International Workshop on

Challenges and Visions in the Social Sciences

ETH Zurich, Aug 18-23, 2008

Measuring Science
Fears, Challenges and Opportunities

Luis A. N. Amaral

Northwestern Institute on Complex Systems

Dept of Chemical and Biological Engineering

Northwestern University

To measure or not to measure

- Is Science (i.e., the impact of papers and scientists) measurable?
- Are the measures meaningful?
- Can use a scalar or do we need a vector?
- Do these measures have characteristic scales?

Some measures

- Papers
  - Impact factor of journal
  - Number of citations
  - Betweenness of paper in citation network
  - Page ranking (a la Google)
- Scientists
  - Number of papers
  - Number of citations
  - Average number of citations
  - h-index
  - ...

What we (think we) know

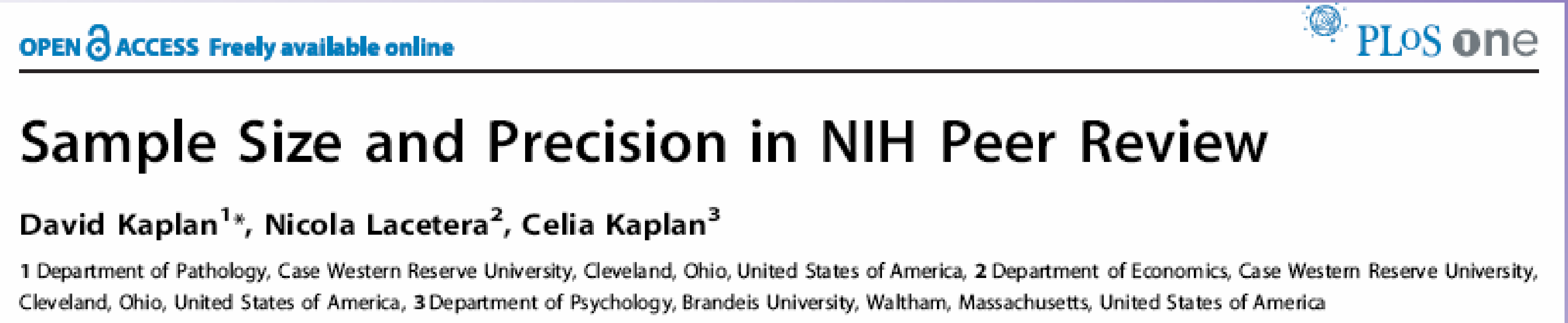
- Many measures have highly-skewed distributions
  - Most papers cite 30-100 other papers
    - ✓ Some, however, cite thousands
  - Distribution of number of citations is said to decays as a power law
    - ✓ About 30% of the papers published each year never get cited
    - ✓ Median number of citations is about 5
    - ✓ Highest-cited paper has nearly 200 thousand citations
  - Some researchers publish thousands of papers
    - ✓ Most researchers publish < 200 papers

What we (think we) know

- The values of these measures appear to be strongly correlated for both researchers and organizations
  - This is not true for researchers in the early stages of their career
  - Results are primarily obtained for highly-successful researchers
  - Broad analysis is difficult because of issues with name disambiguation (who is Smith J?)
- What is correct procedure when they are not strongly-correlated?
  - What is better 1 paper with 1000 citations or 10 papers with 100 citations?

The fears

- Do scientific papers have an intrinsic quality, i.e., would a set of qualified independent reviewers provide an asymptotically unbiased score?
  - How many reviewers would we need for accurate scoring?



OPEN ACCESS Freely available online

PLoS one

Sample Size and Precision in NIH Peer Review

David Kaplan[1*], Nicola Lacetera[2], Celia Kaplan[3]

1 Department of Pathology, Case Western Reserve University, Cleveland, Ohio, United States of America, 2 Department of Economics, Case Western Reserve University, Cleveland, Ohio, United States of America, 3 Department of Psychology, Brandeis University, Waltham, Massachusetts, United States of America

$$L = \sqrt{\frac{(z_{\alpha/2}\sigma)^2}{n}} = \sqrt{\frac{(1.96{*}1)^2}{4}} \approx 1$$

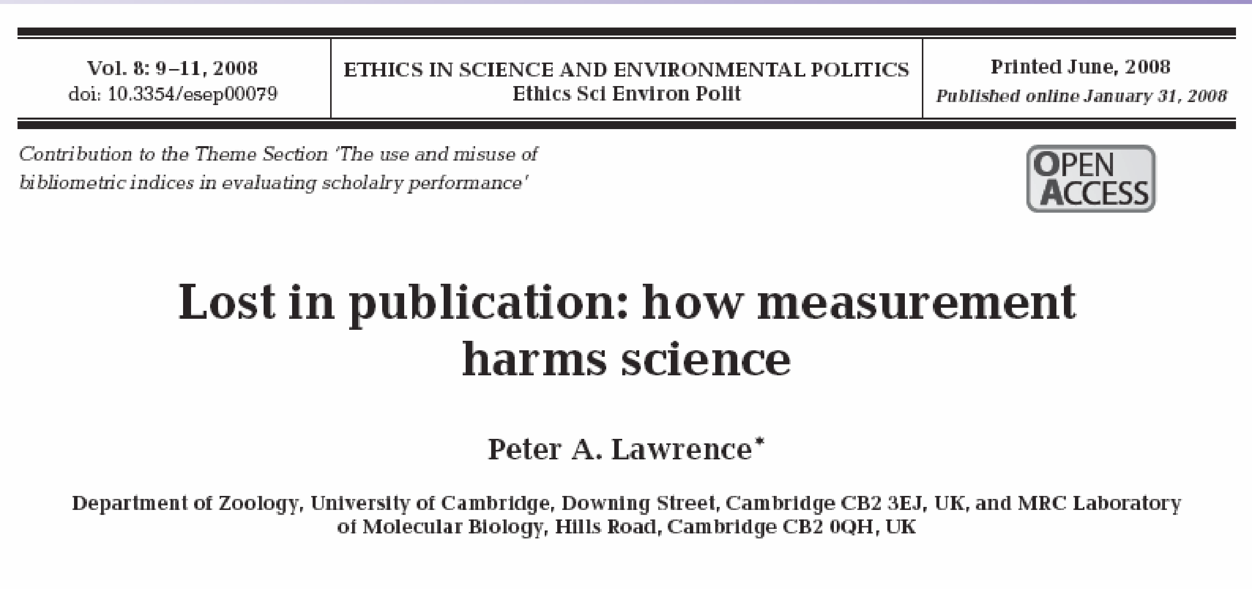
$$n \geq \left(\frac{z_{\alpha/2}\sigma}{L}\right)^2 = \left(\frac{1.96{*}1}{0.01}\right)^2 = 38{,}416$$

The fears

- Do scientific papers have an intrinsic quality, i.e., would a set of qualified independent reviewers provide an asymptotically unbiased score?
  - How many reviewers would we need for accurate scoring?

- Even if answer is YES, will measurement lead to “throwing out the baby with the bath water”, i.e., is measurement harmful to science?

QuickTime™ and a
decompressor
are needed to see this picture.



Vol. 8: 9–11, 2008 doi: 10.3354/esep00079	ETHICS IN SCIENCE AND ENVIRONMENTAL POLITICS Ethics Sci Environ Polit	Printed June, 2008 *Published online January 31, 2008*

Contribution to the Theme Section 'The use and misuse of bibliometric indices in evaluating scholalry performance'

OPEN ACCESS

Lost in publication: how measurement harms science

Peter A. Lawrence*

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK, and MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

The fears

- Do scientific papers have an intrinsic quality, i.e., would a set of qualified independent reviewers provide an asymptotically unbiased score?
    - How many reviewers would we need for accurate scoring?

- Even if answer is YES, will measurement lead to “throwing out the baby with the bath water”, i.e., is measurement harmful to science?

- Assuming that papers can be scored and that “good” measurement is not harmful, can measurement be useful, i.e., are there low-dimensional summary measures that quantify the quality of a set of papers?

Why we need to ignore the fears

- Millions of new papers are published every year

- Hundreds of thousands of full time scientists are engaged in a broad range of fields of science

- Many scientific questions require multidisciplinary tools, techniques, or concepts

- Accumulated knowledge makes is impossible for a single person to:
  - have expertise in all the areas needed to address many interesting scientific questions
  - be able to evaluate expertise by others in all areas needed to address many interesting scientific questions

The data

- Thomson Reuters' *Web of Science* database (to Dec 31st 2006)
  - **20 million** scientific articles (published 1955 to 2006)
  - **5,800** science and engineering journals
  - **1,700** social science journals
  - **1,100** arts and humanities journals
  - Over **1 million** researchers

- Faculty list of top 30 US chemical engineering departments
  - PhD date
  - Publications with number of citations up to 2006

Scientific papers

Clearly, we cannot read them all.

We cannot even read all the papers published in our areas of interest.

Even if we could read them all, should we?

How can we find what to read as soon as it is published?

Heuristics to the rescue

We can read from *good* journals. (part 1)

We can read from *good* authors. (part 2)

We can read what others read/recommend/cite.

How can we identify if a newcomer
(scientist or journal) is any good?

Journal Impact Factor (JIF)

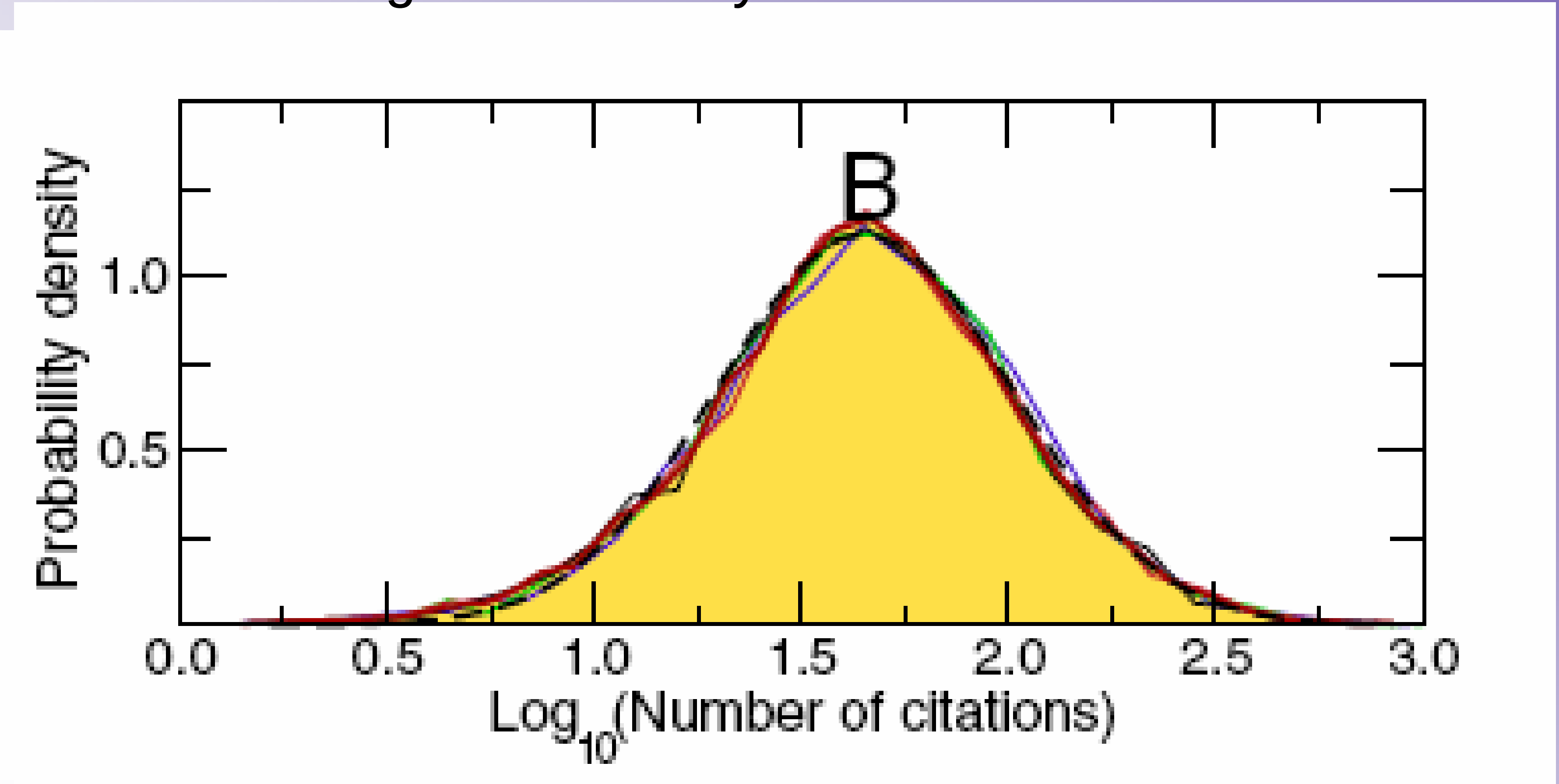
JIF is broadly (mis)used.

It cannot be pure crap, otherwise it would not survive as a heuristic.

Can we do better, though?

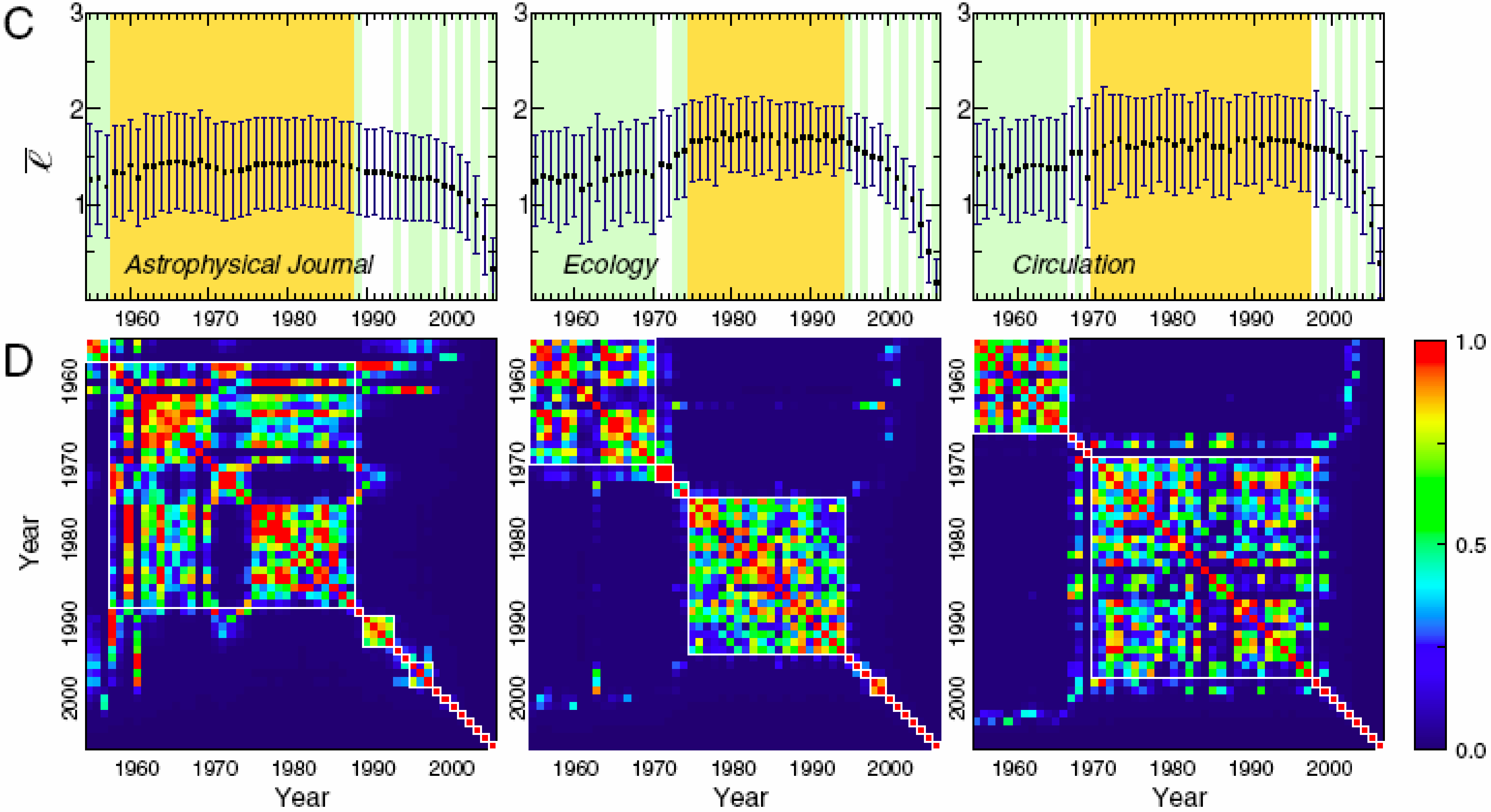
Journal citation distribution

Journal of Biological Chemistry



Stringer, Sales-Pardo & Amaral, PLoS One **3**, e1683 (2008)

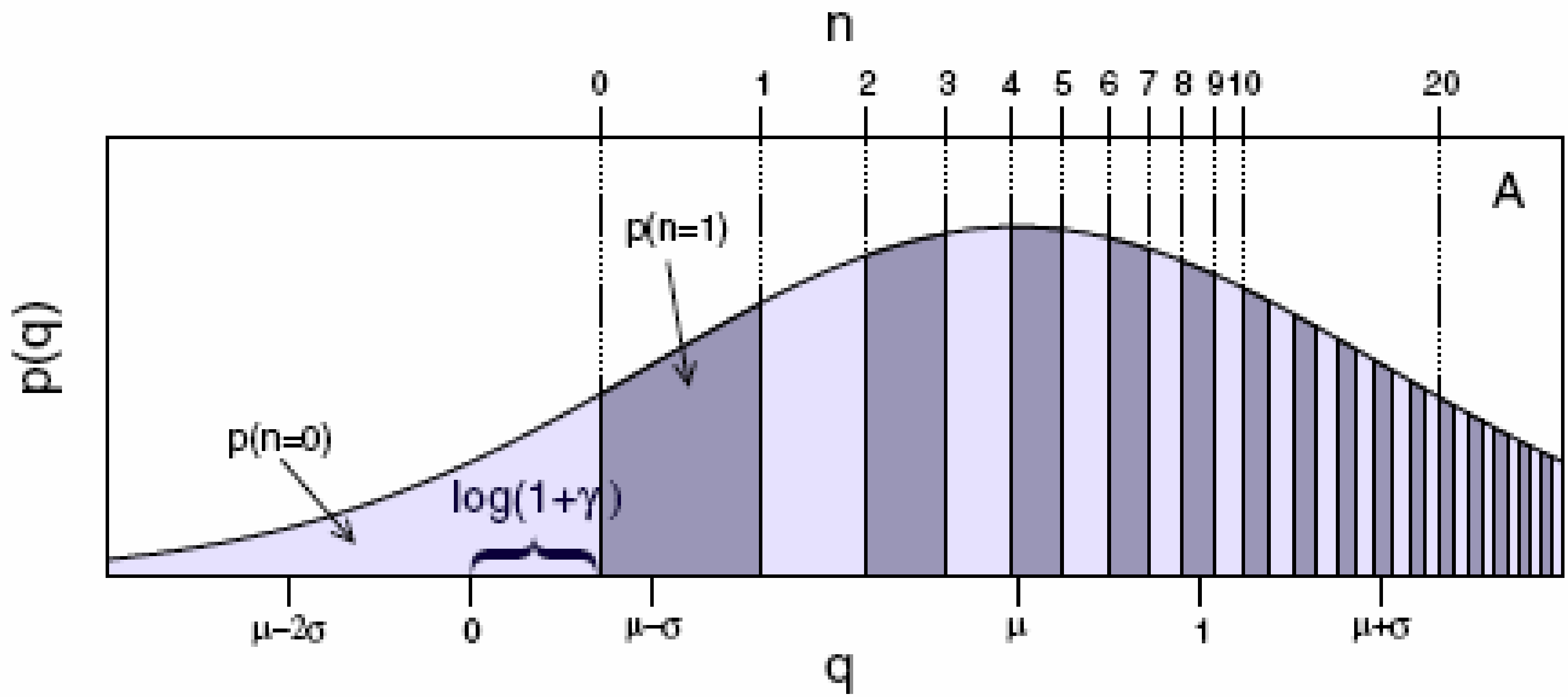
Distribution convergence



Stringer, Sales-Pardo & Amaral, PLoS One **3**, e1683 (2008)

Model

$$n = \max(0, \mathrm{floor}[\exp(q) - \gamma])$$

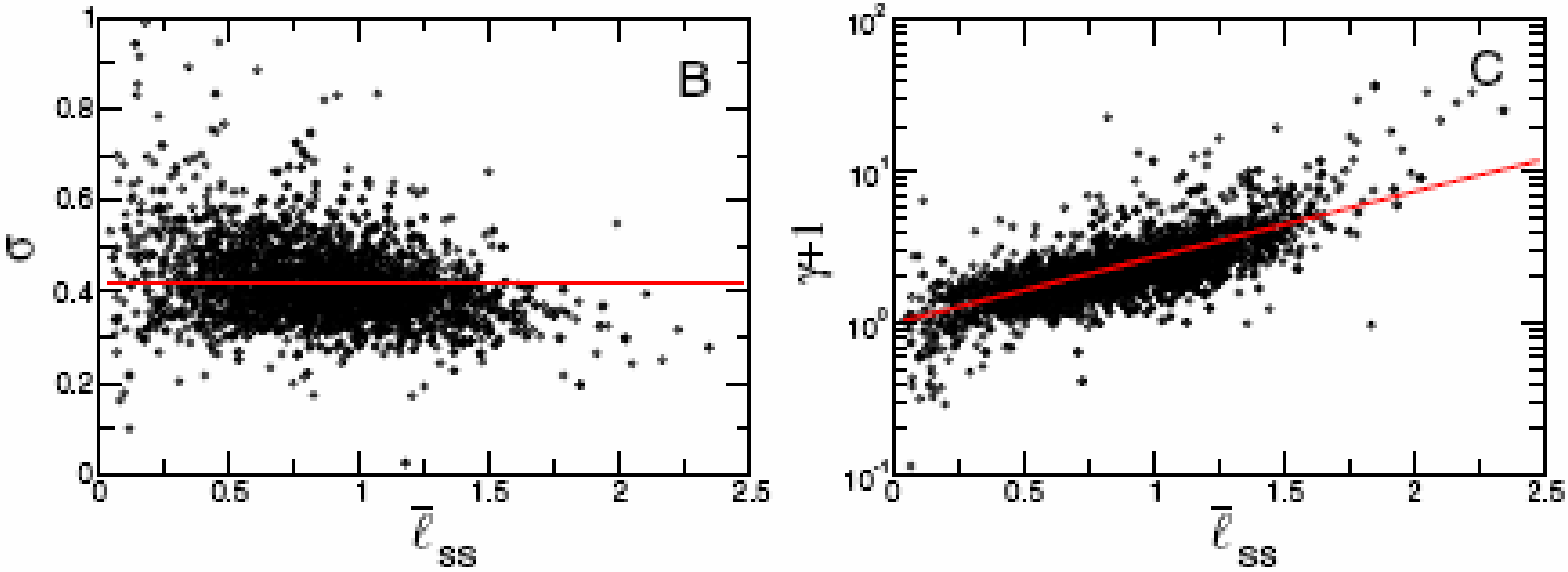


q is normally distributed

so ***mean*** and ***standard deviation*** provide all information

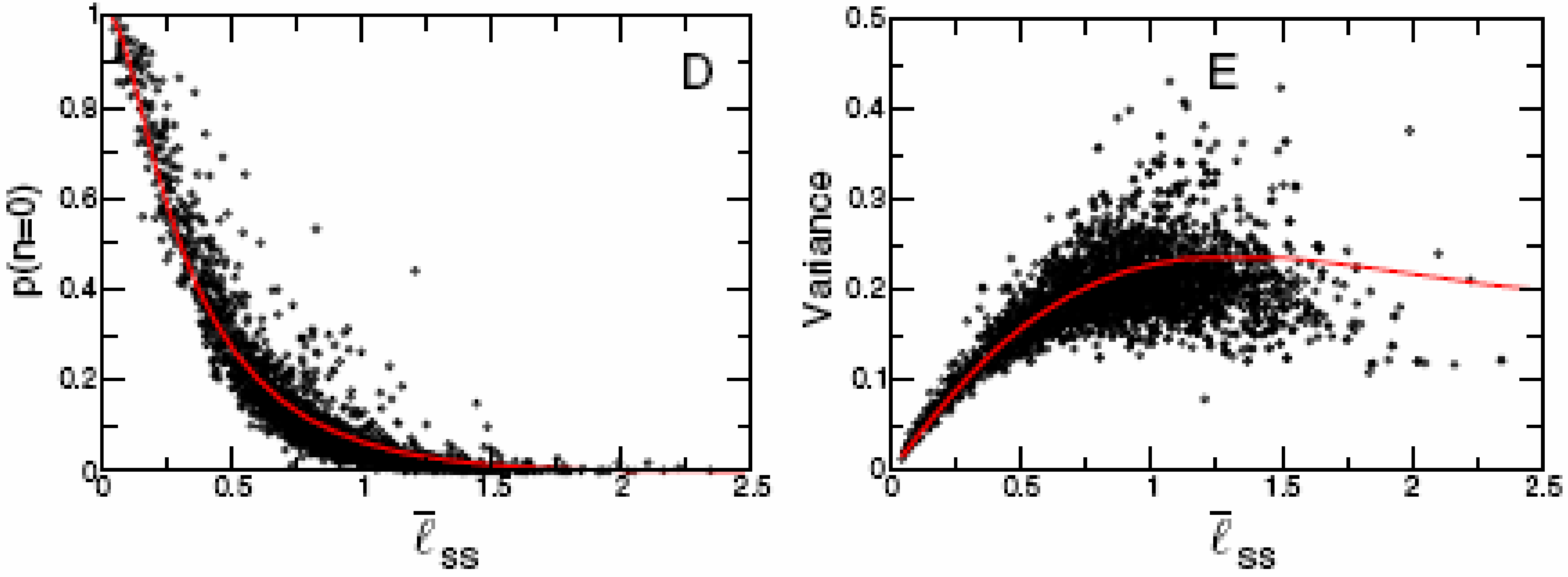
Parameter estimation

Minimization of χ^2 statistic: bin data, require about 10 data points per bin



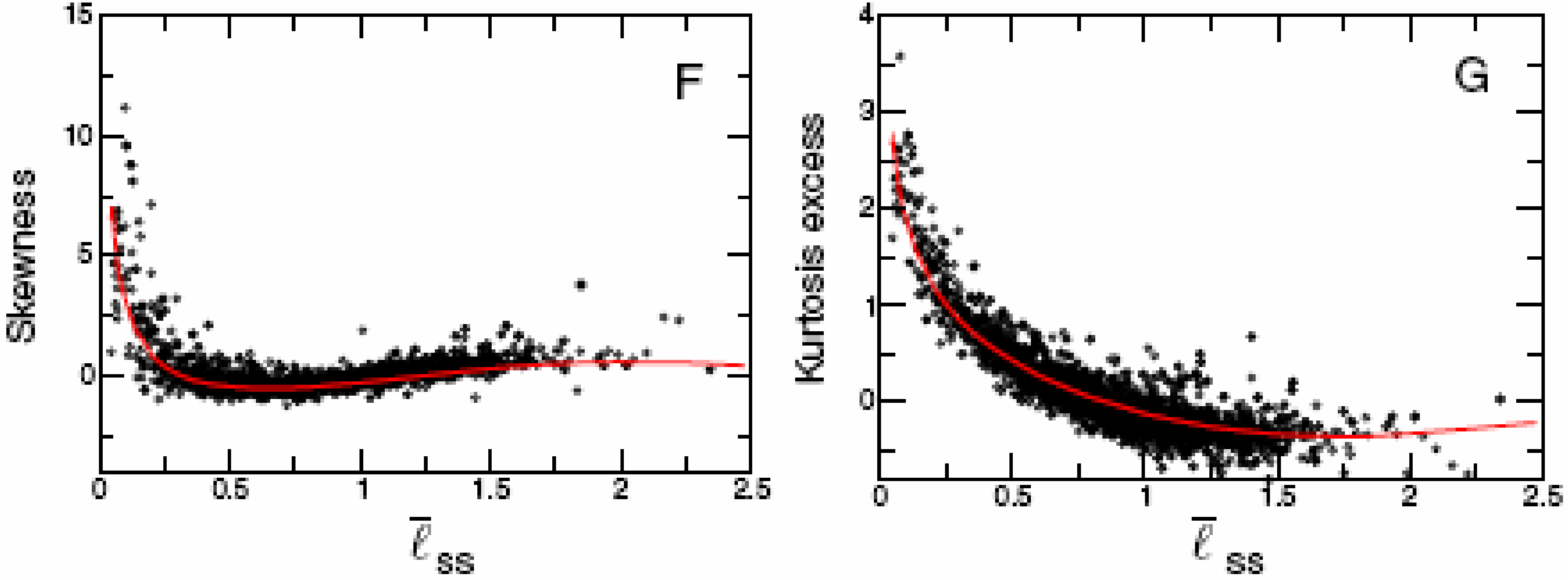
$$n = \max(0, \text{floor}[\exp(q) - \gamma])$$

Model validation



$$n = \max(0, \text{floor}[\exp(q) - \gamma])$$

Model validation



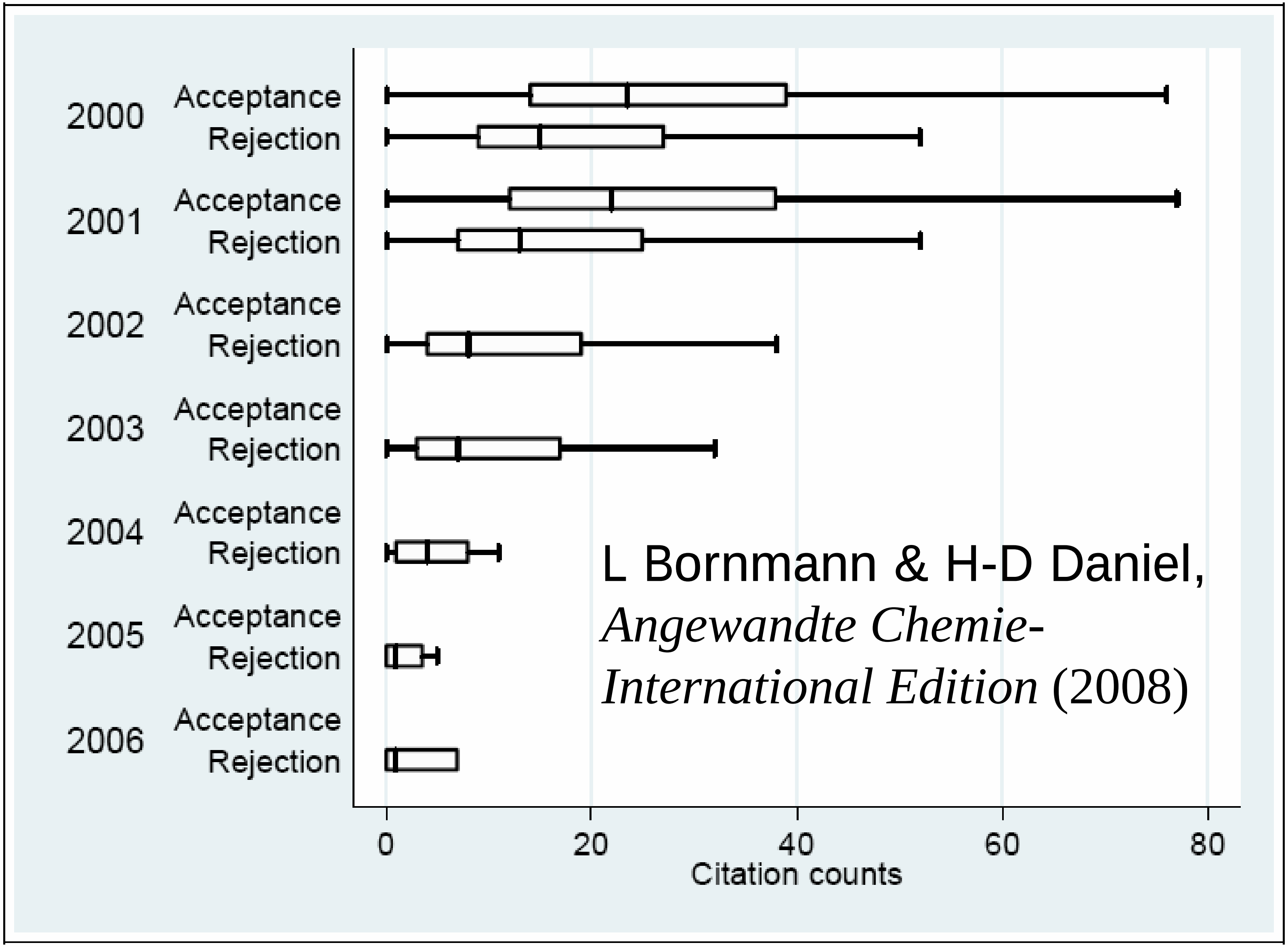
$$n = \max(0, \mathrm{floor}[\exp(q) - \gamma])$$

What we have learned so far

- *q* of papers published in a journal is normally-distributed. **Outcome is not!**
  - Think of Rosen's *Economics of Superstars* or Watts' *Inequality and….*

- What makes up *q*?
  - **noise** + intrinsic quality + journal effect
  - noise + **intrinsic quality** + journal effect
  - noise + intrinsic quality + **journal effect**
  - noise + intrinsic quality + journal effect

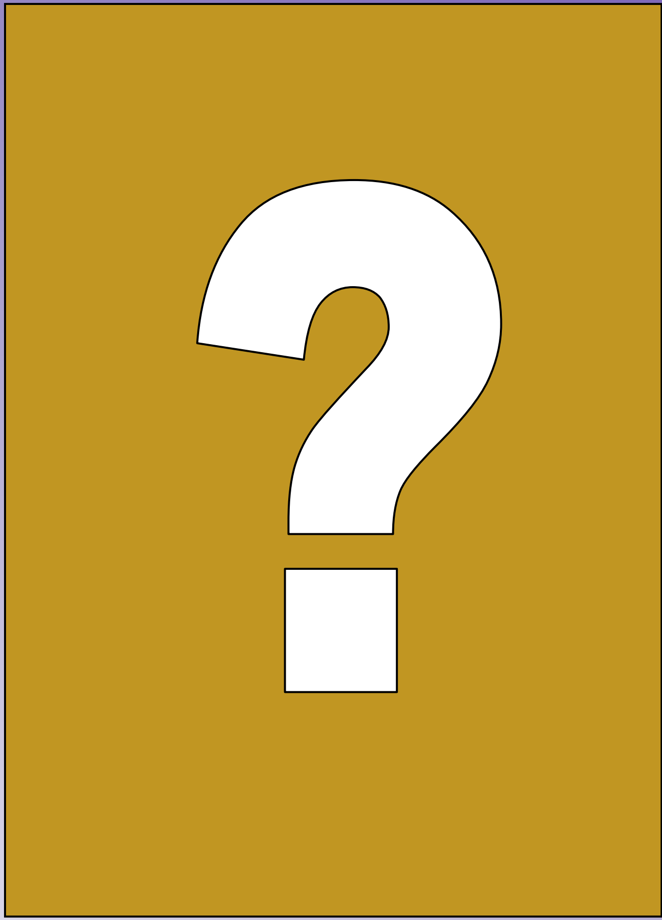
Some more on *q*



L Bornmann & H-D Daniel, *Angewandte Chemie-International Edition* (2008)

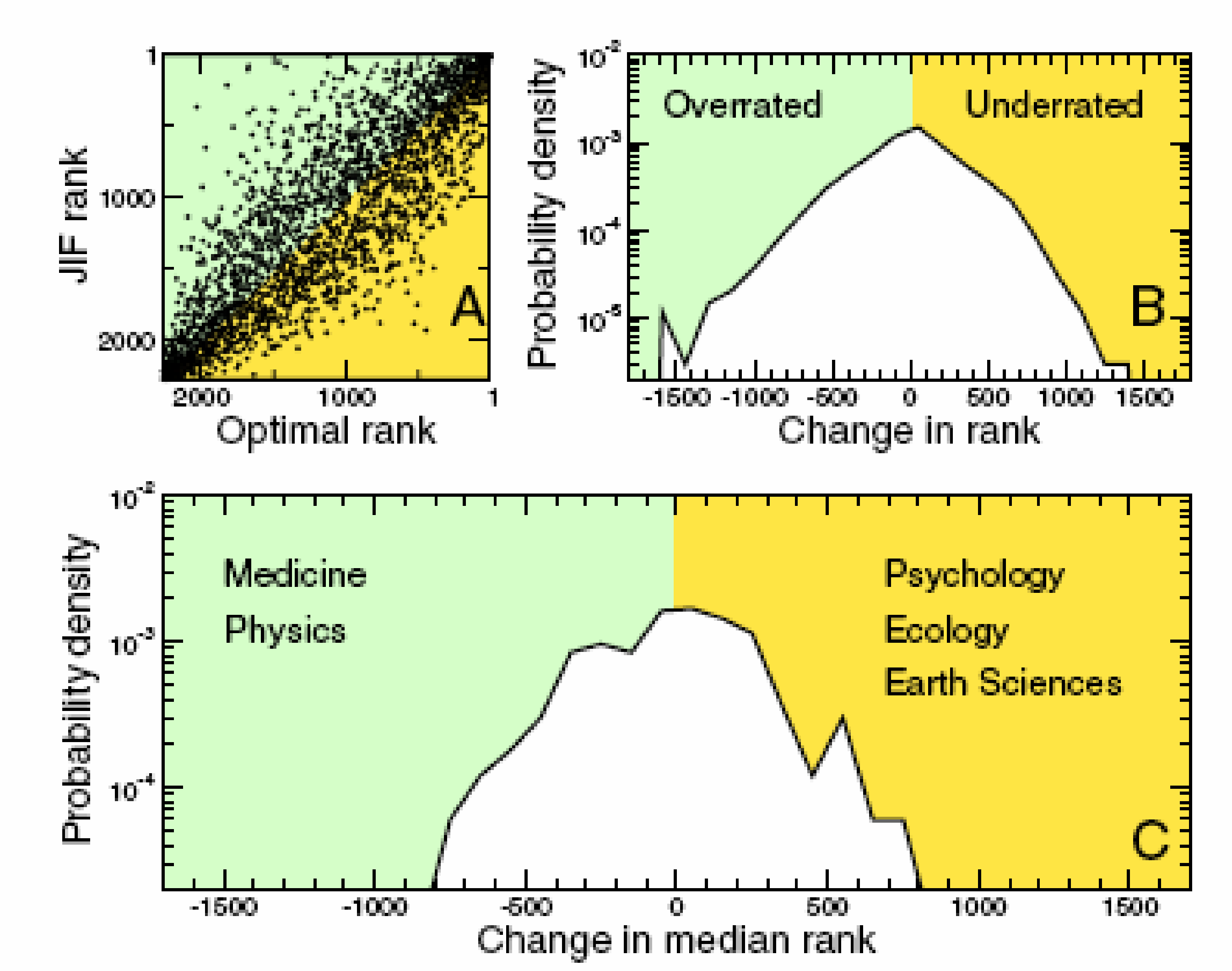
A little quiz

Journal	Stat. Period	JIF
Phys Rev Lett	1966-2006	7.072
Ecology	1974-1994	4.782
Am Sociol Rev	1955-1996	3.205
Econometrica	1980-1996	2.402
Ann Math	1955-1995	2.406



Stringer, Sales-Pardo & Amaral, PLoS One **3**, e1683 (2008)

A little surprise



What we have learned so far

- *q* of papers published in a journal is normally-distributed. **Outcome is not!**
  - Think of Rosen's *Economics of Superstars* or Watts' *Inequality and….*

- What makes up *q*?
  - **noise** + intrinsic quality + journal effect
  - noise + **intrinsic quality** + journal effect
  - noise + intrinsic quality + **journal effect**
  - noise + intrinsic quality + journal effect

What can we use these results for?

Information retrieval

Pick at random:

Paper from *Journal 1* published in year *Y*

Paper from *Journal 2* published in year *Y*

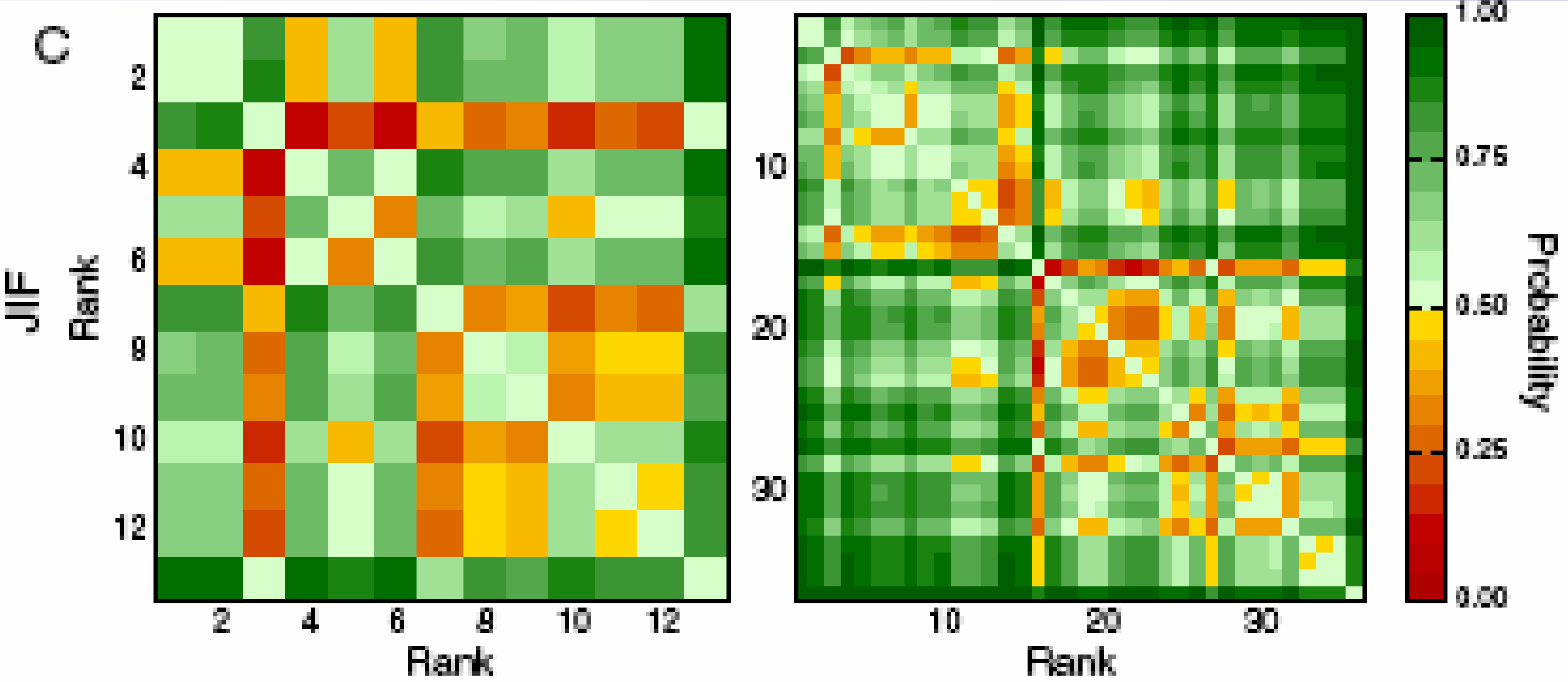
What is ***probability*** that paper from higher-ranked journal has accrued more citations by year *Y*+15?

Stringer, Sales-Pardo & Amaral, PLoS One **3**, e1683 (2008)

Information retrieval (JIF)

Experimental psychology

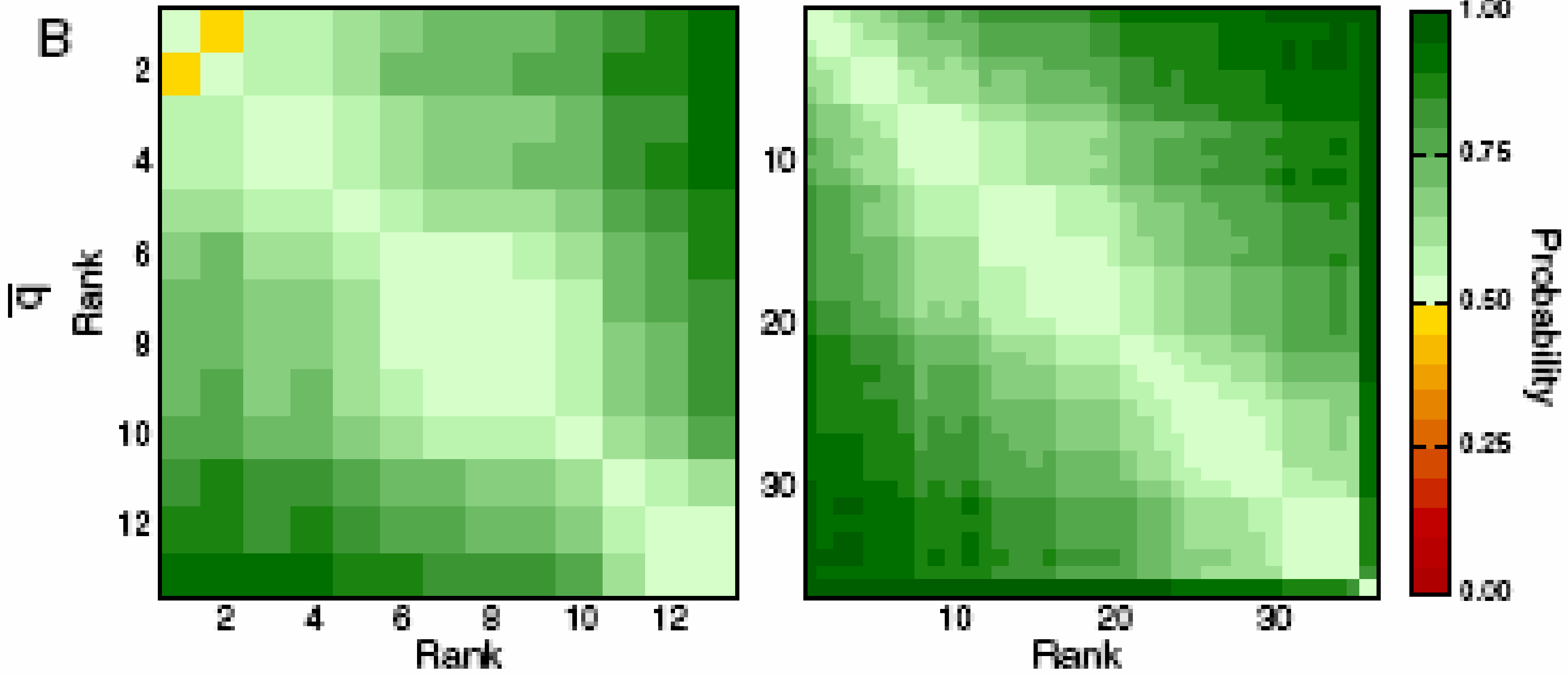
Ecology



Information retrieval (q)

Experimental psychology

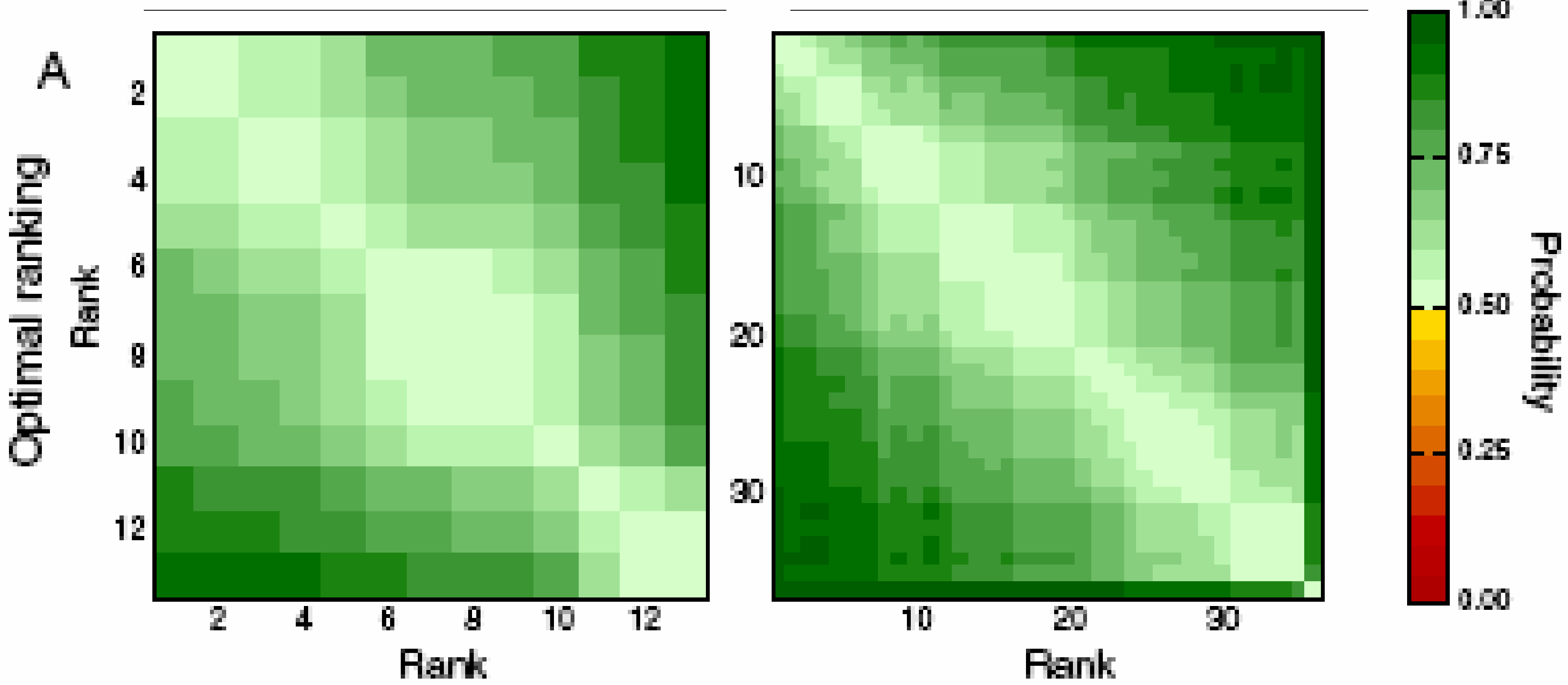
Ecology



Information retrieval (AUC)

Experimental psychology

Ecology



Food for thought

- If ultimate number of citations (i.e., 20 years after publication) is a good proxy for intrinsic quality, then journal ranking can be used to
  - select where and how to read
  - predict ultimate impact of recent publications

- If ultimate number of citations is a good proxy for intrinsic quality and one can predict one of them, then one could devise informative measures of quality of the work of an individual researcher

Acknowledgements

James S. McDonnell Foundation

NIGMS

KECK DISTINGUISHED YOUNG SCHOLARS

NSF

Roger Guimera

Marta Sales-Pardo

Mike Stringer