The functional relation between the impact factor and the uncitedness factor revisited

Peer-reviewed author version

EGGHE, Leo (2013) The functional relation between the impact factor and the uncitedness factor revisited. In: Journal of Informetrics, 7(1), p. 183-189..

DOI: 10.1016/j.joi.2012.10.007
Handle: http://hdl.handle.net/1942/14330

The functional relation between the impact factor and the uncitedness factor revisited

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek,
Belgium[1]

and

Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen,
Belgium

leo.egghe@uhasselt.be

ABSTRACT

We give a heuristic proof of the relation between the impact factor (IF) and the uncitedness factor (U), the fraction of the papers that are uncited:

$$U = \frac{1}{1+IF}$$

[1] Permanent address

Key words and phrases: impact factor, uncitedness factor, universal informetric law, distribution dependent law, parameter dependent law.

Acknowledgement: The author is grateful to J.-w. Hsu and D.-w. Huang for the details of their proof of the relation between the impact factor and the uncitedness factor.

This generalizes the proof of Hsu and Huang [Physica A 391, 2129-2134, 2012] who obtain the same result but based on the assumption of the validity of the Matthew-effect. This new informetric function opens the discussion on universal informetric laws, distribution dependent laws and parameter dependent laws of which examples from the informetrics literature are given.

Introduction

Let us have a set of papers (e.g. papers in a journal). Then we can determine a fixed publication period and a fixed citation period where citations to the publications in the publication period are studied. Given these time periods we can determine the average number μ of citations per article (which is in fact a generalized impact factor - generalized because the publication and citation period are not specified). We can also study the fraction of uncited papers U.

Here μ can vary between 0 and ∞ and U can vary between 0 and 1 (being a fraction). An intuitive idea tells us that μ is large if and only if U is small (close to 0) and that μ is small if and only if U is large (close to 1). In the next section we will prove the second assertion but we will present citation size-frequency distributions that contradicts the first assertion.

Yet in van Leeuwen and Moed (2005) one shows by examples that both assertions on U and μ are true. There one studies the U(μ) functionality in a semi-logarithmic scale (logarithmic in μ), see also Egghe (2008, 2010).

Further, in Hsu and Huang (2012), the following simple relation between μ and U is proved (heuristically)

$$U = \frac{1}{1+\mu} \tag{1}$$

The heuristic proof is based on the assumption of the Matthew effect (also called cumulative advantage or success-breeds-success – see e.g. Egghe (2005)): the more citations an article has, the larger the probability that this paper will receive a new citation. An outline of the simple heuristic proof is based on Hsu and Huang (2012) and on written communication between these authors and the present author and goes as follows.

Denote by $f\left(n\right)$ the fraction of papers with *n* citations $\left(n \geq 0\right)$. The Matthew effect is expressed by

$$\frac{df(n)}{dn} = cf(n) \tag{2}$$

where $c < 0$ is a constant. Hence

$$\frac{df(n)}{f(n)} = cdn$$

or

$$\int_{f(0)}^{f(n)} \frac{df}{f} = c\int_0^n dn$$

or

$$\ln \frac{f(n)}{f(0)} = cn$$

implying

$$f(n) = f(0)e^{cn} \tag{3}$$

Expressing that

$$1 = \sum_{n=0}^{\infty} f(n) = \frac{f(0)}{1-e^c} \tag{4}$$

and

$$\mu = \sum_{n=0}^{\infty} nf(n) = \frac{f(0)e^c}{\left(1-e^c\right)^2} \tag{5}$$

(see Gradshteyn and Ryzhik (1980), p.7, formulae 1 and 2 and since $e^c < 1$) yields

$$f(n) = \left(\frac{1}{1+\mu}\right)\left(\frac{\mu}{1+\mu}\right)^n \tag{6}$$

for $n = 0,1,2,\ldots$. But $U = f(0)$. So (6) yields (1) by putting n=0.

Note that the above proof is heuristic since, in (2), the discrete $n = 0,1,2,...$ is approximated by the continuous $n \geq 0$. Function (1) has the properties that

$$\lim_{\mu \to \infty} U = 0 \tag{7}$$

and that

$$\lim_{\mu \to 0} U = 1 \tag{8}$$

In Hsu and Huang (2012), it is shown by examples that the convexly decreasing function (1) fits practical data very well. In the next section we present a heuristic proof of (1) which is not based on the Matthew principle, showing that (1) is a really universal informetric function.

Based on this fact, we present in the third section a classification of several informetric functions (known in the literature) into universal functions, distribution dependent functions and parameter dependent functions.

The paper closes with some concluding remarks.

Study of the functional relation between μ and U

Let us have a general citation size-frequency function $f(n)$, $n = 0,1,2,...$. Here $f(n)$ denotes the fraction of the papers which received $n = 0,1,2,...$ citations (we do not specify the publication and citation periods since this is not important here). In general we have

$$\sum_{n=0}^{\infty} f(n) = 1 \tag{9}$$

$$\sum_{n=0}^{\infty} nf(n) = \sum_{n=1}^{\infty} nf(n) = \mu \tag{10}$$

$$U = f\left(0\right) = 1 - \sum_{n=1}^{\infty} f\left(n\right) \tag{11}$$

We have the following proposition.

<u>Proposition 1:</u> $\mu = 0$ if and only if $U = 1$

By continuous heuristic extension: μ is small if and only if U is large.

Proof: $\mu = 0$ implies, by (10)

$$\sum_{n=1}^{\infty} nf\left(n\right) = 0$$

So

$$0 \le \sum_{n=1}^{\infty} f\left(n\right) \le \sum_{n=1}^{\infty} nf\left(n\right) = 0$$

So

$$\sum_{n=1}^{\infty} f\left(n\right) = 0$$

and hence, by (11), $U = 1$.

$U = 1$ implies, by (11)

$$\sum_{n=1}^{\infty} f\left(n\right) = 0$$

Hence $f\left(n\right) = 0$ for all $n = 1, 2, ...$

Hence

$$\mu = \sum_{n=1}^{\infty} nf\left(n\right) = 0$$

□

The next “assertion” is not true in general

Assertion 1: $\mu = \infty$ if and only if $U = 0$.

By continuous heuristic extension: μ is large if and only if U is small.

Counterexample to Assertion 1.

Take (Lotka’s law)

$$f\left(n\right) = \frac{C}{n^{\alpha}} \tag{12}$$

with $\alpha > 2$ for $n = 1, 2, ...$ and $f\left(0\right) = U = 0$ and where $C > 0$ is chosen so that

$$\sum_{n=1}^{\infty} f\left(n\right) = 1$$

Since $\alpha > 2$, we know that

$$\mu = \sum_{n=1}^{\infty} nf\left(n\right) = \sum_{n=1}^{\infty} \frac{C}{n^{\alpha-1}} < \infty$$

This example shows that $U = 0$ can occur together with $\mu < \infty$.

An example where $\mu = \infty$ can occur and where $U \neq 0$ is given now. Let $f\left(0\right) = U \neq 0$ and

$$f\left(n\right) = \frac{C}{n^{\alpha}}$$

for $n = 1, 2, ...$ with $1 < \alpha < 2$ and where C is taken such that

$$\sum_{n=1}^{\infty} f\left(n\right) = 1 - U$$

This can indeed be accomplished since

$$\sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} < \infty$$

since $1 < \alpha < 2$. But now we have

$$\mu = \sum_{n=1}^{\infty} nf(n) = \sum_{n=1}^{\infty} \frac{C}{n^{\alpha-1}} = \infty$$

since $\alpha - 1 < 1$.

Another example where $\mu = \infty$ can occur where $U \neq 0$ is given now (in continuous variables).

Let

$$f(n) = \frac{C}{(n+1)^{\alpha}} \tag{13}$$

, $C > 0$, $\alpha > 1$ for $n = 0,1,2,...$ be the shifted Lotka distribution (cf. Egghe and Rousseau (2012a)). We have to take

$$C = \alpha - 1 \tag{14}$$

in order to have

$$\int_0^{\infty} f(n)dn = 1 \tag{15}$$

Then

$$\mu = \int_0^{\infty} nf(n)dn = \frac{1}{\alpha - 2} \tag{16}$$

for $\alpha > 2$ as is readily seen (see also Egghe and Rousseau (2012a)).

The uncitedness factor U is approximated by

$$U = \int_0^1 f(n)dn$$

$$U = 1 - 2^{1-\alpha} \tag{17}$$

by (14). Using (16) we see that

$$\alpha - 1 = \frac{1}{\mu} + 1$$

from which we derive from (17) that

$$U = 1 - \frac{1}{2^{\frac{1}{\mu}+1}} \tag{18}$$

We now see that U can never be 0. Its minimal value is for $\mu = \infty$ (for $\alpha \to 2$): $U = \frac{1}{2} \neq 0$.

Now we turn our attention to a heuristic proof of equation (1)

Heuristic proof of (1):

Let us have a set of T articles and their citations and let μ be the average number of citations per article. Of course in such a general situation we cannot prove (1). Therefore, during this proof, we will make a logical assumption (see further). Let U be the fraction of articles with 0 citations (hence the uncitedness factor as above).

Now we add to each article one citation. Hence now U is the fraction of articles with 1 citation and in this system the average number of citations per article is $\mu + 1$. Since there are T articles, the total number of citations is $T(\mu+1)$ and they are spread out over T articles. Hence

$$\frac{T}{T(\mu+1)} = \frac{1}{\mu+1} \tag{19}$$

stands for the fraction of T representing 1 citation. By the above,

$$U = \frac{1}{\mu + 1} \tag{20}$$

completing this argument, assuming that the fraction of T representing 1 citation is also the fraction of articles with 1 citation (hence is the original fraction U of articles with 0 citations).

Note that formula (20) satisfies Proposition 1 (as it should) but also Assertion 1 (which is not always satisfied).

Formula (20) is a universal function, independent of any informetric distribution. This and related phenomena will be discussed in the next section.

Universal and “semi-universal” laws in informetrics

In this section we will generalise articles to “sources” and their citations to “items” so that we work with general information production processes (IPPs).

As said above, formula (20) is a universal function, not dependent on any parameter. In Hsu and Huang (2012) it was proved (see the proof in the previous section) assuming the validity of the Matthew effect (or “Success-Breeds-Success” – see e.g. Egghe (2005)). In this paper we gave a heuristic (algebraic) proof without supposing any distribution.

In the sequel we will consider the following “classification” of informetric laws.

- (I) Universal laws, not dependent on any distribution or parameter,
- (II) Universal laws, dependent on a distribution but not on the distribution’s parameters. We can call these “distribution dependent laws”.
- (III) Universal laws, dependent on the parameters of a distribution (and hence also of the distribution itself). We can call these “parameter dependent laws”.

Several examples will be given.

Function (20), based on the proof of Hsu and Huang (2012) can be considered to belong to category (II) (since it depends on the exponential distribution (6)) but because of the heuristic proof given in this paper, it can be considered to belong to category (I).

Let us denote by T the total number of sources and by A the total number of items in the IPP. In Chen and Leimkuhler (1989) we find the following function:

$$\frac{T}{A}+\frac{\ln T}{\ln A}=1 \tag{21}$$

which is shown to be approximately true, independent of any distribution (see also Egghe (2005), p. 298-300 and Egghe (2007)). This approximate law hence belongs to category (I).

In order to be able to continue, remark that Lotka's law (12) (with $\alpha > 1$) is equivalent with Zipf's law

$$g(r)=\frac{B}{r^{\beta}} \tag{22}$$

with $\beta, B > 0$. Here (22) is the continuous version of the number of items in the source on rank $r = 1,...,T$. In this connection one has shown (see Egghe (2005)) that

$$\beta=\frac{1}{\alpha-1} \tag{23}$$

a universal law, belonging to category (II) since it depends on an informetric distribution (Lotka and its equivalent Zipf). The formula (for $\alpha > 2$)

$$\frac{A}{T}=\mu=\frac{\alpha-1}{\alpha-2} \tag{24}$$

(Egghe (2005)) also belongs to category (II). The same goes for the formulae for percentiles in this Lotkaian framework (see Egghe (2009)). If we combine function (20) with (24) we obtain

$$U = \frac{\alpha - 2}{2\alpha - 3} \tag{25}$$

, also a universal law belonging to category (II) since it depends on the same laws of Lotka and Zipf.

The “generalized 80/20-rule”

$$y = x^{1/\mu} \tag{26}$$

belongs to category (III) (here 100 x % of the most productive sources have 100 y % of the items) (see Egghe (2005)).

Also the formula for the Hirsch-index h (Egghe and Rousseau (2006)) clearly belongs to category (III) since it depends on Lotka’s law and its parameters:

$$h = T^{1/\alpha} \tag{27}$$

with

$$T = \frac{C}{\alpha - 1} \tag{28}$$

(Egghe (2005)).

In Egghe, Liang and Rousseau (2009) and Egghe and Rousseau (2012b) we give relations between the h-index h and the impact factor IF. Supposing Lotka’s law (12) we show

$$h = \left(C \left(1 - \frac{1}{IF} \right) \right)^{\frac{IF-1}{2IF-1}} \tag{29}$$

Since the parameter C is involved, this relation belongs to category (III).

Supposing an exponential aging function

$$c(t) = ca^t \tag{30}$$

for the number of received citations at time $t(0 < a < 1)$ and Lotka's law for the number of papers with *n* citations, we proved in Egghe (2000) that cumulative first-citation distribution $\Phi(t_1)$ is

$$\Phi(t_1) = \gamma\left(1 - a^{t_1}\right)^{\alpha-1} \tag{31}$$

(here γ is the fraction of ever cited papers), clearly belonging to category (III). Also the formulae on the Price-index PI_d (the fraction of references that are less than or equal to *d* years old) in Egghe (1997) belong to category (III).

The remarkable formula of Naranan (1970)

$$\alpha = 1 + \frac{\ln a_1}{\ln a_2} \tag{32}$$

where α is Lotka's law and $a_1 > 1$ is the growth rate of the exponential growth of the sources and $a_2 > 1$ is the growth rate of the exponential growth of the items is a universal relation of category (II).

Finally we consider Benford's law

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{33}$$

(see e.g. Benford (1938) and Egghe (2011)) being the probability to have the digit $d = 1, 2, ..., 9$ as first digit in a number is derived from Zipf's law (22) with $\beta = 1$ and hence belongs to category (II). The generalized law of Benford (Pietronero, Tosatti, Tosatti and Vespignani (2001), Nirgini and Miller (2007), Luque and Lacasa (2009) and Egghe and Guns (2012))

$$P(d)=\frac{1}{10^{1-\beta}-1}\left((d+1)^{1-\beta}-d^{1-\beta}\right) \tag{34}$$

is derived from the general law of Zipf (22) and contains its parameter β. Hence it belongs to category (III).

Concluding remarks

In this paper we studied the relation between the average number of citations per paper (μ) and the uncitedness factor (U), i.e. the fraction of uncited papers. We show that in general $\mu=0$ if and only if $U=1$ is true but that $\mu=\infty$ if and only if $U=0$ is not true in general.

Then we give a heuristic proof of the relation of Hsu and Huang (2012):

$$U=\frac{1}{1+\mu} \tag{35}$$

without supposing the Matthew effect (as they do). Hence the function (35), being independent of a distribution function and of parameters, is a universal law in informetrics.

This finding gave us the idea of classifying informetric laws into three categories:

(I) Universal laws, not dependent on any distribution or parameter,

(II) Universal laws, dependent on a distribution but not on the distribution’s parameters,

(III) Universal laws, dependent on the parameters of a distribution.

We have classified several well-known informetric laws into these three categories and noted that there are very few laws belonging to category (I); most laws are belonging to category (III).

No doubt that there are other laws in informetrics that can be classified this way. This is left to the reader who is hereby invited to present his/her ideas on this matter with the present author.

References

F. Benford (1938). The law of anomalous numbers. Proceedings of the American Philosophical Society 78, 551-572.

Y.-S. Chen and F.F. Leimkuhler (1989). A type-token identity in the Simon-Yule model of text. Journal of the American Society for Information Science 40(1), 45-53.

L. Egghe (1997). Price index and its relation to the mean and median reference age. Journal of the American Society for Information Science 48(6), 564-573.

L.Egghe (2000). A heuristic study of the first-citation distribution. Scientometrics 48(3), 345-359.

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

L. Egghe (2007). Untangling Herdan’s law and Heaps’ law: mathematical and informetric arguments. Journal of the American Society for Information Science and Technology, 58(5), 702-709.

L.Egghe (2008). The mathematical relation between the impact factor and the uncitedness factor. Scientometrics 76(1), 117-123.

L. Egghe (2009) Performance and its relation with productivity in Lotkaian systems. Scientometrics 81(2), 567-585.

L. Egghe (2010). The distribution of the uncitedness factor and its functional relation with the impact factor. Scientometrics 83(3), 689-695.

L. Egghe (2011). Benford’s law is a simple consequence of Zipf’s law. ISSI Newsletter 7(3), 55-56.

L. Egghe and R. Guns (2012). Applications of the generalized law of Benford to informetric data. Journal of the American Society for Information Science and Technology, to appear.

L. Egghe, L. Liang and R. Rousseau (2009). A relation between *h*-index and impact factor in the power-law model. Journal of the American Society for Information Science and Technology 60(11), 2362-2365.

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch-index. Scientometrics, 69(1), 121-129.

L. Egghe and R. Rousseau (2012a). Theory and practise of the shifted Lotka function. Scientometrics, to appear.

L. Egghe and R. Rousseau (2012b). The Hirsch-index of a shifted Lotka function and applications to the relation with the impact factor. Journal of the American Society for Information Science and Technology, to appear.

J.-w. Hsu and D.-w. Huang (2012). A scaling between impact factor and uncitedness. Physica A 391, 2129-2134.

B. Luque and L. Lacasa (2009). The first digit frequencies of prime numbers and Riemann zeta zeros. Proceedings of the Royal Society A 465, 2197-2216.

S. Naranan (1970). Bradford's law of bibliography of science: an interpretation. Nature, 227, 631-632.

M.J. Nigrini and S.J. Miller (2007). Benford's law applied to hydrology data - Results and relevance to other geophysical data. Mathematical Geology 39, 469-490.

L. Pietronero, E. Tosatti, V. Tosatti and A. Vespignani (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. Physica A 293, 297-304.

T. N van Leeuwen and H.F. Moed (2005). Characteristics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics 63(2), 357-371.