SELF-CITATIONS EFFECT ON SCIENTOMETRIC INDEXES

Vladimir Atanassov, Ekaterina Detcheva

Abstract: *Scientometric studies on self-citations reveal variety of aspects, like attempted fraud approaching a crime, a self-advertising tool, a standard scientific publication practice that saves space and time by reducing repetitions, and so on up to self-citations being an important element of scientific networking. It seems, however, that self-citation analysis will remain a hot topic for a long time mainly due to its effect on assessment of scientific impact. Therefore, an obvious question that arises is, to what extent self-citations could modify the basic scientometric indicators and indexes, used to perform such assessment.*

This paper represents an attempt to quantitatively estimate the effect of self-citations on several widely used scientometric indexes (Hirsch's h, Egghe's g and Zhang's e). This has been achieved by incorporating self-citations into various explicitly given citation-paper rank distributions under more or less realistic assumptions. The latter are deduced making use of well known empirical relationships, based on analyzing a considerable amount of scientometric data. The results obtained contain self-citation corrections for the scientometric indexes, as well as some indications for a 'normal' and 'extraordinary' self-citation behavior.

Keywords: *self-citations, citation-paper rank distributions, scientometric indexes, h-index, g-index, e-index, empirical relationships, scientometric 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

Studies on self-citations are an important part of the scientific activity oriented citation analysis. Pros and cons have been reported since many years ([Asknes, 2003], [Glänzel et al, 2004], [Glänzel, 2008], [Costas et al, 2010]). In particular, self-citation considered as a part of referencing to some knowledge that has been already established, thus *reducing repetitions*, is an essential paradigm for the whole science and for a huge part of the human activity as well. Self-citations are also considered as a part of knowledge distribution, and a powerful self-advertising tool, in particular for the Knowledge Markets [Markov et al, 2013]. Self-citations appear to be a substantial part of establishing *scientific networks* [Ausloos et al, 2008]. It seems, however, that self-citation analysis will remain a hot topic for a long time mainly due to the effect of self-citations on scientific activity assessment. This reveals mainly the backside of self-citations, starting with the somewhat naïve idea to pull oneself out of the swamp holding one's own hair just like the well known baron did, by inflating the citation count, up to intentionally manipulating the scientometric indexes ([Asknes, 2003], [Markov et al, 2013], [Bartneck and Kokkelmans, 2011]).

One cannot leave aside also the psychological aspects of self-citing. The almost absolute lack of self-citations over a longer period is just as pathological as an always-overwhelming share [[Glänzel et al, 2004]. A small number of self-citations could be associated with quick change of research fields, jumping from one theme to another and perhaps frivolous attitude towards the scientific problems in general. Large number of self-citations accompanied by a relatively small number of external citations could indicate severe communication problems of an introverted scientist, closed in its shell and living outside the scientific community, producing his/her own *parallel* and *self-sufficient* science. However, it could be also due to a habit of numerous coauthors that form a self-citing circle. Whatever the reason could be, it is worth noting that the real problem for a scientist is the *lack* of external citations rather than the *excess* of self-citations.

The aim of this paper is to obtain quantitative estimates for the effect of self-citations on the three most widely used scientometric indicators, namely Hirsch's *h*-index [Hirsch, 2005], Egghe's *g*-index [Egghe, 2006] and Zhang's *e*-index [Zhang, 2009]. There are already several case studies involved in this topic, that provide a good empirical base for the theory, *e.g.* on the self-citation corrections to Hirsch's ([Schreiber, 2007], [Engqvist and Frommen, 2008], [Ferrara and Romero, 2013]) and Egghe's [Schreiber, 2008] indexes. The paper is organized as follows: in the first section we consider a (single author) discrete self-citation model and suggest a notion for 'normal' *vs.* 'abnormal' (throughout this paper we call it 'extraordinary') self-citation behavior. This concept has been used to construct appropriate models for self-citation corrections to the citation-paper rank distributions. The next two sections are devoted to estimating the effect of the (almost) additive and the multiplicative corrections on the three scientometric indexes. The analytics has been illustrated in a separate section with the results of two case studies on self-citations effect on Hirsch's index, where some comments on possible applications have been made, too.

The discrete model for normal and 'extraordinary' self-citation behavior

Let us consider a sequence of papers $\{P_1, P_2, \ldots, P_N\}$ arranged in order of their appearance (the earliest first). For a *single-authored* papers P_2 may contain a reference to P_1, P_3 to P_2 and P_1 *etc*. The number of self-citations a single author produces in N papers, provided he/she cites *all* preceding papers, is

$$N_{sc}\left(N\right) = \frac{1}{2} N\left(N-1\right). \tag{1}$$

These self-citations are *linearly* distributed to the paper count I:

$$C_{sc}\left(I\right) = N - I,\ I = 1, 2, \ldots, N\ . \tag{2}$$

Eq. (2) specifies an absolute limit for the self-citations of a *single author*. If, however, the author cites (no more than) K of his/her last published papers, we have a *linear* dependence of self-citations number on the total number of papers N:

$$N_{sc}(K) = K\left[N - \frac{1}{2}\left(K+1\right)\right], \tag{3}$$

and (almost) *uniformly* distributed self-citations:

$$C_{sc}\left(K;I\right) = K \text{ for } 1 \le I \le N-K\ ,\ C_{sc}\left(K;I\right) = N - I \text{ for } N-K+1 \le I \le N\ . \tag{4}$$

These relations allow us to distinguish between two types of self-citation behavior (of a single author) provided that K is much smaller than N. It is quite normal that scientists support their publications by citing *some* of their newest own papers ($N_{sc} \sim N$, $C_{sc}\left(I\right) \approx K = const$), while citing *almost all* own previous papers ($N_{sc} \sim N^2$, $C_{sc}\left(I\right) \approx N - I$) demonstrates a rather extraordinary self-citation practice (Fig. 1). It is worth noting that each type of self-citation behavior smoothly evolves to the other when varying K from much smaller values than (for 'normal') to values close to (for 'extraordinary') the total number of papers N, and *vice versa*.

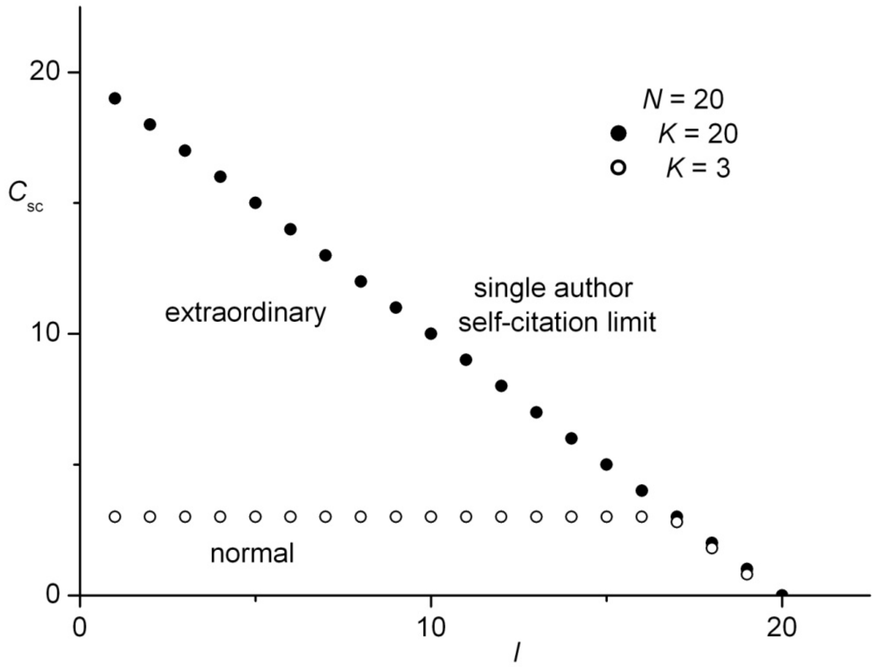

Figure 1. Self-citation distributions for 'normal' and 'extraordinary' self-citation behavior

In order to illustrate these considerations we present results of two case studies (Fig.2a, b and Fig. 3a, b) based on Thomson-Reuters Web of Science scientometric data for two scientists. Citation-paper rank distributions for the *overall* and the *external* citations (*i.e.* with and *without* self citations) are plotted on Fig. 2a and Fig. 3a, while Fig. 2b and Fig. 3b demonstrate the corresponding self-citation distributions. It should be pointed out that papers are ranked separately for each data set; in particular, the self-citation histograms are obtained by computing number of self-citations for each paper, then papers are rearranged to form a rank distribution.

There are several conclusions that could be drawn from this analysis. The 'normal' self-citation behavior (Fig. 2a, b) is characterized with: i) close citation-paper rank distributions and (almost) equal Hirsch indexes for overall and external citations; ii) self-citations data that is approximately uniformly distributed and lies far away from the (single author) self-citation limit. This is in contrast with the 'extraordinary' self-citation practice (Fig. 3a, b), where citations-paper rank distributions and Hirsch indexes with and without counting self-citations are drastically different and the self-citation distribution approaches the linear one of Fig.1. An intuitive support for the (close to) uniform distribution of self-citations over papers and their (relatively) small number could be associated with the observation that self-citation lifetime is *usually* much smaller than the one of the external citations ([Glänzel et al, 2004], [Glänzel, 2008]). There are several explanations for this phenomenon: i) paper's aging gets faster for their

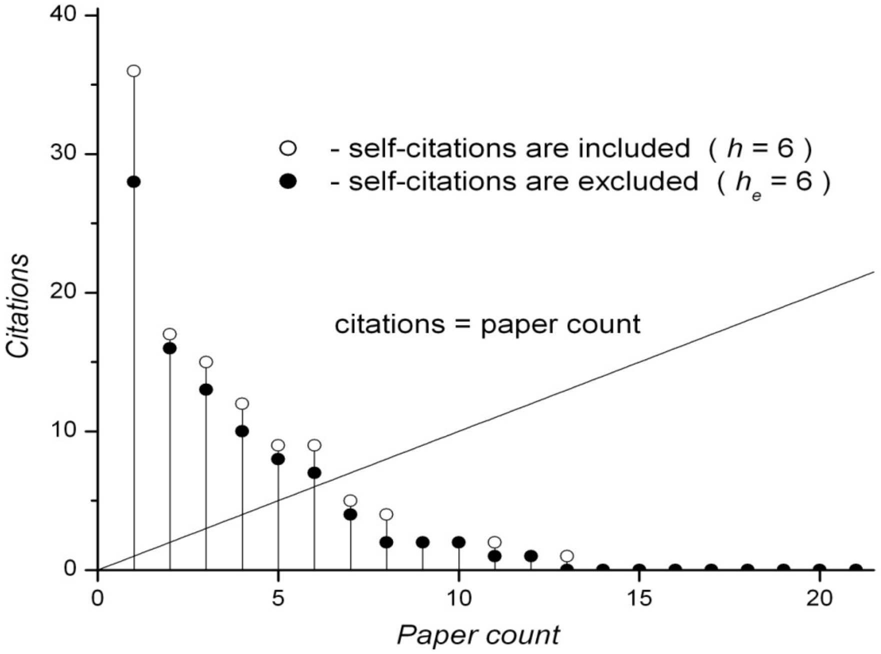

Figure 2a. Citation-paper rank distributions for a case study illustrating 'normal' self-citation behavior

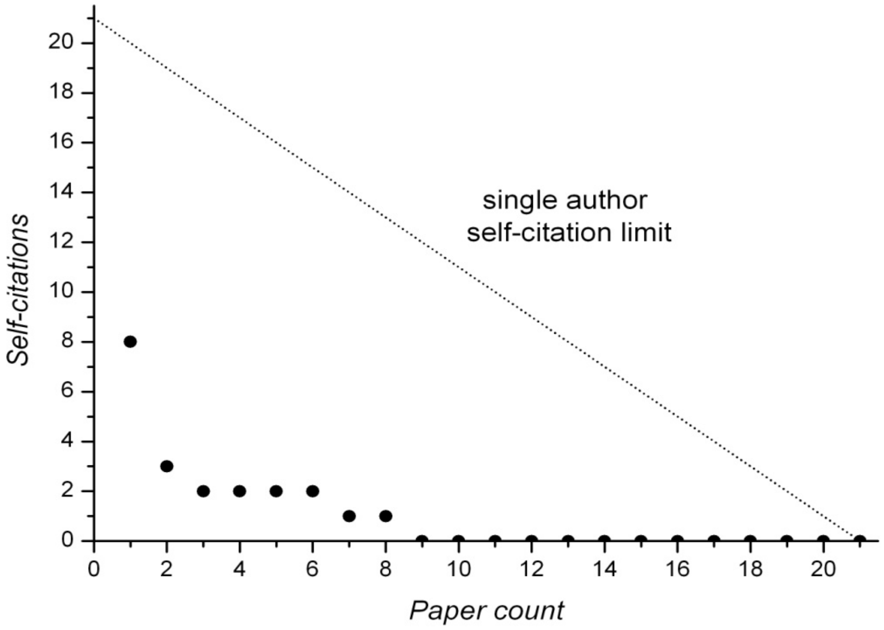

Figure 2b. Self-citation distribution for the case study illustrating 'normal' self-citation behavior

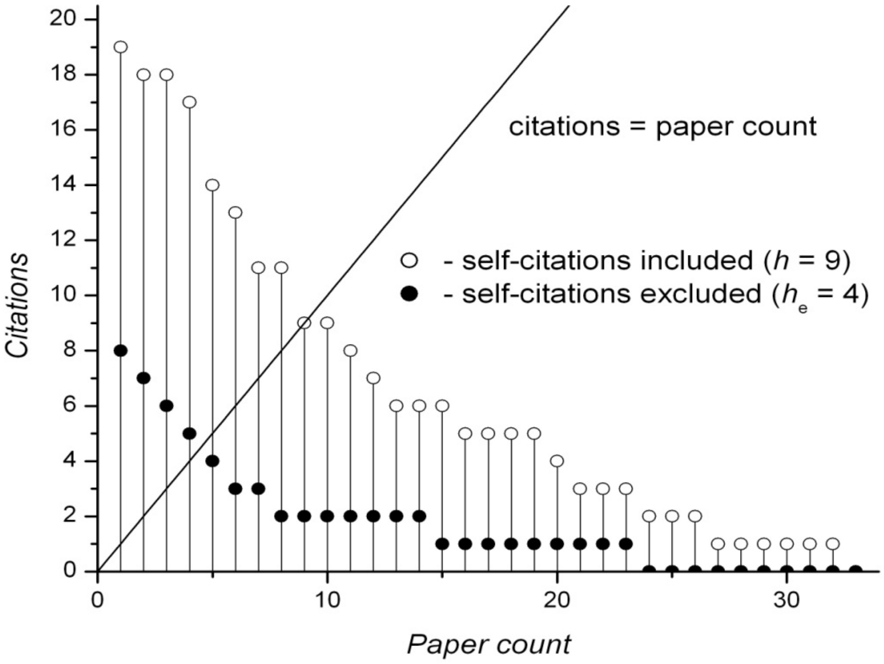

Figure 3a. Citation-paper rank distributions for a case study illustrating 'extraordinary' self-citation behavior

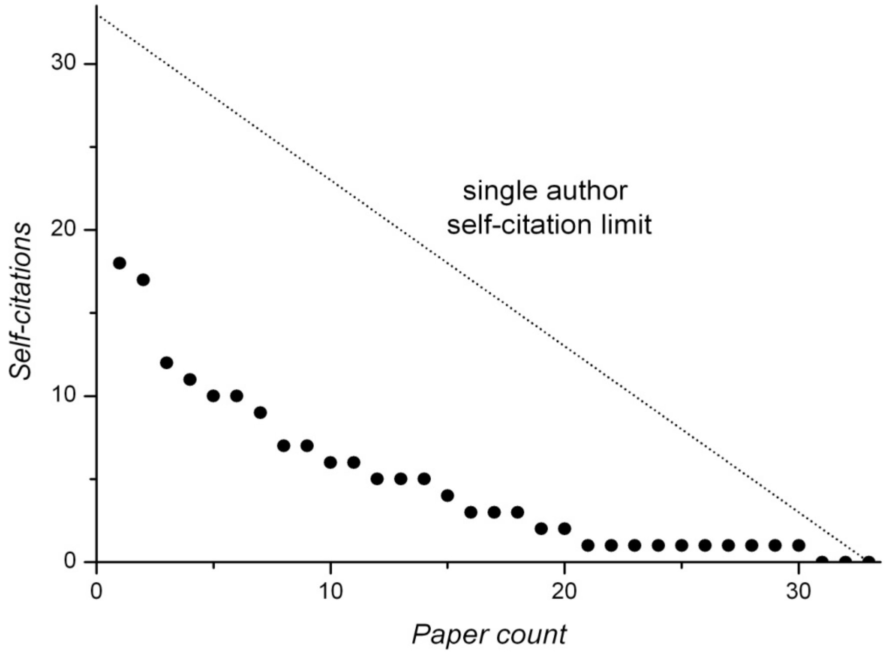

Figure 3b. Self-citation distribution for the case study illustrating 'extraordinary' self-citation behavior

authors than for the other scientists (*e.g.* newcomers in the field); ii) due to the fast and sometimes unexpected developments in science changing research themes and even research fields is rather common for individual scientists and research groups nowadays.

As it can be seen on Figs. 1, 2b and 3b the border between 'normal' and 'extraordinary' self-citation behavior is rather fuzzy. The problem becomes even more complicated, when multiple co-authors are taken into account. Obviously, if a co-authors group acts as one virtual 'author', Eqs. (1) and (2) remain unchanged for each individual member of this group. This situation, however, is far from realistic and the number of self-citations for a given paper could vary from zero to the maximum number of subsequent publications authored by at least one member of the group under consideration. What is known from empirical studies is that multiple authorship boosts the external citations stronger than the self-citations [Asknes, 2003] and the ratio of self-citations to overall citation count for single-authored papers is notably lower than the one for multi-authored papers (no matter how many the co-authors are) [Glänzel and Thijs, 2004b].

Another point that should be addressed is the interdependence between self-citations and external citations. At a first glance, it seems that no clear relation between them exists at all. Glänzel and co-authors have demonstrated, however, that (from statistical point of view) there is nothing arbitrary in this relation and the conditional expectation of self-citation number a paper receives for a given number of external citations depends on the square root of the latter [Glänzel et al, 2004]. It appears that 'the more one cites oneself the more one is cited by the other scholars' [Fowler and Asknes, 2007]. Thus, going a bit ahead, we find some support for a basic assumption used in another part of our analysis, namely considering self-citation distribution as a reduced copy of the overall citation-paper rank one.

Continuous distribution analysis: additive self-citations corrections

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, keeping at the same time most of the properties and peculiarities of the discrete ones. In what follows we extend a previous model study on citation-paper rank distributions and associated scientometric indexes [Atanassov and Detcheva, 2013] to account for self-citations. Our approach consists in computing scientometric indexes for various *continuous* citation-paper rank distributions with and without prescribed self-citation corrections. The latter are specified bearing in mind the considerations concerning the 'normal' and 'extraordinary' self-citation behavior in the previous section. Further on we distinguish between the *overall* citation count C and the *external* (*i.e.* with self-citations excluded) one C_e distributed to (the same) paper ranks P , in descending order to the citations gained (most cited placed first). The paper ranks P are considered within domains $0 \le P \le P_m$ and $0 \le P \le P_{em} \le P_m$ for the overall and external citation distributions, respectively.

By assuming (almost) *uniformly* distributed self-citations one can write down

$$C_e(P) = C(P) - a \ , \ 0 \le P \le P_{em} < P_m , \tag{5}$$

The self-citation *additive* correction level a must obey the inequalities $0 < a < C_m$, where $C_m = C(0)$ is the maximal *overall* citation count (Fig. 4). Obviously, the maximal *external* citation count is $C_{em} = C_m - a$, and P_{em} is obtained as the least of the solutions to $C(P) = a$. Note that the support of the external citation-paper rank distribution remains finite even for $P_m \to \infty$. It should be also emphasized that, in general, the external citation-paper rank distribution $C_e(P)$ does not follow the shape and type of $C(P)$.

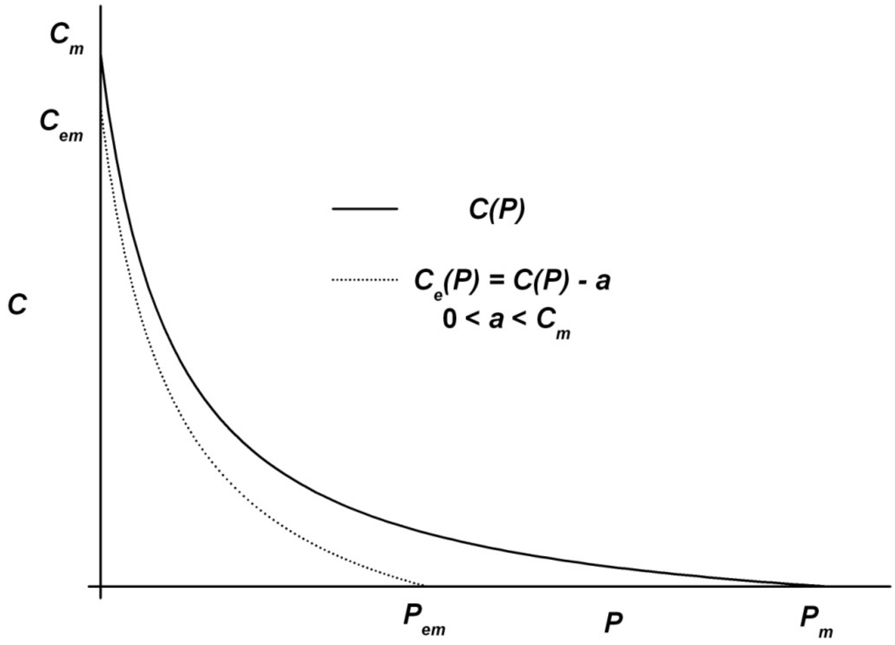

Figure 4. Additive self-citation corrections to citation-paper rank distributions

Let us now consider what happens with Hirsch's index when additive self-citations are taken into account. Denoting *h*-indexes for all citations and external citations with h and h_e, respectively, we have (by definition)

$$C(h) = h\,,\ C_e(h_e) = h_e\,. \tag{6}$$

Further on, by introducing $h_s = h - h_e$ and keeping the first three terms in Taylor series expansion of $C_e(h - h_s)$ one obtains the following quadratic equation for h_s:

$$\frac{1}{2}C''(h)h_s^2 + \left[1 - C'(h)\right]h_s - a = 0\,, \tag{7}$$

where C' and C'' denote first and second derivatives of C, respectively. The only acceptable solution to Eq. (7) is

$$h_s = 2a \,/ \left\{[1 - C'(h)] + \sqrt{[1 - C'(h)]^2 + 2aC''(h)}\right\}, \tag{8}$$

provided that $2aC''(h) \geq -[1 - C'(h)]^2$ (this inequality might fail for extremely *concave* distributions).

By definition, for Zhang's *e*-indexes we have

$$e^2 = \int_0^h C(P)dP - h^2\,,\ e_e^2 = \int_0^{h_e} C_e(P)dP - h_e^2\,. \tag{9}$$

Following the same schema one arrives at

$$e_s^2 = (h - h_s)(a - h_s) - \frac{1}{2}C'(h)h_s^2 + \frac{1}{6}C''(h)h_s^3 . \qquad (10)$$

We treat Egghe's *g*-indexes in the same way, bearing in mind that $g^2 = \int_0^g C(P)dP$ for $g^2 \le \int_0^{P_m} C(P)dP$, otherwise $g = P_m$. Now one obtains a cubic equation for $g_s = g - g_e$:

$$\frac{1}{6}C''(g)g_s^3 + \left[1 - \frac{1}{2}C'(g)\right]g_s^2 + \left[C(g) - 2g - a\right]g_s + ag = 0, \qquad (11)$$

provided that $g^2 \le \int_0^{P_m} C(P)dP$, $g_e^2 \le \int_0^{P_{em}} C_e(P)dP$. If any of these inequalities is not fulfilled, the corresponding *g*-index must be put equal to the corresponding maximal paper count P_m or P_{em}.

We would like to emphasize that the (approximate) relationships (7-8) and (10-11) are *exact* if all C derivatives of order higher than 2 are zero. Such model citation-paper rank distribution appears in [Atanassov and Detcheva, 2013] as the *three-parameter polynomial distribution*.

Further on we consider the self-citation effect on the three scientometric indexes (Hirsch's *h,* Egghe's *g* and Zhang's *e*) for *additive* self-citation corrections to several model citation-paper rank distributions (uniform, linear and Pareto).

- **Uniform distribution:** $C(P) = C_m$, $C_e(P) = C_m - a$, $0 \le P \le P_m$, $0 \le a \le C_m$.

Now $C(P)$, $C_e(P)$ and the self-citation citation-paper rank distribution have the same form. A simple, but somewhat tedious logics yields the following relationships for the self-citation corrections to Hirsch's *h*-index $h_s = h - h_e$, Egghe's *g*-index $g_s = g - g_e$ and Zhang's *e*-index $e_s^2 = e^2 - e_e^2$:

$$h_s = g_s = 0 ,\ e_s^2 = aP_m \text{ for } P_m \le C_m - a,$$

$$h_s = g_s = a + P_m - C_m ,\ e_s^2 = P_m(C_m - P_m) \text{ for } C_m - a \le P_m \le C_m , \qquad (12)$$

$$h_s = g_s = a,\ e_s^2 = 0 \text{ for } P_m \ge C_m .$$

- **Linear negative slope (Hirsch) distribution:** $C(P) = C_m - sP$, supported for $0 \le P \le P_m$, where the distribution slope is $s = C_m / P_m$.

The external citation-paper rank distribution is written as $C_e(P) = C_{em} - sP$ for $0 \le P \le P_{em}$, where $C_{em} = C_m - a$ for $0 \le a \le C_m$ and $P_{em} = C_{em} / s$. The following relationships take place:

$$h_s \equiv h - h_e = a / (1 + s),\ e_s^2 = e^2 - e_e^2 = sa(C_m - \frac{1}{2}a) / (1 + s)^2 . \qquad (13a)$$

Hence for a linear negative slope citation-paper rank distribution the additive self-citation correction to the Hirsch index is constant, depending on the self-citation level a and the slope s only; in particular, it decreases when the distribution gets steeper. This result confirms the intuitively clear notion that a set of a small number of highly cited papers is less sensitive to self-citations than numerous poorly cited ones.

The Egghe's *g*-index correction $g_s = g - g_e$ is

$$g_s = 2a / (2 + s) \text{ for } 0 \le s \le 2 \text{ and } g_s = a / s \text{ for } s > 2 . \qquad (13b)$$

Two different representations of g_s appear in connection with the *g*-saturation phenomenon (see [Zhang, 2009], [Atanassov and Detcheva, 2013]). Note that the self-citation correction to Egghe's index does not depend on the index itself and decreases for steeper distributions, too.

- **Pareto distribution:** $C(P) = C_m P^{-\beta}$, supported for $1 \le P < \infty$ at $\beta > 1$.

For an additive self-citation correction the external citation-paper rank distribution is no more a Pareto one:

$$C_e(P) = C_m P^{-\beta} - a \, , \; 1 \le P \le P_{em} = \left(C_m / a\right)^{\frac{1}{\beta}} , \; 0 < a < C_m \, . \tag{14}$$

Since both $C(P)$ and $C_e(P)$ are supported for $P \ge 1$, the conditions for existence of solutions to Hirsch's index definition equations $C(h) = h$, $C_e(h_e) = h_e$ impose the following inequalities for h, h_e, C_m and a:

$$h \ge 1, \; h_e \ge 1, \; C_m \ge 1, \; a \le C_m - 1.$$

Now Hirsch's index correction $h_s = h - h_e$ satisfies the equation

$$\left(1 - \frac{h_s}{h}\right)^{-\beta} - \left(1 - \frac{h_s}{h}\right) = \frac{a}{h}. \tag{15}$$

An exact solution to this equation at the limit $\beta \to 1 + 0$ is given by:

$$h_s = a \Bigg/ \left[1 + \frac{1}{2}\left(\frac{a}{h}\right) + \sqrt{1 + \frac{1}{4}\left(\frac{a}{h}\right)^2}\,\right]. \tag{16}$$

Eq. (16) describes a smooth transition of h_s from $\frac{1}{2}a$ (at $a/h << 1$) to $h - 1$ (for $a \to C_m - 1 = h^2 - 1$). This transition, akin to the discrete one described by Eq. (4) in the previous section, corresponds to an evolution from 'normal' self-citation behavior (almost uniform self-citation distribution, small compared to and independent on h, corrections to the Hirsch index) to 'extraordinary' one (almost linear self-citation distribution, external Hirsch's index close to 1, h_s linearly increases with h). In particular, for arbitrary $\beta > 1$, assuming that the additive self-citation level is small compared with the Hirsch index, $a/h << 1$, Eq. (16) yields

$$h_s = \frac{a}{1+\beta}\left[1 - \frac{\beta}{2(1+\beta)}\frac{a}{h} + O\left(\frac{a^2}{h^2}\right)\right]. \tag{17}$$

The (exact) self-citation correction to Zhang's *e*-index $e_s^2 = e^2 - e_e^2$ is

$$e_s^2 = \frac{\beta}{\beta - 1}\left[h\left(a - 2h_s\right) - h_s\left(a - h_s\right)\right] - a, \tag{18}$$

which, for $a/h << 1$, is reduced to

$$e_s^2 = a\left(\frac{\beta}{1+\beta}h - 1\right). \tag{19}$$

The *g*-index correction $g_s = g - g_e$ is estimated for $a/g << 1$ as follows:

$$g_s = \frac{(g-1)}{g}\frac{1 - g^{1-\beta}}{\left[2 - (1+\beta)g^{1-\beta}\right]}a\left[1 + O\left(\frac{a}{g}\right)\right]. \tag{20}$$

Thus, for a large enough g the self-citation correction g_s approaches $\frac{1}{2}a$. Note that $g_s = g - \left(C_m / a\right)^{1/\beta}$ must be used instead of Eq. (20) if g_e^2 exceeds the total number of external citations (i.e. a saturation of g_e occurs).

Continuous distribution analysis: multiplicative self-citations corrections

For modeling of self-citation effect on scientometric indexes in the grey zone between 'normal' and 'extraordinary' self-citation behavior it seems appropriate to introduce *multiplicative* self-citation corrections (Fig. 5):

$$C_e\left(P\right) = bC\left(P\right),\ 0 \le P \le P_{em} = P_m \text{, where } 0 < b < 1. \tag{21}$$

Eq. (21) simply states that each paper P has $100b$ percent external citations. Moreover, all three citation-paper rank distributions (overall, external and self-citation ones) have the same form. Following the Taylor expansion procedure, previously used in this paper, we derive a quadratic equation for $h_s = h - h_e$:

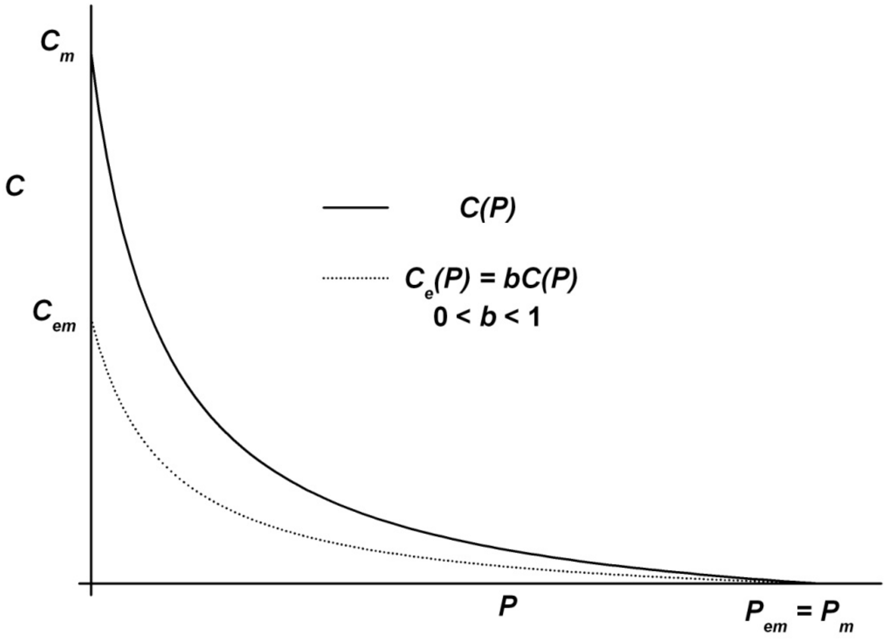

Figure 5. Multiplicative self-citation corrections to citation-paper rank distributions.

$$\frac{1}{2}bC''\left(h\right)h_s^2 + \left[1 - bC'\left(h\right)\right]h_s - (1-b)h = 0\,. \tag{22}$$

The only reasonable solution to Eq. (22) is

$$h_s = 2(1-b)h \,/\, \left\{[1 - bC'(h)] + \sqrt{[1 - bC'(h)]^2 + 2b(1-b)C''(h)}\right\}. \tag{23}$$

The self-citation correction to Zhang's *e*-index is obtained from

$$e_s^2 = e^2 - e_e^2 = (1-b)\left(e^2 + h^2\right) - (2-b)hh_s + [1 + \frac{1}{2}bC'(h)]h_s^2 + \frac{1}{6}bC''(h)h_s^3 . \tag{24}$$

The cubic equation for Egghe's *g*-index self-citation correction $g_s = g - g_e$ now looks as follows:

$$\frac{1}{6}bC''(g)g_s^3 + \left[1 - \frac{1}{2}bC'(g)\right]g_s^2 + \left[bC(g) - 2g\right]g_s + (1-b)g^2 = 0 , \tag{25a}$$

provided that $g^2 \le \int_0^{P_m} C(P)dP$, $g_e^2 \le b\int_0^{P_m} C(P)dP$ (*i.e.* no saturation of *g*-indexes occurs),

$$\frac{1}{6}bC''(P_m)g_s^3 + \left[1 - \frac{1}{2}bC'(P_m)\right]g_s^2 - 2P_m g_s + (1-b)P_m^2 = 0 , \tag{25b}$$

for $g > \int_0^{P_m} C(P)dP$, $g_e^2 \le b\int_0^{P_m} C(P)dP$ and

$$g_s = 0 \tag{25c}$$

for $g > \int_0^{P_m} C(P)dP$, $g_e^2 > b\int_0^{P_m} C(P)dP$ (*i.e.* both *g*-indexes are saturated).

Once again we note that the (generally approximate) relationships (22-25) are *exact* if all C derivatives of order higher than 2 are zero, *i.e.* for the *three-parameter polynomial distribution* studied in [Atanassov and Detcheva, 2013].

In order to better illustrate the self-citation effect on scientometric indexes we consider multiplicative self-citation corrections to several model citation-paper rank distributions (uniform, Hirsch, Pareto).

- **Uniform distribution:** $C(P) = C_m$, $C_e = bC_m$, $0 \le P \le P_m$, $0 \le b \le 1$.

The multiplicative self-citation corrections are obtained by simply replacing a in Eq. (12) with $(1-b)C_m$:

$$\begin{gathered} h_s = g_s = 0 ,\ e_s^2 = (1-b)P_m C_m \text{ for } P_m \le bC_m , \\ h_s = g_s = P_m - bC_m ,\ e_s^2 = P_m(C_m - P_m) \text{ for } bC_m \le P_m \le C_m , \\ h_s = g_s = (1-b)C_m ,\ e_s^2 = 0 \text{ for } P_m \ge C_m . \end{gathered} \tag{26}$$

- **Linear negative slope (Hirsch) distribution:** $C(P) = C_m - sP$, $C_e(P) = b(C - sP)$, both supported for $0 \le P \le P_m$, where $s = C_m / P_m$ and $0 \le b \le 1$.

The relations for the Hirsch's and Zhang's indexes' corrections are as follows

$$h_s \equiv h - h_e = \left[(1-b)/(1+bs)\right]h ,\ e_s^2 = e^2 - e_e^2 = e^2\left\{1 - b^3\left[(1+s)/(1+bs)\right]^2\right\} , \tag{27a}$$

and for Egghe's *g*-index correction $g_s = g - g_e$ one obtains

$$g_s = \frac{2(1-b)}{(2+bs)}g \text{ for } 0 \le s < 2 ,\ g_s = \frac{2-bs}{2+bs}g \text{ for } 2 \le s \le \frac{2}{b} \text{ and } g_s = 0 \text{ for } s > \frac{2}{b} . \tag{27b}$$

- **Pareto distribution:** $C(P) = C_m P^{-\beta}$, supported for $1 \le P < \infty$ at $\beta > 1$.

For a multiplicative self-citation correction the external citation-paper rank distribution remains a Pareto one:

$$C_e(P) = bC_m P^{-\beta} \text{ , } 1 \le P \le P_m \text{ , } 0 < b < 1 \text{ .} \tag{28}$$

Now we have

$$h_s = h\left(1 - b^{1/(1+\beta)}\right), \tag{29}$$

$$e_s^2 = e^2 - e_e^2 = h^2\left[\left(1-b\right)h^{\beta-1} - \beta\left(1 - b^{2/(1+\beta)}\right)\right] / \left(\beta - 1\right) \tag{30}$$

and (for $0 < 1 - b << 1$):

$$g_s = \left(1-b\right)g \,/\, \left\{2 - \left[\left(\beta-1\right) / \left(g^{\beta-1} - 1\right)\right]\right\} . \tag{31}$$

It follows from Eqs. (27), (29) and (31) that the multiplicative self-citation corrections to Hirsch's and Egghe's indexes depend *linearly* on h and g , respectively. For Pareto distributed citations the linear dependence of g_s on g is revealed asymptotically, for a sufficiently large Egghe's index. This situation might appear if a scientist cites more those of his/her papers that gain more external citations. As far as the necessity of such self-citations is rather questionable from a scientific viewpoint, this could be considered as 'extraordinary' self-citation behavior, too.

Examples and applications

In order to illustrate the possible applications of self-citation corrections estimated in the previous two sections we briefly consider the results of two case studies. The first example (Fig. 6) is based on a data set borrowed from M. Schreiber's paper ([Schreiber, 2007], Table 1). The data has been fitted (without much success) to a power-low function. We could, however, clearly distinguish between data points lying close to the line $h_s = h$ (this line corresponds to the self-citation limit on Fig. 1) and other ones that remain constant with increasing h . In this way we could find data suspicious for 'extraordinary' self-citation behavior.

The second example (Fig. 7) represents the self-citation correction $h_s = h - h_e$ to Hirsch's index h for several research groups specialized in the field of high-power gas lasers [Krasteva et al, 2011]. Since co-authorship plays a significant role here, 164 individual scientists have been separated in 9 groups. An approximately linear dependence $h_s(h)$ (hence, an 'extraordinary' citation behavior) could be attributed to groups 2, 4, 7 and 8, while groups 6 and 9 keep the self-citation correction relatively constant and at a low level. An attempt has been also made to fit all data to a power-law function. Although not quite reliable, the result obtained is rather similar to the one presented on Fig. 6. Possible explanation for this coincidence might be that mixing data resulting from several kinds of self-citation behavior causes huge data scatter. In these conditions the fit procedure cannot distinguish between square root and a function varying between linear (at low *h*-values) and a constant (at large *h*-values), as it follows from Eq. (16).

A possible application of the bundle of formulae obtained in the previous two sections is the option to promptly estimate self-citation corrections to scientometric indexes. For this purpose one may assume that the additive self-citation level a is simply the mean self-citations number per paper (i.e. total self-citations number divided by total number of papers). For the multiplicative self-citation approach the self-citation correction factor b represents exactly the ratio of external citation number and the number of all citations. Thus, for $a = 5$ (five self-citations per paper in average) and $\beta = 1.5$, Eq. (17) yields $h_s = 3$, provided that the Hirsch's index obtained by taking into account all citations is much greater than 3. For the same β Eq. (29) states that a scientist that cites 40 percent of his/her own papers (the self-citation world average for natural sciences, [Glänzel and Thijs, 2004a])

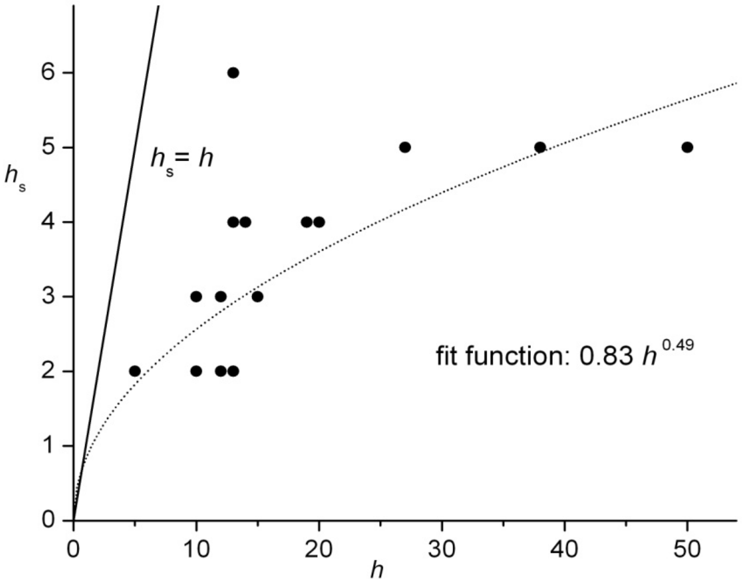

Figure 6. Self-citation correction $h_s = h - h_e$ *vs.* Hirsch's index h (after M. Schreiber's data [Schreiber, 2007])

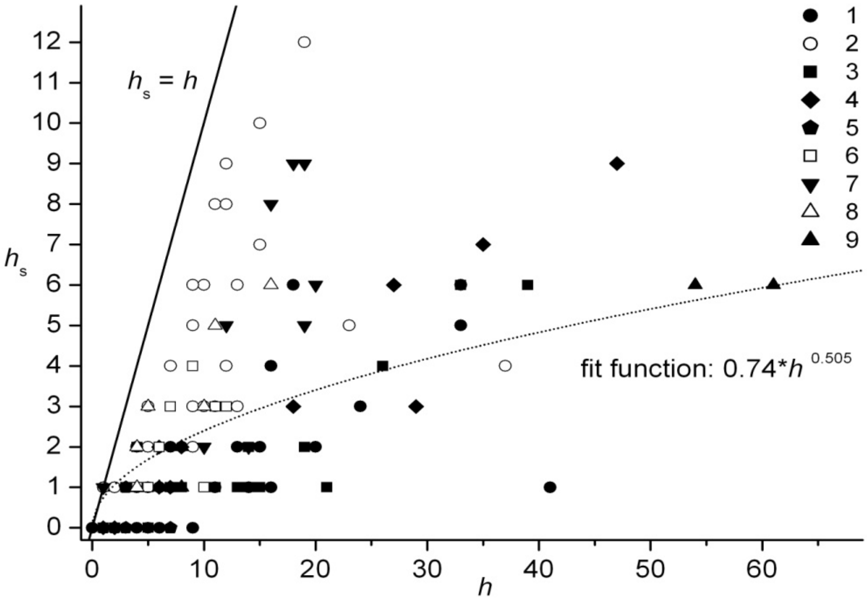

Figure 7. Self-citation correction $h_s = h - h_e$ *vs.* Hirsch's index h (a case study, [Krasteva et al, 2011])

would have (only) about 18 percent of his/her Hirsh index due to self-citations. We also apply Eq. (17) with $a = 3$ and β between 1 and 2 to confirm the results of [Engqvist and Frommen, 2008], where it has been empirically found that, on average, three self-citations per paper yield an increase of the *h* index by one.

The last point we would like to address concerns the applicability of a recently suggested self-citation correction approach for the Hirsch index. Ferrara and Romero [Ferrara and Romero, 2013] have introduced a 'discounting *h*-index' $dh = h\sqrt{R}$, where R is the ratio of external to overall (*i.e.* external plus self) citation numbers (note that $R = b$ for our case of multiplicative self-citation corrections). The authors claim that *no* prior knowledge of the self-citation distribution is required. Let us consider, for simplicity, a situation, where both overall and external citations follow linear (Hirsch) distributions with slopes s and s_e , respectively. The ratio of external and overall citations numbers is [Atanassov and Detcheva, 2013]

$$R = \frac{(C_{em}P_{em}/2)}{(C_m P_m/2)} = \frac{s}{s_e}\frac{(1+s_e)^2}{(1+s)^2}\frac{h_e^2}{h^2}, \tag{32}$$

where s and s_e are the corresponding distribution slopes. Hence $h_d = h - dh = \left(1-\sqrt{R}\right)h$ is equal to the *exact* self-citation correction $h_s = h - h_e$ if and only if $s = s_e$ or $s = 1/s_e$, *i.e.* if the triangles associated with the corresponding linear distributions are geometrically *similar*. In particular, for additive self-citation correction one obtains $h_d = a/(1+s) = h_s$, in agreement with Eq. (13a). In case of multiplicative correction to the linear citation paper rank distribution (Eq. (27a)) we have $h_d = h_s$ if and only if $b = 0$ or $b = 1/s^2 \le 1$ (both relations imply geometrically similar triangles, too). Further on, by comparing h_d with the multiplicative Hirsch's index self-citation correction for a Pareto distribution, given by the *exact* relationship (Eq. (29)), one concludes that $h_d = h_s$ if and only if $\beta = 1$. Hence, although rather convenient, the discounting *h*-index dh has problems with the *shape* of citation-paper rank distributions and should be used with some caution when estimating self-citation corrections to scientific impact.

Summary and conclusions

We have obtained quantitative estimates of self-citations effect on Hirsch's *h*-, Egghe's *g*- and Zhang's *e*-indexes for continuous citation paper rank distributions in general form as well as for especially chosen ones (uniform, linear and Pareto). Prior to this two types of a (single author) self-citation behavior have been considered, in order to provide support for the basic assumptions of additive and multiplicative self-citation corrections.

Our main conclusions are summarized as follows:

- Two types of self-citation behavior can be distinguished. The 'normal' one is characterized with almost uniform self-citation distribution and self-citations number that linearly depends on the number of papers. The 'extraordinary' citation practice is associated with linearly decreasing self-citation distribution and self-citations number depending on the squared number of papers. A smooth transition exists between both self-citation policies, governed by the number of self-cited papers;
- For a linear negative slope citation-paper rank distribution the additive self-citation corrections to the Hirsch's and Egghe's indexes are constants, depending on the self-citation level and the slope only; in

particular, they decrease when the distribution gets steeper. The latter means that a set of small number of highly cited papers is less sensitive to self-citations than a large number of poorly cited ones;

- For a Pareto distribution the additive self-citation correction to Hirsch's index varies from constant, depending on self-citation level and power exponent only, to linear function of the index, depending on whether the additive self-citation level is much smaller than the index, or it approaches its maximal value, indicating an 'extraordinary' self-citation behavior;
- The multiplicative self-citation corrections to Hirsh's and Egghe's indexes depend linearly on the corresponding indexes, which could also be suspicious for 'extraordinary' self-citation behavior;
- Both types of self-citation behavior have been qualitatively detected by analyzing scientometric data. The approximate formulae have been successfully compared to empirically and theoretically obtained self-citations corrections to Hirsch's index.

In conclusion, we believe that the model suggested could prove useful in analyzing self-citation effect on assessment of scientific activity. A list of appropriate topics for future studies include an extension of the discrete model for self-citation behavior by considering two and more co-authors, as well as discrete (Zeta, Zipf) citation-paper rank distributions.

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

[Asknes, 2003] D. W. Asknes. A macro study of self-citation, Scientometrics 56(2) 235-246 (2003)

[Atanassov and Detcheva, 2013] V. Atanassov, E. Detcheva. Citation-paper rank distributions and associated scientometric indicators – a survey, Int. J. Information Models and Analyses 2(1) 46-61 (2013)

[Ausloos et al, 2008] M. Ausloos, R. Lambiotte, A. Scharnhorst, I. Hellsten. Andrzej Pekalski networks of scientific interests with internal degrees of freedom through self-citation analysis, Int. J. Mod. Phys. C 19 371-384 (2008)

[Bartneck and Kokkelmans, 2011] C. Bartneck, S. Kokkelmans. Detecting h-index manipulation through self-citation analysis, Scientometrics 87 85-98 (2011)

[Costas et al, 2010] R. Costas, T. N. van Leeuwen, M. Bordons. Self-citations at the meso and individual levels: effect of different calculation methods, Scientometrics 82 517-537 (2010)

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

[Engqvist and Frommen, 2008] L. Engqvist and J. G. Frommen. The h-index and self-citations, Trends in Ecology and Evolution 23(5) 250-252 (2008)

[Ferrara and Romero, 2013] E. Ferrara and A. E. Romero. Scientific impact evaluation and the effect of self-citations: mitigating the bias by discounting h-index, J. Am. Soc. Information Sci. and Technology 64(11) 2332-2339 (2013)

[Fowler and Asknes, 2007] J. H. Fowler, D. W. Asknes. Does self-citation pay?, Scientometrics 72(3) 427-437 (2007)

[Glänzel and Thijs, 2004a] W. Glänzel and B. Thijs. The influence of author self-citations on bibliometric macro-indicators, Scientometrics 59(3) 281-310 (2004)

[Glänzel and Thijs, 2004b] W. Glänzel and B. Thijs. Does co-authorship inflate the share of self-citations?, Scientometrics 61(3) 395-404 (2004)

[Glänzel et al, 2004] W. Glänzel, B. Thijs, B. Schlemmer. A bibliometric approach to the role of author self-citations in scientific communication, Scientometrics 59(1) 63-77 (2004)

[Glänzel, 2008] W. Glänzel. Seven Myths in Bibliometrics. About facts and fiction in quantitative science studies, Proceedings of WOS 2008 Berlin (Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting), Eds. H. Kretschmer & F. Havemann (2008)

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

[Krasteva et al, 2011] E. Krasteva, V. Atanassov, L. Vulkova, P. Zubov, A. Dikovska and O. Yordanov. Bibliometric indicators and their evolution for a group of scientists in a narrow, highly specialized area, Third Annual Meeting and Conference of the COST Action MP0801 Physics of Competition and Conflicts, May 18-20, 2011, Eindhoven

[Markov et al, 2013] K. Markov, K. Ivanova, V. Velichko. Usefulness of scientific contributions, Int. J. Information Theories and Applications 20(1) 4-38 (2013)

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

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

[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.