Information Sciences 185 (2012) 18–31

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

An Internet measure of the value of citations

Boleslaw K. Szymanski [a,b], Josep Lluis de la Rosa [b,c,*], Mukkai Krishnamoorthy [b]

[a] Społeczna Wyższa Szkoła Przedsiębiorczości i Zarządzania, ul. Sienkiewicza 9, 90-113 Łódź, Poland
[b] Network Science and Technology (NeST) Center, Rensselaer Polytechnic Institute, 118, 8th Street, Troy, NY, USA
[c] EASY Innovation Center, University of Girona, Campus de Montilivi, E17071 Girona, Catalonia (EU), Spain

ARTICLE INFO

Article history:
Received 7 June 2010
Received in revised form 4 February 2011
Accepted 10 August 2011
Available online 26 August 2011

Keywords:
Bibliometrics
Scientometrics
Citation analysis
Author ranking
Impact factor
PageRank

ABSTRACT

A new method for computing the value of citations is introduced and compared with the PageRank algorithm for author ranking. In our proposed approach, the value of each publication is expressed in CENTs (sCientific currENcy Tokens). The publication's value is then divided by the number of citations made by that publication to yield a value for each citation. As citations are the acknowledgements of the work by authors other than oneself (indicating that it has been useful), self-citations count as zero in acknowledged citation value. Circular citations, a generalized type of self-citation, are considered to have a reduced acknowledged citation value. Finally, we propose a modification of the h-index to define it as the largest integer such that the *i-th* publication (on the list of publications sorted by their value in CENTs) is worth more than *i* CENTs. This new index, termed the *i-index* or i^2 in short, appears to be a more precise measure of the impact of publications and their authors' productivity than the h-index.

1. Introduction

Currently, the impact of a scientific publication is often measured by the number of citations it receives. Perhaps we are suffering from an over-analysis of citations for the purposes of assessing scientists and universities productivity, impact, or prestige—the examination of citations of scientific publications has become a cottage industry in higher education. This approach has been taken to extremes both for the assessment of individuals and as a measure of the productivity and influence of entire universities or even academic systems. Pioneered in the 1950s in the United States, bibliometrics was invented as a tool for tracing research ideas, the progress of science and the impact of scientific work. First developed for the "hard" sciences, it was later expanded to include the social sciences and humanities.

The citation system was invented mainly as a way to understand how scientific discoveries and innovations are communicated and how research functions [1]. It was not initially seen as a tool for evaluating individual scientists, entire universities or academic systems. Hence, the citation system is useful for tracking how scientific ideas are propagated among researchers and how individual scientists use and communicate research findings. The use of citation analysis for the assessment of research productivity or impact questionably extends the original reasons for creating the bibliometric system. Evaluators and rankers need to go back to the drawing board in considering a reliable system for the accurate measurement of the scientific and scholarly work of individuals and institutions. The unwieldy and inappropriate use of citation analysis and bibliometrics for the evaluation and ranking of research and researchers does not serve higher education well and it entrenches existing inequalities.

* Corresponding author at: EASY Innovation Center, University of Girona, Campus de Montilivi, E17071 Girona, Catalonia (EU), Spain.
E-mail address: peplluis@eia.udg.edu (J.L. de la Rosa).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.08.005

More recently, a new index based on citations, the h-index, has been proposed as an indicator of overall productivity and impact of the published work of a researcher [15]. The h-index of a researcher is the largest integer h such that at least h publications by this researcher have no less than h citations each. For example, an author with an h-index of 20 must have at most 20 publications with 21 or more citations and at least 20 publications with 20 citations each.[1] This index can easily be determined from the "times cited" in the Thomson ISI Web of Science or Google Scholar and it provides a metric for the author's productivity in terms of citations.

The h-index focuses more on measuring productivity than on measuring the impact and influence of the dissemination of a publication. However, some h-index variations attempt to capture the latter [4,5]. Measuring *impact* by the number of *new* authors who cite a publication appears to be a more accurate measure than measuring it by the h-index because it reflects the utility of an author's work to various individuals rather than only *the same* people. Thus, any type of direct or *indirect* self-citations should be discounted to a certain degree. Moreover, if *impact* signifies the importance of knowledge dissemination in publications citing the given publication, then citing a publication with a greater impact should in turn endow a higher impact to the cited publication.

In this work, we propose a new approach for measuring the impact of publications and compare it with an author ranking computed using the PageRank algorithm [17]. To the best of our knowledge, PageRank was originally inspired by the scientific bibliometric system (citations), but only recently has it been applied to measure the impact of journals, publications and scientists. The success of Google's method of ranking web pages has inspired numerous measures of journal impact that apply social network analysis to citation networks. Pinski and Narin [20] (predating PageRank) proposed ranking journals according to their eigenvector centrality in a citation network. Extending this idea, we propose a more accurate measure of impact than those based on the h-index. Our measurement is based not on a row citation count but on the impact of the citing publications and their distance from self-citations. Section 2 provides a precise explanation of our approach, introducing scientific currency tokens as a measure of the impact of citations. Section 3 presents an algorithm for estimating this value from a network of publications and authors connected by citations. Section 4 presents an example of how many tokens would be assigned to each citation in a network of nine citations among 6 publications by four authors. Section 5 describes an application of the PageRank algorithm to the same example followed by a comparison of the values of the citations calculated by both algorithms. Section 6 describes a method to compute the citation earnings of each author when there are multiple authors for a publication and shows an example of how to apply the h-index to CENTs instead of to citations. The conclusions and prospective future work are provided in Section 7.

2. CENTs – scientific currency tokens

We first describe the heuristics behind our model. We advocate measuring the value of each citation in sCientific currENcy Tokens (CENTs). The introduction of this currency was inspired by complementary currencies for the scientific communities proposed in [8,11,12] and also conceptualized as tokens or measure of reputation by [10,18]. Scientists are assumed to hold a new scientific currency and to have rational expectations with it [7]. The initial value of a publication is one CENT and then each raw non-self-citation received by the publication increases its value by one CENT. The initial value of each citation in a publication, called the raw value of citation and denoted $r_{i,j}$ when publication i cites publication j, is equal to the inverse of the total number of citations in publication i, denoted R_i, so $r_{i,j} = 1/R_i$ for $j \in [1, R_i]$. The raw value of citation is constant and therefore not affected by future publications.

The value of each citation of a cited publication is proportional to the value of the citing publication. Hence, every citation of a publication has a value in CENTs that is computed by multiplying the value of the publication by the raw value of this citation. Both the value of the publication and the value of citation increase with each publication that cites the publication, either directly or indirectly via a chain of citations from a new publication to the original publication.

Let P_i be the value (in CENTs) of Publication i, and, as previously noted, let R_i be the number of its citations; the value $w_{i,j}$ of citation of publication j by publication i is then:

$$W_{i,j} = P_i/R_i = P_i r_{i,j} \quad \text{CENTs} \tag{1}$$

Eq. (1) captures the notion that a high-impact publication endows its citations with high values. For example, if a publication receives 99 CENTs of citations after its publication, then its value, which includes the initial one CENT, becomes $P_i = 100$ CENTs. If this publication cites $R_i = 10$ other publications, then each of its citations is worth 10 CENTs.

There is a problem with Eq. (1) when all citations in the publication are self-citations. In such a case, the real value of the publication citations should be zero CENTs because there is no real external acknowledgement of the cited work. Another case of overvaluing arises when all 10 raw citations are of the work of another author who always cites back to the author of the citing publication. In other words, these two authors cite each other every time. Thus, they are in fact half self-citations, and intuitively the value of each citation should be halved. The following section shows how to eliminate any type of self-citation prior to the conversion of the raw value of a citation into its value in CENTs.

[1] Interestingly, a researcher with 20 publications with 100 citations each and 20 publications with 19 citations each has the same h-index as a scientist with only 20 publications with 20 citations each or a researcher with 100 publications with 20 citations each.

3. Raw citations and acknowledged citations

In our approach, citations to a publication are distributed among the authors of the publication. This is simple in the case of single-author publications, as the value of the publication is passed onto its author. A straightforward extension to multi-author publications, which we use here, is to divide the value of each publication equally among its authors. A more sophisticated extension could allow the authors to decide among themselves how the credit is divided among them and such a solution can be implemented in the future. In essence, this approach replaces a multi-author publication with a set of single-author publications with the values of these single-author publications summing to the value of the original publication.

A self-citation in a publication is worth zero CENTs because it does not acknowledge any other work but the author's own. The citations of other researchers are discounted in cases of closely co-cited authors, that is, authors with a high dependency on each other's citations. For example, if author A cites a publication by author B that cites a publication by author A, this chain of citations indicates an external impact of the publication nearly as small as if author A had cited himself. In fact, it is just an indirect form of self-citation. Hence, this *raw* citation should not be counted as a full citation of the work by other authors. To address this issue, we propose calculating the ratio of the value of a raw citation to the acknowledged citation value for each author. We call this ratio *the acknowledged citation*. The purpose of this calibration of raw citations is to avoid abuses of the citation system by cliques of authors or cyclic citations.

Consider a publication i written by m_i authors ($m_i = 1$ for a single-author publication) with $R_i$citations, where each citation j has m_j authors ($m_j = 1$ for a single-author citation). The acknowledged citations of Publication i to Publication j with regard to the Author l of Publication i and Author k of Publication j are denoted as a_{ijlk}, where $a_{ijlk} \leqslant 1/(m_i m_j)$. Thus, each raw citation r_{ij} will create the acknowledged citation value for each of its citing Author l and each of its cited Authors k, calculated as follows:

$$\begin{aligned} &a_{ijll} = 0 \text{ if} j \text{ is a self-citation, in other words, it is a citation to a publication written by Author } l.\\ &a_{ijkl} = r_{ij}/(m+m) \text{ if } j \text{ is not a self-citation but Publication } i \text{ cites Publication } j,\\ &\quad \text{which in turn cites another publication written by Author } l. \\ &\ldots \\ &a_{ijkl} = s \cdot r_{ij}/(s \cdot m + m), \end{aligned} \tag{2}$$

where s is the number of intermediate authors who cite the publications written by the next one in the **shortest** path of authors that **cite back** to a publication written by Author l.

The shortest path in a directed graph can easily be calculated using the Floyd–Warshall algorithm [19]; here, assigning unit cost to all edges results in a computation of complexity $\Theta(V^3)$, where V, proportional to the total number of publications (the coefficient of this proportionality is the average number of authors by a publication), denotes the number of nodes. This is a fast approximation of the general problem of estimating the total returning flow of citations from Author l to Author k when the former cites the later, applying the heuristics that the returning citation flow is dominated by the shortest path.

The total value of acknowledged citations A_i of Publication i is:

$$A_i = \sum_{j=1}^{R_i} \sum_{l=1}^{m_i} \sum_{k=1}^{m_j} a_{ijlk} \leqslant R_i \tag{3}$$

The worth of raw citation w_{ij} is then calculated as follows:

$$w_{ij} = P_i \sum_{l=1}^{m_i} \sum_{k=1}^{m_j} \frac{a_{ijlk}}{A_i} \tag{4}$$

where a_{ijlk}/A_i represents the percentage of acknowledgement of the raw citation r_{ij} for the Author l citing Author k. This value not only increases when the value of publication, P_i, grows, but it may also decrease when a new publication shortens the shortest path between two authors involved.

Eqs. (2)–(4) do not fully account for dependencies between authors. For the sake of generality, we next discuss how to precisely capture the authors' interdependence in the model and how such dependence impacts the value of the authors' citations.

3.1. Mathematical Model

We consider two different cases in this publication. Case 1 is that of a Single Author with Multiple Publications (each publication has one author) and Case 2 is that of Multiple Authors with Multiple Publications (each publication may have multiple authors).

Let $U = \{u_1, \ldots, u_m\}$ be the set of publications.
Let $V = \{v_1, \ldots, v_m\}$ be a copy of U referred to as the set of cited publications.
Let $T = \{t_1, \ldots, t_n\}$ be the set of authors of publications in U and V.

In Case 1, there are mappings of Φ_a *and* Φ_c from T to U and V to T, respectively, such that $\Phi_a(t_i) \neq \Phi_a(t_j)$ and $\Phi_c(v_i) \neq \Phi_c(v_j)$ when $i \neq j$. However, this condition does not hold in Case 2 for coauthors of joint publications.

We can model a set of publications (U,T,V) as a directed graph $G = (V,E)$, where $V \subseteq T \cup U \cup V$ is the set of vertices and $E \subseteq T \times U \cup U \times V \cup V \times T$ is the set of edges defined as follows:

- authorship edges $\langle t_i, u_j \rangle$ that exist when Publication u_j is written by Author t_i
- cited author edges $\langle v_i, t_j \rangle$ that exist when cited Publication v_i is authored by t_j
- citation edges $\langle u_i, v_j \rangle$ that exist when Publication i cites Publication j

We will also use a simpler graph, $G_P = (U, E_p \subseteq U \times U)$, where an edge $\langle u_i, u_j \rangle$ exists when Publication i cites Publication j. We note that two publications cannot cite each other when they are published sequentially, so graph G_D is acyclic. In contrast, two authors can cite each other through their publications, so cycles may exist in graph G.

Finally, we also use another graph derived from G, $G_A = (T, E_p \subseteq U \times U)$, where an edge $\langle t_k, t_l \rangle$ exists when Author t_k writes a publication that cites a publication written by Author t_l.

For the acknowledged citations, we use the balance matrix B of authors calculated from a transition matrix of graph G as follows: the balance $b(k,l)$ between two authors k and l is equal to the number of citations that go from the publications of Author l towards the publications of Author k **divided** by the number of citations that go in the opposite direction, as long as both numbers are nonzero. It may happen that $b(k,l) = b(l,k)$, and then of course the balance is equal to 1. Let c_{kl} be the number of citations from Author k to Author l and c_{lk} be the number of citations from Author l to Author k.

In this study, we define:

$$b(k,l) = c_{lk}/c_{kl} \text{ if and only if } c_{lk} \neq 0, \quad c_{kl} \neq 0 \tag{5}$$

We next analyze what the acknowledged citations should be when the balance among two authors is given. When Author l cites Author k much more than k cites $l (c_{kl} \ll c_{lk})$, we should set $a_{mnlk} \approx 0$ (where m is the publication of Author k that cites Publication n of Author l). The reason for this (de)valuation is that k citing l brings no significant *additional* impact beyond the one already captured by the citations received by author l because in fact the work of l is based on the work of k (see Fig. 1). However, each citation of l to k carries full acknowledged of k's work, so $a_{ijkl} \approx 1$ (where Publication i of Author l cites Publication j of Author k). This means that when $c_{kl} = 0$, the balance has no impact (so it may be undefined) on the acknowledged value of citations of l to k. In such case, there are also no citations from k to l that require a value adjustment, so the balance may again remain undefined in such case. When both c_{kl} and c_{lk} are nonzero, the shortest path from l to k is of length *1*, so the balance is of interest only in such cases.

Other approaches for computing the citation balance between two authors might be proposed (but remain to be evaluated), for example, equating $b(k,l)$ with the correlation of the citations among authors k and l or setting $b(k,l) = c_{kl}/(c_{kl} + c_{lk})$. Nonetheless, $b(k,l)$ should reflect the citation dependence between authors k and l. Consider a pair of authors $\langle k,l \rangle$ such that k wrote a publication j that cites a publication i written by l and there is a cycle in graph G that passes through nodes k and l. For such a pair of authors, let $s(k,l)$ denote the inverse of the number of author nodes other than k in the shortest such cycle. For all other pairs of authors $\langle k,l \rangle$, we set $s(k,l) = 0$. To calculate the acknowledged citations of Publication i authored by k regarding the dependence of this author on Author l, we propose the following formula:

$$a_{ijkl} = r_{ij} \frac{1}{1 + b(k,l) \cdot s(k,l)} \tag{6}$$

In simplified notation, $a_{ij} = r_{ij} \frac{1}{1+bs}$, applied to authors k and l. It will be referred also as the *acknowledged value* of citations.

The properties of Eq. (6) at the limits are

$$\lim_{b,s \to \infty} \left(\frac{1}{1+bs} \right) = 0; \quad \lim \left(\frac{1}{1+bs} \right) = 1$$

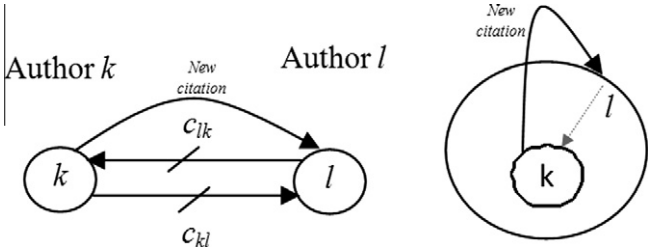


Fig. 1. Author k cites author l c_{kl} times while l cites author k c_{lk} times. When $c_{lk} \gg c_{kl}$, l is *strongly* citing k and k is inside the core work of l, thus any citation from k to l is a self-citation.

Table 1
The acknowledged values for a range of s and b values.

b	s					
	∞	1.00	0.50	0.33	0.10	0.00
0	0.00	1.00	1.00	1.00	1.00	1.00
0.01	0.00	0.99	1.00	1.00	1.00	1.00
0.10	0.00	0.91	0.95	0.97	0.99	1.00
0.33	0.00	0.75	0.86	0.90	0.97	1.00
0.50	0.00	0.67	0.80	0.86	0.95	1.00
1	0.00	0.50	0.67	0.75	0.91	1.00
2	0.00	0.33	0.50	0.60	0.83	1.00
3	0.00	0.25	0.40	0.50	0.77	1.00
10	0.00	0.09	0.17	0.23	0.50	1.00
100	0.00	0.01	0.02	0.03	0.09	1.00

Eq. (6) may be heuristically interpreted as lowering the value of an acknowledgement with an increasing balance between the authors k and l. Similarly, a shorter path to receiving a citation back (thus, a higher s) lowers the value of an acknowledgement. The acknowledged value term $\frac{1}{1+bs}$ goes from 0 (no acknowledgement) to 1 (full acknowledgement) depending on b and s, as shown in Table 1.

Hence, Eq. (6) defines that the acknowledged values of citations as equal to the raw citations when the cyclic citation is of infinite length ($s = 0$) or when there is no dependence between authors k and l ($b = 0$ and $s = 1$). However, they are zero when there is a full dependence (the citing author is always cited by the other author).

Our problem here is as follows: *Given the initial value map P of publications, compute the value map T of authors taking into account their (acknowledged) citations*. This computation is described by the following algorithm.

Algorithm *Impact*:

Input: $D = (d_{ij})$ the adjacency matrix of graph G_P (giving the citations of the publications), the value map P of publications.

Output: final value map (final worth) T of authors according to the acknowledged citations.

Step 1: Compute the matrix $C = (c_{kl})$ of raw citations as follows: for each nonzero entry d_{ij} in the transition matrix D of graph G_P, (i) create each possible pair $\langle k, l\rangle$ of authors such that k is an author of publication i while l is an author of publication j (as defined by edges $\langle v_i, t_k\rangle$ and $\langle v_j, t_l\rangle$ in graph G), and let p denote the number of such pairs ($p = 1$ in the case of a single-author publication); (ii) for each pair $\langle k, l\rangle$ created in Step 1 (i), add the value of the publication u_i/p to c_{kl}.

Step 2: Compute the matrix $B = (b_{kl})$ of balanced citations of authors according to Eq. (5).

Step 3: Compute the matrix $S = (s_{kl})$ of the inverses of the shortest paths between authors k and l in graph G, setting $s_{kk} = \infty$.

Step 4: Compute the matrix $A = (a_{ij})$ of acknowledged citations of authors using Eq. (6).

Step 5: Compute the worth of citations using Eqs. (3) and (4).

Step 6: Compute the value map (final worth) T of authors from the worth of citations.

End of Algorithm

We name our algorithm the *i-algorithm* (impact-algorithm). In the next section, we describe the computations in an example using this algorithm.

4. Example

In this section, we consider the following citation pattern of a sequence of publications and their interrelated citations shown in Fig. 2.

The citation pattern is represented by the transition matrix D of graph G_P, which is shown in Table 2.

This matrix corresponds to the graph of citations shown in Fig. 3:

The functions b and s, computed according to Eqs. (2) and (5), respectively, are shown in Table 3.

Applying Eq. (6), the acknowledged citation matrix $a(k, l)$ shown in Table 4 is generated:

The initial value of each publication is one CENT; thus, applying Eqs. (3) and (4), we obtain the value w_{ij} of each raw citation r_{ij} as shown in Table 5. Notice that a citation by Author 1 of a publication by Author 2 is worth 0.40 CENTs, which is lower than the value of 0.60 CENTs for a citation from Author 1 to Author 3. This is a sign that there is a stronger dependence of Author 1 on Author 2 than of Author 1 on Author 3. Moreover, author 1's self-citation has a value of zero because it brings no independent acknowledgement of the work. There is a case of a low-acknowledgement citation of Publication 2 (by Author 2) in Publication 4 by Author 1; this low valuation is due to the strong balance between Authors 1 and 2. Finally, as long as there is only one citation in Publication 3 by Author 3, all the value of Publication 3 is assigned to that citation, causing it to be valued at 1.00 CENT.

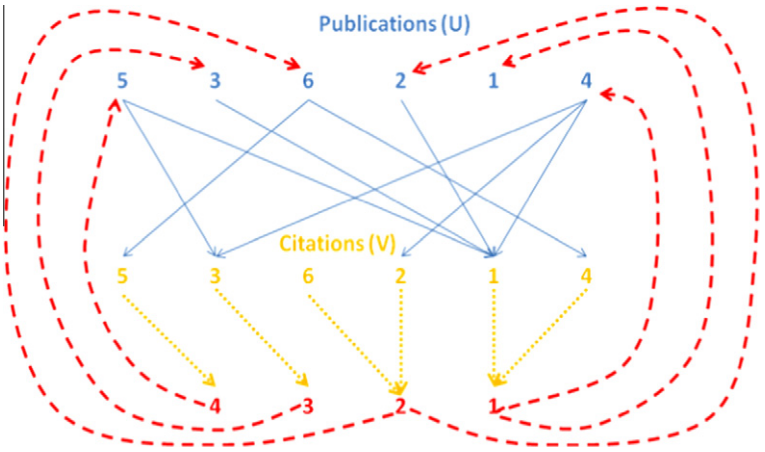


Fig. 2. Graph G of a sequence of six publications and their citations and authors; the authorship edges are dashed red, the cited author edges are dotted yellow and the citation edges are solid blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The transition matrix of graph G_P.

$D(i,j)$ publications	Outbound						
	1	2	3	4	5	6	R_i
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1
3	1	0	0	0	0	0	1
4	1	1	1	0	0	0	3
5	1	0	1	0	0	0	2
6	0	0	0	1	1	0	2
Inbound	4	1	2	1	1	0	9

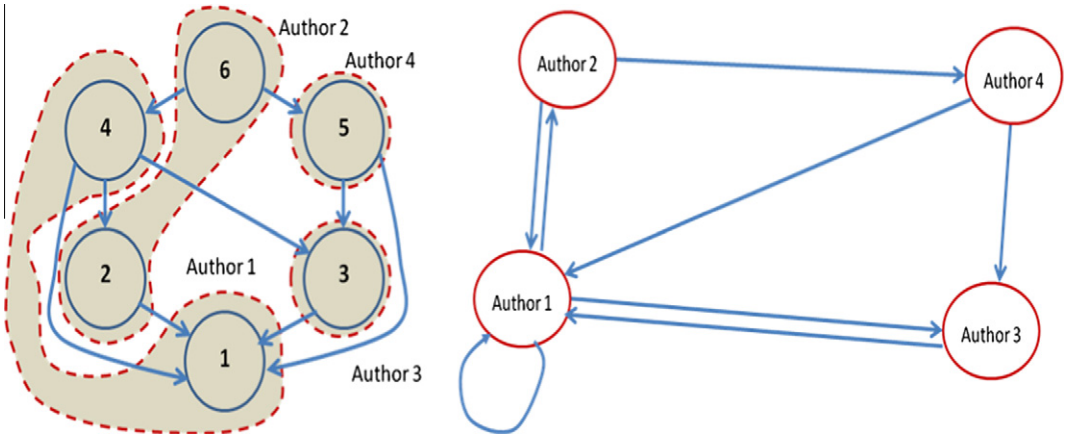


Fig. 3. Six publications shown with their citation patterns and assignments to authors in the acyclic graph G_P and four authors in graph G_A with the same citations containing cycles.

The resulting values of the received citations (in the number of CENTs the authors receive) per author are shown in Table 6.

In summary, five inbound raw citations to Author 1 are worth 3.00 CENTs, whereas one inbound raw citation to Author 2 is worth 0.1 CENTs. The two citations to Author 3 are worth 1.1 CENTs, and the single citation received by Author 4 is worth 0.50 CENTs.

Table 3
The balances and inverses of citation distances.

Authors	$b(k,l)$				$s(k,l)$			
	1	2	3	4	1	2	3	4
1	NA	2.0	1.0	NA	CO	1.00	1.00	0.50
2	0.5	NA	NA	NA	1.00	1.00	0.33	0.50
3	1.0	NA	NA	NA	1.00	0.33	1.00	033
4	NA	NA	NA	NA	0.50	0.50	0.33	0.50

Table 4
The acknowledged citations matrix.

Authors	$a(k,l)$			
	1	2	3	4
1	0.00	0.33	0.50	0.00
2	1.33	0.00	0.00	0.67
3	0.50	0.00	0.00	0.00
4	1.00	0.00	1.00	0.00

Table 5
The raw citations (r_{ij}) and their conversion into acknowledged citations (A_i) and CENTs (w_{ij}).

Citing author	Citing publication	Cited publication	Cited author	r_ij	s/b	a_ij	A_i	w_ij
1	4	2	2	1	0.5	0.33	0.83	0.40
1	4	3	3	1	1.0	0.50	0.83	0.60
1	4	1	1	1	0.0	0.00	0.83	0.00
2	2	1	1	1	2.0	0.67	0.67	1.00
2	6	4	1	1	2.0	0.67	1.33	0.50
2	6	5	4	1	2.0	0.67	1.33	0.50
3	3	1	1	1	1.0	0.50	0.50	1.00
4	5	3	3	1	∞	1.00	2.00	0.50
4	5	1	1	1	∞	1.00	2.00	0.50 CENTs

Table 6
The values of the received citations for each author.

Authors	Inbound	Outbound	Value
1	5	3	3.00
2	1	3	0.40
3	2	1	1.10
4	1	2	0.50
Total	9 Citations	9 Citations	5.00 CENTs

5. An Internet measure of the citation value

As said in Section 1, PageRank [17] can also be used to estimate the value of citations. The success of Google's method of ranking web pages has inspired numerous measures of journal impact that apply social network analysis to citation networks. Pinski et al. [20] (predating PageRank) proposed a journal ranking based on their eigenvector centrality in a citation network. They suggested the use of a recursive impact factor to give citations from high-impact journals greater weight than citations from low-impact journals. This impact factor is based on a "trade balance" similar to our Eq. (5) approach, in which journals score highest when they are often cited but rarely cite other journals. Such a recursive impact factor resembles the Internet-born PageRank algorithm introduced several years later. Bollen et al. [2] and Dellavalle et al. [13] proposed ranking journals according to their citation PageRank (an approximation of Pinski's eigenvector centrality), followed by the launch of innovative ranking services such as http://eigenfactor.org that started publishing journal PageRank rankings in 2006. The Scimago group (http://www.scimagojr.com) publishes the Scimago Journal Rank (SJR) that ranks journals based on a principle similar to that used to calculate citation PageRank. PageRank has also been proposed as a basis for ranking individual articles [9].

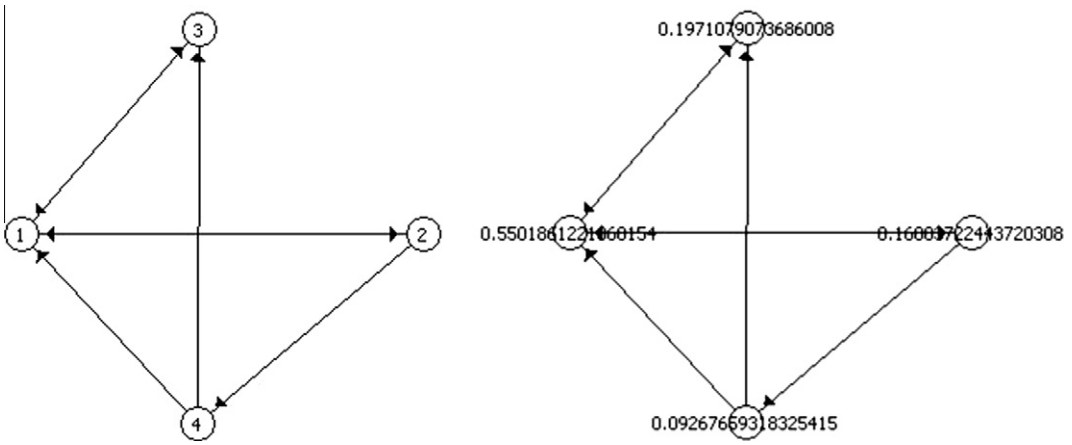


Fig. 4. PageRank applied to four authors and their six publications.

As PageRank was inspired by the scientific publication model, one might expect that it can be used naturally as a method for measuring the impact of scientific research, after certain improvements are made by discarding any self-citation loops; however, there remain challenges to this application. The first challenge is in deciding what values to assign for the cost of links (the citations). It may even be the case that the costs assigned to links are not constant but a function of time. Another challenge is that in Internet/web graphs where page ranks are calculated, both outgoing and incoming edges (links) can be deleted and/or added at any time. In the citation graphs of publications, we can only add outgoing edges to existing nodes (publications) and there can never be incoming edges to newly created nodes. In the citation graphs of authors, we can only add edges.

The basic PageRank algorithm does not converge when applied to publications but does when applied to the citations of authors. Fig. 4 shows the results of applying the basic PageRank algorithm[2] to a graph of four authors using the same citations given in Fig. 2. PageRank only applies to authors instead of publications because its inability to deal with weighted links is not a problem in the case of authors. To illustrate the idea, we use the simple example from Fig. 1; by applying to it the basic PageRank algorithm, we obtain the graph shown in Fig. 4.

The PageRank algorithm concludes that Author 1 has the highest impact, with a 0.5502 probability of his publications being noticed, while Authors 2 and 3 have probabilities of 0.1971 and 0.16, respectively. Author 4 is the least likely to be noticed, with a probability of being cited equal to 0.0927.

We now compare several different author rankings for the same set of authors and publications in Table 7. The first column contains the results using the proposed CENTs approach, and the values in the second column were computed using the basic PageRank algorithm. The third column lists the number of citations after discounting self-citations, while the fourth one is the number of raw citations. Column 1 is normalized with respect to the total number of CENTs, while columns 3 and 4 are normalized in terms of the total number of citations. This normalization is done to facilitate the comparison with the probabilistic results produced by PageRank, which yield values between 0 and 1.

The first observation is that both our approach (column 1) and PageRank (column 2) assigned the highest value to citations received by Author 1. Our approach gave more credit to Author 1 than the raw rankings with (column 3) and without (column 4) the self-citation discount. Our approach and PageRank both discounted the double credit that Author 3 received versus Authors 2 and 4 in the raw ranking (columns 3 and 4). Our approach gave the lowest credit to Author 2, while PageRank gave it to Author 4, yet these authors received equal credits in the raw rankings. The reason for these differences is that Author 2's only citation comes from Author 1, and there is a stronger dependence between these two authors than between Author 4 and Author 2, who provided the only citation published by Author 4.

PageRank algorithms do not consider citations of the authors' own publications differently from other citations; thus they consider all the citations to have equal values. Page rank computation is an iterative computation. However, the h-index is a static local algorithm computing the in-degree of each publication in the citation graph. Our impact algorithm described herein differs significantly from both of these computations.

5.1. Considerations of the dynamics of the i-algorithm

The former ranking was obtained when the publications' values were computed in a single iteration. Next, we examined the effect on the rankings when the same algorithm was applied iteratively and whether or not it converged in that case.

[2] Using software from Preston and Krishnamoorthy, "*GraphDraw*: A Graph Drawing System to Study Social Networks". Unpublished manuscript, Rensselaer Polytechnic Institute, Troy, NY, 2004.

Table 7
Comparison of the different measures of author ranking (h-index is not shown because of small number of publications considered).

Authors	Col. 1 Value		Col. 2 Page rank		Col. 3 Raw-self		Col. 4 Raw	
		%1		%1		%1		%1
1	3.000	0.60	0.550	0.55	4.000	0.50	5.000	0.56
2	0.400	0.08	0.160	0.16	1.000	0.13	1.000	0.11
3	1.100	0.22	0.197	0.20	2.000	0.25	2.000	0.22
4	0.500	0.10	0.093	0.09	1.000	0.13	1.000	0.11
	CENTs		PROB.		CIT.		CIT.	

Table 8
Converged values for the publications (above) and the authors (below).

Publication	Converged values	%1
1	1.6	0.36
2	0.32	0.07
3	0.88	0.20
4	0.80	0.18
5	0.807	0.18
6	0.0000	0.00
Author	Initial value	%1
1	3.00	0.60
2	0.40	0.08
3	1.10	022
4	0.50	0.10
Author	Final value	%1
1	2.40	0.55
2	0.32	0.07
3	0.88	0.20
4	0.80	0.18

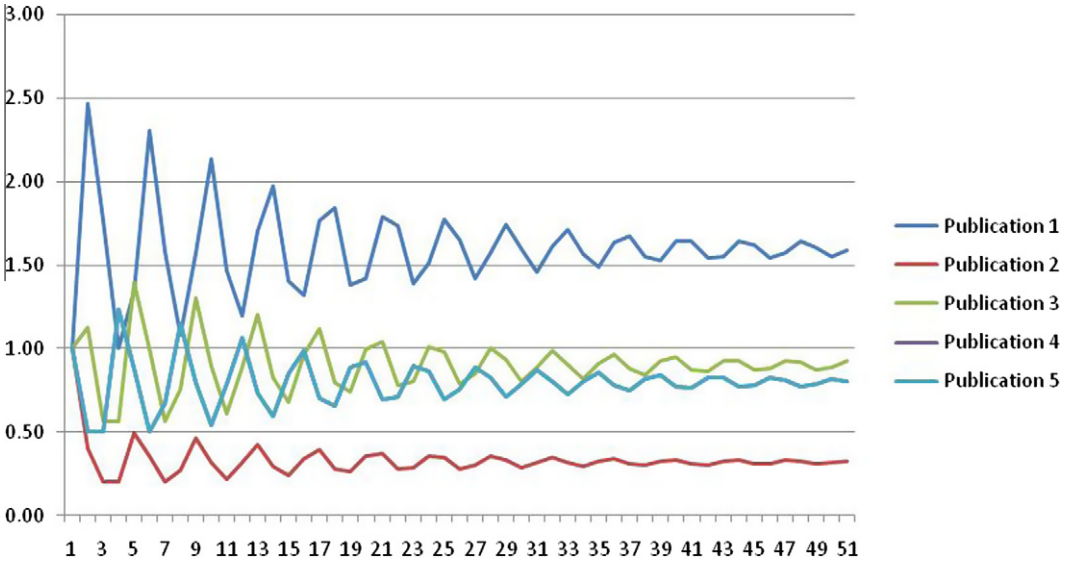


Fig. 5. Convergence of the algorithm.

As is the case with PageRank when ranking publications, there were no cycles within the citation graph based on publications. The corresponding solution has Publication 1 collecting all CENTs that flow out into the system from Publication 6. With this assumption, we obtained the following convergences of the values of the publications (see Fig. 5).

The final values after 500 iterations and the author final values are depicted in Table 8.

Table 9
A comparison of the different measures of author ranking. columns 1 and 5 are the static and dynamic measure of the value of citations in CENTs, column 2 is the probability generated by PageRank and columns 3 and 4 are the citations without and with self-citations.

Authors	Col. 1 Value static		Col. 2 Page rank		Col. 3 Raw-self		Col. 4 Raw		Col. 5 Value dynamic	
		%1		%1		%1		%1		%1
1	3.000	0.60	0.550	0.55	4.000	0.50	5.000	0.56	2.40	0.55
2	0.400	0.08	0.160	0.16	1.000	0.13	1.000	0.11	0.32	0.07
3	1.100	0.22	0.197	0.20	2.000	0.25	2.000	0.22	0.88	0.20
4	0.500	0.10	0.093	0.09	1.000	0.13	1.000	0.11	0.80	0.18
	CENTs		PROB.		CIT.		CIT.		CENTs	

With this dynamic publication value algorithm (labeled "Value Dynamic" in Table 9, column 5), we observed that the value of Author 4 increased to nearly equal that of Author 2.

The dynamic publication value algorithm retains the same ranking order as the static one. Perhaps it is more precise, but its principal drawback is that it requires an update of the entire graph of citations and acknowledged citations every time a new publication and its citations join the graph of citations. We propose to limit the citation loops to five authors to limit the complexity of the acknowledgement calculus. Techniques for the efficient implementation of our algorithm will be the subject of future study.

Another restriction of this approach is the assumption that publications cannot cite each other reciprocally. In the future, authors may be able to continuously update citations in their publications and thus negate this assumption. In such a case, our dynamic algorithm will be applicable and may rank publications more accurately than the static algorithm.

6. Multiple authors

The model with unique authors presented earlier can easily be generalized to publications with multiple authors. Currently, a citation of a multi-authored publication implies a multiplication of this citation by the number of authors of the cited publication. If a publication P_i has k_i authors, one inbound raw citation r_i to the publication generates k_i inbound raw citations. Coauthors share the ownership of a publication, and historically they are assumed to have equal shares. In reality, authors often have different levels of contribution to the publication, sometimes reflected in the nonalphabetic order of the authors. In patents, there are often explicit definitions of the ownership share that are potentially different for each coinventor, but this is not the case for scientific publications. Here, we define e as the share of ownership for each author of a publication. If publication P_i has k_i authors, then we calculate e as:

$$e_i = 1/k_i \tag{7}$$

that is, two authors each have 50% of the ownership share, three authors 33.3% each, and so on.

Using the same graph as in Fig. 3, let us suppose that Author 2 coauthors Publications 3 and 4. The multiple raw citations change the graph as in the following Fig. 6, where the dotted arrows are the new raw citations resulting from the new authorship:

The result is that Author 2 increases his "purse" by three raw citations, one of them a new self-citation and one associated with the already existing edge from Author 1 to Author 2, as depicted in Table 10.

In his example, the shares of Publications 1.4 and 3.3 are set for each coauthor at 50%, as a default share. Inputting this information into the *i-algorithm* results in the earnings shown in Table 11.

The corresponding iterative calculation of the algorithm and the stabilized results are also shown in Table 11, demonstrating that, predictably, Author 2 increased his earnings, and Authors 1 and 3 reduced their earnings. Author 3 suffered the largest loss because of the new authorship share structure, whereas Author 4 suffered no reduction in his earnings with the new citation graph, although a new cycle of citations appeared among Authors 4 and 2. In summary, after preprocessing

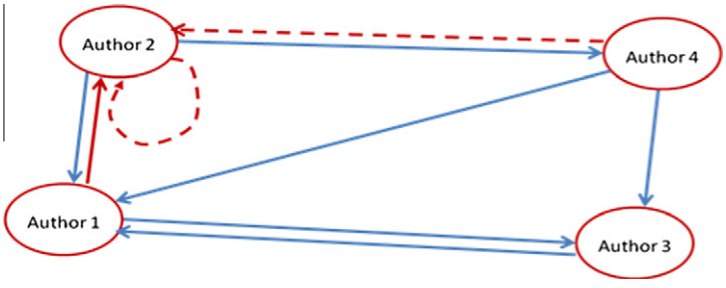


Fig. 6. New raw citations (dotted arrows) arise when author 2 coauthors publications 3 (with author 3) and 4 (with author 1).

Table 10
Comparison of the different measures of author ranking.

Authors	Single author	Multiple authors
Inbound raw citations		
1	5	5
2	1	4
3	2	2
4	1	1
Total	9.00	12.00

Table 11
Static calculation (above) and dynamic calculation after 500 iterations (below).

Authors	Single author	Multiple authors
Value		
1	3.00	2.75
2	0.40	1.17
3	1.10	0.58
4	0.50	0.50
Total CENTs	5.00	5.00
Earnings		
1	1.99	2.05
2	1.21	1.35
3	0.41	0.40
4	0.8	0.63
Total in CENTs	4.41	4.4

Table 12
Comparison of rankings with PageRank (column 2), raw citations (columns 3 and 4) and the *i-algorithm* (columns 1 and 5).

Authors	Col. 1 Value (static)		Col. 2 PageRank		Col. 3 Raw-self		Col. 4 Raw		Col. 5 Value (dynamic)	
		%1		%1		%1		%1		%1
1	2.750	0.55	0.550	0.55	4.000	0.5	5.000	0,.56	1.992	0.45
2	1.167	0.23	0.160	0.16	1.000	0.13	1.000	0.11	1.207	0.27
3	0.583	0.11	0.192	0.20	2.000	0.25	2.000	0.22	0.410	0.09
4	0.500	0.10	0.093	0.09	1.000	0.13	1.000	0.11	0.797	0.18
	CENTs	PROB.		CIT.		CIT.		CENTs		

the shares of the authors involved in multi-author publications, the number of CENTs created was the same as with the single-author publications, in this case, exactly five CENTs (Table 11).

We next examined the new rankings, shown in Table 12. The update of the PageRank calculation with the new graph resulted in the following rankings with the updated bidirectional citation between Authors 4 and 2. The table shows how the *i-algorithm* (column 1) better reflects the relative values of the authors than the raw citations (columns 3 and 4) and shows some correlation with the PageRank values, although the i-algorithm is more sensitive to the different impacts of Authors 2 and 3 than PageRank. i-index as an h-index calculated with CENTs (see Fig. 7).

Next, we evaluated how determining the value of citations impacts other bibliometric measures, such as the h-index. The h-index[15] represents a breakthrough in the bibliometric measure of the impact of scientific research [3]. Today, the most-important databases use it, and several sites have recently implemented h-index computations based on Google Scholar (see, for example http://interaction.lille.inria.fr/~roussel/projects/scholarindex/). The h-index is now widely used, even though it has several weaknesses that have been pointed out by Bornmann et al. in [4,5]. Other works have proposed a generalized h-index for disclosing latent facts in citation networks like Sidiropoulos et al. [23]. However, the h-index has several weaknesses. The first is that it assigns the same importance to all citations and, like most pure citation measures, it is field-dependent and may be influenced by self-citations. The second is that the number of coauthors may influence the number of citations received. Finally, the third weakness is that the h-index, in its original setting, puts newcomers at a disadvantage because both their publication output and observed citation rates will be relatively low. The h-index lacks sensitivity to changes in number of publications written and citations received by an author: it can never decrease and is only weakly sensitive to the total number of citations received [21].

All empirical studies to date that have tested the various indices used for scientists or journals have reported high correlations between these coefficients. This may indicate a redundancy among the various indices in measuring achievement [16]. However, the results of two studies by Bornmann and Daniel [3,6] stated more precisely that the h-index and its

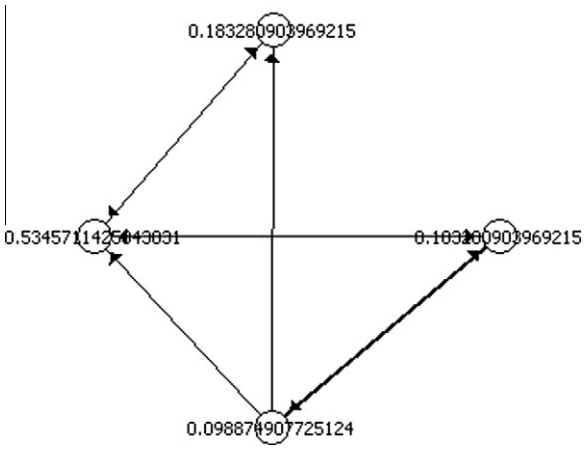


Fig. 7. New PageRank results.

Table 13
Rankings with *h* and *a* indexes.

Authors	Raw-self	*h*	*a*
1	4	1	4.00
2	3	1	3.00
3	2	1	2.00
4	1	1	1.00

variants are, in effect, a mixture of two types of indices. "The one type of indices [...] describes the most productive core of the output of a scientist and tells us the number of publications in the core. The other indices [...] depict the impact of the publications in the core" [6]. To measure the quality of scientific output, it would therefore be sufficient to use just two indices: one that measures productivity and one that measures impact, for example, the h-index and the a-index.

Like the previous authors (e.g., see [21]), we prefer to develop a further-refined impact factor that combines the two measures mentioned above. Specifically, in this study, we attempted to combine productivity and impact into one index, the *i-index*, the *impact index* or i^2 in short, of a set of publications, which is defined as follows:

i is the largest number such that there are at least *i* publications in the set with the value in CENTs higher than or equal to *i*.

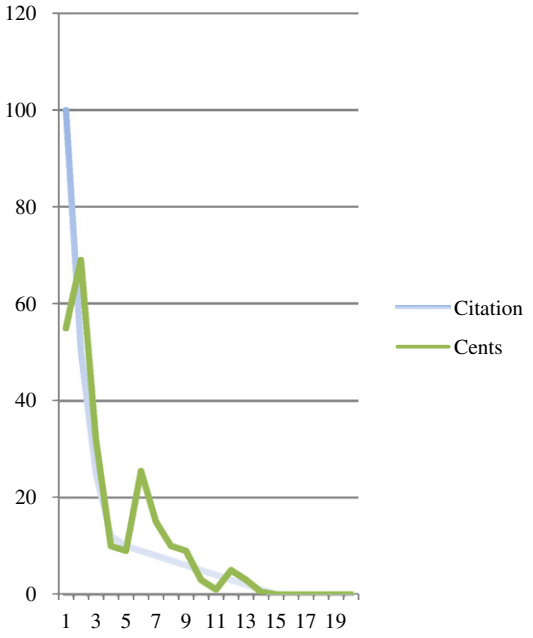
Thus, the *i-index*, unlike the h-index, takes into account the relevance of the authors who are citing a work to capture the fact that not every citation is of the same value. The h-index undervalues those scientists whose publications have a highly unbalanced impact, as measured by the formula $a \cdot h^2$. For example, a scientist with one highly cited publication and a number of unnoticed publications with one or no citations will have $h = 1$, even though the ratio of citations per publication may be high (the a-index would be high in this case).

The case described above will yield different result when the *i-index* is used and if some of the relatively "unnoticed" publications are cited by influential scientists who published highly-cited publications in which they cite that scientist's

Table 14
Rankings with raw citations, h-index and i-index.

Publication	Authors	Co-author	Value	Raw-self
1	1		2.50	3
2	2		0.33	2
3	3	2	1.17	2
4	1	2	0.50	1
5	4		0.50	1
6	2		0.00	0
			CENTs	Citations

Authors	Value	Raw-self	*h*	*i*
1	2.75	4	1	1
2	1.17	5	2	1
3	0.58	2	1	0
4	0.50	1	1	0
Total	5.00 CENTs	12 Citations		



Publication	Citations	Cents
1	100	55
2	50	69
3	25	32
4	12	10
5	10	9
6	9	25.5
7	8	15
8	7	10
9	6	9
10	5	3
11	4	1
12	3	5
13	2	3
14	1	0.5
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0
TOTAL 20	242	247
	h	***i***
	7	***9***

Fig. 8. Example of a comparison of the h-index and *i-index* for an author.

publications. In such a case, their citations will be credited with many CENTs, yielding $i > h$, and thus better representing the impact that was missed by the h-index.

To further illustrate the differences between the *h* and *i* indices, we use the graph of citations from Fig. 6, in which there is a universe of four authors and six publications with 12 citations and only 5.50 acknowledged citations. *h* is 1 for all authors, even though they have different citation patterns and different number of publications published, as represented in the following Table 13 using the case of multiple authors from Fig. 7.

In contrast, the *i-index* is 0 for Authors 3 and 4 because each has a publication with citations that are worth less than one CENT, as depicted in the "value" column in Table 14. Taking into account that Author 2 has coauthored Publications 3 and 4, this means that Author 2 has four publications published, while Author 1 has two and Authors 3 and 4 have only one each.

Hence, the *i-index* measures the impact of an author's publications more precisely than the h-index. A more elaborate example is given below with an author of 20 publications, each with a different number of citations. This author's $h = 7$ while $i = 9$, as shown in Fig. 8. It is worth noting that Publications 2 and 6 are valued much higher than the number of citations they received, while Publication 1 is worth much less than the number of citations that it received.

7. Final discussion

This is an approach for measuring the impact of authors by calculating first (with some heuristics) the estimated value of acknowledged citations and then converting this value into CENTs as a measure of the value of a publication. The value of publications in CENTs is propagated through the outgoing citations of a publication. We compared the resulting rankings of authors generated by our approach with that of PageRank and two citation-based rankings. More accurate algorithms for acknowledged citations based on optimal flow through the network of citations will be explored in our future work.

The advantage of using CENTs to measure the worth of citations is that it creates a basis for a variant of the h-index, termed the *i-index* (or impact index), which is able to measure the impact of publications more precisely than the original h-index. The *i-index* can increase or decrease according to changes in the value of the citing authors, unlike the h-index, which is a nondecreasing function of time. The *i-index* maintains the benefit of the h-index of balancing the measure of the quantity of publications and their impacts. An example of this benefit is shown in Fig. 8, where an author has an h-index = 7 and *i-index* = 9. To further illustrate this, consider a collection of 100 publications, each written by a distinct single author. Let each publication cite the other 99 publications. This is clearly a clique (cluster) of authors jointly citing each other's work. The h-index value of each of these publications will be 99, whereas the *i-index* will just be 1, demonstrating that, unlike the h-index, the i-index cannot be artificially increased by a clique of authors.[3]

[3] See the case of the fictitious author Ike Antkare with h-index 100 in http://scholar.google.com/scholar?q=Antkare&hl=en&btnG=Search&as_sdt=1%2C33&as_sdtp=on, http://membres-lig.imag.fr/labbe/Publi/IkeAntkareV2.pdf, http://rachelgliese.wordpress.com/2010/12/15/le-h-index-d%E2%80%99ike-antkare/.

Our approach is conceptually simpler than the one proposed in Papavlasopoulos et al. [22] that tries to measure the impact of citations using a number of correlations of impact factor with immediacy index, cited half-life, and citing half-life. The strengths of our approach are that it is synergistic with Pagerank inspired algorithms [9,16], unique in its analysis of the web of citations and avoids an inflation of those citations by publications with multiple authors. Further implementations of the *i-index* algorithms with the systems described by Guo in [14] will be considered for future work.

Acknowledgments

This research was funded by the European Union Project No. 238887, *A unique European citizens' attention service* (iSAC6+) IST-PSP, the ACC1Ó grant *ASKS – Agents for Social Knowledge Search* – Catalan Government, the Spanish MCI project *TIN2010-17903. Comparative approaches to the implementation of intelligent agents in digital preservation from a perspective of the automation of social networks*, 2009 BE-1-00229 of the AGAUR awarded to Josep Lluis de la Rosa, and the CSI-ref.2009SGR-1202.

References

[1] P.G. Altbach, The Tyranny of Citations, Inside Higher Ed, 2006. <http://insidehighered.com/views/2006/05/08/altbach>.
[2] J. Bollen, M.A. Rodriguez, H. Van den Sompel, Journal status, Scientometrics 69 (3) (2006). <http://www.arxiv.org/abs/cs.GL/0601030>.
[3] L. Bornmann, H-D. Daniel, What do we know about the h index?, Journal of American Society Information Science and Technology 58 (2007) 1381–1385
[4] L. Bornmann, R. Mutz, H-D. Daniel, Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine, Journal of the American Society for Information Science and Technology 59 (2008) 830–837.
[5] L. Bornmann, R. Mutz, H-D. Daniel, G. Wallon, A. Ledin, Are there really two types of h index variants? A validation study by using molecular life sciences data, in: J. Gorraiz, E. Schiebel (Eds.), Proceedings of 10th International Conference on Science and Technology Indicators, Austrian Research Centers, Vienna, Austria, 2008, pp. 256–258.
[6] L. Bornmann, H.-D. Daniel, The state of h index research. Is the h index the ideal way to measure research performance, EMBO Reports 10 (1) (2009) 2–6.
[7] G.A. Calvo, C.A. Rodríguez, A model of exchange rate determination under currency substitution and rational expectations, Journal of Political Economy 85 (1977) 617–625.
[8] C. Carrillo, J.Ll. de la Rosa, A. Canals, Towards a knowledge economy, International Journal of Community Currency Research, 1325-9547 11 (2007) 84–97.
[9] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with google's pagerank algorithm, Journal of Informetrics 1 (1) (2007) 8–15.
[10] J. Crowcroft, S. Keshav, N. McKeown, Scaling the academic publication process to internet scale, Communications of the ACM 52 (1) (2009) 27–30.
[11] J.Ll. de la Rosa, B.K. Szymanski, Selecting scientific papers for publication via citation auctions, IEEE Intelligent Systems 22 (6) (2007) 16–20.
[12] J.L. de la Rosa, B.K. Szymanski, Towards symbiosis between the scientific community and the internet with peer review as one of the core scientific processes, in: B.G. Kutais (Ed.), Internet Policies and Issues, vol.9, 2011, pp. 75–79.
[13] R.P. Dellavalle, L.M. Schilling, M.A. Rodriguez, H. Van de Sompel, J. Bollen, Refining dermatology journal impact factors using pagerank, Journal of the American Academy Dermatology 57 (2007) 116–119.
[14] G.M. Guo, A computer-aided bibliometric system to generate core article ranked lists in interdisciplinary subjects, Information Sciences, 0020-0255 177 (17) (2007) 3539–3556, doi:10.1016/j.ins.2007.02.04.
[15] J.E. Hirsch, An index to quantify an individual's scientific research output, Proceedings of the National Academy of Sciences 102 (46) (2005) 16569–16572. <http://www.pnas.org/content/102/46/16569.full.pdf+html>.
[16] B. Jin, L. Liang, R. Rousseau, L. Egghe, The R- and AR-indices: complementing the h-index, Chinese Science Bulletin 52 (2007) 855–863.
[17] A.N. Langville, C.D. Meyer, Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, 2006.
[18] S. Mizzaro, Quality control in scholarly publishing: a new proposal, Journal of American Society Information Science and Technology 54 (11) (2003) 989–1005.
[19] C. Papadimitriou, M. Siderib, On the Floyd–Warshall algorithm for logic programs, The Journal of Logic Programming, 0743-1066 41 (1) (1999) 129.
[20] G. Pinski, F. Narin, Citation influence for journal aggregates of scientific publications: theory with application to literature of physics, Information Processing and Management 12 (1976) 297–312, doi:10.1016/0306-4573(76)90048-0.
[21] R. Rousseau, Reflections on recent developments of the h-index and h-type indices, in: Proceedings of WIS 2008, Berlin, Fourth International Conference on Webometrics, Berlin, Germany: Informetrics and Scientometrics & Ninth COLLNET Meeting, 2008.
[22] S. Papavlasopoulos, M. Poulos, N. Korfiatis, G. Bokos, A non-linear index to evaluate a journal's scientific impact, Information Sciences, 0020-0255 180 (11) (2010) 2156–2175, doi:10.1016/j.ins.2010.01.018.
[23] A. Sidiropoulos, D. Katsaros, Y. Manolopoulos, Generalized Hirsch h-index for disclosing latent facts in citation networks, Scientometrics 72 (2) (2007) 253–280. Akadémiai Kiadó, co-published with Springer Science + Business Media B.V., Formerly Kluwer Academic Publishers B.V, ISSN 0138-9130 (Print) 1588-2861 (Online), vol. 72, Number 2/August, (2007), doi:10.1007/s11192-007-1722-z.