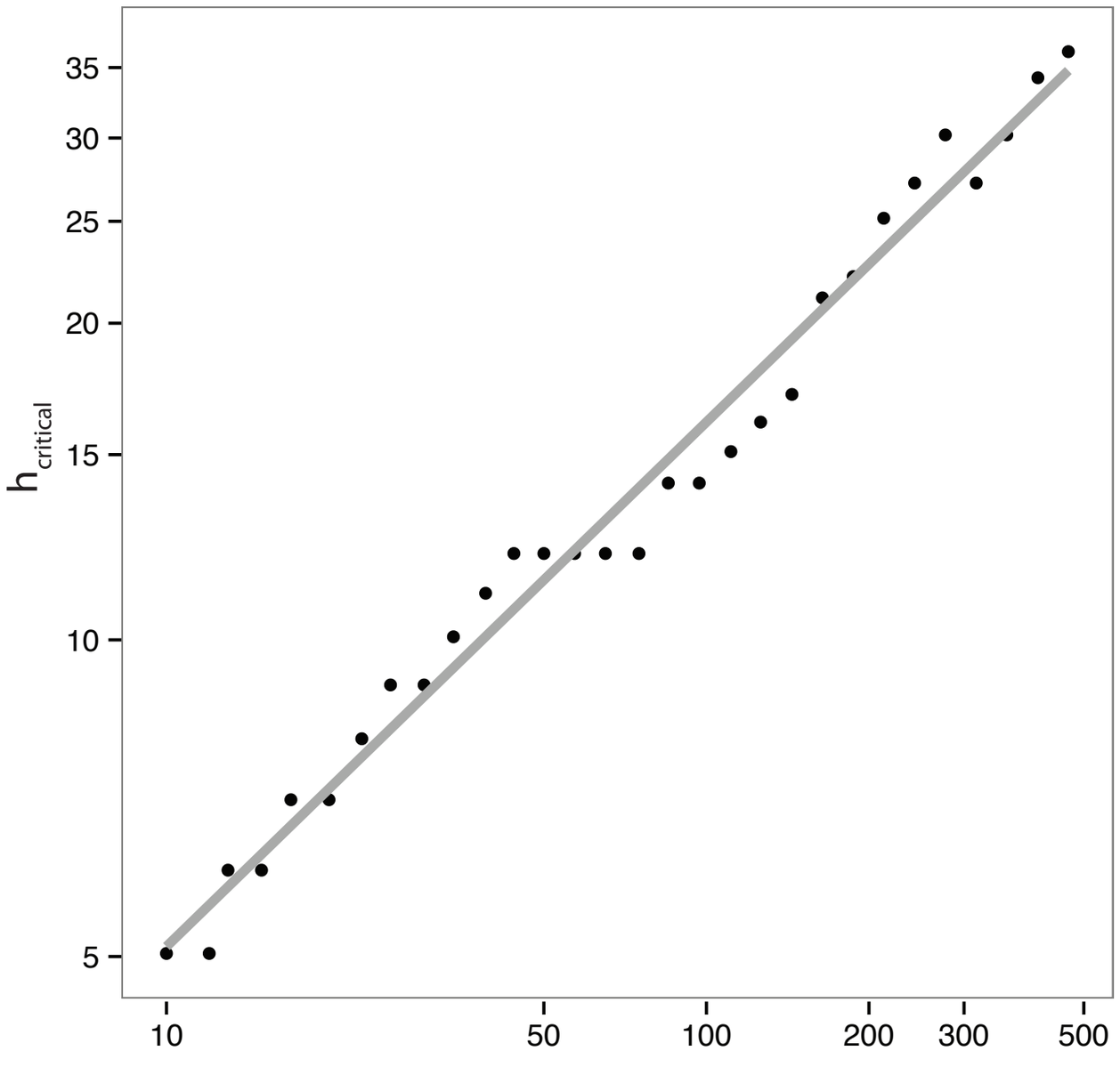

Probability distribution
0.5
0.4
0.3
0.2
0.1
0.0
3
6
9
12
13
18
23
28
33
38
2
5
8
11
49
63
78
92
107
h

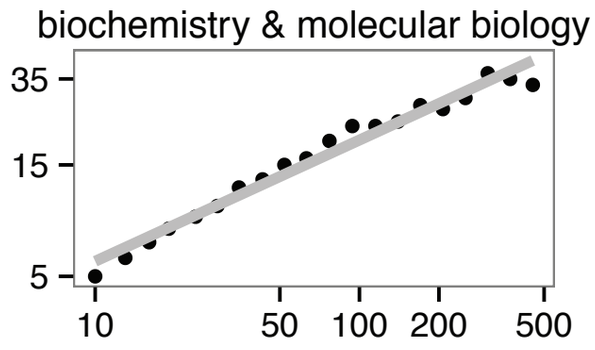
q-score and h-index

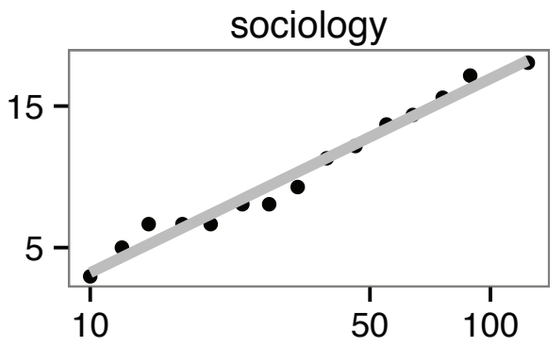


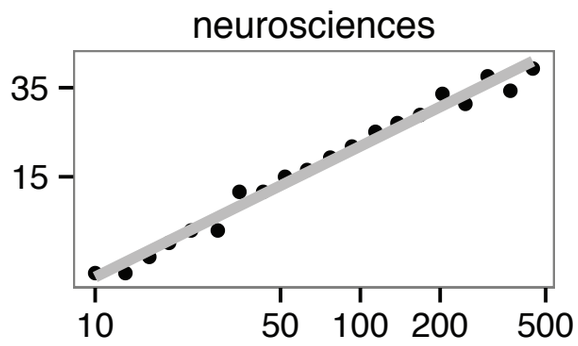
only authors with exactly N = 50 publications

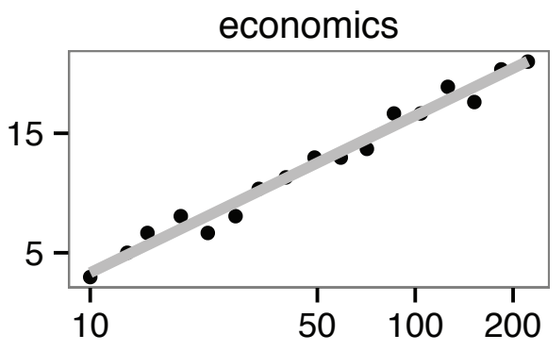
critical h-index

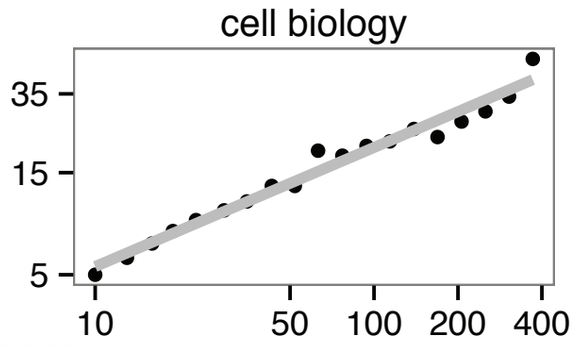


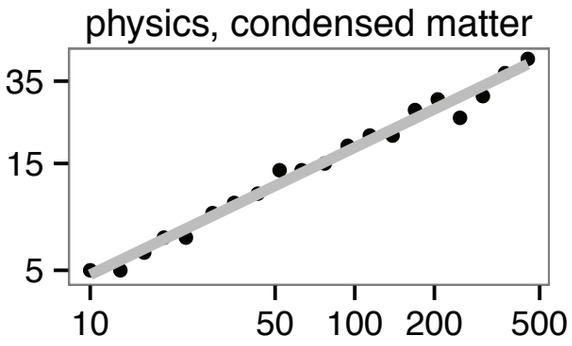












$$h_{\mathrm{critical}} \simeq 1.2\sqrt{N}$$

extension to journals