On the mathematics of ranking universities and scientific products

Andrei Mărcuș

Babeș–Bolyai University Cluj–Napoca

http://math.ubbcluj.ro/~marcus

Cluj-Napoca, September 19, 2009

Contents

- Quantitative metrics are rather superficial choices for assessing the research output of an individual scholar.
- Methods like counting papers and citations, calculating impact factors and h-indices, or looking at Eigenfactor Scores are not adequate compared with what should be the standard: reading the scholars publications and talking to experts about his/her work.
- But many scholars, librarians, historians of science, editors, and other individuals are also interested in larger scale questions that require assessing thousands of articles.
- As the number of scientific journals and papers is increasing at an almost exponential rate, and if a library can afford only a limited number of subscriptions, which journals should the librarian choose?
- The burden of similar decisions affects researchers, funding agencies, university administrators, reviewers.
- It can be very difficult and costly to give an in-depth evaluation of the research, therefore aggregate bibliometric statistics, regarded as indirect indicators of quality, can be useful.

We start with an example which should be much easier that ranking universities, that is, ranking athletes in combined events.

The **decathlon** is an olympic event with an arbitrary scoring system.

- Under the original 1912 scoring tables, *Akilles Järvinen* of Finland finished second in both the 1928 and 1932 Olympics.
- The new scoring system introduced in 1934 gave Jarvinen higher converted totals than both the men he lost to.
- The table were updated again in 1050 and in 1962.
- In 1984, at the Los Angeles Olympic Games, Great Britain's *Daley Thompson* missed the world record by one point on the 1962 tables.
- The tables were changed later in 1984 and Thompsons score in Los Angeles converted to a best-ever mark.
- The current 1984 table shows the following benchmark levels needed to earn 1000, 900, 800, and 700 points in each sport.

Decathlon Scoring Table

Event	1000 pts	900 pts	800 pts	700 pts	Units
100m	10.395	10.827	11.278	11.756	Seconds
Long Jump	7.76	7.36	6.94.1	6.51	Meters
Shot Put	18.4	16.79	15.16	13.53	Meters
High Jump	2.20	2.10	1.99	1.88	Meters
400m	46.17	48.19	50.32	52.58	Seconds
110m Hurdles	13.8	14.59	15.419	16.29	Seconds
Discus Throw	56.17	51.4	46.59	41.72	Meters
Pole Vault	5.28	4.96	4.63	4.29	Meters
Javelin Throw	77.19	70.67	64.09	57.45	Meters
1500m	233.79	247.42	261.77	276.96	Seconds

Even more puzzling is the **modern pentathlon**, which includes five events:

- 10 metre air pistol shooting,
- épée fencing,
- 200 m freestyle swimming,
- show jumping over a 350 to 450 m course with 12 to 15 obstacles,
- 3 km cross-country run.

Why is it so difficult to aggregate a number of indicators and to establish "the best" or "the most complete" athlete in the world?

Choosing a winner is a problem in the *social choice theory*.
It is usually formulated in terms of voting systems.

The need to aggregate preferences occurs in many different disciplines:

- in welfare economics, where one attempts to find an economic outcome which would be acceptable and stable;
- in decision theory, where a person has to make a rational choice based on several criteria;
- in voting systems, which are mechanisms for extracting a decision from a multitude of voters' preferences.

Arrows impossibility theorem, or Arrows paradox, tells that no voting system can convert the ranked preferences of individuals into a community-wide ranking while also meeting a certain set of reasonable criteria with three or more options to choose from.

Kenneth Arrow - 1972 Nobel Prize in Economics.

Arrow's theorem is related to the *Condorcet voting paradox*.

- Assume that we need to extract a preference order on a given finite set $A = \{a, b, c, \dots\}$ of at least three options.
- Each individual consisting of at least two members gives a particular order of preferences on the set of options.
- We are searching for a preferential voting system, called a *social welfare function*, which transforms the set of preferences into a single global societal preference order.

The following assumptions are considered to be reasonable requirements of a *fair voting method*:

Non-dictatorship. The social welfare function is a *dictatorship* by individual n, if for every pair a and b, society strictly prefers a whenever n strictly prefers a to b.

Unrestricted domain (or universality). The social welfare function should account for all preferences among all voters to yield a unique and complete ranking of societal choices.

Independence of irrelevant alternatives. The social welfare function should provide the same ranking of preferences among a subset $\{a, b\}$ of options as it would for a complete set of options. Changes in individuals' rankings of irrelevant alternatives (ones outside the subset $\{a, b\}$) should have no impact on the societal ranking of the subset $\{a, b\}$.

Pareto efficiency or unanimity. Society puts alternative a above b whenever every individual puts a above b.

Arrow's theorem

If the decision-making body has at least two members and at least three options to decide among, then it is impossible to design a social welfare function that satisfies all these conditions at once.

The study of aggregation procedures is one of the tasks of the Multi-Criteria Decision Making (MCDM).

The website devoted to the Leiden Ranking 2008 says:

"The Centre for Science and Technology Studies, Leiden University, has developed a new ranking system entirely based on its own bibliometric indicators. ... on the basis of the same data and the same technical and methodological starting points, different types of impact-indicators can be constructed, for instance one focusing entirely on impact, and another in which also scale (size of the institution) is taken into account. Rankings based on these different indicators are not the same, although they originate from exactly the same data. Moreover, rankings are strongly influenced by the size-threshold used to define the set of universities for which the ranking is calculated."

This is exactly what one can expect in the light of Arrow's theorem.

- Gaufriau and Larsen show that the rankings of countries research output based on number of publications or citations heavily depend on the counting method, and that rankings based on different counting methods cannot be compared.
- They compared the *Whole Counting* method, when full credit for a publication is given to a country when at least one of the authors is from that country, and *Fractional Counting*, when a country receives a fraction of full credit for a publication equal to the fraction of authors to the publication coming from that country.
- Gaufriau and Larsen also recommend that all rankings of countries and institutions in the should be based on Fractional Counting.

Shanghai ARWU ranking

- A study done by Dehon, McCathie and Verardi with the aid of the so-called *principal component analysis* (a mathematical technique that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components) of the Shanghai ARWU ranking reveals that two different and uncorrelated aspects of academic research are measured and aggregated: *overall research output* and *top researchers*.
- This implies that the relative weight given by the rankers to these two factors determines to a large extent the final ranking.

The same Shanghai ARWU ranking is analyzed by Billaut, Bouyssou and Vincke from the perspective of Multiple Criteria Decision Making.

Their conclusions are quite critical:

- criteria used by authors of ARWU are only loosely connected with what they intended to capture;
- the ranking involves several arbitrary parameters and many micro-decisions that are not documented (the raw data is not publicly available, hence it cannot be checked);
- the aggregation technique used is flawed;
- in general, weighted sum would is a poor way to aggregate criteria;
- the aggregation technique that is used is nonsensical.

- Dependence on the methodology is not such a huge problem if the rankings are made by or for newspapers.
- It becomes one when governments and funding agencies take them into account.
- Not even Arrow's theorem, says that every ranking or voting system is totally flawed. Often the differences are visible.
- The universities of Bucharest, Iași and Cluj, have quite similar performance. Most likely, for any prescribed order of them, a reasonable ranking method which produces that order can be devised. This means that somebody has to take a political decision .
- The problem persists even if one ranks departments or programs instead universities.

- Most of automatic methods rely on citation analysis in order to rank academics, journals, and scholarly papers.
- Powerful measures of journal influence and journal value may usefully supplement expert opinion and other sources of information in making difficult decisions about journal holdings.
- There are plausible assumptions underlying the use of citation analysis as a heuristic.
- Databases which offer citation statistics, such as ISI's JCR, Scopus, American Mathematical Society's Mathematical Reviews, Citeseer, etc.
- Recent ranking methods based on Google's PageRank are used by EigenFactor, SCImago, RedJasper, and it is based on the idea that not all the citations are equal.

- A journal's impact factor (Eugene Garfield) counts the number of times that articles published in a census period cite articles published during an earlier target window.
- *Thomson Scientific* JIF has a one year census period and uses the two previous years for the target window. Let n_t^i be the number of times in year t that the year $t-1$ and $t-2$ volumes of journal i are cited. Let a_t^i be the number of articles that appear in journal i in year t. The impact factor IF_t^i of journal i in year t is

$$\mathrm{IF}_t^i = \frac{n_t^i}{a_{t-1}^i + a_{t-2}^i}.$$

- The time window of two years is too small for many disciplines, especially in the case of social sciences, but also mathematics.

Counting citations gives the possibility to define various measures, such as the h-index, the m-index, the g-index, the $g1$-index etc.

These are intensely studied from a statistical point of view.

Althouse et al. study how impact factors vary across fields and over time.

Vanclay argues that the journal IF does not deal evenly with journals to which citations accumulate quickly over a confined period and journals to which citations accumulate slowly over an extended period, because of the 2-year target window.

Google's PageRank and the Eigenfactor algorithm

Bergstrom, West and Wiseman describe a new metric for the assessment of journal quality that is based on the use of Google's PageRank. PageRank uses:

Theorem (Perron-Frobenius)

A real square matrix with positive entries has a unique largest real eigenvalue and that the corresponding eigenvector has strictly positive components.

This theorem is also used used

- to rank football teams and tennis players;
- to generate schedules for sports teams;
- to rank academic doctoral programs based on their records of placing their graduates in faculty positions.

- The PageRank algorithm computes the status of a web page based on a combination of the number of hyperlinks that point to the page and the status of the pages that the hyperlinks originate from.
- By taking into account both the popularity and the prestige factors of status, Google avoids assigning high ranks to popular but otherwise irrelevant web pages.
- PageRank was inspired by old papers on citation analysis (de Solla Price).

- Eigenfactor uses citations in the academic literature as provided by ISI's Journal Citation Reports (JCR).
- The aim is to identity the most "influential" journals, where a journal is considered to be influential if it is cited often by other influential journals.
- One citation from a high-quality journal may be more valuable than multiple citations from less important publications
- A citation from a review article that has references to large numbers of papers counts for less than a citation from a research article that cites only papers that are relevant for the argumentation.
- This seems circular, but one can iteratively calculate the importance of each journal in the citation network by a simple algorithm.

The computation the 2006 Eigenfactor scores

- Let $C = (c_{i,j})$ be the *citation matrix* for the 7611 ISI-linked journals, where $c_{i,j}$ is the number of citations from 2006 articles in journal j to articles in journal i published in 2001-2005.
- Fill in with zeros the diagonal of C to ignore journal self-citations.
- Let $H = (h_{i,j})$, where
$$h_{i,j} = \frac{c_{i,j}}{\sum_k c_{k,j}}.$$
- Compute an **article vector** a, where a_i is the number of articles published by journal i in 2001-2005, divided by the total number of articles published by all journals in 2001-2005.
- Some of the journals listed in the matrix H do not cite any other journals. Any column of the matrix H that has all 0 entries is replaced with the vector a to produce a new modified matrix H'. This is a stochastic matrix by construction.

The computation the 2006 Eigenfactor scores

- From this, we construct a new stochastic matrix, P:

$$P = \alpha H' + (1 - \alpha) a \mathbf{e}^T,$$

  where $\mathbf{e}^T$ is a row vector of all 1's (and ^{T} means transposition), and thus $a\mathbf{e}^T$ is a matrix with identical columns a.
- This corresponds to a process which follows the literature with probabilities $1 - \alpha$ and "teleports" to a random journal with weights proportional to the number of articles published by a journal.
- As in the case of Google's PageRank, the value $\alpha = 0.85$ is used.
- Define the vector π^* as the leading eigenvector of P, which corresponds to the fraction of time spent at each journal in P. These fractions serve as weights of journal influence.

The *Eigenfactor* score, EF, is defined as

$$\mathrm{EF} = 100 \frac{H\pi^*}{\sum_i [H\pi^*]_i}.$$

The *Article Influence* score AI_i journal i is a measure of the per-article citation influence of the journal:

$$\mathrm{AI}_i = 0.01 \frac{\mathrm{EF}_i}{a_i},$$

where EF_i is the *Eigenfactor* score for journal i, and a_i is the i-th entry of the normalized article vector.

- A main assumption is that receiving a citation is always good!
- A weakness of *Eigenfactor* is that all citations from articles in a given journal j to articles published in j during the preceding five years are discarded as "self-citations".

Bini, Del Corso, and Romani have proposed an integrated ranking of authors, journals, papers, areas, and institutions.

This is a flexible method based again the PageRank algorithm.

The general principle is the "mutual reinforcement between papers, journals, authors":

- A paper is important if published in an important journal but also if cited by important papers and authored by important authors.
- An author is important if she has important co-authors and has written important papers published in important journals.
- A journal is important if collects citations from important journals, publishes important papers by important authors.

There are parameters that have to be agreed upon at a "political" level.

Advantages of automated rankings: easy to calculate, time aware, objective, etc. (although "objectivity is disputed).
Weaknesses:

- a citation is not always a trusting vote;
- data source and coverage are subject of dispute;
- citation gathering can be a very slow process;
- the way authors cite depends on the area of research;
- in the same journal there are articles with different citation rates;
- the ranking doesn't always agree with the widely accepted journal's reputation.

PageRank:

- is an improvement of IF when one is interested less in "popularity" and more in "value";
- because of complexity the final results are harder to understand;
- Albert Einstein: "*Everything in life should be as simple as possible, but no simpler*".

- Nobody claims that PageRank is the best possible algorithm for ranking web pages. Most users wild find relevant results for most of their searches at the top of the list.
- Other search engines use different algorithms; Google's domination of the search market has many reasons.
- The details of the ranking methodology are kept secret, to avoid manipulation by creating link farms, or by other methods.

Berlin principle no. 6 (2006 IREG meeting)

Thou shalt be transparent regarding the methodology used for creating the rankings.

- Experiments like Thomson JIF and Shanghai ARWU ranking appear to be irreproducible, and much of their raw data is not publicly available.
- The current methods of evaluation of scientific papers and journals are undergoing re-evaluation.

Concluding remarks

- By human nature, "rankings are here to stay", despite the fact that the very notion of "the best university" is illusory.
- Rankings are important marketing tools, but can also have various unintended and perverse consequences.
- Arrow's theorem tells us that there is no such thing as "the best ranking method". Different mathematical models may be useful in different circumstances.
- Ranking universities is not a "scientific" exercise, but rather a journalistic one, as it can be motivated by the right of public to be informed.

Concluding remarks

- Automated rankings can be manipulated.
- Ranking methodologies, even if carefully devised, are at most statistically relevant. (For instance, if a university library subscribes to journals by selecting them acording to a PageRank type algorithm, then most of the faculty will be satisfied.)
- When evaluating individual papers or researchers, or even individual journals or universities, there is no substitute for reading and understanding the work.
- Official rankings made for the purpose of differentiated funding are the result of political decisions.

Concluding remarks

- Automated rankings can be manipulated.
- Ranking methodologies, even if carefully devised, are at most statistically relevant. (For instance, if a university library subscribes to journals by selecting them acording to a PageRank type algorithm, then most of the faculty will be satisfied.)
- When evaluating individual papers or researchers, or even individual journals or universities, there is no substitute for reading and understanding the work.
- Official rankings made for the purpose of differentiated funding are the result of political decisions.

Einstein: *Not everything that can be counted counts, and not everything that counts can be counted*.

- K.J. Arrow. *A Difficulty in the Concept of Social Welfare*. The Journal of Political Economy, **58** (1950), 328–346.
- C.T. Bergstrom, J.D. West and M.A. Wiseman. *The Eigenfactor metrics*, Journal of Neuroscience 28(45) (2008), 11433–11434.
- J.-C. Billaut, D. Bouyssou and Ph. Vincke. *Should you believe in the Shanghai ranking? An MCDM view*. preprint, 15 July 2009.
- D.A. Bini, G.M. Del Corso and F. Romani. *Evaluating scientific products by means of citation-based models: a first analysis and validation.* Electronic Trans. on Numerical Analysis **33** (2008), 1–16.
- S. Brin and L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Comput. Networks ISDN Systems, **30** (1998), 107–117.
- J. Ewing, R. Adler and P. Taylor. Joint IMU/ICIAM/IMS-Committee on Quantitative Assessment of Research. *A report on Citation Statistics*, 2008.