Scientometrics via Low-Rank Approximation and Completion

Ivan Markovsky

Vrije
Universiteit
Brussel

erc

Plan

Introduction

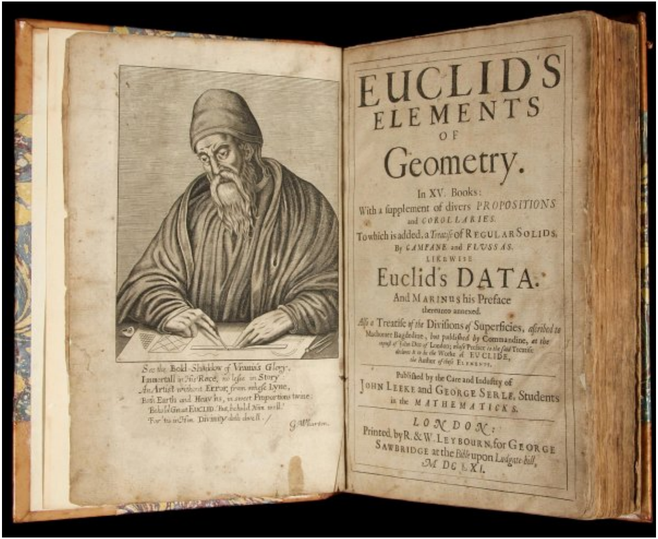
Science evaluation as data processing

Challenges

The rationale for publishing in peer reviewed journals is ranking

1. dissemination
2. quality check
3. prioritization of the literature

Before the existence of journals scientists published in books



See the Bold Shaddow of Vrania's Glory,
Immortall in His Race, no leſse in Story:
An Artist without Error, from whoſe Lyne,
Both Earth and Heav'ns, in sweet Proportions twine:
Behold Great EUCLID. But, behold Him well!
For 'tis in Him Divinity doth dwell.

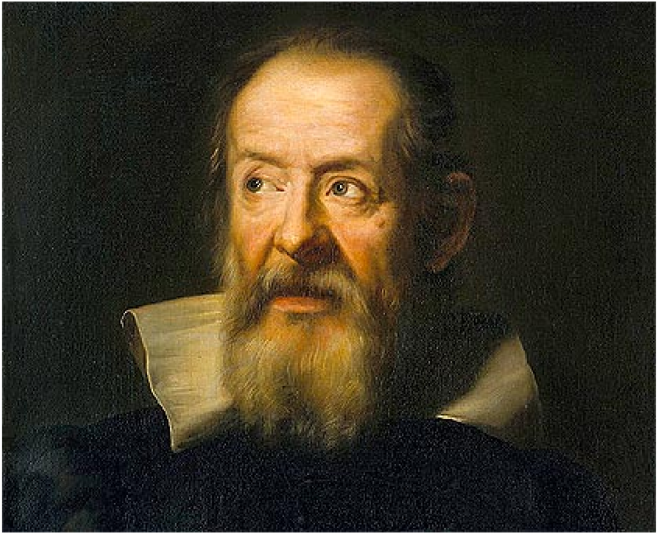
G. Wharton.

EUCLID'S
ELEMENTS
OF
Geometry.
In XV. Books:
With a ſupplement of divers PROPOSITIONS
and COROLLARIES.
To which is added, a Treatiſe of REGULAR SOLIDS,
By CAMPANE and FLUSSAS.
LIKEWISE
Euclid's DATA.
And MARINUS his Preface
thereunto annexed.
Alſo a Treatiſe of the Diviſions of Superficies, aſcribed to
Machomet Bagdedine, but publiſhed by Commandine, at the
requeſt of John Dee of London; whoſe Preface to the ſaid Treatiſe
declares it to be the Worke of EUCLIDE,
the Author of theſe ELEMENTS.

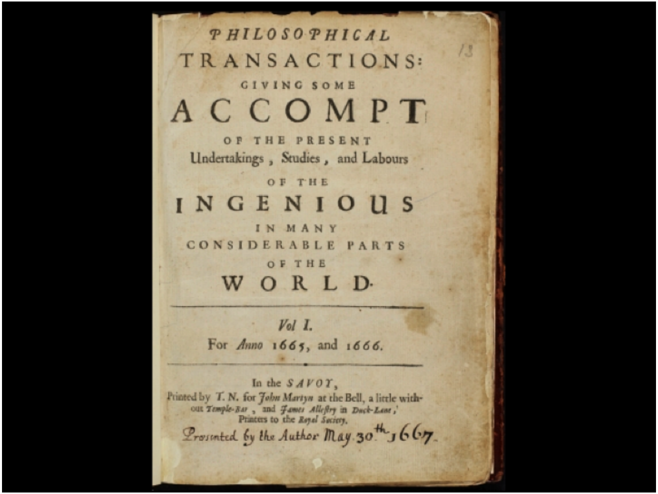
Publiſhed by the Care and Induſtry of
JOHN LEEKE and GEORGE SERLE, Students
in the MATHEMATICKS.

LONDON:
Printed by R. & W. LEYBOURN, for GEORGE
SAWBRIDGE at the Bible upon Ludgate-hill,
M DC LXI.

. . . or encoded their findings in anagrams



1752–1960 peer review was done by editors



13

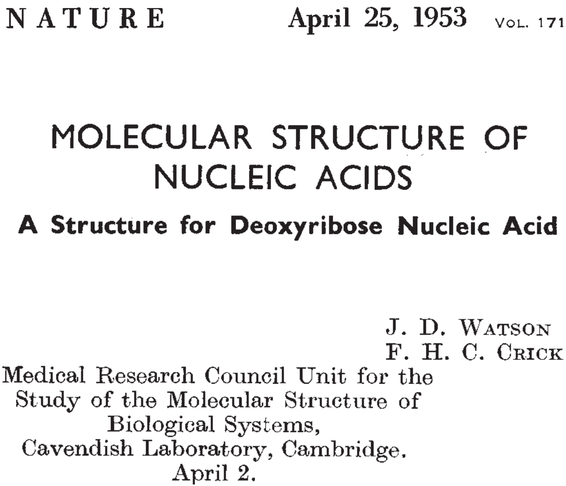
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Alleſtry* in *Duck-Lane*,
Printers to the *Royal Society*.

Presented by the Author May 30th 1667.

The discovery of the DNA structure was published based on editorial reviews only



NATURE April 25, 1953 Vol. 171

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the
Study of the Molecular Structure of
Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

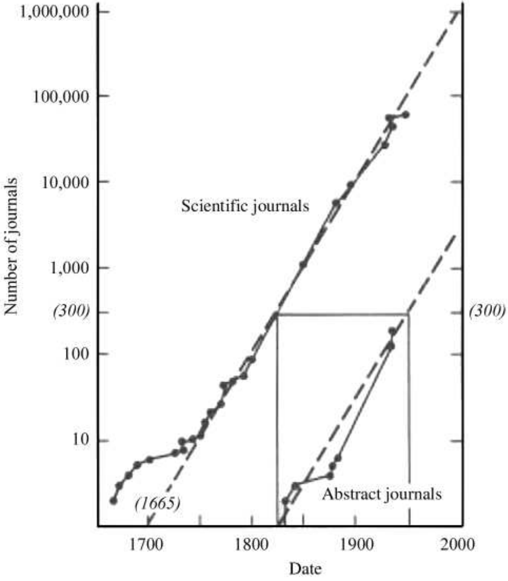
Only one of Einstein’s 300 papers was peer reviewed

Only one of Einstein’s 300 papers was peer reviewed

Dear Sir,

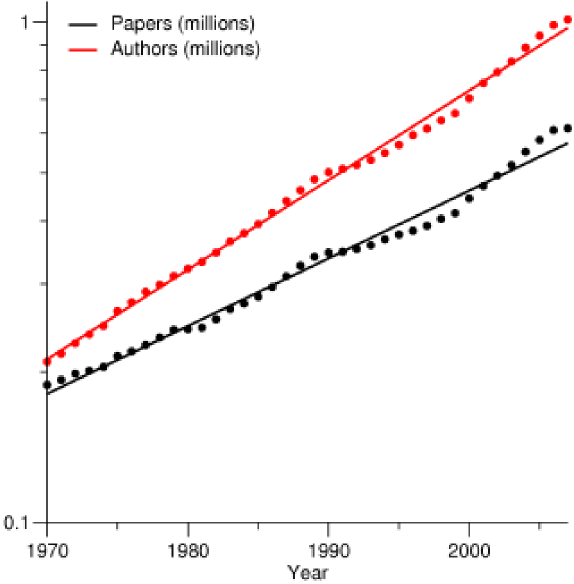
We (Mr. Rosen and I) had sent you our manuscript for publication and had not authorized you to show it to specialists before it is printed. I see no reason to address the in any case erroneous comments of your anonymous expert. On the basis of this incident I prefer to publish the paper elsewhere.

The rate of publication increases exponentially



Source: D. de Solla Price, Science since Babylon, Yale, 1961.

The rate of publication increases exponentially



Source: Medline database

Increased number of submissions requires external reviewers

From Computer Desktop Encyclopedia
Reproduced with permission.
© 1996 Xerox Corporation

Research becomes more and more interdisciplinary

finding new research topics becomes harder

combination of topics is a way to gain novelty

example:

> *Non-fragile reduced-order dynamic output feedback H-infinity control for switched systems with average dwell-time switching*

reviewers need to know four different research topics

Finding suitable reviewers is challenging also due to conflict of interests and bias

well chosen peers are coworkers or competitors

some reviewers are more critical

how to calibrate?

The two faces of scientists and the two sides of the scientific output

	visible	hidden
people	authors	reviewers
output	papers	reviews
reward	yes	no

data/software are not always visible and rewarded

Redundancy of the literature: a more difficult problem than plagiarism

can be intentional or unintentional

checking novelty is no longer feasible for human

automatic methods need to do semantic search

Current trend: open science

post publication open review (discussions)

Wikipedia: an example of collaborative effort

Why the legacy peer review practice persist?

financial interest of publishers

inertia of the scientific community

Plan

Introduction

Science evaluation as data processing

Challenges

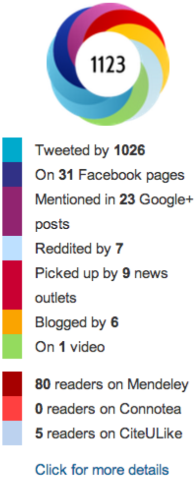
How to measure the impact of a paper?

citations

ratings

alt-metrics

- views and downloads
- bookmarks
- conversations



1123

🟦 Tweeted by **1026**
🟦 On **31** Facebook pages
🟪 Mentioned in **23** Google+ posts
🟦 Reddited by **7**
🟥 Picked up by **9** news outlets
🟧 Blogged by **6**
🟩 On **1** video

🟥 **80** readers on Mendeley
🟥 **0** readers on Connotea
🟦 **5** readers on CiteULike

Click for more details

Particularities of the data collected in the peer review process

multivariate

ordinal

missing values

Posing the problem as matrix completion

$$\begin{array}{rccccc}
 & \text{paper 1} & \text{paper 2} & \text{reviewer 3} & \text{paper 4} & \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \cdots \\
\text{reviewer 1} \rightarrow & * & ? & ? & * & \cdots \\
\text{reviewer 2} \rightarrow & ? & * & ? & ? & \cdots \\
\text{reviewer 3} \rightarrow & * & ? & * & ? & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

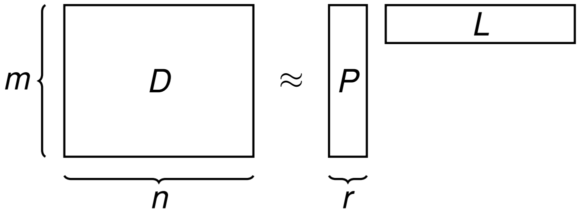
$*$ — available rating

? — missing rating

Analogy with data modeling in engineering

engineering	scientometrics
phenomenon	papers
sensors	reviewers
experiment design	reviewer selection
measured data	collected reviews
dynamical model	low-rank model

Differences between engineering applications and scientometrics



engineering	scientometrics
D given	D has missing elements
real data	ordinal data
n grows	n grows
m fixed	m grows

In system identification D is structured

D Hankel $\quad\Longrightarrow\quad$ P and L (generlized) Vandermonde

$\leadsto$ subspace identification methods

kernel representation:

$$\operatorname{rank}(D) \leq r \quad\iff\quad \text{there is f.r.r. } R \in \mathbb{R}^{m-r \times m}, \text{ such that } RD = 0$$

R is a model parameter (unstructured)

Usage of the model

ranking of papers

paper recommendations

reviewer assignment

Low-rank approximation problem

$$\begin{array}{lll} \text{minimize} & \text{over } \widehat{D} & \sum w_{ij}(d_{ij} - \widehat{d}_{ij})^2 \\ \text{subject to} & \text{rank}(\widehat{D}) \leq r & \end{array}$$

$w_{ij} = 0$ if d_{ij} is missing (and 1 otherwise)

additional constraints on $\widehat{D}$:

- non-negativity
- upper bound
- integer valued

Weights incorporate prior knowledge

w_{ij} reflects the "trustworthiness" of the rating d_{ij}

errors-in-variables model:

$$D = \text{"true value"} + \text{"noise"}$$

assume zero mean, independent, Gaussian noise

let v_{ij} be the variance of the noise on d_{ij}

then, $w_{ij} = 1/v_{ij} \quad \leadsto \quad$ maximum-likelihood estimate

$w_{ij} = 0$ — infinite noise (completely untrustworthy rating)

Recursive update of the model

by adding papers

by adding reviewers

rank adaptation

Plan

Introduction

Science evaluation as data processing

Challenges

Challenges

is the low-rank assumption relevant?

effect of quantization on the singular values?

effect of truncation on the singular values?

how to exploit the multidimensional aspect?

how to merge data from multiple sources?

Closing

peer review needs revision

issue 1: more and higher quality data

issue 2: advanced data processing

low-rank approximation is promising approach

we can shape the future of peer review