Journal of Informetrics 8 (2014) 824–839

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Distributions for cited articles from individual subjects and years

Mike Thelwall*, Paul Wilson

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK

ARTICLE INFO

Article history:
Received 16 June 2014
Received in revised form 28 July 2014
Accepted 6 August 2014
Available online 3 September 2014

Keywords:
Citation distribution
Power law
Hooked power law
Lognormal distribution
Citation analysis

ABSTRACT

The citations to a set of academic articles are typically unevenly shared, with many articles attracting few citations and few attracting many. It is important to know more precisely how citations are distributed in order to help statistical analyses of citations, especially for sets of articles from a single discipline and a small range of years, as normally used for research evaluation. This article fits discrete versions of the power law, the lognormal distribution and the hooked power law to 20 different Scopus categories, using citations to articles published in 2004 and ignoring uncited articles. The results show that, despite its popularity, the power law is not a suitable model for collections of articles from a single subject and year, even for the purpose of estimating the slope of the tail of the citation data. Both the hooked power law and the lognormal distributions fit best for some subjects but neither is a universal optimal choice and parameter estimates for both seem to be unreliable. Hence only the hooked power law and discrete lognormal distributions should be considered for subject-and-year-based citation analysis in future and parameter estimates should always be interpreted cautiously.

1. Introduction

Citation counts are known to be highly skewed in the sense that, after a few years, many articles in a typical collection will have received few citations and a few articles will have received many citations (Price, 1965). Other phenomena also exhibit similar behaviour. For example, web hyperlink counts are also highly skewed, with a small number of pages attracting huge numbers of hyperlinks whereas many pages attract few (Barabási & Albert, 1999). A mathematical distribution that has the same property, and which has been argued to model citations, either for all papers with at least one citation or for all papers with at least a moderate number of citations, is the power law (e.g., Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2011; Clauset, Shalizi, & Newman, 2009; Newman, 2001; Redner, 1998; Wen & Hsieh, 2013). This is the simple formula $1/x^{\alpha}$ with one parameter, α. For example, if a power law is applied to the number of citations to a set of articles, then for some fixed value α it would be true that the probability that a randomly selected article had received x citations would be proportional to $1/x^{\alpha}$. Nevertheless, other mathematical distributions have also been proposed, such as a discrete version of the lognormal distribution (Radicchi & Castellano, 2012) with continuous probability density function $(1/x\sqrt{2\pi}\sigma)e^{(-((\ln x-\mu)^2/2\sigma^2))}$, where μ and σ are its two parameters. For other types of data, investigations have shown that, despite initial claims, power laws

* Corresponding author. Tel.: +44 1902 321000.
E-mail addresses: m.thelwall@wlv.ac.uk (M. Thelwall), PaulJWilson@wlv.ac.uk (P. Wilson).

http://dx.doi.org/10.1016/j.joi.2014.08.001
1751-1577/

Table 1
The 20 Scopus subjects selected and the number of articles (excluding reviews and documents that are not articles) extracted from them from the end of 2004 (maximum 5000 articles). The data, including citations, was collected in May 2014.

Scopus subject	Abbreviation	Articles	General subject area
Accounting	Accounting	1178	Business, Management and Accounting
Algebra and Number Theory	Algebra	528	Mathematics
Applied Mathematics	AppliedMaths	5000	Mathematics
Biochemistry	Biochem	5000	Biochemistry, Genetics and Molecular Biology
Dermatology	Dermatology	3184	Medicine
Developmental Biology	Developmental	4541	Biochemistry, Genetics and Molecular Biology
Ecology, Evolution, Behaviour and Systematics	Ecology	5000	Agricultural and Biological Sciences
Genetics	Genetics	5000	Biochemistry, Genetics and Molecular Biology
History	History	5000	Arts and Humanities
Horticulture	Horticulture	3009	Agricultural and Biological Sciences
Literature and Literary Theory	Literature	5000	Arts and Humanities
Logic	Logic	4547	Mathematics
Marketing	Marketing	1550	Business, Management and Accounting
Oncology	Oncology	4646	Medicine
Rehabilitation	Rehab	5000	Medicine
Soil Science	Soil	4347	Agricultural and Biological Sciences
Statistics and Probability	StatsProb	5000	Mathematics
Tourism, Leisure and Hospitality Management	Tourism	608	Business, Management and Accounting
Urology	Urology	5000	Medicine
Visual Arts and Performing Arts	Visual	4096	Arts and Humanities

are less appropriate than the lognormal distribution (e.g., Downey, 2001; Mitzenmacher, 2004) and so it is important to assess whether power laws, or variants, are appropriate for citation distributions.

An alternative possible distribution, based on the formula $1/(B+x)^{\alpha}$, where $B > -1$, and called here the *hooked power law*, is an extension of the power law where B is a second parameter. This distribution was derived for web links (Pennock, Flake, Lawrence, Glover, & Giles, 2002 – see below) but seems to have not been used for citations although it is a logical alternative distribution, given the acknowledged parallels between hyperlinks and citations. This article assesses whether the hooked power law is a better fit than the lognormal and power law distributions for citation data for a single year and subject. Although a previous study has suggested that citation distributions from all fields can be scaled to give the same shape (Radicchi, Fortunato, & Castellano, 2008), this has subsequently been shown not to be the case (Waltman, van Eck, & van Raan, 2012) and so the issue is still unresolved. Moreover, a citation distribution that fits for long periods of time will not necessarily fit time-sliced data, such as that for all citations within a given year. Scientometric evaluations that use citations often compare articles separately that are published within a single year or within a short time window (e.g., for the main UK research evaluation, expert panels consider articles from several years but, if they use citations, are given field-based citation averages separately for each year under consideration) and so it is important to identify distributions that fit the type of data that is used in practice.

The question of which statistical model best fits citations is an abstract mathematical one, but the answer has both theoretical and practical implications (e.g., Glanzel, 2007; Ruiz-Castillo, 2013). It has practical implications for the use of citation analysis in research evaluation because power laws can be generated by feedback loops, and this appears to be the most common explanation for natural phenomena that obey power laws. In bibliometrics, the feedback loop is known as the Matthew effect (Merton, 1968) and elsewhere it is known as rich-get-richer (e.g., Adamic & Huberman, 2000). In other words, if citations obey a power law then, once cited, articles generate citations primarily because of their existing citations rather than because of their intrinsic value. If true, then this would be an argument against using citations in research evaluation. In contrast, a hooked power law is consistent with a process in which articles attract new citations partly due to their existing citations and partly due to other factors (Pennock et al., 2002), such as their intrinsic worth. This would be more consistent with citations being used in research evaluation but would nevertheless be an argument against counting the citations to particularly highly cited articles (i.e., citation classics: Garfield, 1987) at face value because a high proportion may be due to existing citations rather than the intrinsic worth of the article. In bibliometrics, this may be related to the previously observed phenomena of the perfunctory citations given to works that appear to be merely concept markers (Case & Higgins, 2000). Although the hooked power law behaves like a power law when x is much greater than B, the additional parameter B suggests that another factor is also at work (see Appendix A).

It is also important to assess which distributions fit citation data best because this can help when conducting theoretical studies of factors that affect citation counts, such as the role of collaboration and countries (Didegah & Thelwall, 2013; Peters & van Raan, 1994) as well as when using citation counts for research evaluation purposes – typically for articles within a single subject and a single year or small time window. Such studies can potentially be more powerful if they use statistical methods that are best tailored to the appropriate citation distribution.