Lotkaian informetrics and applications to social networks

EGGHE, Leo (2009) Lotkaian informetrics and applications to social networks. In: BULLETIN OF THE BELGIAN MATHEMATICAL SOCIETY-SIMON STEVIN, 16(4). p. 689-703.

Handle: http://hdl.handle.net/1942/9281

Lotkaian informetrics and applications to social networks

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium
E-mail address: leo.egghe@uhasselt.be

ABSTRACT. Two-dimensional informetrics is defined in the general context of sources that produce items and examples are given. These systems are called "Information Production Processes" (IPPs). They can be described by a size-frequency function f or, equivalently, by a rank-frequency function g. If f is a decreasing power law then we say that this function is the law of Lotka and it is equivalent with the power law g which is called the law of Zipf. Examples in WWW are given.

Next we discuss the scale-free property of f also allowing for the interpretation of a Lotkaian IPP (i.e. for which f is the law of Lotka) as a self-similar fractal.

Then we discuss dynamical aspects of (Lotkaian) IPPs by introducing an item-transformation φ and a source-transformation ψ. If these transformations are power functions we prove that the transformed IPP is Lotkaian and we present a formula for the exponent of the Lotka law. Applications are given on the evolution of WWW and on IPPs without low productive sources (e.g. sizes of countries, municipalities or databases).

Lotka's law is then used to model the cumulative first citation distribution and examples of good fit are given.

Finally, Lotka's law is applied to the study of performance indices such as the h-index (Hirsch) or the g-index (Egghe). Formulas are given for the h- and g-index in Lotkaian IPPs and applications are given.

Key words and phrases: law of Lotka, law of Zipf, information production

process, IPP, fractal, dynamics, cumulative first-citation distribution, h-index, Hirsch-index, g-index.

AMS classification codes:
Primary: 94A15

Acknowledgement: The author is grateful to Profs. Dr. A. Bultheel and F. Dumortier for the invitation to write this article.

1 Introducing Lotkaian informetrics

1.1 Information Production Processes (IPPs)

Every system in which there are sources that "produce" items can be considered as an IPP (in a generalized meaning). Examples:

Sources	$\rightarrow$	Items
Authors	$\rightarrow$	Articles
Journals	$\rightarrow$	Articles
Articles	$\rightarrow$	Citations (to/from)
Articles	$\rightarrow$	Co-authors
Books	$\rightarrow$	Borrowings
Words (= types)	$\rightarrow$	Use of words in a text (=tokens)
Web sites	$\rightarrow$	Hyperlinks (in-/out-)
Web sites	$\rightarrow$	Web pages
Cities/villages	$\rightarrow$	Inhabitants
Employees	$\rightarrow$	Their production
Employees	$\rightarrow$	Their salaries
	...	

All social networks are examples of (extended) IPPs: WWW, Intranets, Internet, citation networks, collaboration networks, ...

1.2 Informetrics

Every IPP can, mathematically, be described by a size-frequency function f:

$$f(n) = \#\text{ sources with } n \text{ items} \tag{1}$$

($n = 1, 2, 3, \ldots$). Equivalently, every IPP can be described by a rank-frequency function g: order the sources in decreasing order of their number of items. Then g is defined as

$$g(r) = \#\text{ items in the source on rank } r \tag{2}$$

($r = 1, 2, 3, \ldots$). We have, clearly

$$r = g^{-1}(n) = \sum_{k=n}^{\infty} f(k) \tag{3}$$

For calculatory reasons we will work with continuous variables, i.e. with item and source densities: $r \in [0, T]$, $j \in [1, +\infty[$,

$$r = g^{-1}(j) = \int_{j}^{\infty} f(j')dj' \tag{4}$$

$$f(j) = -\frac{1}{g'(g^{-1}(j))} \tag{5}$$

1.3 Lotkaian informetrics

The most important type of informetrics is Lotkaian informetrics, i.e. where $f(j)$ has the form:

$$f(j) = \frac{C}{j^{\alpha}} \tag{6}$$

$C > 0$, $\alpha > 1$ (cf. Lotka (1926)). Even when this is disputed by some or when better fits are possible, using more intricate models (with more parameters), the Lotka function (6) yields an easy tool to explain many regularities (see e.g. further) and hence, (6) could be considered as a kind of "axiom" within this theory. So f is a decreasing power function. The value $\alpha = 2$ is a turning point in informetrics: many informetric properties change when going from $\alpha < 2$ to $\alpha > 2$ (see also further).

It is easily seen that Lotka's law is equivalent with Zipf's law: $g(r)$ has the form

$$g(r) = \frac{B}{r^{\beta}} \tag{7}$$

B, $\beta > 0$. Important here is the formula

$$\beta = \frac{1}{\alpha - 1} \tag{8}$$

Zipf's law originates from linguistics (Zipf 1949) (but see already Estoup (1916) and Condon (1928)). The type of law (7) also appears in econometrics and is there called the law of Pareto. That the same law was invented in these diverse disciplines underlines its importance.

All mentioned IPPs (incl. all social networks) agree reasonably with Lotka's law (see examples in Figs.1-4, for more examples, see the book Egghe (2005b)). The source-item relation is based on the principle "Success-Breeds-Success" (cf. Simon (1957), Price 1976)). Random networks (the so-called Erdös-Rényi networks) are different and agree with exponentially decreasing size-frequency functions f (Erdös and Rényi (1960)).

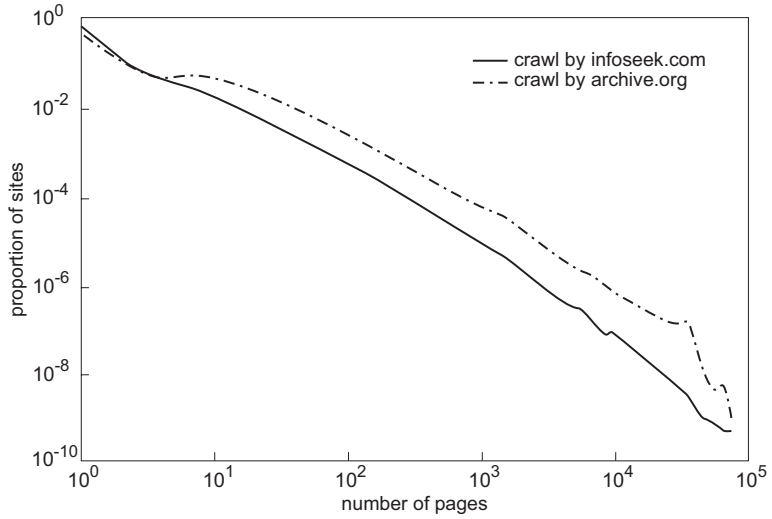

Fig.1 Rank-frequency distribution of web sites versus # web pages (log-log scale)

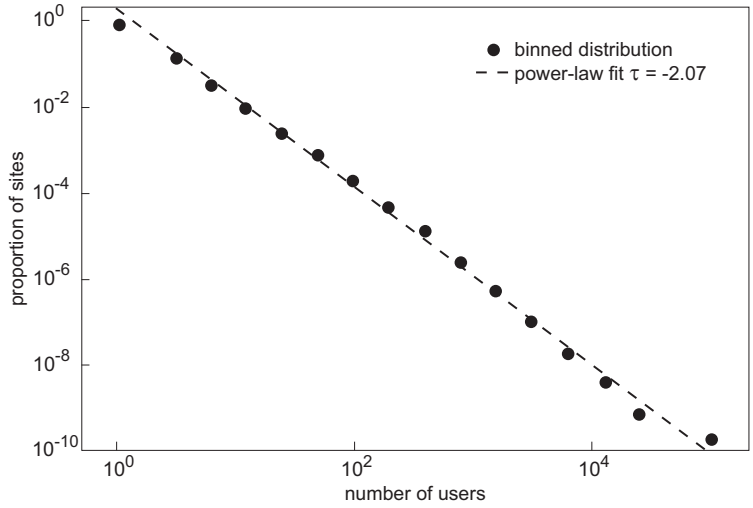

Fig.2 Rank-frequency distribution of web sites versus # users (log-log scale)

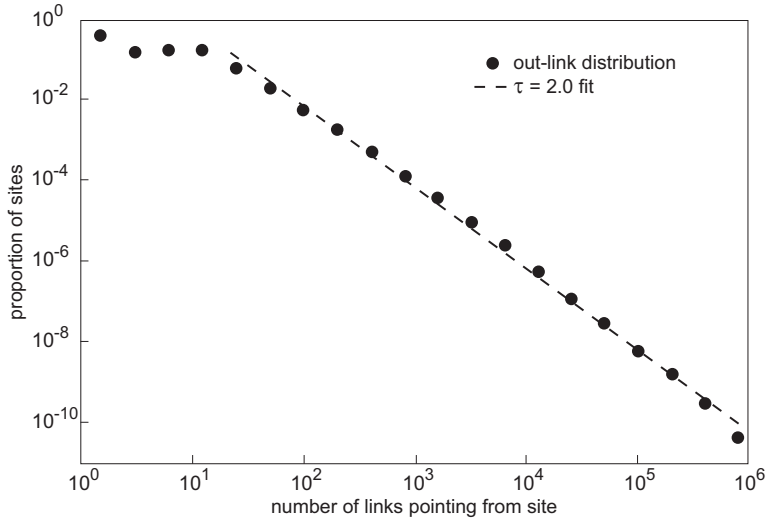

Fig.3 Rank-frequency distribution of web sites versus # out-links (log-log scale)

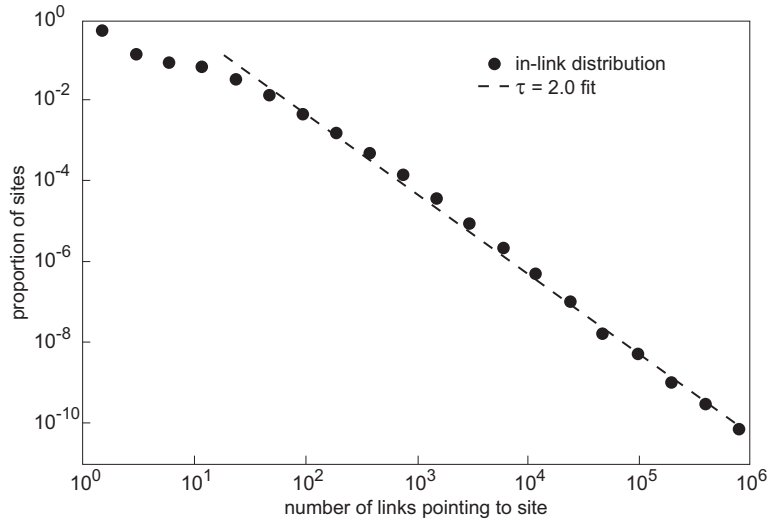

Fig.4 Rank-frequency distribution of web sites versus # in-links (log-log scale)

2 Fractal aspects of Lotkaian IPPs

The law of Lotka - being a power law - has the so-called scale-free property for functions defined on $[1, +\infty[$: $\forall C > 1$, $\exists D > 0$ such that $f(Cx) = Df(x)$ for all x. In fact, as is well-known, this property characterizes power laws (Roberts (1979)) for functions defined on $\mathbb{R}^+$. This property, indirectly, is the reason why we can interpret Lotkaian IPPs as self-similar fractals. The following theorem is of Naranan (1970).

Theorem: Suppose that

(i) the number of sources grows exponentially in time t:

$$N(t) = c_1 a_1^t \tag{9}$$

(ii) the number of items in each source grows exponentially in time t and the growth rate is the same in every source:

$$P(t) = c_2 a_2^t \tag{10}$$

Then this IPP is Lotkaian: if $f(p)$ denotes the number of sources with p items, then

$$f(p) = \frac{C}{p^\alpha} \tag{11}$$

where

$$\alpha = 1 + \frac{\ln a_1}{\ln a_2} \tag{12}$$

Naranan did not see that this result makes it possible to interpret a Lotkaian IPP as a self-similar fractal. Indeed, take e.g. the triadic Koch curve (Feder (1988), p.16): start with a line piece, divide its length by 3 and form a "hat" by using 4 of these line pieces. This is then repeated infinitely. Hence we have

(i) The number of line pieces grows exponentially in time t, proportionally to 4^t

(ii) 1/length of each line piece grows exponentially in time t with the same growth rate of 3: we have growth proportional to 3^t.

Hence by the theorem of Naranan we can consider a Lotkaian IPP as a self-similar fractal. The fractal dimension of the Koch curve is $\frac{\ln 4}{\ln 3}$. So a Lotkaian IPP is interpreted as a self-similar fractal with fractal dimension

$$D = \frac{\ln a_1}{\ln a_2} \tag{13}$$

By (12) and (13) we see that

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

We see here the crucial role of the Lotka exponent α. Result (14) was proved earlier by Mandelbrot but only for a special IPP: random texts (linguistics) (Mandelbrot (1967, 1977). Egghe (2005a, 2005b)) extended this to general Lotkaian IPPs. Note that $\alpha = 2 \Leftrightarrow D = 1 \Leftrightarrow a_1 = a_2$ so that sources and items (in sources) grow at the same rate in this case.

3 Dynamical aspects of Lotkaian IPPs

Dynamical aspects of Lotkaian IPPs can be described using transformations on the sources and on the items (Egghe (2007): an article in the new "Journal of Informetrics"). Let us have a general IPP in which item densities j and rank densities r are described by the size-frequency function $j \to f(j)$ and the rank-frequency function $r \to g(r)$. We apply the following transformations on j and $r : j \to \varphi(j) =$

j^*, $r \to \psi(r) = r^*$ such that the new rank-frequency function, denoted $r^* \to g^*(r^*)$, is given by

$$g^*(r^*) = g^*(\psi(r)) = \varphi(g(r)) \tag{15}$$

Theorem (Egghe (2007)): The new size-frequency function $j^* \to f^*(j^*)$ is given by

$$f^*(j^*) = f(j)\frac{\psi'(g^{-1}(j))}{\varphi'(j)} \tag{16}$$

Theorem (Egghe (2007)): In case our IPP is Lotkaian:

$$f(j) = \frac{C}{j^\alpha} \tag{17}$$

$(C > 0, \alpha > 1)$ and in case we apply power transformations:

$$r^* = \psi(r) = Ar^b \tag{18}$$

$$j^* = \varphi(j) = Bj^c \tag{19}$$

$(A, B, b, c > 0)$, we have

$$f^*(j^*) = \frac{G}{j^\delta} \tag{20}$$

($G > 0$ a constant) and where δ is given by

$$\delta = 1 + \frac{b(\alpha - 1)}{c} \tag{21}$$

Note that δ is only dependent on b/c due to the scale-free nature of Lotkaian systems.

Corollary (Egghe (2007)):

$$\delta < \alpha \Leftrightarrow b < c \tag{22}$$

$$\delta > \alpha \Leftrightarrow b > c \tag{23}$$

$$\delta = \alpha \Leftrightarrow b = c \tag{24}$$

The results above where checked in Cothey (2007) in the connection of the evolution of a part of WWW: the above theory was confirmed except in one case where non-Lotkaian evolution was found, probably due to "automatic" creation of web pages (deviation from a social network).

A further application is given in Egghe and Rousseau (2006a), based on a special case of the results in, Egghe (2007) which were already published in Egghe (2004): $\psi = Id =$ the identity function and $\varphi(j) = Bj^c$, B, $c > 1$: sources remain the same but they grow in number of items. Now (21) gives

$$\delta = 1 + \frac{\alpha - 1}{c} \tag{25}$$

and (22) gives: $\delta < \alpha$ and, since $j \geq 1 : \varphi(j) \geq B > 1$. Repeated application of this transformation yields that IPPs where there are no low productive soures ($\varphi(j) >> 1$) have small Lotka exponents δ. This is confirmed in all the cases which Egghe and Rousseau investigated:

1. Country sizes: data from www.gazetteer.de (July 10, 2005): 237 countries: $\delta = 1.69$
2. Municipalities in Malta (1997 data): 67 municipalities: $\delta = 1.12$
3. Database sizes: on the topic "fuzzy set theory" (20 largest databases on this topic - Hood and Wilson (2003)): $\delta = 1.09$
4. Unique documents in the 20 databases above (Hood and Wilson (2003)): $\delta = 1.33$.

4 Lotka's law and the modelling of the cumulative first-citation distribution

The cumulative first-citation distribution is the cumulative distribution over time at which an article receives its first citation. The time t_1 at which an article receives its first citation is an important indicator of the visibility of research: at the time t_1, the article switches its status from "unused" to "used". t_1 is a measure of "immediacy" but, of course, different from the immediacy index (instant impact factor) of Thomson Scientific.

So let $\Phi(t_1)$ denote the cumulative fraction of all papers that have, at t_1, at least 1 citation. In the literature one finds two different typical shapes of $\Phi(t_1)$: a concavely increasing one (e.g. Motylev (1981)) and an S-shaped one (first convexly and then concavely increasing) (e.g. Rousseau (1994)).

Rousseau (1994) uses two different differential equations to model the two different shapes. However, these equations are not explained and are not linked with any informetric distribution. In Egghe (2000) we applied the law of Lotka to solve this problem and one model (of course involving the Lotka exponent α) will explain both the concave shape and the S-shape of Φ according to $\alpha < 2$ or $\alpha > 2$. We see here that $\alpha = 2$ is a turning point in Lotkaian informetrics. We obtained:

Theorem (Egghe (2000)): Let

(i) $c(t) = ba^t$ = the density function of citations to an article, t time after its publication (exponential function, $0 < a < 1$)

(ii) $\varphi(A) = \dfrac{D}{A^\alpha}$ = the density function of the number of papers with A citations (received) in total (Lotka, $\alpha > 1$), $A \geq 1$ (only ever cited papers are used here)

Let $\gamma \in]0, 1[$ be the fraction of ever cited papers (we use γ in order to include also the never cited articles). Then

$$\Phi(t_1) = \gamma(1 - a^{t_1})^{\alpha - 1} \tag{26}$$

which is concave if $1 < \alpha \leq 2$ and is S-shaped if $\alpha > 2$.

For the Motylev (1981) data we have a fit as in Fig. 5. For the Rousseau (1994) data we have a fit as in Fig. 6, conform with the above theory.

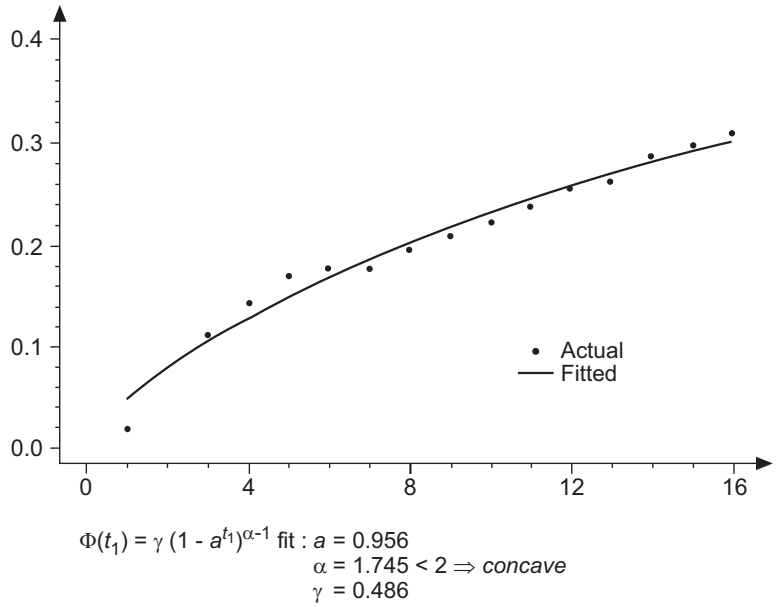

Fig. 5 Cumulative first-citation distribution: case of Motylev data.

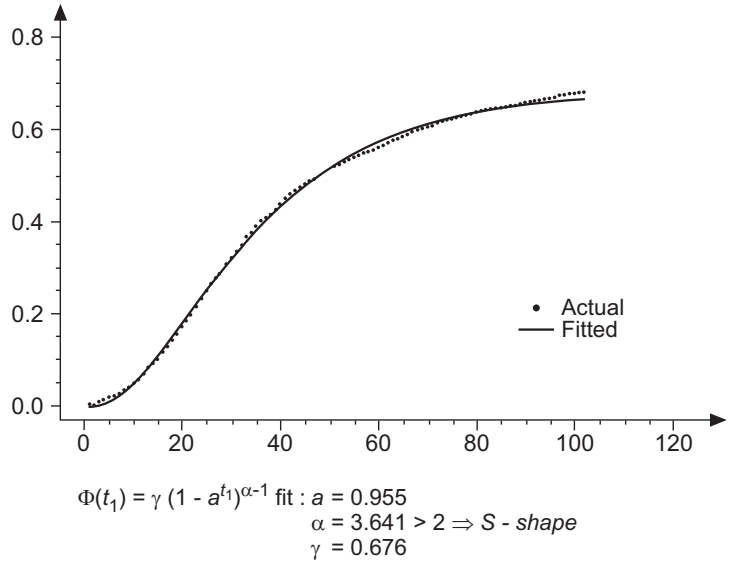

Fig. 6 Cumulative first-citation distribution: case of Rousseau data.

All the results on Lotka's law described here (and a lot more) are described in the book Egghe (2005b).

We now come to an application of Lotka's law, established in the post-2005 period.

5 Lotka's law and its applications to the modelling of the *h*-index and the *g*-index

The h-index, defined by Hirsch in 2005 (Hirsch (2005)), is a new performance indicator for the visibility and quality of the publication output of a researcher. However, the h-index can be calculated for any set of articles.

Definition (Hirsch (2005)): Let us rank the publications of an author in decreasing order of the number of citations they received. The h-index is then the largest rank such that the paper on this rank (and hence also all papers on rank $1, \ldots, h$) has h or more citations.

Example: the h-index of this author (based on data of the Web of Science), date July 24, 2008. We give the number of citations to the first 23 papers (for reasons which will become clear later) in decreasing order of citations received

r	# citations
1	56
2	44
3	43
4	36
5	27
6	22
7	21
8	20
9	18
10	17
11	17
12	16
13	16
14	16
15	16
16	16
17	15
18	15
19	14
20	14
21	14
22	13
23	13

It is clear that $h = 16$ for this author at this time.

The h-index can be defined in any IPP by ranking the sources in decreasing order of their number of items. In Lotkaian IPPs we have the following result on the h-index.

Theorem (Egghe and Rousseau (2006b)) Let the IPP be Lotkaian with Lotka exponent $\alpha > 1$. Denote by T the total number of productive sources (i.e. with at least one item). Then

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

This basic result, together with the transformation theory for IPPs (Section III) gives us the possibility, at least theoretically, to see how h changes when we change

the number of items (citations) or sources (papers). Let us have a Lotkaian IPP such that (27) is valid. Suppose we double the number of citations per paper. Then Egghe (2008) has shown that, for this new situation (denoting its h-index by h^*)

$$h < h^* = 2^{\frac{\alpha-1}{\alpha}} h < 2h \tag{28}$$

Suppose that we start with (27) and that we double the number of papers (# citations remain the same). Then this situation has a h-index (denoted h^*)

$$h < h^* = 2^{1/\alpha} h < 2h \tag{29}$$

Suppose we start with (27) and now we double the number of papers and that we divide the number of citations by 2. The new h-index h^* is now

$$h^* = 2^{2/\alpha - 1} h > h \tag{30}$$

iff $\alpha < 2$. So $h^* > h \Leftrightarrow \alpha$ low $\Leftrightarrow h$ high (by (27)) $\Leftrightarrow$ prolific author. A situation like this occurs in non-controlled listings (splitting articles) or in case of "publicitis": publishing the "least publishable unit".

The h-index is robust in the sense that it is insensitive to the existence of a set of lowly cited articles. However, a disadvantage of h-index is that it is also insensitive to the number of citations of the highly cited articles. Indeed, in the previous example of citation data of this author, the first article has 56 citations but h would remain the same ($h = 16$) if this article had any number of citations larger than or equal to 16. The same goes for the other articles.

In Egghe (2006), we tried to present a new index which takes into account the number of citations of the "most" cited articles. Note that the h-index satisfies the property: the first h articles have, **together**, at least h^2 citations. Egghe (2006) defines the g-index as the largest rank g with this property. Otherwise formulated we have the following definition.

Definition (Egghe (2006)): Let us rank the publications of an author in decreasing order of the number of citations they received. The g-index is the largest rank such that the first g papers have, **on average**, g citations.

In the example of this author we have that the total number of citations of the

first 22 papers equals $486 > 22^2$ while the total number of citations of the first 23 papers equals $499 < 23^2$, hence $g = 22$. Note that always, by definition and the property of h, $g \geq h$. In practise it can be that the sum of all citations of an author is higher than T^2, the square of the total number of papers. In this case add fictitious papers with zero citations, enough to calculate the g-index as indicated. Schreiber (2008a, b) and Tol (2008) agree with the fact that the g-index characterizes the data set better than the h-index and that the g-index has greater discriminating power than the h-index.

In Lotkaian IPPs we have the following result, analogous to (27), for the g-index.

Theorem (Egghe (2006)): Let the IPP be Lotkaian with Lotka exponent $\alpha > 2$. Denote by T the total number of sources. Then

$$g = \left(\frac{\alpha - 1}{\alpha - 2}\right)^{\frac{\alpha-1}{\alpha}} T^{1/\alpha} \tag{31}$$

$$g = \left(\frac{\alpha - 1}{\alpha - 2}\right)^{\frac{\alpha-1}{\alpha}} h \tag{32}$$

The h-index is calculated in Scopus and the Web of Science. The h- and the g-index are calculated in http://www.harzing.com/pop.htm which is a Google Scholar related site. Note, however, that the h- and g-index values depend on the used database.

Note: In Jin, Liang, Rousseau and Egghe (2007) one defines the R-index which has also the purpose to improve the h-index by taking actual citation scores of highly cited papers. Denote by c_i the number of citations received by the ith paper (as always, papers are ranked in decreasing order of the number of received citations). Then R is defined as

$$R = \sqrt{\sum_{i=1}^{h} c_i} \tag{33}$$

where h is the h-index. It is trivial that $R \geq h$ and in Jin, Liang, Rousseau and Egghe (2007) we have shown that in Lotkaian IPPs with exponent $\alpha > 2$, we have

$$R = \sqrt{\frac{\alpha - 1}{\alpha - 2}}\, T^{1/\alpha} \tag{34}$$

$$R = \sqrt{\frac{\alpha - 1}{\alpha - 2}}\, h \tag{35}$$

Note: The reader might think that, from (32) and (35), g and R are linear functions of h and hence one could wonder what is the specific value of g and R above h. However, g and R are not linear functions of h. Indeed, by (27), if we keep T, the total number of sources, constant then h can only vary when α varies. But then we see from (32) and (35) that g and R do not vary linearly on h.

References

[1] E.U. Condon (1928). Statistics of vocabulary. Science 67 (1733), 300.

[2] V. Cothey (2007). Applying Egghe's general theory of the evolution of information production processes to the World Wide Web. Proceedings of ISSI 2007 (D. Torres-Salinas and H. Moed, eds.), 231-240, CSIC, Madrid, Spain.

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

[4] L. Egghe (2004). Positive reinforcement and 3-dimensional informetrics. Scientometrics 60(3), 497-509. Correction. Scientometrics 61(2), 283, 2004.

[5] L. Egghe (2005a). The power of power laws and the interpretation of Lotkaian informetric systems as self-similar fractals. Journal of the American Society for Information Science and Technology 56(7), 669-675.

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

[7] L. Egghe (2006). Theory and practise of the g-index. Scientometrics 69(1), 131-152.

[8] L. Egghe (2007). General evolutionary theory of IPPs and applications to the evolution of networks. Journal of Informetrics 1(2), 115-122.

[9] L. Egghe (2008). Examples of simple transformations of the h-index: qualitative and quantitative conclusions and consequences for other indices. Journal of Informetrics 2(2), 136-148.

[10] L. Egghe and R. Rousseau (2006a). Systems without low productive sources. Information Processing and Management 42(6), 1428-1442.

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

[12] P. Erdös and A. Rényi (1960). On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5, 17-61.

[13] J.B. Estoup (1916). Gammes Sténographiques, 4th Edition, Institut Sténographique, Paris.

[14] J. Feder (1988). Fractals. Plenum, New York, USA.

[15] J.E. Hirsch (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America 102, 16569-16572.

[16] W. Hood and C.S. Wilson (2003). Overlap in bibliographic databases. Journal of the American Society for Information Science and Technology 54, 1091-1103.

[17] B. Jin, L. Liang, R. Rousseau and L. Egghe (2007). The R-and AR-indices: complementing the h-index. Chinese Science Bulletin 52(6), 855-863.

[18] A.J. Lotka (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences 16(12), 317-324.

[19] B. Mandelbrot (1967). How long is the coast of Britain ? Statistical self-similarity and fractional dimension. Science 156, 636-638.

[20] B. Mandelbrot (1977), The Fractal Geometry of Nature. Freeman, New York, USA.

[21] V.M. Motylev (1981). Study into the stochastic process of change in the literature citation pattern and possible approaches to literature obsolescence estimation. International Forum on Information and Documentation 6, 3-21.

[22] S. Naranan (1970). Bradford's law of bibliography of science: an interpretation. Nature 227 (5258), 631-632.

[23] D. De Solla Price (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27, 292-306.

[24] F.S. Roberts (1979). Measurement Theory with Applications to Decisionmaking, Utility and the social Sciences. Addison-Wesley, Reading (MA), USA.

[25] R. Rousseau (1994). Double exponential models for first-citation processes. Scientometrics 30, 213-227.

[26] M. Schreiber (2008a). The influence of self-citation corrections on Egghe's g-index. Scientometrics 76(1), 187-200.

[27] M. Schreiber (2008b). An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the A-index, and the R-index. Journal of the American Society for Information Science and Technology 59(9), 1513-1522.

[28] H.A. Simon (1955). On a class of skew distribution functions. Biometrika 42, 425-440.

[29] R.S.J. Tol (2008). A rational successive g-index applied to economics departments in Ireland. Journal of Informetrics 2(2), 149-155.

[30] G.K. Zipf (1949). Human Behavior and the Principle of least Effort. Addison-Wesley, Cambridge (MA), USA. Reprinted: Hafner, New York, USA, 1965.