Bibliometrics

Werner Nutt

Bibliometrics: Motivation

Are there objective (= quantitative) ways to measure

- the productivity of a researcher?
- the quality of journals?
- the productivity of institutions?

Bibliometrics: Ideas

- Researchers write papers
  - ➔ Analyse the paper output
- Researchers build upon the work of other researchers, which they cite
  - ➔ Analyse how often papers have been cited

Bibliometric indices

N_p: total number of papers published

N_c: total number of citations received

n_c: mean number of citations per paper (= N_c/N_p)

IF: impact factor, calculated for journals (E. Garfield, 1955)

h-index: Hirsch's h-index (Jorge Hirsch, 2005)

g-index: Response to h-index (Leo Egghe, 2006)

*i*10-index: number of papers with ≥ 10 citations (Google Scholar)

Impact factor of a journal

Published yearly in the *Journal Citation Reports (JCR)*

- by Clarivate Analytics, USA (owns also Web of Science)
- formerly by the ISI (= Institute for Scientific Information), founded by E. Garfield (1960)
  which then became part of the *Science and Scholarly Research* division of Thomson-Reuters (1992)
  which then was sold off under the name Clarivate Analytics (2016)
- based on two databases with scientific articles, one for science and another one for humanities

Definition of IF_n (= impact factor for year *n*)

- IF_n = average number of citations in year *n*
  for articles published in years *n-1* and *n-2*
- IF_n is published in year *n+1*

Impact factors of some journals in 2013/14/15/16/17

Nature:	42.3/41.5/41.5/43.8/41.6	(5year index, 2yr index is similar)
Science:	31.5/33.6/?/37.2/41.1	

JACM:	2.939/1.394/1.803/1.855/1.744
ACM TODS	0.750/0.684/0.633/1.517/1
VLDB Journal	1.701/1.568/1.744/4.269/2.689
IEEE TOSE	2.292/1.614/1.516/3.272/n.a.
TCS	0.516/0.657/0.643/0.698/0.772

JACM	= Journal of the Association of Computing Machinery
ACM TODS	= ACM Transactions on Database Systems
IEEE TOSE	= IEEE Transactions on Software Engineering
TCS	= Theoretical Computer Science

What the IF is supposed to tell us

According to ISI/Clarivate:

- Helps librarians and researchers to decide which journals to subscribe to
- Provides sales arguments for journal publishers
- Supports academic evaluation
  (of researchers, institutions, etc.)

Criticism of IF

- IF can be increased by playing games
  - invite senior researchers
  - reduce types of articles that attract fewer citations
  - publish articles likely to attract citations early in the year

  …
- IFs differ across disciplines
- Articles in a journal with high IF are not necessarily frequently cited
  - Distribution of citations over articles follows a power law
    e.g., 90% of citations in Nature come from 25% of the papers
  - Correlation between paper citation frequency and impact factor is diminishing, since papers are available electronically

    See: Lozano, Larivière, and Gingras. The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology, 2012

The relation between IF and citation frequency of papers is diminishing

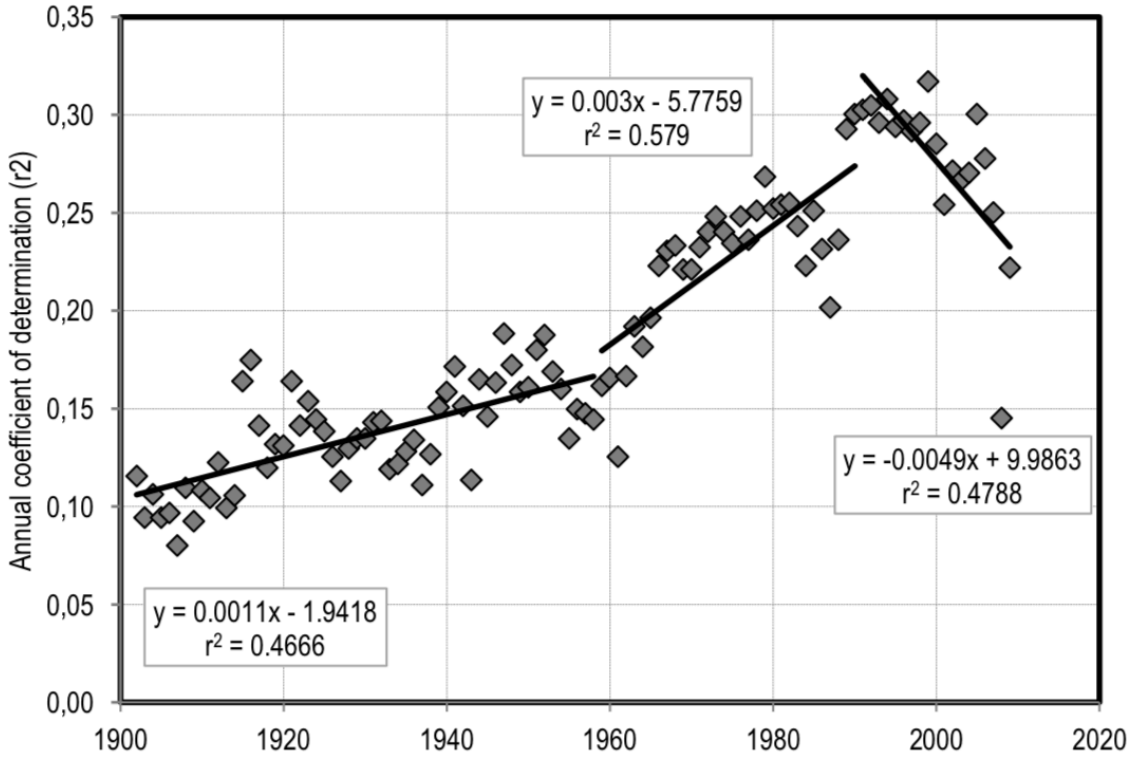
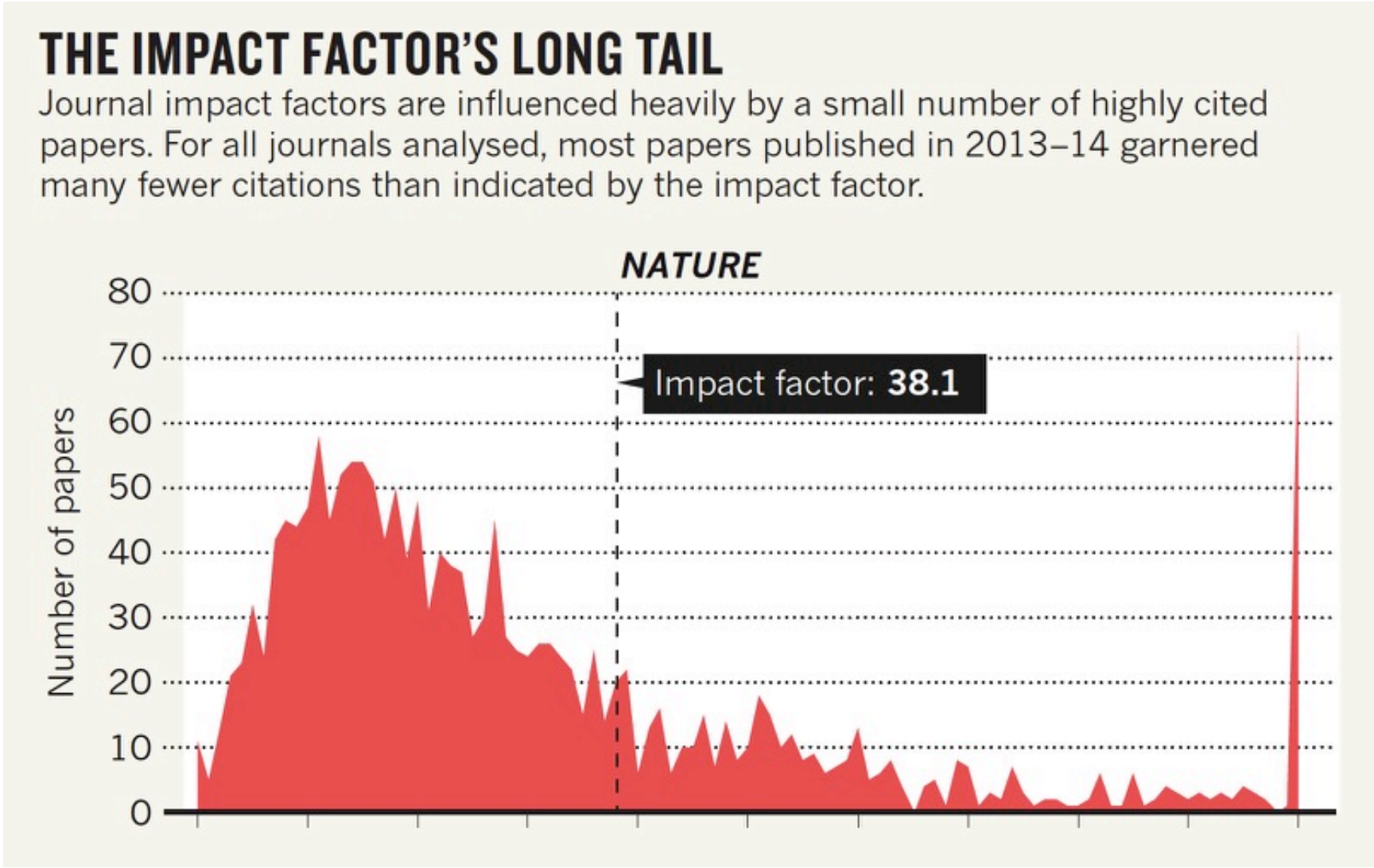
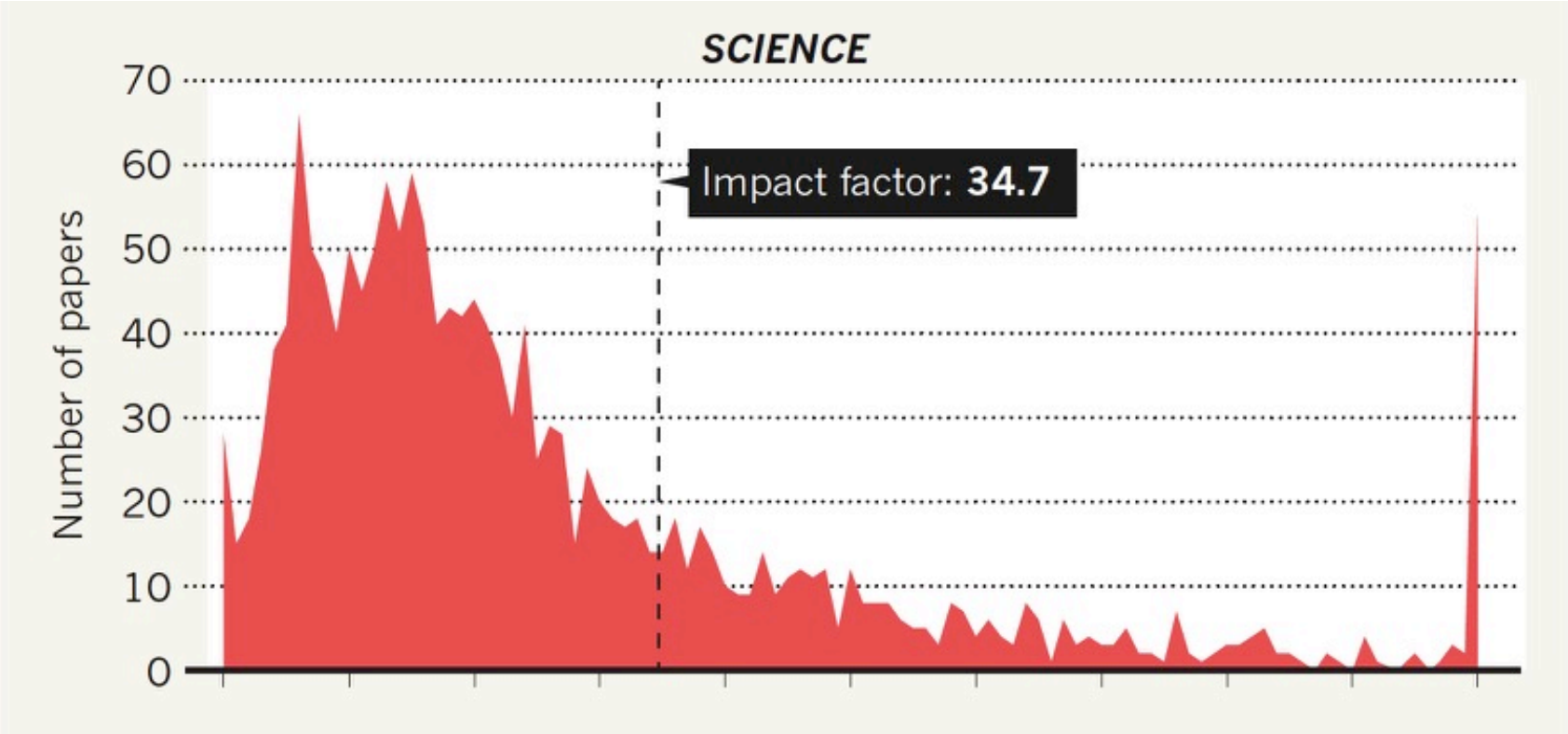


FIG. 1. Coefficient of determination (r^2) between the impact factor of journals and the 2-year citation rate of their papers from 1902 to 2009, for all natural and medical sciences journals.

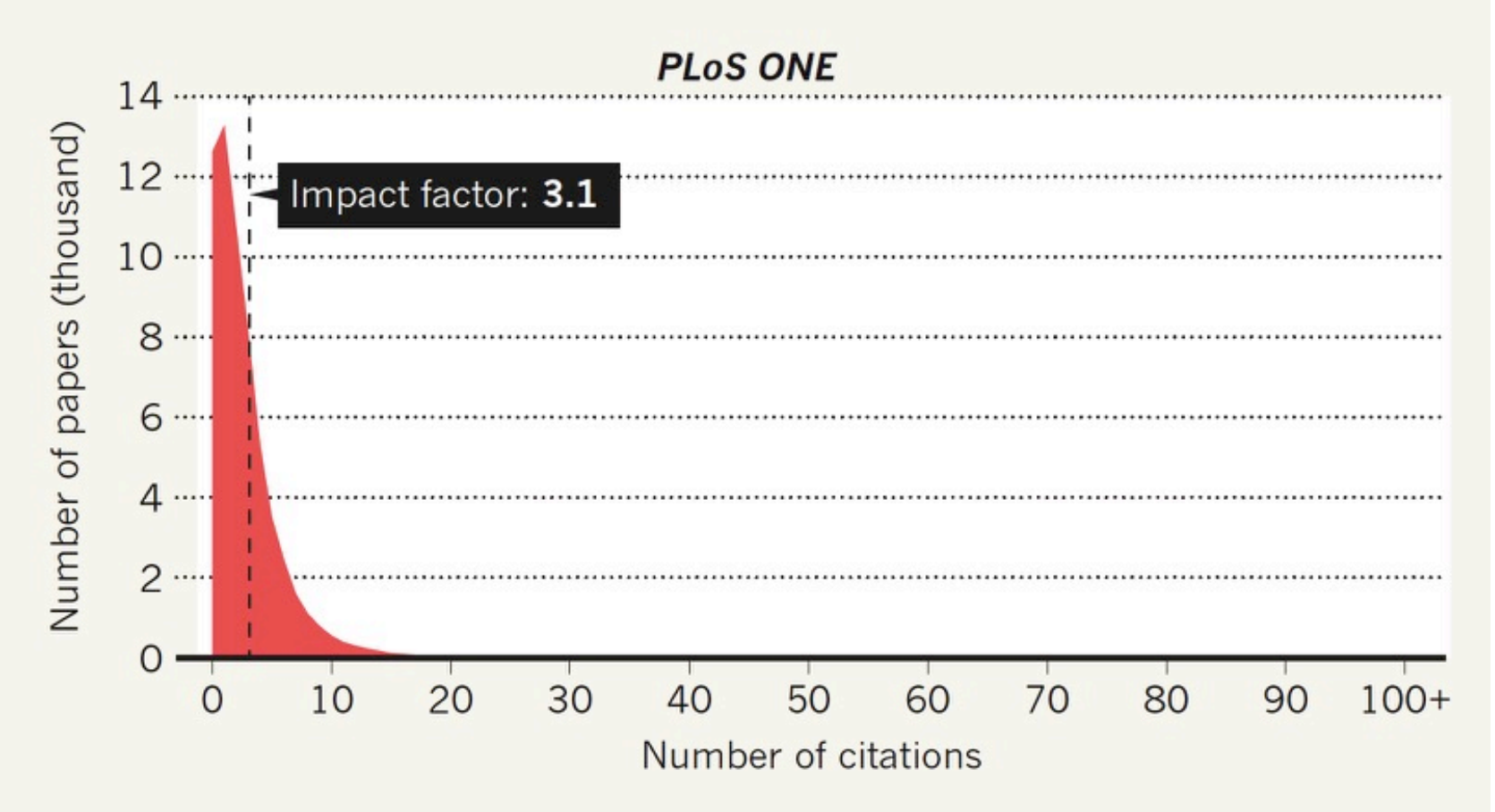
From: Lozano, Larivière, and Gingras. The weakening relationship between the impact factor and papers' citations in the digital age. Journal of the American Society for Information Science and Technology, 2012



Nature, 14 July 2016, Vol 535



Nature, 14 July 2016, Vol 535



Nature, 14 July 2016, Vol 535

Proposal

- Don’t publish the IF,

  but

- publish the distribution of citation frequency

Larivière et al., A simple proposal for the publication of journal citation distributions, bioRxiv preprint first posted online Jul. 5, 2016; doi: http://dx.doi.org/10.1101/062109

More criticism of the IF (Lozano et al.)

1) Some types of publications within journals, such as letters and commentaries, are used to count citations (the nominator), but do not themselves count as "papers" (the denominator), and hence inflate the journal's IF
2) the IF depends on the number of references, which differs among disciplines and journals
3) the inclusion of journals in the database depends solely on Thomson Reuters, a private company, and not on the fields' practitioners,
4) the exact IF published by Thomson Reuters cannot be replicated using publicly available data,
5) the distribution of citations/paper is not normal, so at the very least the mode or median ought to be used instead of the mean,
6) the 2-year span for papers followed by one year for citations is completely arbitrary and favours high-turnover over long-lasting contributions,
7) journal editors can manipulate and artificially inflate their IFs

More criticism of the IF (Lozano et al.) /2

- It does not make sense to use the IF of a journal as a proxy for paper quality, since today the number of citations of a paper can be accessed itself
- Even more troubling, is the 3-step approach of using the IF to infer journal quality, extend it to the papers therein, and then use it to evaluate researchers.

The h-index

Suggested in 2005 by Jorge E. Hirsch, a physicist at UCSD (= University of California, San Diego), as a single index tool for determining a researcher's productivity

Definition: For a multiset (bag) of natural numbers M, we say that

$$h(M) := \max \{ n \mid \text{there are } n \text{ elements } c \in M \text{ with } c \geq n \}$$

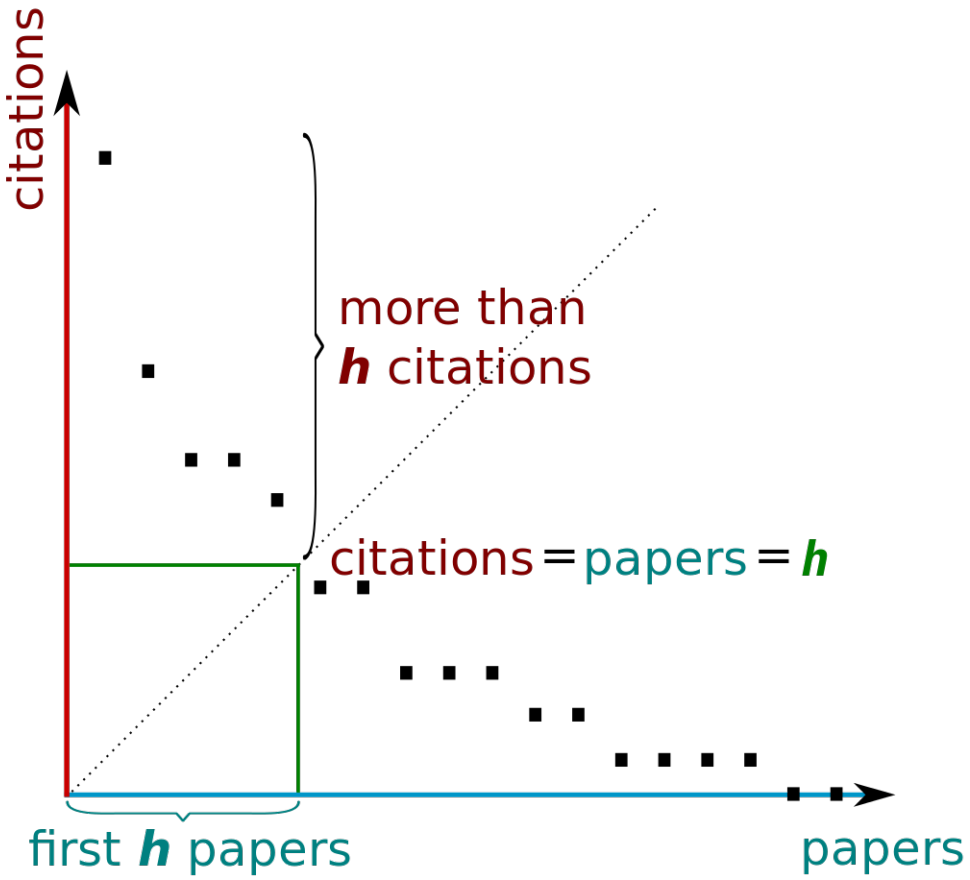
Examples:

$M_1 = \{\{100, 99, 80, 8, 4, 4, 1\}\}$ ➔ $h(M_1) = 4$

$M_2 = \{\{5, 5, 5, 5, 4, 4\}\}$ ➔ $h(M_2) = 4$

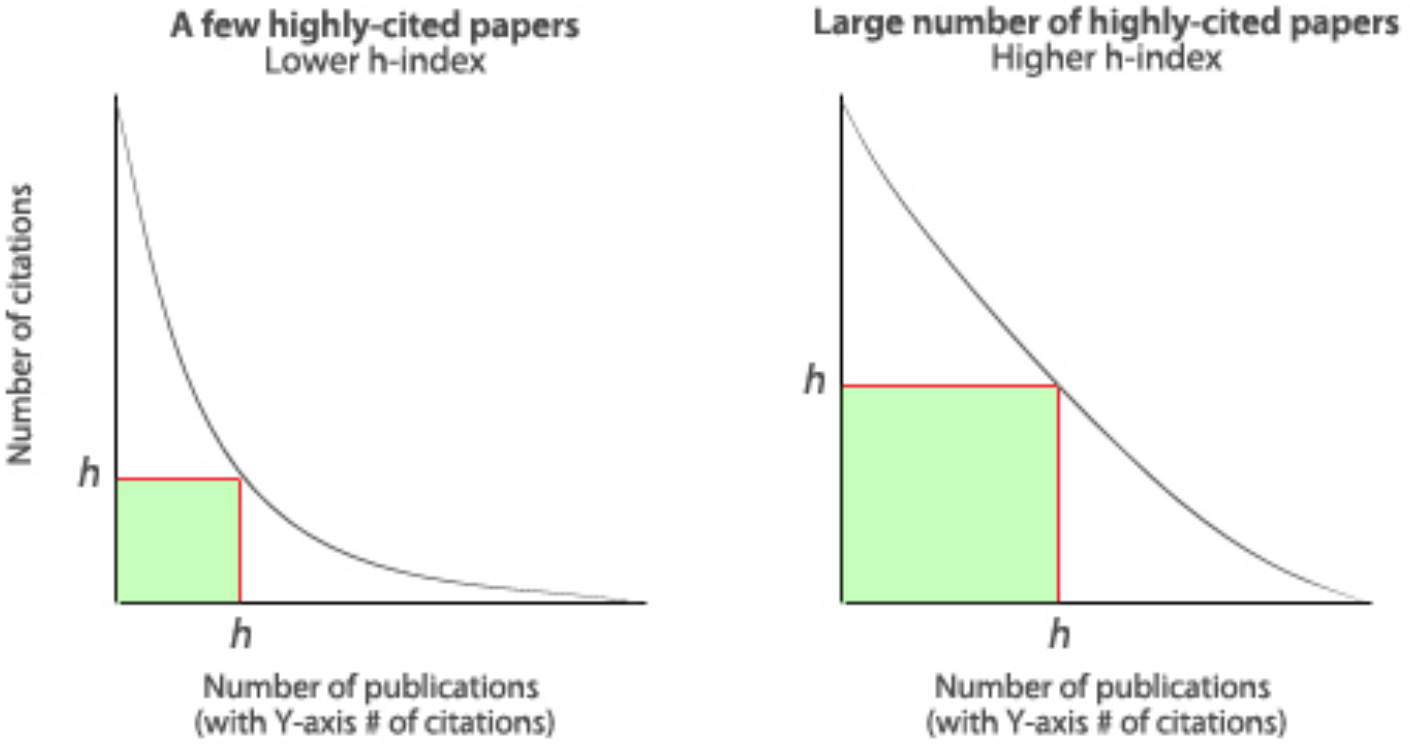
In bibliometrics, the multiset M of an author a is the collection of citation counts of a's papers

The h-index: graphically



Source: Wikipedia

The h-index: graphically (2)



Source: http://www.benchfly.com/blog/h-index-what-it-is-and-how-to-find-yours/

Growth of the h-index

If a researcher publishes at a steady rate, and produces papers of similar quality, then

- each paper collects a constant number, say c, of citations per year
- then, the total number of citations grows as y^2 over the number of years y

➔ the h-index grows linearly with y, i.e., $h \sim m\, y$

- the coefficient m depends
  - on the researcher
  - on the discipline

Advantages of the h-index

- It relies on citations to papers themselves, not the journals
- It is not dramatically skewed by a single well-cited, influential paper (unlike total number of citations would be)
- It is not increased by a large number of poorly cited papers (unlike total number of papers would be)
- It minimizes the politics of publication. A high-impact paper counts regardless of where it was published
- It's good for comparing scientists within a field at similar stages in their careers
- It may be used to compare not just individuals, but also departments, programs or any other group of scientists.

Cited from: http://www.benchfly.com/blog/h-index-what-it-is-and-how-to-find-yours/
Blog by Alan Marnett

Criticism of the h-index

The h-index

- favours papers with many authors
- discards information in author placement on author lists
- does not take into account the context of a citation

(e.g., favorable vs. critical,
fleshing out an introduction vs. result or method
enabling work in current paper)

Cited from: http://www.benchfly.com/blog/h-index-what-it-is-and-how-to-find-yours/
Blog by Alan Marnett

h-index: Summary

"In summary, I have proposed an easily computable index, h, which gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions. I suggest that this index may provide a useful yardstick with which to compare, in an unbiased way, different individuals competing for the same resource when an important evaluation criterion is scientific achievement."

Hirsch, J. E. (15 November 2005). "An index to quantify an individual's scientific research output". PNAS 102 (46): 16569–16572

The g-index

Idea: Give more credit to researchers with a few (or just one) landmark paper.

Definition: For a multiset (bag) of natural numbers M, we say that

g(M) is the largest number n such that

there are *n* numbers $c_1,...,c_n \in M$ with

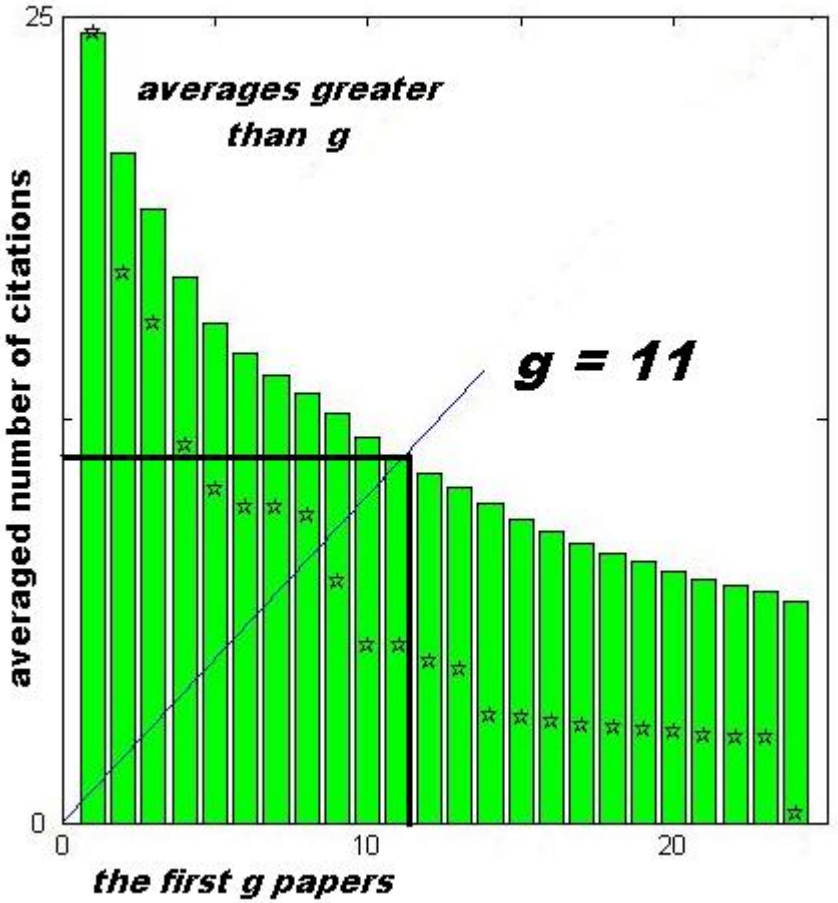
$(c_1+...+c_n)/n \geq n$

In other words, the average of the n numbers $c_1,\ldots,c_n$ is at least n.

Equivalently, we can require that n is the largest number such that we find counts $c_1,\ldots,c_n$ with $c_1+...+c_n \geq n^2$

Egghe, Leo (2006) Theory and practise of the g-index, Scientometrics, vol. 69, No 1, pp. 131–152

The g-index: graphically



An example of a g-index.

- The raw citation data are plotted with stars.
- The height of the *i*-th green bar shows the average citation number of the first *i* publications.
- While the g-index here is 11, the h-index is only 8.

Source: Wikipedia

Predictive power

Hirsch J. E. (2007). “Does the h-index have predictive power?”.
PNAS 104 (49): 19193–19198. Also: http://arxiv.org/abs/0708.0646

Compared four indices: N_p, N_c, n_c, h

Test on a sample of 50 physicists, each with a 24 years career

- divided career in 2 halves of 12 years
- computed the indices for the first half
- computed the indices for
  - all publications in the entire career (cumulative future performance)
  - only publications in 2nd half of career (exclusive future performance)
- found out:
  - h-index best at predicting itself and other indices (correlation of .91 and .89, resp., for self-prediction)

Hirsch: h-index favors productive authors

For papers with several authors:

- less prolific and junior authors benefit less from the number of citations to a paper

Assume, paper p has N_p citations.

It only helps authors to increase their h, if $h < N_p$

➔ if it helps a productive author (big H),
then its value is H for that author

➔ if it helps a less productive author (little h),
then its value is h for that author

However, other indexes (N_c, g) give the same credit for all authors of a paper

In reality, there is not “the” h-index

The h-index is always computed over some database:

- Google Scholar
- Scopus (http://scopus.com by Elsevier)
- Clarivate – Web of Knowledge (http://www.webofknowledge.com)

Databases differ wrt to the publications they cover …
(Sometimes it’s only a view of a db, e.g., based on the publications covered by your Web of Knowledge subscription)

Variants of the h-index:

- over the last x years
- w/ or w/o self-citations
- for institutions/journals/conferences/countries …

Challenges in creating a citation db: …

Who is Jing Wang?

search for publications

[+] [–] Jing Wang

> Home > Persons

by year Trier 1

This is just a *disambiguation page*, and is not intended to be the bibliography of an actual person. The links to all actual bibliographies of persons of the same or a similar name can be found below. Any publication listed on this page has not been assigned to an actual author yet. If you know the true author of one of the publications listed below, you are welcome to contact us.

[–] **Other persons with the same name**

- Jing Wang 0001 — Tsinghua University, Beijing, National Laboratory for Information Science and Technology
- Jing Wang 0002 — Chinese Academy of Sciences, Institute of Computing Technology, China
- Jing Wang 0003 — Chinese Academy of Sciences, Institute of Psychology, China (and 1 more)
- Jing Wang 0004 — Harbin Medical University, College of Bioinformatics, MA, USA
- Jing Wang 0005 — Bethume Cookman University, Daytona Beach, Florida
- Jing Wang 0006 — Texas A&M University, TX, USA
- Jing Wang 0007 — University of Houston, TX, USA
- Jing Wang 0008 — Rutgers University, New Brunswick, NJ, USA
- Jing Wang 0009 — Delft University of Technology, The Netherlands
- Jing Wang 0010 — University of Texas at Arlington, TX, USA

How can one distinguish authors?

- Keep track of author affiliation
- Analyze co-author graph
- Let authors maintain profiles, e.g., on Google Scholar; authors identify themselves as authors, add new papers
- Introduce author ids, e.g.

  ORCID-ID (Open Researcher and Contributor ID)

  ORCID is an initiative by publishing companies (Elsevier, Nature, Springer, …) and research organisations (CERN, …)

  Journals, research funding agencies, etc. require researchers to identify themselves by their ORCID-ID

Ranking of Conferences and Journals

CORE (= Computing Research and Education Association of Australasia) is an association of university departments of CS in Australia and New Zealand

- ranks CS conferences
- ranks CS journals

Based on introduction and reranking requests (with detailed arguments) by researchers

- decision about change are also based on Google Scholar and ArnetMiner/Aminer

See CORE conference ranking
(http://www.core.edu.au/index.php/conference-rankings)

Scimago

- Published by Scimago Lab, Granada, Spain
- Analyses science based on journal publications and their citations (uses Elsevier's Scopus database)
- Ranks journals according to SCImago Index, which is a refined version of PageRank;
- Idea: do not only count citations, but also the prestige of the journals that cites the work

http://www.scimagojr.com/

Google PageRank: Idea

Recursively define the weight of a node in a directed graph:
The weight of a node equals
the sum of the fractions of weights of the nodes
pointing to it
where the fraction is the weight divided by
the number of outgoing edges

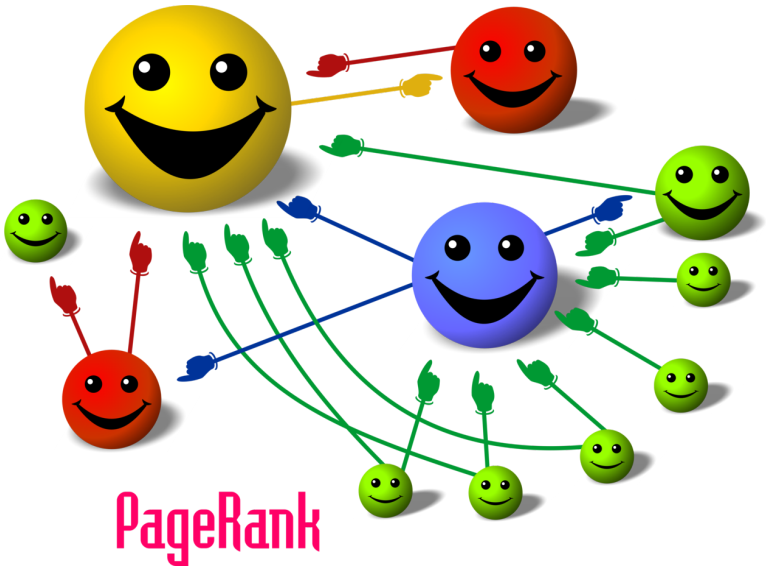


Illustration from Wikipedia

PageRank in a Formula (1st Try)

The rank of p equals the sum over the nodes with edges coming into p, where we sum the ranks of each node divided by the outdegree of a the node q.

$$R(p) = \sum_{q \in In(p)} \frac{R(q)}{out(q)}$$

What happens to

- sources
- sinks?

PageRank in a Formula (2nd Try)

View R(p) as the probability to visit p during during a random walk (that is, R(p) in [0,1]).

Introduce a damping factor d in (0,1).

The factor d captures the idea that during the random walk, we jump to a random page with probability d.

$$R(p) = \frac{1-d}{N} + d \sum_{q \in In(p)} \frac{R(q)}{out(q)}$$

where N is the number of nodes.

For a large graph, R(p) can be computed by iteration pretty quickly (~ 100 iterations)

Other Resources: Guide2Research

- Guide2Research (http://www.guide2research.com) maintained by Imed Bouchrika,
  University of Souk Ahras, Algeria
  lists journals, conferences, researchers according to
  - h-index
  - impact factor
  - SCImago Journal Rank

Other Resources: GII-Grin-SCIE Conference Ranking

- Joint initiative of the organisations of Italian (GII, GRIN) and Spanish (SCIE) computer science academics
- Creates a basis for national research evaluation exercises (VQR in Italy) and evaluation of researchers for promotion (abilitazione nazionale)
- Based on 3 sources
  - CORE evaluation from Australia
  - Microsoft Academic collection of publications
  - LiveSHINE (Google Scholar-based conference ranking)
- H-index and lifetime IF are computed, using MS Academic and LiveSHINE, and combined into one ranking
- Finally, the three classifications are combined in to one

http://valutazione.unibas.it/gii-grin-scie-rating/conferenceRating.jsf

Other Resources: U Michigan

Research Guides: Research Impact Metrics : Citation Analysis
(http://guides.lib.umich.edu/citation)

- provided by the University library
- gives an overview of
  - bibliographical databases (Web of Science, Scopus, Google Scholar, ACM DL, …)
  - journal rankings
  - h-index