Temporal Evolution of Universal Performance Indicators for Academic Publication

Imperial College London

T. S. Evans, N. Hopkins and B. S. Kaube

How can we compare different papers?

How do you compare a three year old mathematics paper to a seven year old publication in medicine? Whether trying to find the best papers to read or assessing the performance of a University, bibliometric measures are widely used. The simplest measure of impact is the number of citations $c(t)$ but direct comparisons do not account for the age of a paper nor the variation in citation practices between disciplines.

A better measure is $c_f(t)$, the citation count normalised by the average number of citations for papers in the same field f and year t, $\langle c(t)\rangle_f$:

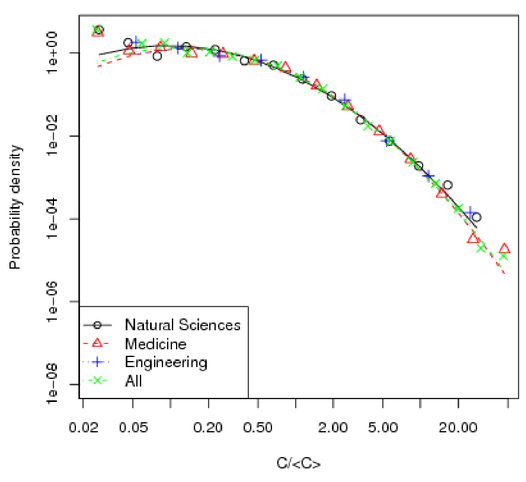
$$c_f(t) = \frac{c(t)}{\langle c(t)\rangle_f} \quad \text{where } \langle c(t)\rangle_f = \frac{\sum_{p\in f} c(t)}{N_f(t)} \quad \text{and } N_f(t) = \sum_{p\in f} 1 \tag{1}$$

Using citation data and field classifications from Web of Science, Radicchi et al. [2] showed that the distribution of relative citation counts $p(c_f)$ across many distinct fields and at different times followed a universal lognormal distribution with $\sigma \approx 1.3$:

$$p(c_f) = \frac{1}{\sigma c_f \sqrt{2\pi}} \exp\left[\frac{-(\ln c_f - \mu)^2}{2\sigma^2}\right] \quad \text{with } \sigma^2 = -2\mu \text{ as } \langle c_f\rangle = 1 \tag{2}$$

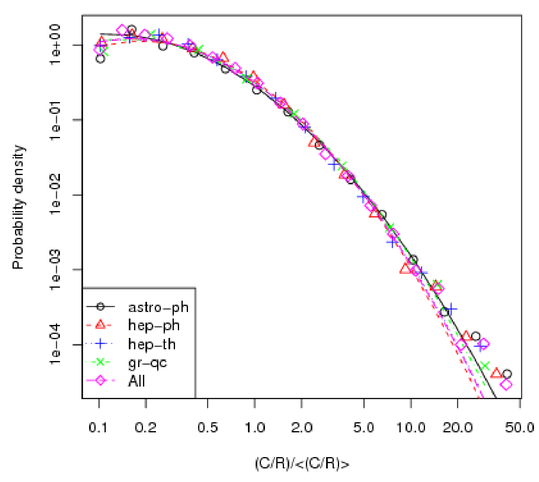
We report results for data from one Institute and from arXiv. We note that simple models are not consistent with the results [3].

Single Institution



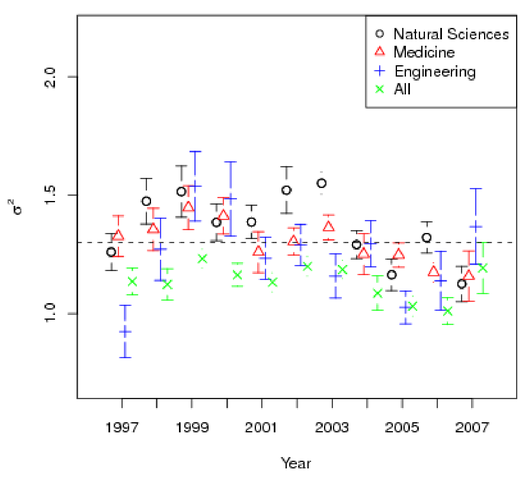
We found the same universality in author approved publications from a single institution for the year 2001 binned by author faculty. A one-parameter lognormal fit was applied to all data with $c_f > 0.1$ resulting in the curves shown. The fit was verified using a χ^2 goodness of fit test and resulted in σ^2 values in agreement with those found by Radicchi et al. [2].

arXiv



Similar analysis was applied to publications contained in arXiv deposited between 1991 and 2006 inclusive, when assigned fields according to the arXiv sub-archives. In this case only citations originating from eight selected arXiv sub-archives were counted. The citation data for the year 2002 is shown above. Again, values of σ^2 were found to be consistent with a value of 1.3.

Best Models



The variance σ^2 of lognormals fitted to single institute citation data show no systematic variation with time.

A simple model showing no change in variance over time is

$$c = qg(t)\prod_t \xi(t) \tag{3}$$

where $g(t)$ is the growth in mean citation count, q measures the intrinsic quality of the paper and ξ is drawn from a universal distribution. The contribution to the variance arising from ξ dies off in $\ln(c)$ as $1/\sqrt{t}$, giving rise to the time invariance of the universality.
Only $g(t)$ can vary between fields in order to explain the universality when the citation count is divided by the field mean. However, this does not explain why the intrinsic quality of publications follows a lognormal distribution. One explanation is that the overall quality is comprised of a *product* of (independent) factors, $q = \prod_a q_a$, where each factor q_a is the effect of some attribute of the publication, e.g. quality of publishing journal, home institution prestige, sub-discipline specific differences and some measure of the true quality of the work. Such a model can help explain the general features of citation patterns, although other effects may also be important. Articles with few citations are less well-described as other processes like self-citation and approximating a discrete process by a continuous distribution appear to be significant.
See `http://goo.gl/9ubQy` for details.

The Failure of Simple Models

No detailed model has yet been proposed to explain the origin of this universality. Variations on the Price model [1], in which citations are preferentially accrued by papers in proportion to the existing number of citations, invariably result in power law behaviour. This is not consistent with observed citation patterns which are well described by lognormal distributions, at least for reasonably highly cited publications.

Lognormal distributions are typically the hallmark of a multiplicative growth process. A simple stochastic model assumes that citations evolve independently with the citation count at time t, $c(t)$, evolving according to, $c(t+1) \to c(t)\xi(t)$, where ξ is chosen from a distribution with mean $1 + \lambda\,(c(t))^{\beta}$, where λ denotes a field-specific citation growth rate and β is an adjustable parameter close to zero. The growth rate is effectively cancelled out when dividing by the mean. A lognormal distribution is reached after 25 iterations for a range of parameters. However the resulting variances, σ^2, are too small and, more fundamentally, the temporal evolution of σ^2 is also incorrect. Over the periods studied, neither the single institute nor the arXiv data showed any significant variation with time. According to the central limit theorem, the variances of multiplicative processes scale inversely with time as $\sigma^2 \sim 1/t$.

References

[1] D.S. Price. "A general theory of bibliometric and other cumulative advantage processes". *Journal of the American Society for Information Science*, **27**, 292–306, 1976.

[2] F. Radicchi, S. Fortunato, C. Castellano. "Universality of citation distributions: Toward an objective measure of scientific impact". *Proceedings of the National Academy of Sciences*, **105**, 17268, 2008.

[3] T.S. Evans, N. Hopkins, B.S. Kaube. "Universality of Performance Indicators based on Citation and Reference Counts" *Scientometrics*, 2012 (to appear).