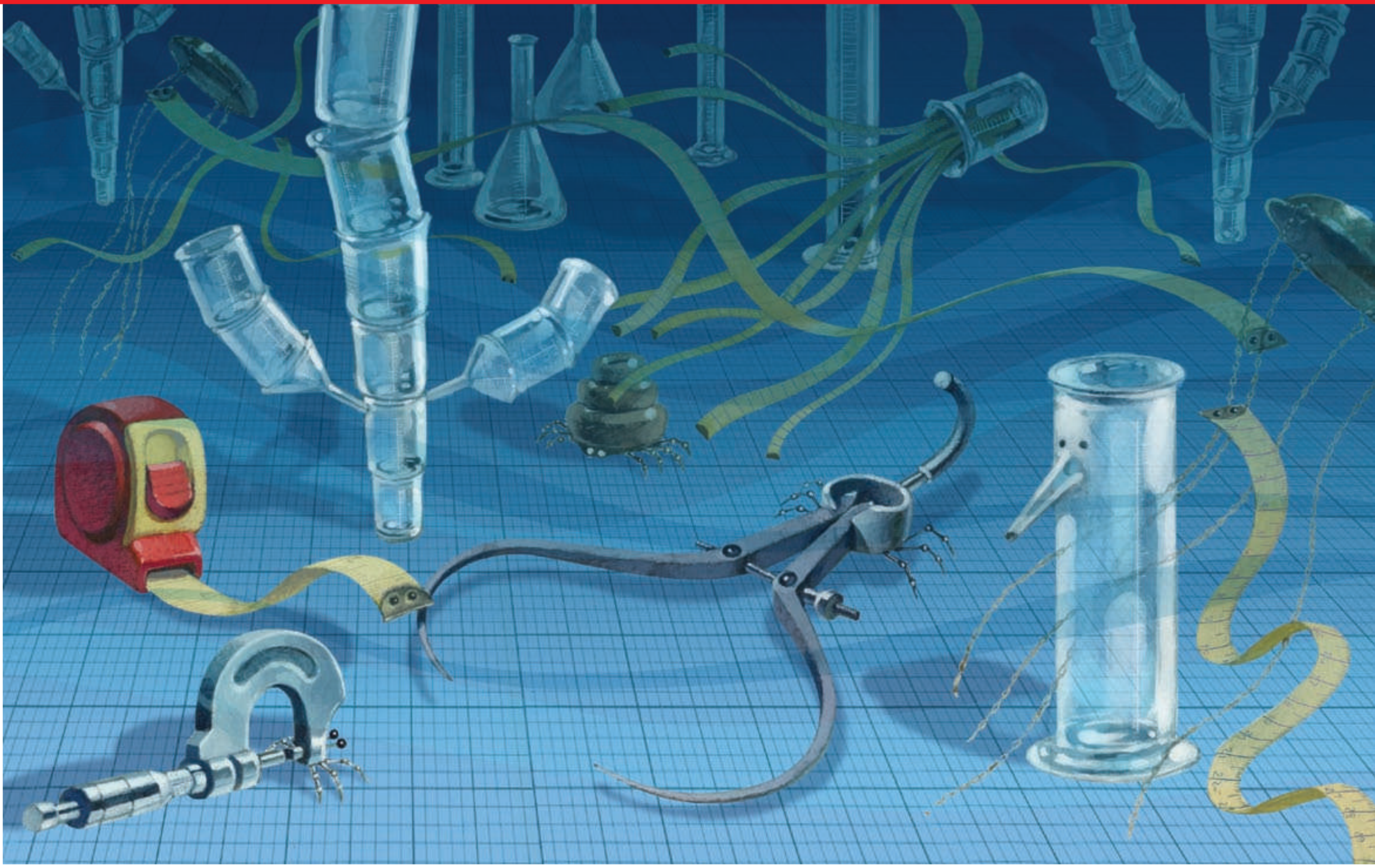


ILLUSTRATION BY DAVID PARKINS

A profusion of measures

Scientific performance indicators are proliferating — leading researchers to ask afresh what they are measuring and why. **Richard Van Noorden** surveys the rapidly evolving ecosystem.

Scientists have been sizing up their colleagues since science began. But it was American psychologist James McKeen Cattell who first popularized the idea that systematically ranking scientists by 'performance' could provide benefits beyond scratching the itch of professional envy. In the 1910 second edition to his 1906 work, *American Men of Science: A Biographical Directory*, he argued that tracking performance over time could assist the progress of research. "It is surely time for scientific men to apply scientific method to determine the circumstances that promote or hinder the advancement of science," he wrote.

That rationale for systematic evaluation hasn't changed much in 100 years, but the evaluation techniques have evolved dramatically. Where Cattell simply asked experts to rank the star performers in a field by merit — "Expert judgment is the best, and in the last resort the only, criterion of performance," he wrote — a host of objective indicators, or metrics, are now used to quantify nebulous notions of scientific quality, impact or prestige.

Within the past decade, the development of ever more sophisticated measures has accelerated rapidly, fuelled by the ready availability of online databases such as the Web of Science from Thomson Reuters, Scopus from Elsevier and Google Scholar.

"Right now we're going through a Cambrian explosion of metrics," says Johan Bollen, an informatics scientist at Indiana University in Bloomington. It has become all but impossible even to count today's metrics. Bibliometricians have invented a wide variety of algorithms, many of them unknown to the everyday scientist, some mistakenly applied to evaluate individuals, and each surrounded by a cloud of variants designed to help them apply across different scientific fields or different career stages. (See 'Metrics explosion', page 866.)

"Right now we're going through a Cambrian explosion of metrics."

Here, *Nature* categorizes metrics old and new, lays out their strengths and weaknesses — and examines a growing feeling among researchers that it is time to slow down and discuss what these measures are actually for.

The era of quantitative, computer-tabulated science metrics dates back to the 1950s, when linguist Eugene Garfield began indexing the scientific literature using punch cards. A company in Philadelphia, Pennsylvania, that Garfield founded in 1955 was renamed the Institute for Scientific Information (ISI) in 1960, the same year it began to publish the Science Citation Index. This was a systematic effort to track citations — the footnotes by which journal authors acknowledge their intellectual debts. (ISI is now owned by the publishing firm Thomson Reuters.) In 1965, Garfield and his colleagues used ISI's databases to show that Nobel laureates published five times the average number of papers, and that their work was cited 30 to 50 times the average — a finding that for decades established citations as the pre-eminent quantitative measure of a scientist's influence[1].

For all that, Garfield's best-known and most widely used citation-based metric, the 'impact factor' (see 'Field guide to metrics'), which he developed in 1963, is of little use in measuring an individual's performance; it applies only to the popularity of the journal. "If there is one thing every bibliometrician agrees, it is that you should never use the journal impact factor to evaluate research performance for an article or for an individual — that is a mortal sin," says Anthony van Raan, director of the Centre for Science and Technology Studies at Leiden University in the Netherlands.

Big impact

A better metric for assessing an individual by their citations is the *h*-index, which has been swiftly adopted by major online databases since being introduced in 2005 by physicist Jorge Hirsch of the University of California in San Diego. According to Hirsch's definition, someone who had written, say, 50 papers that had each been cited at least 50 times would have an *h*-index of 50.

An author's *h*-index has the virtue of measuring his or her article productivity and citation-based impact simultaneously. But it does have flaws, including the fact that an author's *h*-index can reflect longevity as much as quality — and can never go down with age, even if a researcher drops out of science altogether.

To combat this, "there have probably been more than a dozen variants of the *h*-index suggested since 2005, and even scholars in the field of bibliometrics have still not established which are the best ones to use," says Anne-Wil Harzing, a professor of international management at the University of Melbourne in Australia. For that reason, she adds, most scientists stick to the original *h*-index, whatever its limitations.

> "You should never use the journal impact factor to evaluate research performance for an article or for an individual — that is a mortal sin."

A third, increasingly popular, class of measure is the 'evaluative informetric', which gives heavier weight to citations from papers that are themselves highly cited. The principle is much the same as the PageRank algorithm that Google uses to order its search results: a link from a popular page is more highly weighted than one from a not-so-popular page. Both Thomson Reuters and Elsevier now offer to compute this kind of metric for journals — the companies refer to the result as the Eigenfactor and the SCImago Journal Rank (SJR), respectively.

Unlike the resolutely journal-oriented impact factor, the page-rank concept has been usefully applied to individuals by some researchers. Filippo Radicchi, a researcher in complex networks at the Institute for Scientific Interchange in Turin, Italy, and his colleagues have used weighted citations to derive a network of links between more than 400,000 papers published between 1893 and 2006 in the Physical Review journals. By slicing through the network year by year, the researchers then showed how the influence of each scientist's articles spread through a community over time — which they in turn used to produce a quantitative ranking of physics authors[2].

For all their popularity, however, citation-based metrics share some fundamental weaknesses when it comes to evaluating individual researchers. One is that research papers commonly have multiple authors — "possibly hundreds of them", says Henk Moed, a senior science adviser at Elsevier in Amsterdam. A number of corrections can be applied to give the various authors fractional credit. But in some research fields, such as high-energy physics, there can be so many co-authors that assigning credit to individuals makes little sense, Moed says: "Here one seems to reach the limits of the bibliometric system."

Another weakness is that the scores depend on the database being used. Thomson Reuters's science, social science and arts and humanities databases — accessible through its Web of Knowledge interface — include data from about 11,500 journals. Elsevier's Scopus, introduced in 2004, includes abstracts and references from 16,500 peer-reviewed journals. And the free automatically indexed database Google Scholar, also introduced in 2004, includes details of patents as well as scientific papers, and covers many more journals in engineering, social sciences and the humanities than either of the others. A search in May showed that papers in international management by Harzing had been cited 815 times according to Thomson Reuters, 952 times according to Scopus and 2,226 times according to Google Scholar.

Push for normality

For bibliometricians, the most daunting problem with citation-based metrics is getting the 'normalization' right: if molecular biologists tend to cite more often than physicists, then molecular biologists will have higher *h*-indices or citation counts, making it difficult to compare individuals from the two fields.

In principle, such variations can be evened out by dividing a researcher's citation rate by the average citation count for his or her field. But in practice, any attempt to do so swiftly gets bogged down in categorization: what constitutes a 'field'? A stem-cell researcher, for example, may bridle at being normalized by the average citation rate of cell biologists in general. "Everyone has made a contribution to their particular granular subject area. If you define the area too broadly, you miss subtleties; too narrowly and you get nothing useful out of

Field guide to metrics

Number of citations
Number of times a researcher or research paper is cited by others.
- Simple way to denote influence.
- Hard to compare between fields or career stages.
- Variants include citations in top journals only; citations per publication; or citations normalized to scientific field.

***h*-index**
A researcher with an *h*-index of 50 has 50 publications each cited at least 50 times.
- Introduced in 2005, measures productivity and impact.
- Varies with scientific fields and cannot decline with age.
- More than a dozen variants, ranging from the contemporary *h*-index giving more weight to recent articles, to the *g*-index, giving more weight to highly cited articles.

Impact factor
The frequency with which an average article in a journal gets cited. For 2010, it equals the total number of citations in 2010 to items that the journal published in 2009 and 2008, divided by the number of 'citable items' it published during those same two years.
- Highly standardized, introduced in 1963.
- Only indicates impact of journals, not of individual researchers or papers.
- In 2005, 89% of *Nature*'s impact factor was generated by 25% of the articles[5].

Weighted citations
A link or citation from a popular article or researcher is weighted more heavily, as in Google's PageRank algorithm.
- Large database providers offer journal-focused examples (Thomson Reuters's Eigenfactor, or Elsevier's SJR).
- No standard yet for application to individuals; hard to compare between fields.

Online accesses
Number of times a research paper is accessed or downloaded online.
- Focuses on individual articles; more up-to-date than citations; captures online attention from all viewers, not just scientists.
- Global standards for reporting not yet in place; only captures online audience; may prioritize attention over scientific quality.
- Often accompanied by social bookmarking recommendations and comments.

Betweenness centrality
Measures how a network node (such as a paper) is positioned on all pathways between nodes.
- One of a number of metrics trying to quantify the interconnectedness of researchers or research papers.
- Rarely used in practice by evaluators; reduces a useful map to a single number.

it," says Charles Oppenheim, emeritus professor of information science at Loughborough University, UK.

One way to get around that problem is to let the citations define the categorization. This is the idea behind various attempts to construct 'maps of science', using networks of interconnecting citations to spot discrete research fields or intellectual environments. The process is hard to standardize, says van Raan. Nonetheless, he says, "for individual scientists, mapping is the most interesting development in bibliometrics today". Bollen agrees: such maps often show how research papers or novel disciplines lie at the centre of particular fields of activity, he says — which could allow a scientist to assert, "my work connected nanotechnology to archaeology", or "if I hadn't published this paper, these domains would never have been connected".

Bibliometricians have suggested a host of measures to quantify such statements. These include 'betweenness centrality' — how often a paper in the network lies on the shortest path between any other two papers — and 'closeness centrality': the average number of connections required to get from a paper to any of the other papers. What aspects of scientific impact these measure is not entirely clear, but they probably give an indication of interconnectedness and interdisciplinarity.

Cyberstalking

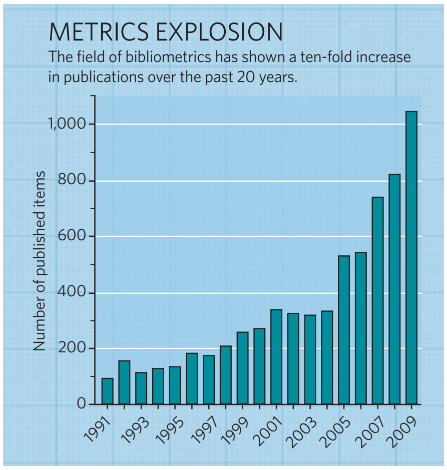
Meanwhile, some metrics researchers are looking to make a break from citations. As most scientific articles are now accessed and read online, why not just track the readers' actions in cyberspace through article or journal page views or downloads?

Publishers such as the Public Library of Science already offer download statistics for their articles, together with social-bookmarking tools that allow scientists to flag papers that they find particularly useful. (Similar tools are offered by the online services Mendeley and Faculty of 1000.)

"These people should know better than to think that there is a single measure you can use."

The disadvantage of this approach is that it apportions the impact of a research paper according to all public views, not just those by scientists. But that can also be seen as an advantage, in that it expands the idea of scientific impact. For example, medical researchers might find that doctors, nurses and public-health policy-makers frequently view articles online, although the researchers never receive a traditional citation from these end-users.

One early hurdle for this nascent field is that there are not yet global standards for journals to report data files of user activity. But COUNTER (Counting Online Usage of Networked Electronic Resources), a consortium of librarians and publishers based in Oxford, UK, is working to reach agreement on such a standard by 2012.



METRICS EXPLOSION

The field of bibliometrics has shown a ten-fold increase in publications over the past 20 years.

THOMSON REUTERS, WEB OF SCIENCE

Bollen's team is exploring whether online-usage data might help funding agencies to pick out fast-moving areas of innovation before citation-based statistics have a chance to catch up. The researchers have obtained a database of 1 billion usage events — records of users accessing scientific articles, newspapers and magazines in the years 2002 to 2007. They can also see in what order a user in any one session clicks through resources, allowing them to track the general flow of activity and produce maps showing which articles are central to which networks of activity. There are now maps that show how work in the social sciences and humanities formed bridges between scientific disciplines[3].

"In principle you could use these records to track scientific activity in real time, and to follow science taking place on Twitter, blogs, or through online software, none of which can be recorded by citation data," says Bollen. Before this vision of instant influence-tracking can become solid, however, data on the Internet need to be organized and referenced in more consistent ways, and publishers need to agree to release information on usage statistics.

Even as they push forward innovative ideas, many researchers in the metrics field say that it is high time for some reflection and consolidation. Little, if any, of the recent buzz has made it past the pages of scholarly journals into regular use on scientists' CVs, and, says Peter Binfield, publisher of *PLoS ONE*, "it feels like the field is going off in multiple directions".

More widely, says Bollen, although bibliometricians know that the idea of measuring scientific performance shares a fuzziness with the idea of measuring intelligence, many are too keen to promote their own innovations rather than focus on what they actually measure. "The point should not be to come up with a new metric. It should be to explain what metrics represent, and why we want them," says Bollen. "Can we come back to the scientific community and say 'if this is what you want to measure, then this is a good way to do that'?" Much of the next few years of clearing through the rubble of metrics will involve this kind of process, he says.

Similarly, although using a variety of metrics gives the clearest picture of scientific impact, some published research demonstrates that many people still desire a single index. "There is some mind-numbing detail on how 'my version is better than yours'; all these people should know better than to think that there is a single measure you can use," says David Pendlebury, a consultant for Thomson Reuters based in Bend, Oregon.

Many metrics correlate strongly with one another, suggesting that they are capturing much of the same information about the data they describe. Bollen's team last year published a study[4] comparing correlations between 39 measures of scientific impact for journals, attempting to tease out what different aspects of scholarly impact they captured. For example, the most important factor seems to be whether a metric measures 'rapid' or 'delayed' impact.

Meanwhile, modern metrics are slowly finding users outside the traditional groups: journals hoping to promote their products or research-performance managers who, like Cattell, hope to boost research. Individual researchers are beginning to explore how new tools such as network mapping and online usage data could help them to identify other scientists who are close to their special interests, deliver relevant papers to literature searches more speedily or to pinpoint emerging innovative fields. Soon they could start to claim bibliometrics for themselves — assisting research in ways that Cattell never envisioned. ■

Richard Van Noorden is assistant news editor for *Nature*.

1. Garfield, E. & Sher, I. H. in *Research Program Effectiveness* (eds Yovits, M. C., Gilford, D. M., Wilcox, R.H., Staveley, E. & Lemer, H. D.) 135–146 (Gordon and Breach, 1966).
2. Radicchi, F., Fortunato, S., Markines, B. & Vespignani, A. *Phys. Rev. E* **80,** 056103 (2009).
3. Butler, D. *Nature* **458,** 135 (2009).
4. Bollen, J., Van de Sompel, H., Hagberg, A. & Chute, R. *PLoS ONE* **4,** e6022 (2009).
5. *Nature* **435,** 1003–1004 (2005).

See Editorial, page 845, and metrics special at www.nature.com/metrics.