Scientometrics & Altmetrics

Dr. Peter Kraker

VU Science 2.0, 20.11.2014
Why Metrics?
“One of the diseases of this age is the multiplicity of books; they doth so overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world.“

Attributed to Barnaby Rich in 1613 (Price 1963)
Information Overload in Science

Information overload is NOT a contemporary problem in science.

Science has been growing exponentially for the last 400 years (Price 1961, 1963)

- Number of papers (Larsen/von Ins 2010)
- Number of researchers (NSB 2010)

Instruments to deal with the overload

- Journals and conferences
- Peer review
- Quantitative analysis → Scientometrics

Price (1963)
Pathways through Science

Science Citation Index (Garfield 1955) → Web of Science

An index of incoming citations

Purpose

- Discovery of literature that is not linked thematically
- Increased collaboration between researchers
- Evaluation of science

Relational scientometrics
Evaluative scientometrics

Garfield et al. (1964)
Relational Scientometrics

Example: Genetics research (Garfield et al. 1964)

From the beginnings in the 1800s to the discovery of DNA

Relationships given by history of science (red), citations (yellow), and both (blue)

Garfield et al. (1964)
Map of Information Science

Van Eck and Waltman (2010)

<table>
<thead>
<tr>
<th>No. of pub.</th>
<th>Important authors</th>
<th>Main topics</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Rousseau, R.; Glänzel, W.; Moed, H.F.; Van Raan, A.F.J.</td>
<td>Citation analysis; research evaluation; general scientometric topics</td>
<td>Red</td>
</tr>
<tr>
<td>101</td>
<td>Thelwall, M.; Vaughan, L.; Bar-Ilan, J.; Wilkinson, D.</td>
<td>Webometrics</td>
<td>Blue</td>
</tr>
<tr>
<td>73</td>
<td>Leydesdorff, L.; Chen, C.M.; White, H.D.; Small, H.</td>
<td>Mapping and visualization of science</td>
<td>Green</td>
</tr>
<tr>
<td>53</td>
<td>Egghe, L.; Burrell, Q.L.; Daniel, H.D.; Glänzel, W.</td>
<td>h-index; citation distributions; Google Scholar</td>
<td>Orange</td>
</tr>
<tr>
<td>48</td>
<td>Glänzel, W.; Cronin, B.; Bozeman, B.; Shaw, D.</td>
<td>Scientific collaboration; co-authorship</td>
<td>Brown</td>
</tr>
<tr>
<td>46</td>
<td>Meyer, M.; Leydesdorff, L.; Tijssen, R.J.W.; Zimmermann, E.</td>
<td>Science and technology studies; patent analysis</td>
<td>Pink</td>
</tr>
<tr>
<td>26</td>
<td>Nisonger, T.E.; Cronin, B.; Shaw, D.; Wilson, C.S.</td>
<td>Studies of the library and information science field</td>
<td>Purple</td>
</tr>
</tbody>
</table>
Knowledge Domain Visualization Process (Börner et al. 2003)

1. Selection of an appropriate data source

2. Definition of unit of analysis
   - Words, articles, authors, journals, categories...

3. Determination of measures & calculation of similarities
   - Linkages, co-occurrences, Vector Space Model...

4. Ordination and/or detection of sub-areas
   - Dimensionality reduction (e.g. multidimensional scaling), cluster analysis, spatial configuration (e.g. force-directed placement)

5. Visualization and interaction design
Citations in Retrieval: Google Scholar

Google

social network analysis methods

Scholar

About 4,850,000 results (0.15 sec)

[BOOK] Social network analysis: Methods and applications
S Wasserman - 1994 - books.google.com
Social network analysis is used widely in the social and behavioral sciences, as well as in economics, marketing, and industrial engineering. The social network perspective focuses on relationships among social entities and is an important addition to standard social and ... Cited by 19605

[BOOK] Models and methods in social network analysis
PJ Carrington, J Scott, S Wasserman - 2005 - books.google.com
Models and Methods in Social Network Analysis, first published in 2005, presents the most important developments in quantitative models and methods for analyzing social network data that have appeared during the 1990s. Intended as a complement to Wasserman and ... Cited by 1009

[ CITATION] The SAGE handbook of social network analysis
J Scott, P.L.Carrington - 2011 - SAGE publications
Cited by 8205

Social network analysis for organizations
NM Tichy, ML Tushman, C Fombrun - Academy of management review, 1979 - amr.aom.org
... Table 1 presents a set of indices describing the structural characteristics of networks. These properties can be used at each of the three levels of analysis. ... In sum, the social network perspective views organizations as consisting of social units with relatively stable patterns of ... Cited by 996
Citation-based Metrics: h-Index

An metric to quantify the scientific output of an individual scientist

“A scientist has index $h$ if $h$ of his or her $N_p$ papers have at least $h$ citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.” (Hirsch 2005)
## Citation-based Metrics: h-Index

<table>
<thead>
<tr>
<th>Paper</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper 1</td>
<td>33</td>
</tr>
<tr>
<td>Paper 2</td>
<td>20</td>
</tr>
<tr>
<td>Paper 3</td>
<td>10</td>
</tr>
<tr>
<td>Paper 4</td>
<td>9</td>
</tr>
<tr>
<td>Paper 5</td>
<td>9</td>
</tr>
<tr>
<td>Paper 6</td>
<td>9</td>
</tr>
<tr>
<td>Paper 7</td>
<td>8</td>
</tr>
<tr>
<td>Paper 8</td>
<td>8</td>
</tr>
<tr>
<td>Paper 9</td>
<td>7</td>
</tr>
<tr>
<td>Paper 10</td>
<td>7</td>
</tr>
<tr>
<td>Paper 11</td>
<td>6</td>
</tr>
<tr>
<td>Paper 12</td>
<td>6</td>
</tr>
<tr>
<td>Paper 13</td>
<td>6</td>
</tr>
<tr>
<td>Paper 14</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Paper 86</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Scopus

![Citation graph](image-url)
Citation-based Metrics: Impact Factor

An measure to quantify the relative importance of a scientific journal

The average number of citations in a given year y to papers of a journal in the years y-1 and y-2

\[ \text{IF}_{2013} = \frac{\text{Citations in 2013}}{\text{Number of articles published by Journal Y in 2011 and 2012}} \]
Citation-based Metrics: Impact Factor

**Journal Information**

**Full Journal Title:** ACM SIGCOMM Computer Communication Review  
**ISO Abbrev. Title:** ACM SIGCOMM Comp. Commun. Rev.  
**JCR Abbrev. Title:** ACM SIGCOMM COMP COM  
**ISSN:** 0146-4833  
**Issues/Year:** 5  
**Language:** ENGLISH  
**Journal Country/Territory:** UNITED STATES  
**Publisher:** ASSOC COMPUTING MACHINERY  
**Publisher Address:** 2 PENN PLAZA, STE 701, NEW YORK, NY 10121-0701  
**Subject Categories:** COMPUTER SCIENCE, INFORMATION SYSTEMS

**Journal Rank in Categories:**

**Journal Impact Factor**

Cites in 2013 to items published in:  
- 2012 = 89  
- 2011 = 182  
Sum: 271

Number of items published in:  
- 2012 = 126  
- 2011 = 120  
Sum: 246

Calculation: 
- Cites to recent items = 271  
- Number of recent items = 246  
**Impact Factor** = 1.102

**Source:** Thomson Reuters
Citation-based Metrics: Exercise

- Get together in groups of 2 or 3
- Calculate the impact factor for 2013 for the two journals below and create a ranking.

Journal X: Published 6 articles in 2011 and 2012.

<table>
<thead>
<tr>
<th>Article ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations in 2013</td>
<td>15</td>
<td>17</td>
<td>14</td>
<td>18</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Journal Y: Published 6 articles in 2011 and 2012.

<table>
<thead>
<tr>
<th>Article ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations in 2013</td>
<td>100</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

- Discuss the results: how justified is the ranking? Where do you see problems?
Citation-based Metrics: Exercise

Solution

<table>
<thead>
<tr>
<th>Name</th>
<th>IF 2013</th>
<th>Rank</th>
<th>Median</th>
<th>Rank</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal X</td>
<td>15.5</td>
<td>2</td>
<td>15</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>Journal Y</td>
<td>18</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>36.7</td>
</tr>
</tbody>
</table>

![Graphs showing citation distribution for Journal X and Journal Y]
Criticisms of the Impact Factor

The IF is volatile as it uses the arithmetic mean, even though citation distributions usually follow a power law

- „Blockbuster“ papers can skew the IF

A change in the number of „citable“ papers can influence the IF considerably

The IF is field dependent – publication and citation behavior varies wildly between fields
Criticisms of Citation-based Metrics

Citations take very long to appear in meaningful quantities

Source: Amin & Mabe (2000)
Criticisms of Citation-based Metrics

Citations take very long to appear in meaningful quantities

Citation metrics are dependent on the corpus that is used for calculation

A single indicator is not sufficient to assess impact
Setting the Stage for Alternative Metrics

Increased use of online services in the scientific community

- E-Journals and pre-print/data archives
- Collaborative reference management systems
- (Micro-)blogs & social networks

Seeing academic literature through the eyes of the readers (Rowlands & Nicholas 2007)

- Usage data (downloads, readership)
- Links, likes and shares
Altmetrics

Altmetrics: alternative metrics based on data generated in online systems

Promises of altmetrics

- Assess publications quicker and on a broader scale
- Consider all outputs of research, not just papers

The altmetrics manifesto:
http://altmetrics.org
Example: PLOS Article-Level Metrics (ALM)

Viewed

<table>
<thead>
<tr>
<th>Total Article Views</th>
<th>HTML Page Views</th>
<th>PDF Downloads</th>
<th>XML Downloads</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>16,833</td>
<td>16,332</td>
<td>626</td>
<td>97</td>
<td>16,257</td>
</tr>
<tr>
<td>PLOS</td>
<td>457</td>
<td>110</td>
<td>n.a.</td>
<td>576</td>
</tr>
<tr>
<td>Totals</td>
<td>16,789</td>
<td>947</td>
<td>97</td>
<td>16,833</td>
</tr>
</tbody>
</table>

6.00% of article views led to PDF downloads

Discussed

<table>
<thead>
<tr>
<th>Source</th>
<th>Twitter</th>
<th>Facebook</th>
<th>Comments</th>
<th>Google Blogs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>109</td>
<td>20</td>
<td>0</td>
<td>Search</td>
</tr>
</tbody>
</table>

Saved

<table>
<thead>
<tr>
<th>Source</th>
<th>Citeulike</th>
<th>Mendeley</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>166</td>
</tr>
</tbody>
</table>

Cited

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scopus</td>
<td>17</td>
</tr>
<tr>
<td>Crossref</td>
<td>16</td>
</tr>
<tr>
<td>PMC</td>
<td>6</td>
</tr>
<tr>
<td>ISI Web of Science</td>
<td>17</td>
</tr>
<tr>
<td>Europe PubMed Central</td>
<td>8</td>
</tr>
</tbody>
</table>

Source:
http://www.plosone.org/article/metrics/info%3Adoi%2F10.1371%2Fjournal.pone.0047523#close
Examples: Altmetric.com

Worldwide health effects of the Fukushima Daiichi nuclear accident

Twitter attention

The data shown below were collected from the profiles of tweeters who shared this article. Click here to find out more about how the information was compiled.

Geographical breakdown

<table>
<thead>
<tr>
<th>#</th>
<th>Country</th>
<th>As %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>JP</td>
<td>42%</td>
</tr>
<tr>
<td>2</td>
<td>US</td>
<td>7%</td>
</tr>
<tr>
<td>3</td>
<td>ES</td>
<td>1%</td>
</tr>
<tr>
<td>3</td>
<td>GB</td>
<td>1%</td>
</tr>
<tr>
<td>3</td>
<td>FR</td>
<td>1%</td>
</tr>
<tr>
<td>6</td>
<td>MX</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>6</td>
<td>CA</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>6</td>
<td>IT</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>6</td>
<td>NZ</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>-</td>
<td>Other</td>
<td>1%</td>
</tr>
<tr>
<td>-</td>
<td>Unknown</td>
<td>39%</td>
</tr>
</tbody>
</table>

Example: ImpactStory

Selected works

- **Is your phylogeny informative? Measuring the power of comparative methods**
  (2012) Boettiger, Coop, Ralph. *Evolution*
  ![read fulltext](https://impactstory.org/)
  ![highly cited](#) +1  ![highly saved](#) +1  ![highly discussed](#) +4  ![highly viewed](#) +1

- **rfishbase: exploring, managing, and sharing data from R**
  (2012) Boettiger, Lang, Wainwright.
  ![read fulltext](https://impactstory.org/)
  ![highly cited](#) ![highly saved](#) ![highly discussed](#) ![highly viewed](#) +1

- **Tipping points: From patterns to predictions.**
  ![read fulltext](https://impactstory.org/)
  ![highly cited](#) +1  ![highly discussed](#) +536  ![highly viewed](#) +1  ![saved](#) +1

- **Lab Notebook, 2012**
  (2013) figshare.
  ![highly discussed](#)  ![highly viewed](#) +2  ![highly discussed](#)  ![highly viewed](#) -1

Key profile metrics

- **1.4M** tweets on
  - 16 articles
- **765** saves on
  - 16 articles
- **99.5k** impressions on
  - 23 datasets
- **33.2k** impressions on
  - 16 slide decks
- **94** forks on
  - 49 software products

Source:
https://impactstory.org/CarlBoettiger
Relational Altmetrics and KDViz

Based on implicit and explicit links created in altmetrics sources

Example: Bollen et al. (2009)

- Based on user clickstreams in digital libraries and bibliographic databases
- Co-occurrence matrix of journals in clickstreams
- Force-directed placement applied to the matrix
- Produces an overview map of all of science
Relational Altmetrics

Example: Head Start (Kraker 2013)

- Based on Mendeley readership
- Co-readership as a measure of subject similarity
- Matrix of document co-occurrences in user libraries
- Multidimensional scaling and hierarchical clustering applied to the matrix; force-directed placement applied to the resulting map; naming heuristic for labels
- Produces an overview map of a research field
# Popular Altmetrics Data Sources

## APIs

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Indicators</th>
<th>License</th>
<th>Open Data</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>figshare</td>
<td>Repository</td>
<td>Views/Downloads</td>
<td>CC0</td>
<td>Yes</td>
<td><a href="http://api.figshare.com">http://api.figshare.com</a></td>
</tr>
<tr>
<td>PLOS ALM</td>
<td>Publisher</td>
<td>Various</td>
<td>CC0</td>
<td>Yes</td>
<td><a href="http://api.plos.org">http://api.plos.org</a></td>
</tr>
</tbody>
</table>

## SDKs

<table>
<thead>
<tr>
<th>Name</th>
<th>Language</th>
<th>License</th>
<th>Data sources</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>rAltmetric</td>
<td>R</td>
<td>CC0</td>
<td>Altmetric.com</td>
<td><a href="http://ropensci.org/packages">http://ropensci.org/packages</a></td>
</tr>
<tr>
<td>alm</td>
<td>R</td>
<td>MIT</td>
<td>PLOS ALM</td>
<td><a href="http://ropensci.org/packages">http://ropensci.org/packages</a></td>
</tr>
</tbody>
</table>
Relationship between different indicators

- **downloads vs. readers**
  - \(r=0.73, n=150\)

- **downloads vs. cites**
  - \(r=0.77, n=150\)

- **readers vs. cites**
  - \(r=0.51, n=150\)

- **downloads vs. readers**
  - \(r=0.66, n=528\)

- **downloads vs. cites**
  - \(r=0.76, n=528\)

- **readers vs. cites**
  - \(r=0.59, n=528\)

Source: Schlögl et al. (2014)
Altmetrics: Exercise

Discuss the two examples below: what are possible reasons for these high altmetrics scores?
Problems of Altmetrics

Intention unknown: What does it mean to download/save/tweet a paper?

What does it mean to aggregate these numbers?

Reliability and validity of altmetrics

Altmetrics are prone to sample biases (Bollen & van de Sompel 2008, Kraker et al. 2014)

Gaming is a potential threat

→ There is a need for a better understanding of altmetrics

→ Altmetrics data needs to be open and reproducible
References


References


Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603.


Images on slides 7 and 25 by Maxi Schramm
Thank You For Your Attention!

Questions?

Dr. Peter Kraker
Know-Center

pkraeker@know-center.at
http://twitter.com/PeterKraker
http://science20.wordpress.com