Bibliometrics and Citation Analysis

Mikael Graffner & Fredrik Åström
mikael.graffner@ub.lu.se
fredrik.astrom@ub.lu.se
”the application of mathematical and statistical methods to books and other media of communication” (Pritchard, 1969)
The Basics: Skewed Distributions

- Words in text, Zipf’s Law (1935)
- Articles in journals, Bradford’s Law (1934)
- Articles by authors, Lotka’s law (1926): scientific productivity
- Citations to articles/authors/journals, The Matthew Effect (Merton, 1968): scientific impact

Three different expressions:
- 20/80 distribution
  - 20% of discrete words, 80% of the text
- Grouping journals according to the distribution of articles
  - 1:n:n^2
  - 1st 400 articles, 10 journals
  - 2nd 400 articles, 60 journals
  - 3rd 400 articles, 260 journals
- Ranking authors by number of articles
  - 1, 1/2^2, 1/3^2... 1/n^2
  - 1 author, 10 articles
  - 100 authors, 1 article
The Matthew Effect Observed in LTH & Faculty of Science Publications (2011-2012)
Publications & Publication Types

LTH documents 2002
References & Citations

• Links between documents
  – Citing-Cited
  – Shared references
  – Co-citations

• References turning into citations
  – Aggregation of occurrences in reference lists
  – Abstraction produced by citation index producers

• Indicator of scientific impact (quality?)

References


What is it we’re counting; and how?

- Number of authors/institutions
  - 1 author vs 4,500 authors
  - 1 institution vs 300 institutions

- Order of authors
  - PI vs alphabetical

- Counting
  - 1st author count
  - Full count
  - Fractionalized count
ISI -> Web of Science

• Garfield (1955)
  – Shepard’s Citation Index (1873): indexing US court cases
• Institute for Scientific Information (ISI): Science Citation Index (1963)
  – Social Science Citation Index (1973)
  – Arts & Humanities Citation Index (1978)
• Journal Citation Reports (JCR) (1976)
  – Journal Impact Factor (JIF) (Garfield, 1963)
• Thomson Reuters: Web of Science
Scientific growth (Price, 1963)

Networks of scientific articles (Price, 1965)
Journal Co-citation analysis
Not just citations...
Research evaluation

• Bibliometric indicators for research evaluation were first developed during the 1970s
• Response to a growing need for transparent measures which can be used for allocation of research funds, strategic planning and appointments.
• Different levels of study: author, research group, department, universities, countries
• Different indicators for different purposes: JIF, h-index, crown indicator, top 5 %
  – More reliable on higher aggregation levels
  – Different indicators should be combined to illuminate different aspects of the research
Assumptions when using indicators in evaluations

• Basic assumptions:
  – Number of publications is an indicator of productivity
  – If many people cite a publication this is an indication of scientific impact
  – Science is a cumulative process, i.e. research is based on prior findings
• Works well for natural sciences, medicine and some areas of technology
• Less suitable for humanities and social science
Common bibliometric indicators...
Journal impact factor

• Shows how much the average publication within a journal is cited
• Published annually in the Journal Citation Report (JCR)
• In a given JCR year (e.g. 2010), the impact factor of a journal is the average number of citations to those papers that were published during the two preceding years (2008 & 2009)
• Criticism:
  – Too short citation window (2 years)
  – No field normalization
  – No normalization of document types
  – Incorrect calculation (total number of citations/”citeable documents”)
  – Skewed distributions of highly cited publications within journals
Differences in citation traffic

<table>
<thead>
<tr>
<th>Journal category</th>
<th>Highest JIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oncology</td>
<td>101.8 (CA-CANCER J CLIN)</td>
</tr>
<tr>
<td>Immunology</td>
<td>52.8 (ANNU REV IMMUNOL)</td>
</tr>
<tr>
<td>Genetics &amp; Heredity</td>
<td>38.1 (NAT GEN)</td>
</tr>
<tr>
<td>Biology</td>
<td>11.5 (PLOS BIOL)</td>
</tr>
<tr>
<td>Engineering, mechanical</td>
<td>14.2 (PROG ENERG COMBUST)</td>
</tr>
<tr>
<td>Law</td>
<td>4.3 (STANFORD LAW REV)</td>
</tr>
<tr>
<td>Political Science</td>
<td>3.1 (AM POLIT SCI REV)</td>
</tr>
<tr>
<td>Mathematics</td>
<td>3.8 (J AM MATH SOC)</td>
</tr>
</tbody>
</table>

Source: JCR 2011
New journal metrics

- **SJR** – SCImago Journal Rank
- **SNIP** – Source normalized Impact Per Paper
- **Eigenfactor and Article Influence Score**
Your (real) Impact Factor

Impact Factor (corrected) = 

# times your work is cited - # citations that actually trash your work - # times you cited yourself (nice try) - # times you were cited just to pad the introduction section - # citations the editor pressured the author to include to increase the journal's impact factor

# original articles you've written + # articles you were included in out of pity or politics + # not-so-original articles you've written copied and pasted
H-index

- H-index for an author is the number of publications \( (h) \) from this author which have been cited at least \( h \) times

- Example: An author has published 25 articles out of which 12 have been cited at least 12 times, H-index=12

- Advantage:
  - A measure of both production and citations
  - Easy to calculate

- Criticism:
  - Favours older authors with long publication lists/unfair to young researchers at the beginning of their careers
  - Unfair to researchers with few but very highly cited articles
  - Cannot be used for comparisons across different research fields
Normalized indicators

• Citation rates vary depending on field, age and document type
• Normalization is needed to make comparisons
• The citation score for every publication are normalized according to field/journal, document type and age
• If the mean field normalized score is 1.0, this means that the analyzed articles have the same citation rate as the world average
• 1.16 = 16% more citations than the world average
Citation rates for different document types

<table>
<thead>
<tr>
<th>Rank</th>
<th>Document Type</th>
<th>Times Cited</th>
<th>Web of Science Documents</th>
<th>Average Cites per Document</th>
<th>Journal Actual/Expected Citations</th>
<th>Category Actual/Expected Citations</th>
<th>Average Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ARTICLE</td>
<td>215,316</td>
<td>16,172</td>
<td>13.31</td>
<td>1.21</td>
<td>1.47</td>
<td>44.25</td>
</tr>
<tr>
<td>2</td>
<td>REVIEW</td>
<td>31,461</td>
<td>887</td>
<td>35.47</td>
<td>1.51</td>
<td>1.74</td>
<td>31.31</td>
</tr>
<tr>
<td>3</td>
<td>PROCEEDINGS PAPER</td>
<td>10,307</td>
<td>2,053</td>
<td>5.02</td>
<td>1.27</td>
<td>1.04</td>
<td>54.50</td>
</tr>
<tr>
<td>4</td>
<td>LETTER</td>
<td>902</td>
<td>321</td>
<td>2.81</td>
<td>1.14</td>
<td>1.46</td>
<td></td>
</tr>
</tbody>
</table>
Metrics mania

• Google Scholar Profile
• Microsoft Academic Search (beta)
• ImpactStory
• And many more: ReaderMeter, ScienceCard, PLoS Impact Explorer, PaperCritic, CrowdoMeter
Discussion

• Do bibliometrics in any way affect the way you do your research and publish your results?
• Do you include your h-index or other metrics in your CV and grant applications?
• Have you been affected by decisions based on bibliometric analysis?