CITATION-PAPER RANK DISTRIBUTIONS AND ASSOCIATED SCIENTOMETRIC INDICATORS – A SURVEY

Vladimir Atanassov, Ekaterina Detcheva

Abstract: *This paper is devoted to studying the interrelations between most widely used scientometric indicators (in particular, Hirsch's h-, Egghe's g- and Zhang's e- indexes) for several more or less realistic citation-paper rank distributions. The analysis is provided for both continuous and discrete representations and is illustrated further on with examples for simultaneous time evolution (during a scientific career) of these indicators, computed by using real life scientometric data. The aim of the study is to illuminate specific properties of the indicators, the pros and cons of their use in various situations (citation-paper rank distributions) and (hopefully) to contribute for a fair and better scientific assessment.*

Keywords*: citation-paper rank distributions, scientometric indicators, h-index, g-index, e-index, approximate relations, data analysis*

ACM Classification Keywords*: H. Information Systems, H.2. Database Management, H.2.8. Database applications, subject: Scientific databases; I. Computing methodologies, I.6 Simulation and Modeling, I.6.4. Model Validation and Analysis*

Introduction

During the past decade the citation-based assessment of scientific activity has been essentially refined by considering details contained in *citation-paper rank distributions* and by suggesting various *scientometric indexes*. The first and most popular of them – the **Hirsch index** ([Hirsch, 2005], [Hirsch, 2007]) – has been introduced for a simple citation-paper rank distribution resulting from an extremely simplified model of a publication-citation process. Being a compromise between productivity and impact, this index ensures the opportunity for scientific assessment by a single number – a dream for many who are involved in various aspects of managing science. Although welcomed by most scientists, Hirsch's index has been criticized for underestimating the score of the most cited papers. The ***g*-index** [Egghe, 2006], constructed from informetric point of view for a Lotka (papers *vs.* citations) or Zipf (citations *vs.* paper ranks) distributions has been suggested as an alternative at least for two reasons. The first one is better accounting for the most cited papers, while the second (and in our opinion, more important) one is, that *g* as a true integral characteristic of the distribution is less subjected to statistical variability. On its turn the *g*-index has been criticized for an effect (we refer to it as *saturation* of *g*) which takes place when the total number of citations exceeds the square of the total number of publications. This criticism has lead to the appearance of the **e-index** [Zhang, 2009], also an integral characteristic that accounts for the excess of citations ignored in Hirsch's index estimation and at the same time free from this drawback. This, however, could not stop the explosion of improvements and nowadays we have several tens of indexes and numbers for scientific activity assessment [Schreiber, 2010], featuring its various aspects, like citation-paper rank distribution details (*e.g.* [Bornmann *et al*, 2010],[Cabrerizo et al, 2010]), accounting for the number of authors [Schreiber, 2008], the effect of self-citations and scientific fields specifics (*cf.* [Schreiber, 2007], [Iglesias and Pecharroman, 2007], [Alonso et

al, 2009], [Ferrara and Romero, 2012]). As an example one could point out that Harzing's *Publish or Perish* tool, based on Google Scholar database [A.-W. Harzing, 2012] estimates about 16 indicators and indexes for individual scientist's evaluation. However, these indicators (representing information squeezed from the citation-paper rank distribution) have their common origin and hence are mutually related.

The aim of this paper is to study the interrelations between most widely used scientometric indicators as Hirsch's *h*-, Egghe's *g*- and Zhang's *e*- indexes for several model continuous and discrete citation-paper rank distributions. The results obtained might be helpful to realize the pros and cons of the use of these indicators in various situations of scientific assessment. In particular, we address problems as, to what extent the indexes are robust (i.e. distribution independent), how many citations of the most cited papers are ignored by the *h*-index, which index – *g* or *e* – performs better in different cases, as well as at what conditions a saturation of *g* occurs.

The paper is organized as follows: in the first section we define the citation-paper rank distributions in both discrete and continuous representations. The second section introduces the scientometric indexes in the form they appear in the original papers ([Hirsch, 2005], [Egghe, 2006], [Zhang, 2009]). Next two sections consider the relations among the scientometric indicators for various discrete and continuous citation-paper rank distributions. Some of the theoretical conclusions are illustrated with examples for time evolution of scientometric indexes during real scientific careers.

Citation-paper rank distributions

Citation-paper rank distribution is defined as the sequence $\{I_c(I_p); I_p = 1, N_p\} : I_c(I_p) \geq I_c(I_p + 1); I_p = 1, N_p - 1$ of citations $I_c(I_p)$ to the paper I_p, where the set of N_p papers has been arranged in descending order to the number of citations gained, *i.e.* most cited placed first. We emphasize that the native, real life distributions (see Fig. 1) are discrete and consist of nonnegative integers. In this study, however, we consider also (as approaches to reality) continuous versions $C(P)$ of the discrete distributions $I_c(I_p)$, as well as discrete model distributions that consist of nonnegative real numbers. The first approach is justified when a large amount of data is analyzed (Fig. 2), while the second one appears in a natural way when approximating integer data with real-valued functions and *vice versa*.

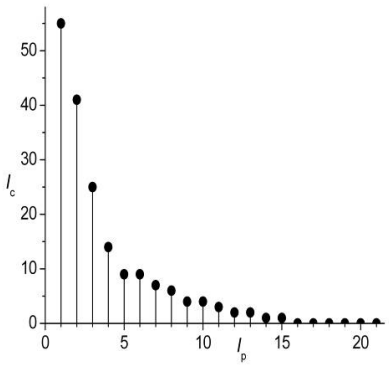

Figure 1. Real life citation-paper rank distribution example

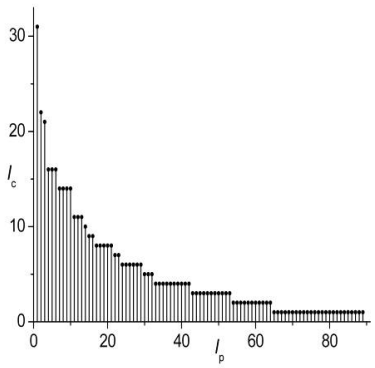

Figure 2. Citation-paper rank distribution example for a large amount of data

In this study, everywhere except explicitly stated, the continuous distributions will be considered as defined on a finite interval of papers $P \in (0, P_m], P_m \geq 1$ and varying between the maximal citation count $C_m = C(0) \geq 0$ and zero (some examples are shown on Fig.3). We analyze a class of distributions such that d^2C / dP^2 does not

change its sign in the interval under consideration. Integration in the continuous case is performed (with one exception) with lower bound equal to zero – one could imagine it as summing the area of 'stripes' (0,1], (1,2],… for the first, second *etc.* papers respectively.

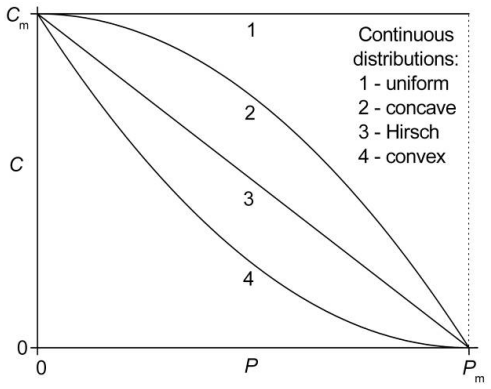

Figure 3. Continuous citation-paper rank distributions (1 – uniform, 2 – concave, 3 – negative slope linear (Hirsch) and 4 – convex). C_m and P_m are citation count of the most cited paper and number of papers, respectively.

Scientometric indicators

The scientometric indicators are considered to be a convenient measure to assess and compare scientific activity, *e.g.* in situations where the use of citation-paper rank distribution is not possible, or reduction of scientometric information is necessary due to time and effort considerations. Scientometric indicators are closely associated with citation-rank distributions. The most widely used are as follows:

- total number of papers N_p (discrete case) or P_m (continuous case);
- total number of citations $N_c = N_c(N_p) \equiv \sum_{Ip=1}^{Np} I_c(I_p)$ or $N_c = N_c(P_m) \equiv \int_0^{Pm} C(P)dP$ for the discrete or continuous case, respectively; it should be noted that (in order to have a distribution) N_c must remain finite, even when considering distributions with *infinite* number of papers, *i.e.* N_p (or P_m) $\to \infty$.;
- average number of citations per paper (N_c / N_p) and average number of citations per year;
- scientometric *indexes*: h -index, g -index, e -index.

Further on we recall the definitions of the scientometric indexes in the way they appear in the original papers and comment some general, more or less distribution-independent properties and relations.

- ***h*-index:**

A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have $\le h$ citations each [Hirsch, 2005], *i.e.* $h : \{I_c(I_p) \ge h$ for $I_p \le h$ and $I_c(I_p) \le h$ for $I_p > h\}$ (discrete case) and $C(h) = h$ (continuous case, Fig. 4). Obviously, h cannot exceed neither N_p (or P_m), nor $I_c(1)$ (or C_m), for discrete (or continuous) distributions. This index has been constructed assuming approximately linear negative-slope citation-paper rank distribution and it equals twice the harmonic mean of impact and productivity. Hirsh's index is the most popular among all other indexes (and the oldest one). The criticism against its use (apart

from the general criticism against scientometrics itself) is due to the fact that usually h ignores a large amount of citations to the first h most cited papers.

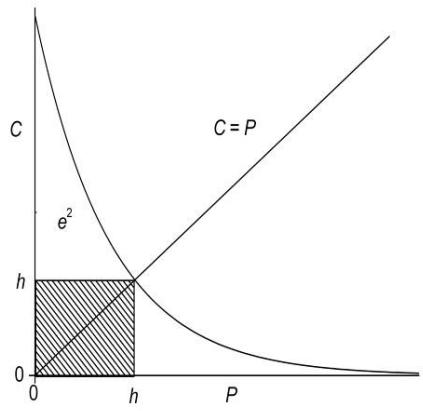

Figure 4. Illustrating Hirsh's h and Zhang's e definitions ([Hirsch, 2005], [Zhang, 2009])

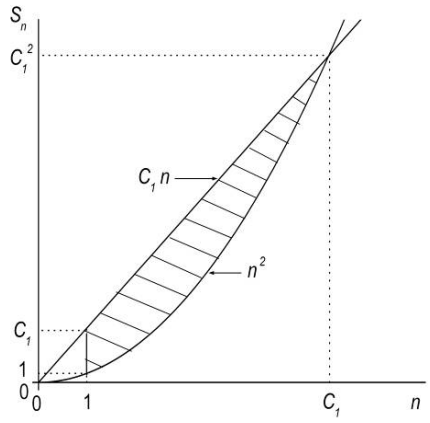

Figure 5. On the existence and uniqueness of g

$$(S_n = \sum_{i=1}^{n} C_i, C_1 \geq C_2 \geq \ldots \geq C_n \geq 0, n \geq 1)$$

- ***g*-index:**

The g -index is introduced as an improvement of the h -index to measure the global citation performance of a set of articles. If this set is ranked in decreasing order of the number of citations that they received, the g -index is the (unique) largest number g such that the top g articles received (together) at least g^2 citations [Egghe, 2006]. Let $N_c(G)$ is the number of citations gained by the first G most cited papers, then $g = \max\{G : N_c(G) \geq G^2\}$. The inequality $g \geq h$ follows immediately from the definition of g ; L. Egghe has also proved its existence and uniqueness for arbitrary citation-paper rank distribution (see also Fig. 5). The g -index is considered to represent the most cited papers better than h does. However, since g is associated with papers in the set, it cannot exceed N_p (or P_m) and remains constant ($g = N_p$ or P_m) if $N_c \geq N_p^2$ or, for a continuous distribution, $N_c \geq P_m^2$. This drawback (illustrated on Fig. 6) has been discussed in [Zhang, 2009], where a possible solution to the problem in the form of introducing virtual papers of zero citation count has been found unacceptable. An additional problem of g is the following: a scientist's saturated (i.e. limited by the number of papers) g -index could be increased by simply publishing (until the saturation level is exceeded) additional papers of mediocre quality, that probably will not be cited at all.

- ***e*-index:**

The *e*-index accounts for the excess citations (represented by e^2) in addition to the h^2 citations of the *h*-core papers [Zhang, 2009]. It is defined as $e^2 = N_c(h) - h^2$ and is free from the constraints on h and g (Fig. 4). The following inequalities take place independently on the citation-paper rank distribution:

$$I_c(g) \leq h \leq g \leq \min(I_c(1), N_p) \text{ or } C(g) \leq h \leq g \leq \min(C_m, P_m), \quad h^2 + e^2 \leq N_c(g),$$

as well as

$$e^2 \geq -\frac{1}{2}h^2(dC/dP)_{P=h} \text{ for a } \textit{convex} \text{ distribution } (d^2C/dP^2 \geq 0),$$

$$e^2 \leq -\frac{1}{2}h^2(dC/dP)_{P=h} \text{ for a } \textit{concave} \text{ distribution } (d^2C/dP^2 \leq 0).$$

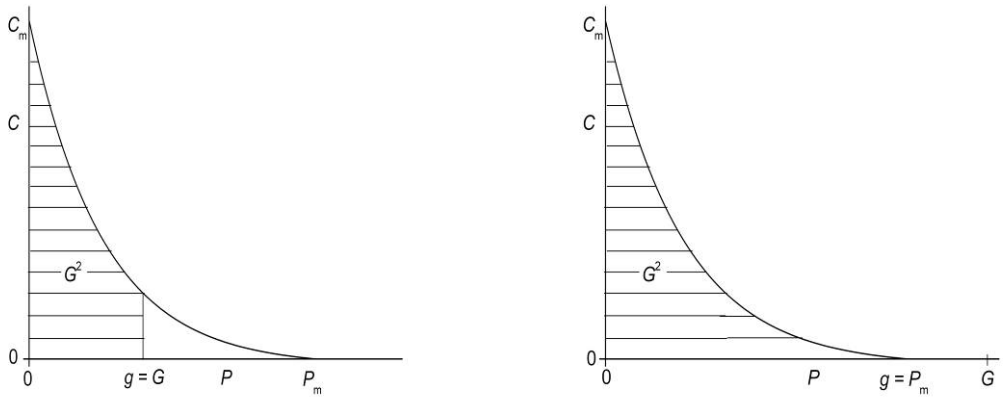

Figure 6. Illustrating Egghe's g index definition:

(left – far from saturation, $g^2 = G^2 \leq N_c \leq P_m^2$; right – saturated, $G^2 \geq N_c \geq P_m^2 = g^2$)

Relations between scientometric indicators (continuous case)

Continuous citation-paper rank distributions are considered as an approach to the real life discrete integer-valued ones. Their advantages include the opportunity to *analytically* compute scientometric indexes and to derive explicitly more or less exact relationships between them. At the same time continuous distributions keep most of the properties and peculiarities of the discrete ones that make them suitable for miscellaneous model studies. Further on we consider two groups of continuous distributions: finite sized ($P_m < \infty$ – uniform, linear negative-slope (or Hirsch), three-parameter polynomial and three-parameter positive exponent power-law distributions) and infinite size ($P_m \to \infty$ – exponential and Pareto distributions).

- **Uniform distribution**

This simple but not quite realistic distribution (Fig. 3, curve 1) reveals clearly the constraints on the h- and g-indexes imposed by the finite number of papers P_m. It is defined as $C(P) = C_m$ for $0 < P \leq P_m$ and

$$N_c = C_m P_m,\ h = \min(C_m, P_m),\ g = h,\ e^2 = (C_m - h)h \tag{1}$$

In this special case h and g coincide and the number of h-core citations is $e^2 + h^2 \geq g^2$. The three indexes h, g and e account for number of citations that represent $r_h = h^2 / N_c$, $r_g = N_c(g) / N_c$ and $r_{h-core} = (h^2 + e^2) / N_c$ parts of all citations N_c, as follows:

$$r_h = \min(s, 1/s),\ r_g = r_{h-core} = \min(1, s),\ s = C_m / P_m. \tag{2}$$

- **Linear negative-slope distribution**

This distribution (Fig. 3, curve 3) has been obtained in [Hirsch, 2005] by assuming constant publication rate (number of papers per year) and constant citation productivity (number of citations per paper per year. It is defined as $C(P) = C_m - sP$ for $0 < P \le P_m$, where the (constant) slope is $s = C_m / P_m$. We have

$$N_c = \frac{1}{2} C_m P_m,\ h = \left(\frac{1}{C_m} + \frac{1}{P_m} \right)^{-1},\ g = \min\left(\left(\frac{1}{C_m} + \frac{1}{2P_m} \right)^{-1}, P_m \right),\ e^2 = \frac{1}{2}(C_m - h)h. \tag{3}$$

The following relations take place:

$$N_c = \frac{(1+s)^2}{2s} h^2, \tag{4}$$

$$g / h = \begin{cases} 2(1+s)/(2+s), & 0 < s \le 2 \\ (1+s)/s, & s \ge 2 \end{cases} \tag{5}$$

$$e^2 = \frac{1}{2} s h^2. \tag{6}$$

Now the ratio g / h reaches its maximum $\frac{3}{2}$ for $s = 2$, where the total number of citations $N_c = g^2 = P_m^2$. Further on, for the relative citation count associated with the indexes one obtains

$$r_h = 2s/(1+s)^2,\ r_{h-core} = s(s+2)/(s+1)^2,\ r_g = \begin{cases} 8s/(2+s)^2, 0 < s \le 2 \\ 1, \quad s \ge 2 \end{cases} \tag{7}$$

Hence h^2 contains no more than 50 percent of all citations (a minimum of $r_h = 0.5$ at $s = 1$). We note that for $s \ge 2$ the g -index remains constant (saturated on a level $g = P_m$) and accounts for <u>all</u> citations.

- **Three-parameter polynomial distribution**

Let us denote $x = P / P_m, y = C / C_m$, then we could consider

$$y(x) = 1 - (1 + \frac{1}{2}\rho)x + \frac{1}{2}\rho x^2 \text{ for } 0 \le x \le 1, 0 \le y \le 1 \tag{8}$$

as a three-parameter polynomial distribution (Fig. 3, curves 2-4), where the third parameter is the constant second derivative $d^2y / dx^2 = \rho$. It covers the linear negative distribution ($\rho = 0$, Fig. 3 - curve 3) as well as convex ($0 < \rho \le 2$) and concave ($-2 \le \rho < 0$) distributions. Both limiting cases ($\rho = 2$, $y = (1-x)^2$) and ($\rho = -2$, $y = 1 - x^2$) are displayed on Fig. 3 (curves 2 and 4, respectively). Now the slope depends on x :

$$dy / dx = -\left(1 + \frac{1}{2}\rho\right) + \rho x,\ (dy/dx)_{x=0} = -1 - \frac{1}{2}\rho,\ (dy/dx)_{x=1} = -1 + \frac{1}{2}\rho. \tag{9}$$

The scientometric indicators are listed as follows:

$$N_c = \frac{1}{2} C_m P_m \left(1 - \frac{1}{6}\rho\right), \tag{10}$$

$$h / P_m = 4 / \left[2\frac{1+s}{s} + \rho + \sqrt{\left(2\frac{1+s}{s} + \rho\right)^2 - 8\rho} \right], \tag{11}$$

$$g = \min(G, P_m),\ G / P_m = 2 / \left[\frac{1}{s} + \frac{1}{2} + \frac{1}{4}\rho + \sqrt{\left(\frac{1}{s} + \frac{1}{2} + \frac{1}{4}\rho \right)^2 - \frac{2}{3}\rho} \right], \quad (12)$$

$$e^2 / h^2 = \frac{1}{3}\left\{ \frac{1}{2} s\left(1 + \frac{1}{2}\rho\right) - 1 + \sqrt{\left[s\left(1 - \frac{1}{2}\rho\right) + 1 \right]^2 + 2s\rho} \right\}, \quad (13)$$

where $s = C_m / P_m$. Egghe's g is saturated (*i.e.* $G \geq P_m$, which corresponds to $N_c \geq P_m^2$) for $s \geq 12 / (6 - \rho)$. The latter inequality implies that for the limiting case of concave distribution ($\rho = -2$) saturation occurs for $s \geq 3 / 2$, while for the limiting convex distribution ($\rho = 2$) saturation takes place for $s \geq 3$. Fig. 7 illustrates the behavior of g / h and e / h for $\rho = 0, \pm 2$. Obviously, Egghe's index g better accounts for the excess of citations in the h-core papers for s below the saturation point (between 1.5 and 3, depending on the second derivative ρ). Above this limit, however, g / h rapidly decreases, while e / h keeps on growing. Table 1 gives a notion of how the scientometric indicators and citation partition for this distribution looks like (since $s = 1$ no saturation of g occurs). We note the robust behavior of e and the large amount of citations accounted by g .

Table 1. Indicators and citation partition for three-parameter polynomial distribution with $C_m = 100$, $P_m = 100$.

	N_c	h	g	e	r_h	r_g	r_{h-core}
$\rho = 2$	3333	38	55	33	0.44	0.91	0.76
$\rho = 0$	5000	50	66	35	0.50	0.89	0.75
$\rho = -2$	6667	61	79	40	0.57	0.94	0.81

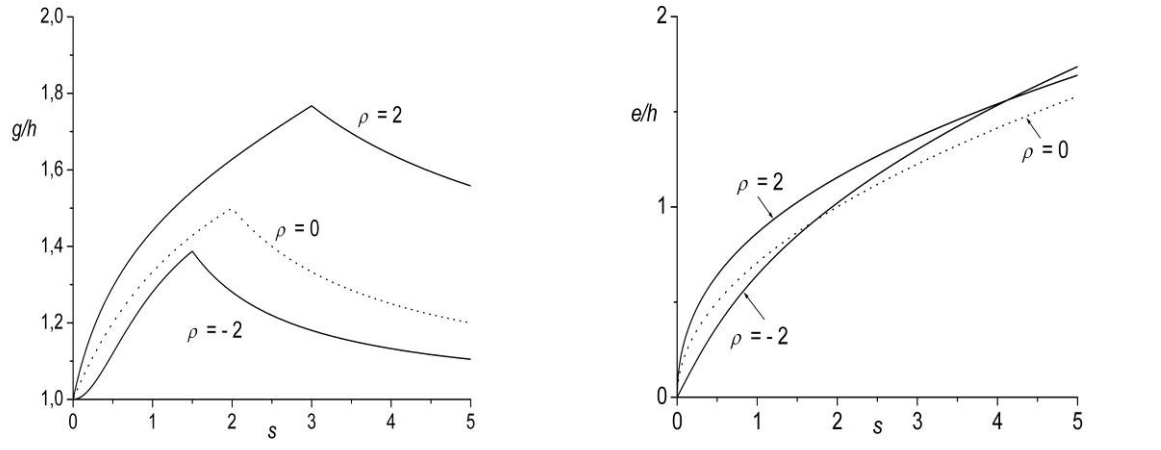

Figure 7. Index ratios g / h (left) and e / h (right) dependence on $s = C_m / P_m$, for the limiting cases of convex ($\rho = 2$), and concave ($\rho = -2$) three parameter polynomial distribution, as well as for the negative slope linear one ($\rho = 0$, dot line)

- **Three-parameter (positive exponent) power-law distribution**

This distribution is defined as $y = 1 - x^{\alpha}$, where $x = P / P_m, y = C / C_m$, $0 < x \le 1, 0 \le y \le 1$ and $\alpha > 0$. It is *convex* for $0 < \alpha < 1$ and *concave* for $\alpha > 1$. Its slope is $dy / dx = -\alpha x^{\alpha-1}$, $(dy / dx)_{x=0} = -\infty, 1$ or 0 for $0 < \alpha < 1, \alpha = 1$ or $\alpha > 1$ respectively and $(dy / dx)_{x=1} = -\alpha$. The total number of citations is $N_c = C_m P_m \alpha / (\alpha + 1)$ and the scientometric indexes are obtained as (unique) solutions to the equations:

$$(h / P_m)^{\alpha} + \frac{1}{s}(h / P_m) - 1 = 0 \tag{14}$$

$$(G / P_m)^{\alpha} + \frac{\alpha + 1}{s}(G / P_m) - (\alpha + 1) = 0, \tag{15}$$

$$\frac{e^2}{h^2} = \frac{\alpha}{\alpha + 1}\left(\frac{C_m}{h} - 1\right), \tag{16}$$

where, as usual, $s = C_m / P_m$ and $g = \min(G, P_m)$. The saturation of g occurs for $s \ge 1 + (1 / \alpha)$. The cases $\alpha = 1$ and $\alpha = 2$ reproduce the linear (negative slope) and the $\rho = 2$ concave distributions considered previously in the paper. Equations (14-16) can be explicitly solved for a *convex* distribution with $\alpha = 1 / 2$ (Fig. 8):

$$\frac{h}{P_m} = \frac{4}{\left(1 + \sqrt{1 + (4 / s)}\right)^2}, \quad \frac{G}{P_m} = \frac{9}{\left(1 + \sqrt{1 + (9 / s)}\right)^2}, \quad \frac{e^2}{h^2} = \frac{1}{6}\left(1 + \sqrt{1 + \frac{4}{s}}\right). \tag{17}$$

For this particular case we have $g = G$ for $s \le 3$ and $g = P_m$ above this limit.

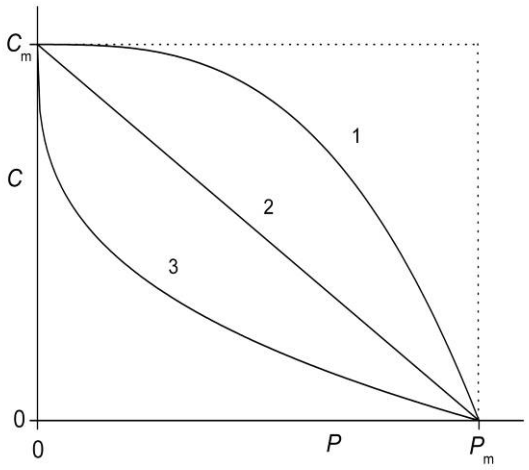

Figure 8. Three-parameter (positive exponent) power-law distribution (1 – concave with $\alpha = 2$, 2 – negative slope linear, 3 – convex with $\alpha = 1 / 2$)

Further on we consider two examples of continuous citation-paper rank distributions with infinite number of papers ($P_m \to \infty$). There are several peculiarities associated with these distributions. The obvious one is absence of g-index saturation for arbitrary large maximal number of citations C_m. This leads to an important inequality, $e^2 + h^2 \le g^2$ and consequently, $e \le g$. Another peculiarity is the fact that, paradoxically, sometimes

the maximal number of citations appears to be greater than their total number ($C_m > N_c$). However, due to some of their properties (of interest for deeper studies, *cf.* [Egghe, 2005]) we give results that might be compared with those of other distributions.

- **Exponential distribution**

is defined as $C(P) = C_m \exp(-\beta P)$, for $0 \le P < \infty$ and $\beta > 0$. Let us introduce $\underline{C} = \beta C$, $\underline{C}_m = \beta C_m$, $\underline{P} = \beta P$, $\underline{h} = \beta h$, $\underline{g} = \beta g$ and $\underline{e} = \beta e$, then we have $\underline{C}(\underline{P}) = \underline{C}_m \exp(-\underline{P})$. The total number of citations is $N_c = C_m / \beta = \underline{C}_m / \beta^2$ and the re-scaled scientometric indexes are obtained as solutions to:

$$\underline{h} \exp(\underline{h}) = \underline{C}_m,\ \underline{g}^2 / [1 - \exp(-\underline{g})] = \underline{C}_m,\ \underline{e}^2 = \underline{C}_m - \underline{h}(\underline{h} + 1). \quad (18)$$

As it could be seen from (Fig. 9), $\underline{g}$ and $\underline{e}$ lay close to each other; both of them are much greater than $\underline{h}$ and hence, represent better the effect of most cited papers than $\underline{h}$ does. Note that $N_c < C_m$ for $\beta > 1$.

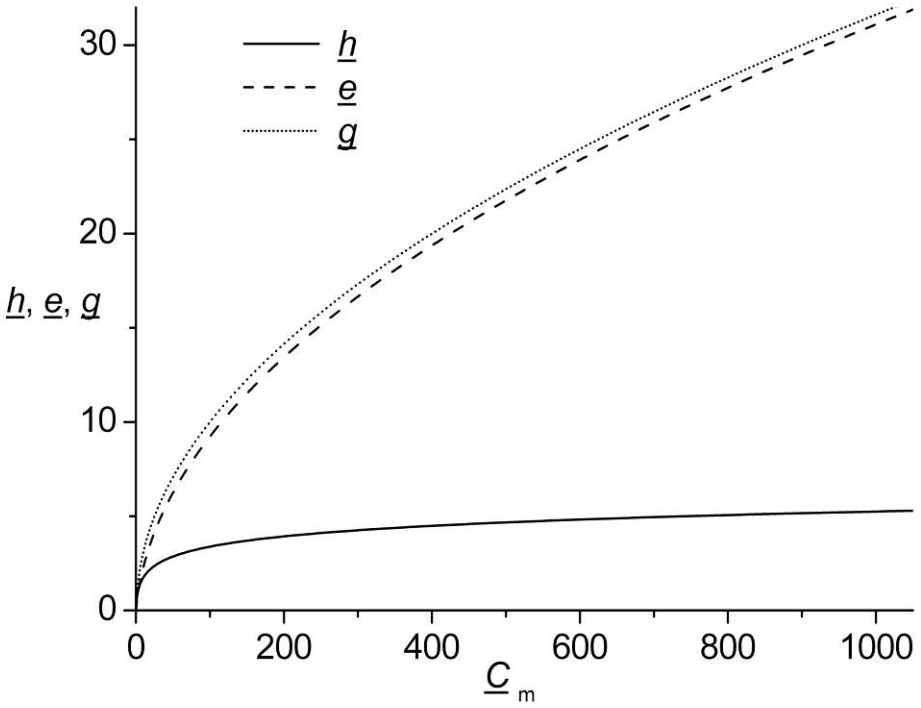

Figure 9. Re-scaled indexes $\underline{h} = \beta h$, $\underline{g} = \beta g$, $\underline{e} = \beta e$ versus re-scaled maximum citation count $\underline{C}_m = \beta C_m$.

- **Pareto distribution**

The convex distribution $C(P) = C_m P^{-\alpha}, 1 \le P < \infty, \alpha > 1$ (introduced by *Alfredo Pareto* (1848-1923) for other purposes) is one of the most commonly used in scientific fields as informetrics, scientometrics and other 'metrics'. It is *scale-free* and has *the product property* [Egghe, 2005]. The total citation count is $N_c = C_m / (\alpha - 1)$, i.e. $N_c < C_m$ for $\alpha > 2$. The scientometric indexes h and e are

$$h = C_m^{1/(\alpha+1)}, \tag{19}$$

$$e^2 = (h^{1+\alpha} - \alpha h^2)/(\alpha - 1), \tag{20}$$

while g is obtained as a solution to

$$g^2 - N_c(1 - g^{1-\alpha}) = 0. \tag{21}$$

Let $\alpha = 3$, then we have $N_c = \frac{1}{2}C_m$, $h = (C_m)^{1/4}$ and

$$g = \frac{1}{2}\sqrt{C_m\left(1 + \sqrt{1 - 8/C_m}\right)} \text{ for } C_m \geq 8,\ e^2 = \frac{1}{2}\sqrt{C_m}(\sqrt{C_m} - 3) \text{ for } C_m \geq 9. \tag{22}$$

For $C_m \to \infty$ e asymptotically approaches g:

$$g^2 / e^2 = 1 + 3C_m^{-1/2} + 7C_m^{-1} - 6C_m^{-3/2} + O(C_m^{-2}),\ C_m >> 1. \tag{23}$$

There is some concern about use of this distribution in analyzing ranked scientometric data (*cf.* [Atanassov and Detcheva, 2012]), mainly associated with the fact that in most real life cases Pareto exponent α is close to (and many times less than) unity. We also note that $\alpha = 1$ corresponds to Lotka's exponent 2 and fractal dimension 1 [Egghe, 2005].

Relations between scientometric indicators (discrete case)

At a first glance these distributions should better describe the real life citation-paper rank histograms. This is probably true, but one should bear in mind that although the *argument* is a positive integer, the *function* (i.e. the distribution) itself is (generally) a positive real number. This inconvenience is usually overcome by considering the nearest integer part of the result *(cf. e.g.* [Clauset *et al*, 2009]). Therefore, in most cases further on we derive relationships that are in this sense only approximately true (denoted here with '$\approx$').

Although all of the continuous distributions addressed in the previous section have their discrete representations, we restrict ourselves with considering two discrete Pareto distributions, resulting from the continuous one, however, better suited for scientometric data analysis.

- **Zeta distribution**

This distribution is defined for all positive integers:

$$I_c(I_p) \approx I_{cm} I^{-\alpha},\ I_p = 1, 2, \ldots,\ \alpha > 1. \tag{24}$$

The total number of citations is

$$N_c \approx I_{cm}\zeta(\alpha), \tag{25}$$

where $\zeta(\alpha)$ is the Riemann zeta function. By definition, for h we have

$$I_c(h+1) - 1 \leq h \leq I_c(h). \tag{26}$$

or

$$1 \leq \frac{I_{cm}}{h^{\alpha+1}} \leq \left(1 + \frac{1}{h}\right)^{\alpha}. \tag{27}$$

Hence (for $h > 2\alpha$)

$$I_{cm} \approx h^{\alpha+1},\ h \approx (I_{cm})^{\frac{1}{\alpha+1}}, \tag{28}$$

(*cf.* [Egghe and Rousseau, 2006]). Under these assumptions, we can estimate h by knowing I_{cm} and *vice versa*. Since the distribution is defined for an infinite sized set of positive integers we have no saturation effects for the g -index; the latter is obtained as a solution to the equation:

$$g^2 / S(\alpha, g) \approx h^{\alpha+1} \approx I_{cm}, \tag{29}$$

where

$$S(\alpha, N) = \sum_{I=1}^{N} I^{-\alpha} \tag{30}$$

is the *incomplete* Riemann zeta function (Figs. 10 and 11). Further on, for e^2 one obtains

$$e^2 \approx h^2 \left[S(\alpha, h) h^{\alpha-1} - 1 \right]. \tag{31}$$

The dependence of Hirsh's h, Egghe's g and Zhang's e on the maximal citation count I_{cm} for a zeta distribution with power exponent $\alpha = 1.1$ is demonstrated on Fig. 12.

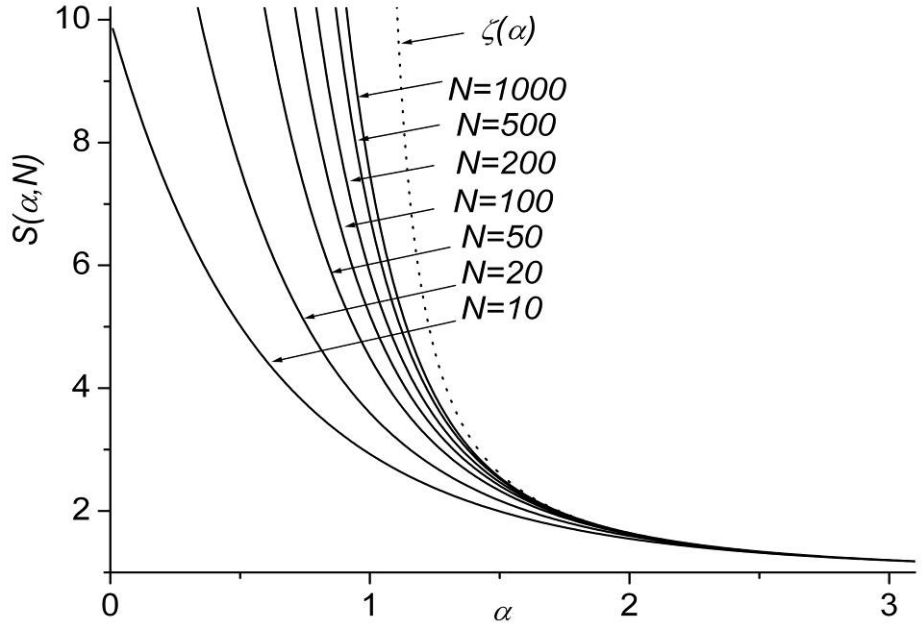

Figure 10. The incomplete zeta-function $S(\alpha, N) = \sum_{I=1}^{N} I^{-\alpha}$ *versus* power exponent α for various N.

Similarly to the continuous case, this *Pareto*-type distribution (defined on *infinite* set of positive integers) has severe problems when the power exponent α approaches, or falls down below unity (*cf.* [Atanassov and Detcheva, 2012]). A possible solution to the problem is the use of

- **Zipf distribution**

named after (the famous law of) *George Kingsley Zipf* (1902-1950), defined for a *finite* set of positive integers:

$$I_c(I_p) \approx I_{cm} I_p^{-\alpha}, \quad I_p = 1, 2, \ldots, N_p, \quad \alpha \geq 0, \tag{32}$$

where N_p is the total number of papers and the total number of citations is

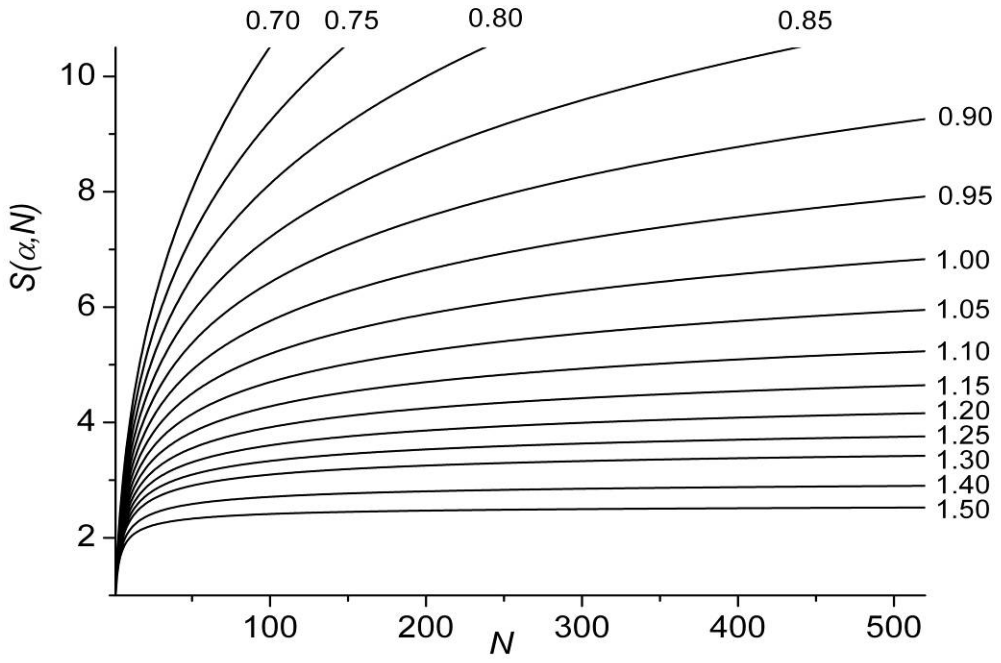

Figure 11. The incomplete zeta-function $S(\alpha, N) = \sum_{I=1}^{N} I^{-\alpha}$ *versus* N for various power exponents α.

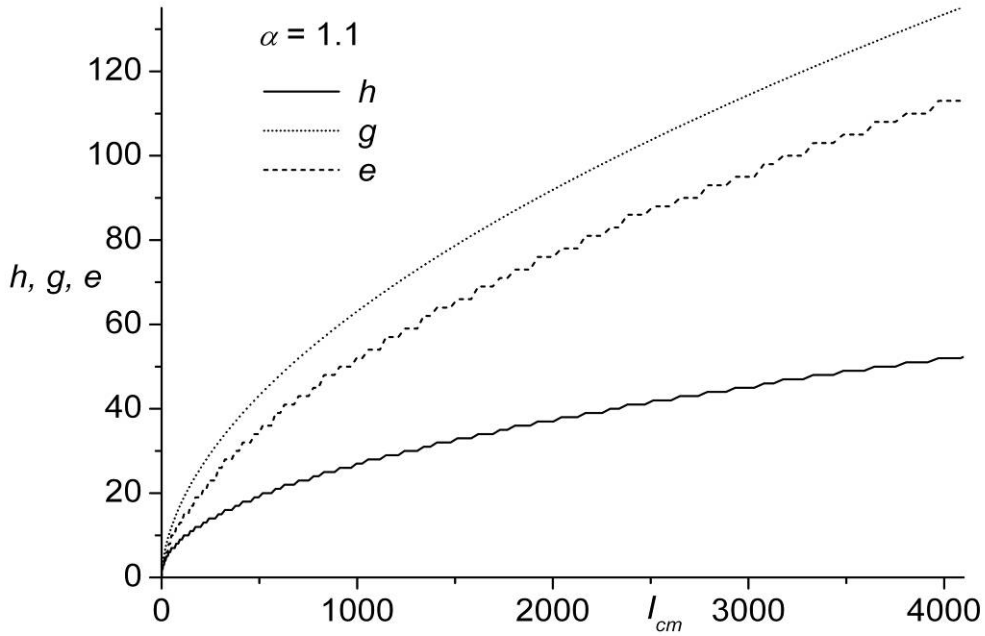

Figure 12. Dependence of h, g and e on maximal citation count I_{cm} for a zeta distribution with $\alpha = 1.1$.

$$N_c \approx I_{cm} S(\alpha, N_p)\,. \tag{33}$$

Now we have saturation effects for both h:

$$h = \min(H, N_p)\,,\ H \approx I_{cm}^{\frac{1}{\alpha+1}}\ , \tag{34}$$

and g:

$$g = \min(G, N_p)\ , G^2 \,/\, S(\alpha, G) \approx I_{cm}\,, \tag{35}$$

while the expression for e^2 (Eq. 31) remains unchanged. This distribution fits quite well many of the citation-paper rank histograms ([Atanassov and Detcheva, 2012], [Atanassov, 2012]); however, one must pay for this with introducing a second parameter, namely the total number of papers N_p. The good news is that, for large enough N and α greater than (and not very close to) unity, $S(\alpha, N)$ is a slowly varying function of N (Fig. 11). Note that in the limit $N \to \infty$ we have $S(\alpha, N) \to \zeta(\alpha)$; for α close to unity this could mean $N > 10^6$ and more. Bearing in mind that the total number of scientific sources that appeared in the whole world history nowadays hardly exceeds several tens of millions, one should be cautious when using zeta distributions, in particular, with lower power exponents. In addition, bearing in mind our *nearest integer* convention one may ask himself what happens when $I_c(I_p) < 1/2$. The answer is (*cf.* [Atanassov and Detcheva, 2012]), that this zeta distribution is indistinguishable from a Zipf 's one with $N_p \approx (2I_{cm})^{1/\alpha}$. In most cases this limit does not exceed 10^3-10^4.

- **Kronecker-type discrete distribution**

might be considered as limiting case of Zipf or zeta distributions with power exponents $\alpha \to \infty$. It is of interest for analyzing cases where a single paper has been highly cited compared to all others; this situation is far not as exotic as it seems (*cf.* next section and the examples in. [Zhang, 2009]). It is defined as $I_c(I_p) = I_{cm}$ for $I_p = 1$ and zero elsewhere. The scientometric indicators are easily obtained to be $N_c = I_{cm}, h = 1$, $g = \min(1, G) = 1$, where $G = \sqrt{N_c} = \sqrt{I_{cm}}$ and $e = \sqrt{G^2 - 1} \approx G - \left(2G\right)^{-1}$ for $I_{cm} >> 1$. We see that a severe loss of citations occurs in the determination of h and g, while e accounts for all of them, as suggested in [Zhang, 2009].

Time evolution examples

One of the advantages in using scientometric indexes is the opportunity to represent in a clear and concise way the scientific activity of a scholar during long periods of his/her academic career. We have chosen two examples to illustrate the main results of the analysis in the previous sections. Both of them are characterized with a Zipf – like citations-paper rank distributions, however, with different power exponents.

The first example (Fig. 13) represents time evolution of the scientometric indexes h, g, e for a twenty-year period of scientific activity in the field of *photonics* (more details can be found in [Atanassov, 2012]). The power exponent α varies gently between 0.8 and 1.0. Due to the high productivity and hence, large number of published papers no saturation effects on g can be observed and this index accounts for the excess of citations in the h-core papers significantly better than e does. This behavior approximately follows the relations between the scientometric indexes for a zeta-distribution with $\alpha = 1.1$ illustrated in Fig. 12.

The second example (Fig. 14) illustrates the effect of saturation on g. This is a case study for a situation where

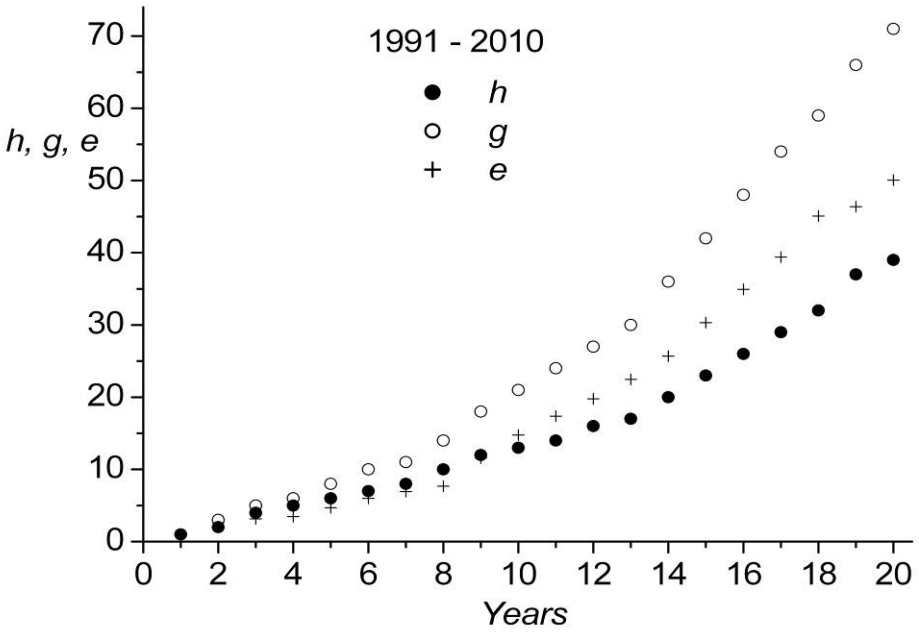

Figure 13. Time evolution of scientometric indexes (case study 1): no saturation of g is observed.

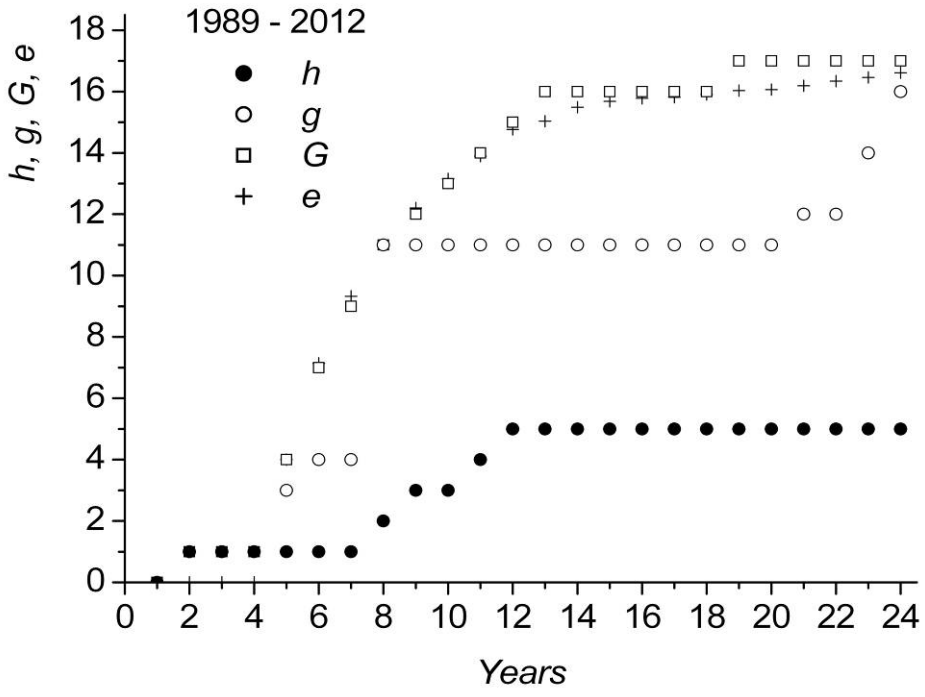

Figure 14. Time evolution of scientometric indicators (case study 2): significant saturation of g takes place.

a scientist has one highly cited paper while the rest of his/her papers is far not so popular. In this real life case g is limited by the number of publications N_p and is noticeably smaller than G, computed by solving Egghe's

relation $G^2 = N_c(G)$. Thus G is approximately equal to Zhang's e – a situation typical for the Kronecker-type distribution considered previously. Similar examples have encouraged Zhang to introduce his e -index. Note that the increase in g observed for $Years$ varying between 20 and 24 is due to papers with negligible citation count.

Summary and conclusions

In this paper we have obtained relations between scientometric indicators like total number of citations, Hirsch's, Egghe's and Zhang's indexes for various model citation-paper rank distributions in continuous and discrete representations. The theoretical considerations have been illustrated with two examples for time evolution of these indicators during real scientific careers.

Our main conclusions are summarized as follows:

- the Hirsch index, compromising between productivity and impact at the same time ignores a considerable amount of citations to the (highly cited) papers; this effect is stronger for distributions of convex type (in particular, the continuous exponential or Pareto distributions) and/or where the ratio of maximum citation count (number of citations gained by the most cited paper) and number of publications significantly deviates from unity;
- a quite general drawback of the *h*-index is its relatively strong dependence on the distribution shape that in the real life could result in statistical instability;
- Egghe's *g*-index , as a true integral characteristic seems to be statistically stable; up to a certain limit it accounts best for the highly cited papers, compared with the other two indexes; however, above this limit (where the total number of citations exceeds the square of the total number of papers) the *g*-index reaches its maximum value (equal to the total number of papers); in such regime of *saturation* the index accounts for all of the citations; it would grow only if the total number of papers (even of zero citation count) is increased; therefore, our conclusion is, that the *g*-index performs best below, and has severe problems above the saturation limit;
- a saturation of the *g*-index takes place for a finite number of papers only; the paper count where saturation occurs for convex distributions is greater than that for concave ones;
- Zhang's *e*-index appears to be quite robust with respect to the distribution shape; however, it accounts for less citations than Egghe's *g*-index (below saturation) does;

Since the *g*-saturation occurrence might be easily overlooked, care is needed when computing and comparing this index. In conclusion, we believe that such model studies could prove useful in analyzing various aspects of scientific activity assessment, in particular, the self-citations effect on the scientometric indexes which seems to be an appropriate topic for future studies.

Acknowledgments

This paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Alonso et al, 2009] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma and F. Herrera. h-index: a review focused in its variants, computation and standardization for different scientific fields, Journal of Informetrics 3 273-289 (2009)

[Atanassov and Detcheva, 2012] V. Atanassov, E. Detcheva. Theoretical analysis of empirical relationships for Pareto-distributed scientometric data, Int. J. Information Models and Analyses 1(3) 271-282 (2012)

[Atanassov, 2012] V. Atanassov. Time evolution of scitation-paper rank distributions and its implications for scientometric models, presented at "Evaluating Science: Modern Scientometric Methods" Conference Sofia May 21-22, 2012,COST Action "Physics of Competition and Conflicts, https://sites.google.com/site/scientometrics2012sofia/presentations (2012)

[Bornmann et al, 2010] L. Bornmann, R. Mutz and H.-D. Daniel. The h-index research output measurement: two approaches to enhance its accuracy, Journal of Informetrics 4 407-414 (2010)

[Cabrerizo et al, 2010] F. J. Cabrerizo, S. Alonso, E. Herrera-Viedma and F. Herrera. q2 –index: quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core, Journal of Informetrics 4 23-28 (2010)

[Clauset et al, 2009] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, J. SIAM Review 51 (4) 661-703 (2009)

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

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

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

[Ferrara and Romero, 2012] E. Ferrara and A. Romero. A scientific impact measure to discount self-citations, submitted to J. Am. Soc. Information Sci and Technology (2012)

[Harzing, 2012] A.-W. Harzing. Publish or Perish v. 3.6.449, http://www.harzing.com (2012)

[Hirsch, 2005] J.E. Hirsch. An index to quantify an individual's scientific research output, Proc. Nat. Acad. Sci. 102 (46) 16569-16572 (2005)

[Hirsch, 2007] J.E. Hirsch. Does the h index have predictive power?, Proc. Nat. Acad. Sci. 104 (49) 19193-19198 (2007)

[Iglesias and Pecharroman, 2007] J. E. Iglesias and C. Percharroman. Scaling the h-index for different ISI fields, Scientometrics 73 (3) 303-320 (2007)

[Schreiber, 2007] M. Schreiber. Self-citation corrections for the Hirsch index, Europhysics Lett. 78 (3) 30002 (2007)

[Schreiber, 2008] M. Schreiber. A modification of the h-index: the h(m)-index accounts for multi-authored manuscripts, Journal of Informetrics 2 (3) 211-216 (2008)

[Schreiber, 2010] M. Schreiber. Twenty Hirsch index variants and other indicators giving more or less preference to highly cited papers, arXiv:1005.5227v1 [physics.soc-ph] (2010)

[Zhang, 2009] C.-T. Zhang. The e-index, complementing the h-index for excess citations, PloS ONE 4(5) e5429 (2009)

Authors' Information

Vladimir Atanassov *– Institute of Electronics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria; e-mail: v.atanassov@abv.bg*

Major Fields of Scientific Research: Plasma Physics and Gas Discharges, Radars and Ocean Waves, Nonlinearity and Chaos, Scientometrics

Ekaterina Detcheva *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,1113 Sofia, Bulgaria; e-mail: detcheva@math.bas.bg*

Major Fields of Scientific Research: Web-based applications, Image processing, analysis and classification, Knowledge representation, Business applications, Applications in Medicine and Biology, Applications in Psychology and Special Education, Computer Algebra.