A Bibliometric Index Based on the Collaboration Distance between Cited and Citing Authors

Maria Bras-Amorós [a] Josep Domingo-Ferrer [a] Vicenç Torra [b]

[a] *Universitat Rovira i Virgili*
Dept. of Computer Engineering and Maths, UNESCO Chair in Data Privacy
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
e-mail {maria.bras,josep.domingo}@urv.cat

[b] *IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia*
e-mail vtorra@iiia.csic.es

Abstract

The popular h-index used to measure scientific output can be described in terms of a pool of evaluated objects (the papers), a quality function on the evaluated objects (the number of citations received by each paper) and a sentencing line crossing the origin, whose intersection with the graph of the quality function yields the index value (in the h-index this is a line with slope 1). Based on this abstraction, we present a new index, the c-index, in which the evaluated objects are the citations received by an author, a group of authors, a journal, etc., the quality function of a citation is the collaboration distance between the authors of the cited and the citing papers, and the sentencing line can take slopes between 0 and ∞. As a result, the new index counts only those citations which are significant enough, where significance is proportional to collaboration distance. Several advantages of the new c-index with respect to previous proposals are discussed.

Key words: Bibliometric indices; h-index; c-index.

[1] This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01/02 "E-AEGIS", TIN2009-11689 "RIPUP", and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia under grants 2009 SGR 1135 and 2009 SGR 7. The second author is partly supported by the Government of Catalonia as an ICREA Acadèmia Researcher. The authors are with the UNESCO Chair in Data Privacy, but their views do not necessarily reflect those of UNESCO nor commit that organization.

1 Introduction

In the field of bibliometrics, the h-index by Hirsch [16] has earned a lot of popularity, being publicized by Ball in *Nature* [2] and implemented in the Web of Knowledge bibliometric database [27]. Previous indicators were the total number of papers or the total number of citations. It is widely accepted that not all papers should count and Hirsch with his index suggested to count only those that were considered significant according to the number of citations.

However, less attention has been paid to citations. Just like not all papers count, neither all citations should count nor those that count should count in the same way. It is true that some attention has been devoted to differentiating self-citations, on the theoretical side by, *e.g.*, [23,9,29], and, on the practical side, most notably by the CiteSeer database [5].

We suggest a new index following Hirsch's idea, but counting only those citations that are considered significant, where the significance of a citation is proportional to the collaboration distance between the cited and the citing authors.

The rest of this paper is organized as follows. Section 2 revisits the h-index and presents an abstraction of it leading to the c-index. Section 3 proposes the new index based on the collaboration distances between citing and cited authors. A discussion of the advantages of the new index w.r.t. previous bibliometric indices is given in Section 4. Section 5 describes a procedure to efficiently compute the new index in practice. Experimental results are reported in Section 6. Section 7 contains some concluding remarks. The appendix contains expanded experimental results.

2 From the h-index to the c-index

The h-index can be reinterpreted as follows. Consider the pool of all papers by an author. Consider the quality $Q(p)$ of a paper p to be the number of its citations. Sort all elements in the pool by decreasing order of their quality. That is, write the elements in the pool as $p_1, \ldots, p_n$, with $Q(p_i) \geq Q(p_{i+1})$. Then, draw the (decreasing) graph with all points $(i, Q(p_i))$. Let the *sentencing line* be the line through the origin with slope 1. The h-index is approximately the ordinate of the intersection of the sentencing line with the graph (see Figure 1 (a)). Formally, h is the maximum of the values $\min(i, Q(p_i))$ for $i \in \{1, \ldots, n\}$ (see Figure 1 (b)). That is,

$$h = \max\{\min(i, Q(p_i)) : i \in \{1, \ldots, n\}\} \tag{1}$$

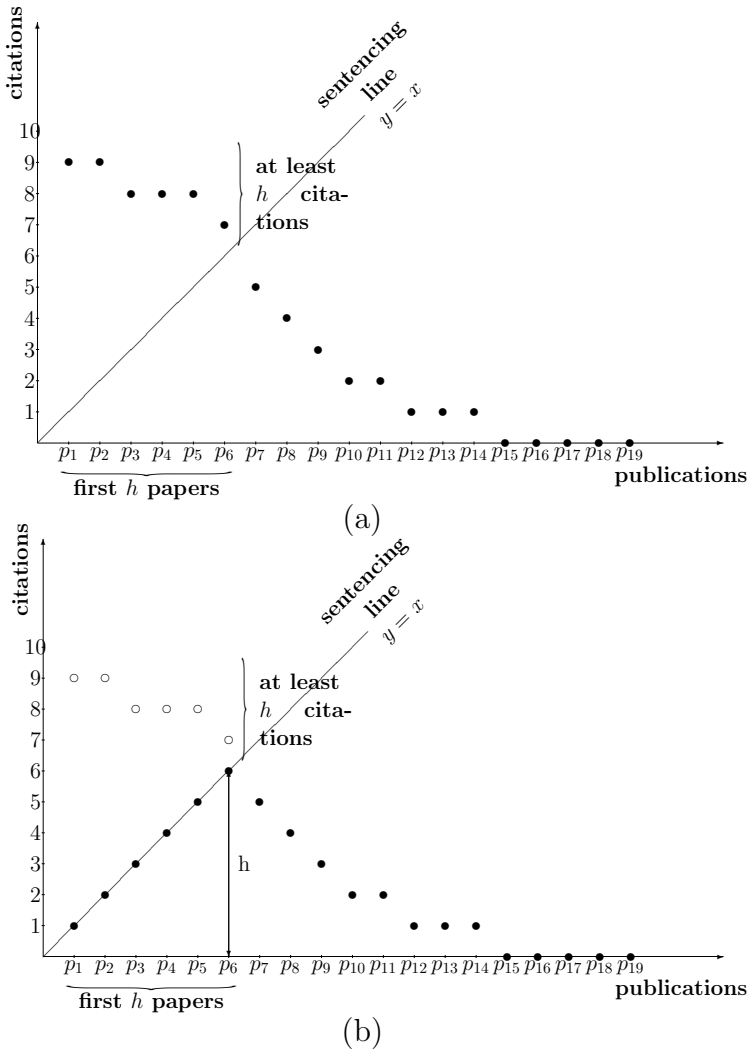

Fig. 1. (a) Reinterpretation of the h-index (inspired in Wikipedia)
(b) Representation of $\min(i, Q(p_i))$

In this way, there are h papers $p_1, \ldots, p_h$ with quality (citations) at least h, and the rest of papers with quality at most h.

From a formal point of view, Equation (1) corresponds (see [26] for details) to the Sugeno integral [25] of function Q with respect to a particular fuzzy measure. Namely, the measure, i in Equation (1), roughly corresponds to the number of elements in the pool such that $Q(p) \geq Q(p_i)$.

Looking at the h-index in this more abstract way, we can see that it is defined by means of these three items:

- Pool of evaluated objects
- Quality function
- Sentencing line

Note that, if we keep the same pool of papers and the quality function given by the number of citations, but we vary the slope of the sentencing line (an idea suggested in [10] and also in [28]), then, as the slope approaches ∞, the index tends to count the quality of the top publication, that is, the number of citations of the publication with most citations, while, as the slope approaches 0, the index (divided by the slope) is an estimate of the number of publications that received at least one citation. Hence, large slopes will benefit authors having a very cited publication, while small slopes will benefit authors having a lot of publications.

While in h-type indices, the objects are always the papers (see [13] for a detailed survey of h-type indices), we suggest that other objects, other quality parameters and other sentencing lines can be considered; this leads to the definition of the c-index.

The c-index of a set of objects Given a slope α and a set of objects $x_1, \ldots, x_n$ sorted by decreasing value of a quality function $Q(\cdot)$ defined on them, define the c-index of the set as

$$c_\alpha = \max\{\min(\alpha i, Q(x_i)) : i \in \{1, \ldots, n\}\} \tag{2}$$

Using the above terminology, α in Expression (2) is the slope of the sentencing line.

Equation (2) also corresponds to a Sugeno integral. In this case, we integrate the quality function Q with respect to a measure that is roughly proportional to the cardinality of the elements in the pool such that $Q(p) \geq Q(p_i)$.

3 The c-index based on citation distances

We propose a new index in which: i) the considered pool is a pool of citations, regardless of the paper to which those citations refer; ii) the quality function is given by the collaboration distance between the citing and the cited papers; iii) the slope of the sentencing line is specified by a parameter α.

We first discuss two ways of computing the collaboration distance and then we present the new index in a more formal way.

Classical collaboration distance We recall in this section the most usual notion of collaboration distance; this "classical" collaboration distance is the one considered by bibliometric databases such as MathSciNet [1].

We say that there is a collaboration path of length l with respect to a database between author a and author a' if there is a sequence of $l+1$ different authors $a_0 = a, a_1, a_2, \ldots, a_l = a'$ such that a_i and a_{i+1} are coauthors of at least one paper in the database for all i between 0 and $l-1$. We say that two different authors a and a' with at least one collaboration path between them are at distance d if d is the minimum of the lengths of the collaboration paths between a and a'. We say that a and a' are at a distance equal to ∞ if there is no collaboration path between a and a' and we say that a is at distance 0 from herself. We say that two sets of authors $\mathcal{A}$ and $\mathcal{A}'$ are at collaboration distance d with respect to a database if d is the minimum of the distances between authors in $\mathcal{A}$ and authors in $\mathcal{A}'$.

Note that all these definitions related to paths and distances depend on the database containing the information. From now on we will omit the database, although it has to be considered when computing particular examples.

Refined collaboration distance The above classical collaboration distance can be refined if the number of coauthored papers is taken into account in the following sense: the more joint papers between two authors, the closer they are.

More formally, consider a collaboration path $a_0 = a, a_1, a_2, \ldots, a_l = a'$ with respect to a database between author a and author a'. The refined length of the path is defined as

$$\sum_{i=0}^{l-1} \frac{1}{|p(a_i) \cap p(a_{i+1})|}$$

where $p(a_i)$ is the set of papers of author a_i and $|\cdot|$ is the cardinality operator.

We say that two different authors a and a' with at least one collaboration path between them are at refined distance d if d is the minimum of the refined lengths of the collaboration paths between a and a'. Also, the refined distance between two sets of authors $\mathcal{A}$ and $\mathcal{A}'$ is defined as the minimum of the refined distances between authors in $\mathcal{A}$ and authors in $\mathcal{A}'$.

Citation distance Given a citation to a certain paper by another paper, the previous definitions allow us to define the distance (classical or refined) of the citation as the collaboration distance (classical or refined) between the group of authors of the cited paper and the group of authors of the citing paper. As for our application, since we do not want this distance to depend on future collaborations, we establish the distance at the precise moment when the citation appears.

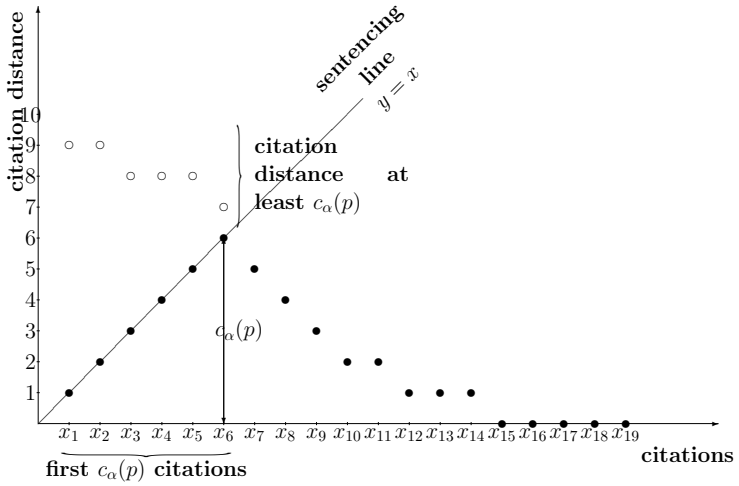

Fig. 2. c-index with slope $\alpha = 1$ for a paper p

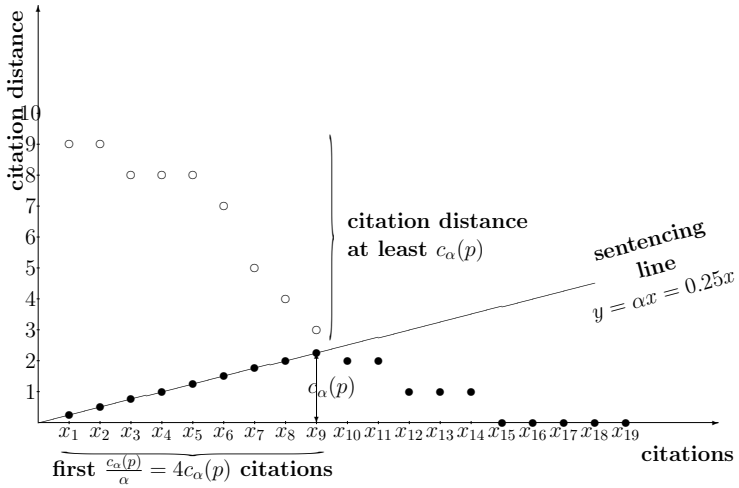

Fig. 3. c-index with slope $\alpha = 0.25$ for a paper p

The c-index of a paper Assume that objects in Expression (2) are citations to a certain paper p and define the quality $Q(x)$ of a citation x to be the citation distance (defined above in terms of collaboration distance). Assume that the set of citations $x_1, \ldots, x_n$ to paper p, is sorted by decreasing $Q(x_i)$. Then the c-index of this paper is defined as

$$c_\alpha(p) = \max\{\min(\alpha i, Q(x_i)) : i \in \{1, \ldots, n\}\} \tag{3}$$

The meaning of the outcome of Expression (3) is that there are $\frac{c_\alpha(p)}{\alpha}$ citations to p at distance at least $c_\alpha(p)$ and the rest of citations at distance at most $c_\alpha(p)$. Figures 2 and 3 illustrate the new index for $\alpha = 1$ and $\alpha = 1/4$, respectively. It is important that *any citation distance be computed at the precise moment in which the citation appears*. Otherwise, the new index would not be monotonic anymore (it could decrease over time), which would discourage cited authors from collaborating with citing authors.

Example 1 With the classical distance and $\alpha = 1$, one single citation by an external coauthor of one of the authors of a paper would mean that the paper has c-index 1, and four citations at distance four or more would mean that the paper has c-index 4, etc.

Example 2 Consider the classical distance and a paper p with two citations at distance 4, one citation at distance 3, two citations at distance 2 and one citation at distance 1. If we sort the citations by decreasing order of distance, the $(i, Q(i))$ graph, where i in the abscissae is i-th citation and $Q(i)$ in the ordinates is the distance of the i-th citation, is as follows:

$$(1, 4), (2, 4), (3, 3), (4, 2), (5, 2), (6, 1)$$

Now, if we take $\alpha = 1$ in Equation (3), we get $c_1(p) = 3$. If we take $\alpha = 1/3$, we get $c_{1/3}(p) = 1.666\ldots$, because there are $5 = 1.666\ldots/\alpha$ citations at distance at least $1.666\ldots$ and the rest is at distance at most $1.666\ldots$ (actually, the remaining citation, the sixth one, is at distance 1).

Example 3 An article p by author a_1 has obtained one citation from an article by author a_2, and another from an article by author a_3. By the time the citing article by author a_2 appears, authors a_1 and a_2 have two joint papers. When the citing article by author a_3 appears, authors a_1 and a_3 have no joint papers, but author a_3 and another author a_4 have four joint papers, and authors a_1 and a_4 have three joint papers. Using the refined distance, the distance of the citation by author a_2 is $1/2$ and the distance of the citation by author a_3 is $1/3 + 1/4 = 7/12$. If we order the two citations by decreasing distance, we get an $(i, Q(i))$ graph $(1, 7/12), (2, 1/2)$. If we now take $\alpha = 0.25$, we get $c_{0.25}(p) = 0.5$, because there are $2 = \frac{0.5}{0.25}$ citations at distance at least 0.5.

Note 1 (The c-index vs the f-index) *The c-index of a paper measures how far from the usual scientific colleagues of the authors is the paper reaching, in other words, how influential the paper is for the scientific community at large (rather than just for the usual colleagues). This aim is also shared by the f-index [18] which, for each paper p of an author, is based on a vector whose i-th component contains the proportion of citing authors for paper p who have cited it i times. For a given number of citations, the f-index of an author is higher if the vectors of his/her papers have larger values in components with low i; in plain words, the less repetitions in the set of citing authors, the higher is the f-index. We claim that the f-index may be unfair to the authors of true seminal, groundbreaking and/or reference papers: such papers are likely to be cited many times by each author, because no one can discuss the corresponding topic without referencing that paper (not just because of any collusion). By evaluating the influence of papers/authors based on the citation distance rather than the citing authors' repetitions, the c-index is fairer.*

The c-index of an author The c-index of an author a is defined by Expression (2) with objects being all the citations received by author a and the quality function being the distance of those citations. That is, if $x_1, \ldots, x_n$ are the citations to *any* of the papers by a, sorted by decreasing $Q(x_i)$, where $Q(x_i)$ is the citation distance between the cited paper and the paper containing citation x_i, then the c-index for author a can be obtained as

$$c_\alpha(a) = \max\{\min(\alpha i, Q(x_i)) : i \in \{1, \ldots, n\}\}$$

While the h-index of an author a is h if a has authored h papers with at least h citations each and the rest of papers with at most h citations, a c-index $c_\alpha(a)$ means that a has received $c_\alpha(a)/\alpha$ citations to the set of her papers (regardless of which citation goes to which paper) at citation distance at least $c_\alpha(a)$ and the rest of citations at distance at most $c_\alpha(a)$.

The c-index can also be extended to a group of authors or a research group by considering the index either over all their joint work or over all the papers authored by at least one group member.

An example to illustrate the characterization of authors by the new index follows.

Example 4 If we take slope $\alpha = 1$ and the classical distance, an author with only self-citations would have index 0, independently of the number of citations, while an author with a single paper with 6 citations at distance 6, 7 citations at distance 1 and 9 self-citations would have index 6. On the other hand, an author who always published without any coauthors would have an index equal to the number of citations from other authors if any: indeed, the distance from that author to all other authors is infinity, so $Q(x_i) = \infty$ for all citations x_i to that lonely author; hence, her index is $\max\{\alpha i : i \in \{1, \ldots, n\}\} = \alpha n$, which, for $\alpha = 1$ is the number of citations n.

An index of a group of papers (journals, proceedings, ...) Given a group of papers $g = \{p_1, \ldots, p_k\}$ corresponding to a conference, a journal volume, etc., we discuss two alternatives for assigning an index to g.

The first alternative is to compute the c-index $c_\alpha(p_i)$ of each paper using Expression (3) for $i = 1, \cdots, k$, and then compute the c-index of the paper c-indices using Expression (2) with objects being papers and the quality function being the c-index of each paper. That is, if the papers in g are sorted by decreasing order of their c-indices $c_\alpha(p_i)$, then define the C-index of the group of papers as

$$C_\alpha(g) = \max\{\min(\alpha i, c_\alpha(p_i)) : i \in \{1, \ldots, k\}\}.$$

For the second alternative, given a group g of k papers, we want a parameter x such that k times x/α citations appear at collaboration distance at least x. In other words, we compute x using Expression (2) with the slope being α/k, objects being all citations received by papers in g and the quality function of a citation x being the citation distance of it. If we denote the outcome of such a computation as $c_{\alpha/k}(g)$, we define the C'-index of the group g of papers as

$$C'_\alpha(g) = c_{\alpha/k}(g)$$

The C- and C'-indices allow assigning a numerical qualification to a journal or conference proceedings. This gives alternatives to the well-known rankings of impact factors of journals or the CORE ranking [6] for conferences in the area of computer science, as well as to the h-like indices for journals [4,24].

The advantage of the C- and C'- indices is that they are a bit more difficult to manipulate than the impact factor or the Braun-Glänzel-Schubert h-index [4]. While a journal can increase its impact factor by requiring authors of accepted papers to cite other articles recently appeared in the same journal, increasing the C-index or the C'-index additionally requires that the fake citations be to authors who are at least at a certain distance.

Small world effect and the need for the slope α By a parallel of the theory of the small world [20], it is likely that any pair (or almost any pair) of authors will be at a short collaboration distance, say, at most at distance 10. By introducing non-integer slopes α, we obtain non-integer indices which can discriminate a wider range of levels. Taking a slope $\alpha < 1$ expands the number of different levels by roughly a factor $1/\alpha$. For example, a slope $\alpha = 1/4$ would expand the number of levels from about 10 to about 40.

Thus, the small world effect upper-bounds the collaboration distance and so the c-index, while by means of the slope α we can refine its granularity. Actually, a bounded range for the c-index is convenient for users of the c-index to recognize whether a certain c-index value is high (close to the range upper bound).

We discuss next the effect of varying α on the c-index of an author. As the slope approaches ∞, the c-index of an author tends to count the largest collaboration distance from which a citation is received while, as the slope approaches 0, the index (divided by the slope) is an estimate of the number of non-self citations of the author. Hence, large slopes will benefit authors having a citation at large distance, while small slopes will benefit authors having a lot of citations.

For the refined collaboration distance, the range decreases with the number of joint papers between authors in the collaboration path. Therefore, it is harder

to give a slope α that will yield a sufficiently high number of levels for the index. At any rate, α should be smaller than for the classical distance.

4 Discussion of the new index

Scientists with few but seminal contributions Hirsch states that *for an author with a relatively low h that has a few seminal papers with extraordinarily high citation counts, the h-index will not fully reflect that scientist's accomplishments.*

The same happens for the h-index for journals. In [4] the authors made an experiment with the journal marks of 2001. They found that the first and second ranked journals of the 2001 impact factor list are not within the 60 journals with highest h-index. As the authors say, this is because of the limited number of articles in these journals. Rousseau [22] proposes an alternative index dividing the h-index by the number of articles to avoid the problem of the small number of articles; however, the problem remains of how to discriminate whether or not a few articles within the top cited h articles have citation counts much higher than h.

Egghe's g-index [11] partially mitigates the latter problem. Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the unique largest number such that the top g articles received (together) at least g^2 citations. Hence, a few very cited articles can "compensate" a number of not very cited articles, and the scientist/journal gets a g-index higher than her h-index.

However, the g-index does not help when there are *only* a few very cited articles (and no other lowly cited articles to compensate), although in [11] it is suggested to circumvent this situation by assuming the existence of fictitious publications with zero citations. The R- and AR-indices [17] tackle this problem by correcting the h-index with the number of citations of the top most cited h papers (h-core). In [12], the idea is further pursued by defining an h-index weighted by citation impact.

Since the new index is based on citations rather than on the papers to which they refer, the above problems of the h-index and the g-index are overcome (in a different way as the A-, AR- and citation-weighted h-indices do).

Self-citations and multiple-author phenomena influence Hirsch says that the h-index is not affected by self-citations. However other researchers disagree with this. It is the case of Derby [9] mentioning an example of *an*

h-index of 12 *based on* > 80% *self-citation* or Zhivotovsky and Krutovsky [29] giving a mathematical explanation of the influence. Purvis [21] criticizes that it is very easy to increase one's own h-index with self-citations: he proves this by citing in his article one of his own works completely unrelated to the article topic. Kelly and Jennions [19] argue:

> The more papers published, the more often someone can self-cite, potentially elevating h. Productive scientists often have more collaborators and students with whom they publish, and if these colleagues cite joint papers this could also elevate the focal researcher's h.

Hirsch [16] adds that

> A scientist with a high h achieved mostly through papers with many coauthors would be treated overly kindly by his or her h.

Schreiber [23] overcomes the problem of self-citations by excluding them in the definition and computation of the h-index, which was already suggested by Hirsch, although Schreiber distinguishes three different kinds of self-citations. The new index based on collaboration distances not only overcomes the problem of self-citations but also the problem of being cited by coauthors in the same small scientific community.

The problem mentioned by Hirsch related to the people having a larger h-index than deserved because of many multiauthored works is partly (although not completely) solved by the new c-index. Indeed, publishing joint papers with many coauthors tends to shorten collaboration distances and thus to worsen the c-index. At the very least, the new index deters the inclusion of gratuitous coauthors.

May the c-index discourage collaboration? Researchers should collaborate if collaboration can be expected to yield useful scientific results, more useful than the results likely to be obtained if the researchers work separately.

We acknowledge that, in some cases, using the c-index as the *only* metric may discourage collaboration. Indeed, suppose that authors A and B are doing related research and, consequently, are citing each other fairly often, although they never have co-authored a paper. By co-authoring a paper, all future citations between A and B, from the coauthors of A to B and from the coauthors of B to A will become less valuable, and this may deter A and B from collaborating.

To the defense of the c-index, we can say that, if the joint paper by A and B is really relevant to people other than themselves, it may attract them a substantial number of citations from other (maybe distant) authors and this

may outweigh the shortcoming mentioned above. However, it could indeed happen that the joint paper between A and B is quite useful and worth writing but it does not attract enough distant citations to outweigh the above negative effect. This suggests that the c-index should not be taken as the only metric of scientific performance: in Section 7 a combination of the h-index and the c-index is proposed. Coming up with a generalization of the c-index which is less sensitive to single collaborations is left as a topic for future research (see Section 7).

There are other cases in which it is not bad that the c-index discourages collaboration. Imagine that A and B mainly "live" on each other's citations and their research essentially matters only to the two of them. The c-index may indeed discourage A and B from publishing a joint paper, but it is unlikely that the lack of that paper will be a serious loss for science.

Finally, the c-index could be accused of stimulating A and B to write joint papers and sign them in turn with only one of their names. A first counterargument is that renouncing coauthorship is a high price to pay. Also, the common practice of excluding self-citations from the citation count of an author (*e.g.* as suggested by Schreiber and implemented in the CiteSeer database) could also be said to encourage this strange and far-fetched behavior, and no one pretends that self-citations ought to be counted.

On the age of scientists One of the most criticized aspects about the h-index, see for instance [3,15] (and the references therein) is that *since h values increase over time, it is apparent that a scientist's h index depends on the person's scientific age (that is, years publishing). Therefore, in ranking scientists, the h-index always puts newcomers at a disadvantage and older, well-established scientists at an advantage.*

Hirsch also mentions in his work a discussion by Redner about the fact that *most papers earn their citations over a limited period of popularity and then they are no longer cited. Hence, it will be the case that papers that contributed to a researcher's h early in his or her career will no longer contribute to h later in the individual's career.*

These facts are related to the behavior of the h-index that, as postulated by Hirsch, increases linearly. The new index does not. The production at the beginning, when the distance to all colleagues is quite large, is the most significant. Then, as time goes, it gets more complicated to obtain citations at a substantial distance. The new index rewards the beginners for their first citations. Similarly, it penalizes the appearance of Ph. D. advisors as (unde-

serving) coauthors of their students:

- The student sees her citation distance shortened, because of her advisor's (probably numerous) coauthors;
- The advisor may earn some citations to her student's work, but she loses the opportunity of earning long-distance citations from her student: a student is likely to cite her advisor, so it is better for the advisor to keep the student at a long citation distance, unless the advisor really collaborated as a joint author in the student's work. Notice that this is a situation very similar to the one explained above when discussing the effect of the c-index on collaboration.

Last but not least, the new index rewards those works that attract the attention of beginners (a citation by a beginner is likely to be at substantial distance), which is like rewarding *modernity*. Yet, a possible downside of attaching more weight to citations by beginners is that, if a beginner lacks good advice, her/his choice of citations may not be the most appropriate one due to lack of knowledge of the field.

On the adequacy of the reference database For an accurate analysis of the scientific production and impact of an author, a set of authors or a set of articles, *independently of the index or the measure*, it is very important to choose an adequate reference database gathering all the information of the related scientific field. However, this is not possible in most of the cases since the fields are not always perfectly delimited.

Computing the c-index for a scientist referred to a database which does not exactly match the scientist's research field causes two effects: a) the scientist is likely to be credited less citations than he actually received; b) the distance of those citations is likely to be higher than it would be if all papers in the literature were in the database. The first effect is negative for the scientist, while the second effect is positive. Hence, although the best option is always choosing the best database as far as possible, the c-index mitigates, to some extent, the shortcoming of using a somewhat inadequate database, should the adequate database not exist. So, we can say that the c-index is more "stable" to changes of databases.

Complexity versus correctness One of the main attractives of the h-index is its apparently fast computation. However, Cronin and Meho [7] report that, in the field of Information Science, when analyzing some important researchers, *it took roughly three hours of searching to generate the h-index for each individual author*. Also, some shortcomings have been observed in [3]

concerning the computation of the h-index. A first one is related to the difficulty of identifying an individual in a database: *unclear citations (for example "to appear" or "forthcoming"), incorrect citations (such as incorrect starting page number by the citing author) are not counted.* A second one is related to multiple invocations of the same item: *strictly speaking, the h-index for a scientist can be found easily in the Web of Knowledge only if the scientist can be identified uniquely by name or if accurate paper lists can be pulled up in Web of Knowledge by using a combination of the author name and address, or affiliation, search fields.* It is concluded that *completing paper lists and manually calculating citation counts require labor-intensive processes.*

Due to distance computation, the new index is computationally more complex than the h-index, even though:

- As argued above, the h-index is not so simple in practice as it appears in theory;
- Computation of the new index can be made more efficient in the way suggested in Section 5 below.

Last but not least, a substantial advantage of the new index is that it improves correctness, by reducing misunderstandings and misclassifications: in the c-index of an author, there is no need to identify all citations to the same paper, because citations are taken into account independently from the work to which they refer.

5 Efficient computation of the new index

It is important to address the distance computation required by the new index, because it is critical to its practical deployability. We assume in this section that the index is to be implemented within a bibliometric database, like MathSciNet [1], the Web of Knowledge [27] or DBLP [8]. To evaluate the new index for papers, authors or groups of papers, we need the collaboration distances between citing and cited authors computed *at the time each citation appeared.*

An efficient solution can be based on Floyd's algorithm [14] which, given a graph with n nodes and edge weights equal to distances between neighbor nodes, finds the shortest distances between *all pairs* of nodes in time $O(n^3)$. The algorithm returns a matrix $\mathbf{P} = \{p_{ij}\}$ of size $n \times n$, where p_{ij} is the shortest distance from node i to node j.

We can view a bibliometric database as a graph where nodes are authors and there is an edge between two authors if they have a joint paper (*e.g.* a collaboration). When the classical distance is used, the edge weight is always 1; when

the refined distance is used, the edge weight is 1 divided by the number of joint papers between the two authors. This graph increases with time, as new authors and new collaborations are added to the database. Now, the database can run Floyd's algorithm (as a batch process) at regular intervals corresponding to the time granularity with which the appearance of publications is recorded: a conservative choice is to take yearly regular intervals, although monthly intervals and even weekly intervals might be considered, because the month of appearance is available for most publications and sometimes the week too. After running the algorithm at time t, the database stores the output distance matrix $\mathbf{P}_t$. If the new index is implemented in the database at time t_0, the above applies to compute Floyd matrices $\mathbf{P}_t$ for $t \geq t_0$. However, we also need historical Floyd matrices since the time of the oldest citation made by a paper indexed in the database. To compute a matrix $\mathbf{P}_t$ with $t < t_0$, use the subgraph resulting after elimination from the database graph of all edges representing collaborations at times $t' > t$.

Assuming that matrices $\mathbf{P}_t$ have been pre-computed and stored at regular time intervals since the time of the oldest citation made by a paper indexed in the database and up to the current time interval, it is easy to compute the new index as illustrated in the example below.

Example 5 Assume that one Floyd distance matrix $\mathbf{P}$ per year is available within a certain bibliometric database. We want to compute the c-index of an article published in 1989 by authors A, B and C. We need to compute the distance of each citation received by this article. Imagine that one of the citations appears in another article published in 1990 by authors E, F and G. To evaluate the distance of this citation, we use the Floyd matrix corresponding to 1990, that is, $\mathbf{P}_{1990}$, to find the shortest distance between the sets of authors $\{A, B, C\}$ and $\{E, F, G\}$ as follows:

(1) From $\mathbf{P}_{1990}$, directly take the 9 shortest distances $A-E$, $A-F$, $A-G$, $B-E$, $B-F$, $B-G$, $C-E$, $C-F$ and $C-G$;
(2) Compute the shortest distance between $\{A, B, C\}$ and $\{E, F, G\}$ as the minimum of the above 9 distances. This minimum is the distance of the 1990 citation to the 1989 article.

Example 6 Using only one Floyd matrix per year as assumed in the previous example may be unnecessarily coarse if the month of appearance for publications is available, as it is the case for journals and some conference proceedings. Imagine an author A published a paper in 1989 which was cited by an author B, with whom A had never collaborated, in a journal article published by B in March 1990. Assume further that A and B published a joint paper in December 1990. Using one Floyd matrix per year, the distance of the March 1990 citation will be computed as 1, whereas it should be > 1. Hence, using one Floyd matrix per month would seem more accurate. Pre-computing one

monthly matrix takes more computation and more storage than one yearly matrix but, since this pre-computation is only performed once, it should be affordable.

6 Experimental results

In this section, we give experimental results on the computation of the c-index of five computer science authors on one side and of three computer science conferences on the other side.

Since there is no current bibliometric database for computer science recording both the citations to papers and the coauthorships, we took the citations from Google Scholar and we computed the collaboration distances from DBLP [8]; the data were taken from the mentioned databases in August 2010 for the experiment on authors and in December 2009 for the experiment on conferences. The DBLP database is specific to computer science and it records all the coauthors of a certain author, so that the author collaboration graph described in Section 5 can readily be obtained. Unfortunately, this is not the case for Google Scholar, so the c-index is very difficult to compute with only this database.

In addition to the problem of having to mix different databases in our computations due to the lack of a database having all the required data, Google Scholar only lists the first 1000 citations of each publication, so that the c-index can not be computed exactly.

Both problems could be overcome if the databases co-operated and provided the c-index as a built-in option.

6.1 Experiments with authors

We considered the next five authors in computer science. The first three among them are outstanding pioneers and we give short qualitative assessments of their importance and contributions based on Wikipedia.

- *Claude Elwood Shannon* (1916-2001), American mathematician and electronic engineering, known as the father of information theory and modern cryptography.
- *Alan Turing* (1912-1954), English mathematician, logician, cryptanalyst and computer scientist, who was influential in the development of computer science by formalizing the concepts of algorithm and computation through

Index	Shannon	Turing	Church	Author X	Author Y
publications in DBLP	9	5	6	4	16
citations in Google Scholar of publications in DBLP	39032	106	2120	26	47
h (DBLP + Google Scholar)	6	2	4	3	4
g (DBLP + Google Scholar)	197	10	46	5	6
cited publications (Google Scholar)	147	90	115	17	17
citations (Google Scholar)	52799	9585	8188	108	94
h (Google Scholar)	42	18	26	7	6
g (Google Scholar)	146	89	90	9	9
c-index (DBLP + Google Scholar)	893	105	693	14	13

Table 1
Numerical comparison of different measures for Shannon, Turing, Church, Author X, Author Y

the Turing machine. Unlike Shannon, though, he is not the only father of a discipline.

- *Alonzo Church* (1903-1995), American mathematician and logician who made major contributions to mathematical logic and the foundations of theoretical computer science. Among theorists of computation, he is probably somewhat less known than Turing, with whom his name is associated through the Church-Turing thesis.
- *Author X*, a young researcher with few publications whose first cited publication in DBLP is from 2006.
- *Author Y*, another young researcher with few publication whose first publication in DBLP is from 2003.

For the five authors we give four well-known measures: the number of publications, the number of citations, the h-index and the g-index. All these measures are computed on one hand using only Google Scholar through the web site `http://interaction.lille.inria.fr/~roussel/projects/scholarindex/`, and on the other hand taking the publications from DBLP and considering the citations from Google Scholar. Finally these measures are compared to the new c-index. The c-index with $\alpha = 1$ is computed taking the publications from DBLP, considering the citations from Google Scholar, and computing the collaboration distances from DBLP.

All the results are numerically shown in Table 1 and graphically shown in Figure 4. In the graphics the scales are different for each measure so that the maximum value attains the maximum height of the graphic. This makes it easier to discern the different values.

It can be observed that:

- Except for the number of publications in DBLP, Shannon is the most outstanding author in all cases.

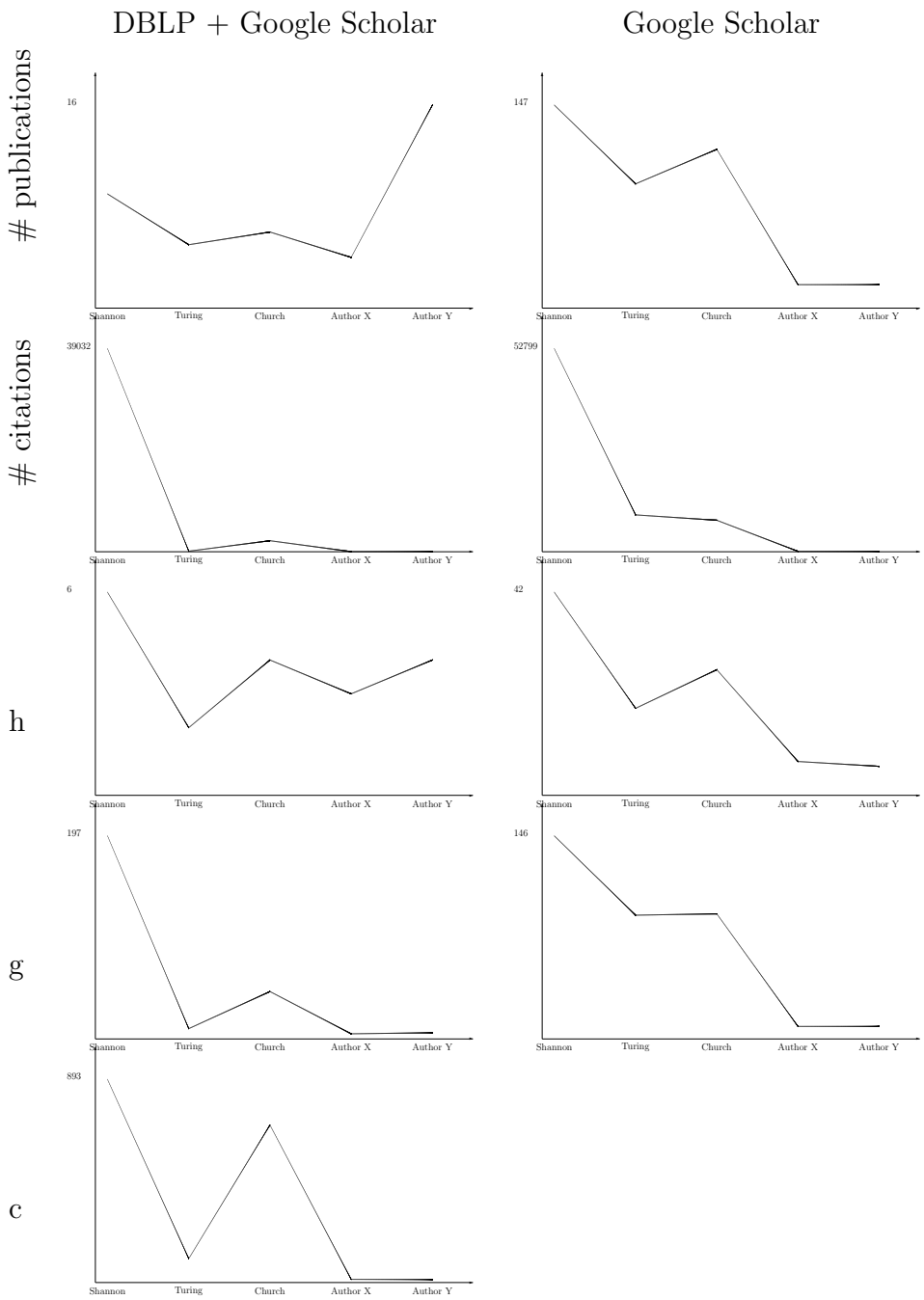

Fig. 4. Graphical comparison of different measures for Shannon, Turing, Church, Author X, Author Y

- Using only Google Scholar does much more justice to authors than using a combination of DBLP and Google Scholar. This is especially obvious for the number of publications and the h-index. Also, Turing's c-index is lower than one would expect: the reason is that Turing is underrepresented in DBLP, with only five publications which are not among his most cited ones.
- Among the results obtained using only Google Scholar, the one that seems to better reflect the above qualitative assessments of authors is the g-index.

- Among the results obtained using both Google Scholar and DBLP, the number of publications as well as the h-index are clearly unfair. As for the other indices, we analyze the ratio of the minimum value attained by the three famous computer scientists and the maximum value obtained by authors X and Y. This gives an idea on how well each index discriminates between pioneering authors with outstanding contributions and standard junior authors. The three ratios are, for the number of citations, the g-index, and the c-index, respectively, 106/47=2.255, 10/6=1.667, and 105/14=7.5. The greatest ratio occurs for the c-index, which therefore seems to have the highest discriminating capability.

6.2 Experiments with conferences

The conferences that we considered were:

- *International Conference on the Theory and Applications of Cryptographic Techniques, EUROCRYPT 2004.* The CORE conference ranking of the EUROCRYPT series was A+ in 2008 and A in 2010, after the new version of the CORE ranking suppressed category A+. EUROCRYPT 2004 accepted 32 papers which, at the time of our experiments, had received 3075 total citations. Among these, 628 citations were at collaboration distance ∞. The most cited paper had received 427 citations.
- *International Information Security Conference, SEC 2004.* The CORE ranking of the SEC series was B both in 2008 and 2010. SEC 2004 accepted 35 papers which, at the time of our experiments, had received 162 total citations. Among these, 35 citations were at collaboration distance ∞. The most cited paper had received 32 citations.
- *International Conference on Information and Communications Security, ICICS 2004.* The CORE ranking of the ICICS series was C in 2008 and B in 2010. ICICS 2004 accepted 42 papers which, at the time of our experiments, had received 365 total citations. Among these, 102 citations were at collaboration distance ∞. The most cited paper had received 46 citations.

For the three conferences we computed the Braun-Glänzel-Schubert h-index [4] and the two proposed C- and C'-indices, where we took $\alpha = 1$. The three indices for the three conferences are numerically shown in Table 2 and graphically shown in Figure 5. We give in the appendix individual graphics showing how each index was computed for each conference.

It can be seen that:

- Among the three conferences, all three indices give the top score to EUROCRYPT 2004, which confirms the CORE qualitative judgment that the EUROCRYPT conference series is "better" than the SEC and the ICICS

	h	C	C'
EUROCRYPT	23	12	17.44
SEC	7	5	3.03
ICICS	11	6	4.79

Table 2
Numerical comparison of Braun-Glänzel-Schubert's h-index, the C-index and the C'-index (the last two with $\alpha = 1$) for the EUROCRYPT 2004, SEC 2004 and ICICS 2004 conferences

series.

- For all three indices, the score of ICICS 2004 is higher than the score of SEC 2004. This contradicts the CORE ranking, but agrees with the data on citations: ICICS 2004 received more citations per paper than SEC 2004, and also the top cited ICICS 2004 paper got more citations than the top cited SEC 2004 paper.
- The difference between ICICS 2004 and SEC 2004 is smaller when collaboration distances are taken into account (C-index and C'-index) than with the h-index. This suggests that, although citations to ICICS 2004 are substantially more numerous than to SEC 2004, when collaboration distances are considered, the superiority of ICICS 2004 becomes narrower: actually, in ICICS 2004 there are 27.9% infinite citation distances against 21.6% for SEC 2004, but, if only finite distances are taken into account, the average citation distance for ICICS 2004 is 3.3, while for SEC 2004 it reaches 3.45. Thus, citations to ICICS 2004 at finite distance would be slightly more local (*i.e.* of less quality according to our quality metric based on collaboration distance) than those to SEC 2004.
- The spread of the C'-index values is greater than the spread of the C-index values (in particular, the departure of the C'-index of EUROCRYPT 2004 is much greater than the departure of its C-index). This is because the C'-index in general has a broader value range due to: i) the C'-index is computed on citations while the C-index is computed on papers, and the number of citations received by a conference is normally greater than the number of papers in it; ii) the quality function for the C'-index can be infinite (distance of citations), whereas for the C-index it is finite (c-index of papers).

7 Conclusions and future research issues

We have presented an abstraction of the h-index (and similar indices) in terms of a pool of objects, a quality function and a sentencing line. By instantiating that abstraction in a novel way, we have presented a new index, the c-index,

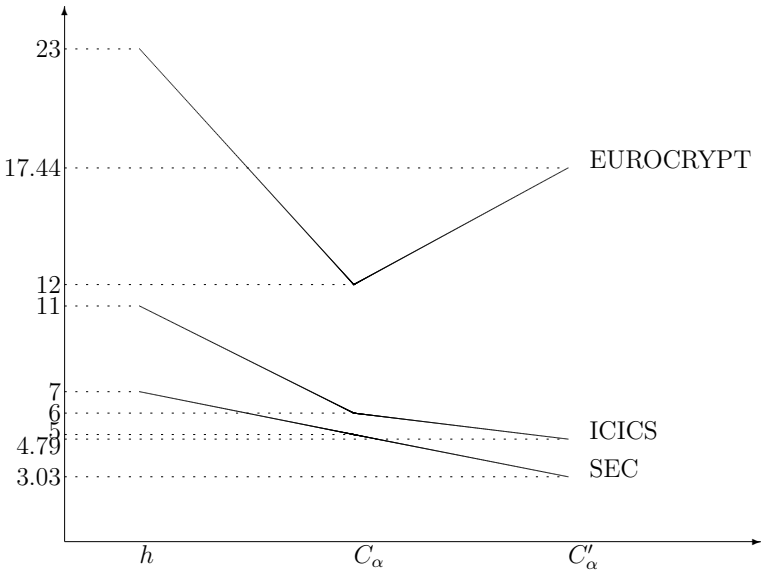

Fig. 5. Comparison of the indices C_α, C'_α and the Braun-Glänzel-Schubert's h index for the conferences EUROCRYPT, SEC, and ICICS.

where the evaluated objects are the citations received, the quality function of a citation is the collaboration distance between the authors of the citing and the cited papers, and the sentencing line can take slopes between 0 and ∞. Two types of collaboration distance have been defined: the classical one and a refined distance reflecting the number of joint papers between authors.

To the best of our knowledge, the c-index is the first one in the bibliometric literature which measures the output of a scientist or a journal based on the quantity and quality of the received citations: the more distant the citing authors, the higher the quality of a citation.

One might criticize that the conceptual advantages of the c-index discussed above (exclusion of self-citations, reward of modernity, encouragement of beginners, fight against systematic and fictitious multi-authoring of papers, etc.) are neutralized by the difficulty of computing the index. To counter this criticism, we have shown that, if the index is referred to a particular bibliometric database (*e.g.* Web of Knowledge, MathSciNet, DBLP, etc.), its computation can be efficiently automated within the database by using Floyd's algorithm for distance computation at regular time intervals. With this approach, the database can offer the c-index with little extra effort vs the h-index, especially because automating the latter index is less simple than it would seem, as it is not easy to uniquely identify each paper and accumulate all citations to it.

Finally, one might argue that the c-index is based only on citations and loses the feature of the h-index of counting how many *papers* among those of an author have had a decent impact; one may also remark that it may discourage collaborations, unless these are groundbreaking. To remedy this, one might

combine the h-index (or an h-type index) and the c-index by providing a pair (h, c) for each author (or journal), just like in [17] it is proposed to use a pair consisting of the h- and the AR-indices.

Future research will attempt to generalize the c-index in such a way that it can be used as a standalone measure of scientific performance. In particular, this involves addressing the two shortcomings mentioned in the previous paragraph. As suggested by one referee, defining the collaboration distance not as the shortest distance in the collaboration graph but as a kind of average distance might yield an index which is less sensitive to a single collaborative publication of two researchers.

Appendix

We expand here Section 6 on empirical work by giving Figures 6 through 14 which graphically show the computation of Braun-Glänzel-Schubert's h-index, the C-index and the C'-index for the EUROCRYPT 2004, SEC 2004 and ICICS 2004 conferences, respectively.

When computing the h-index for EUROCRYPT 2004, there were some papers with a number of citations much higher than the rest of papers, which made graphic depiction in Figure 6 awkward. To remedy this, we increased the scale, which caused papers with many citations to go off-range. We topcoded these very large number of citations as "≥ 50". We took the same scale in the three figures showing the h-indices of the three conferences (Figures 6, 9 and 12).

It is important to notice that, when computing the C'-index, as in the case of Figure 8, not all citation distances are needed. The reason is that, during the computation of the C'-index, an interim index value t in maintained which takes into account the citation distances computed so far. This interim index stays the same or increases with each new computed citation distance; actually, only citation distances greater than the current t may cause the interim index to increase. Hence, if we are computing the minimum collaboration distance between a set of cited authors and a set of citing authors and we find a path between one of the cited authors and one of the citing authors which is not greater than t, then we do not need to compute the minimum distance because it will be at most t and will not change the current interim value for the C'-index. In Figure 8, citations with non-computed citation distances are displayed with citation distance -1.

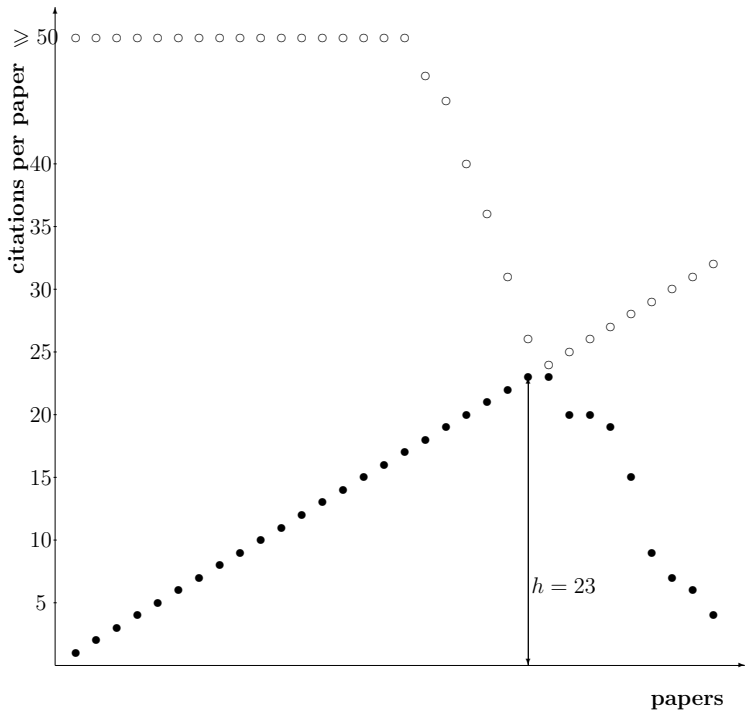

Fig. 6. Braun-Glänzel-Schubert's h index for EUROCRYPT 2004

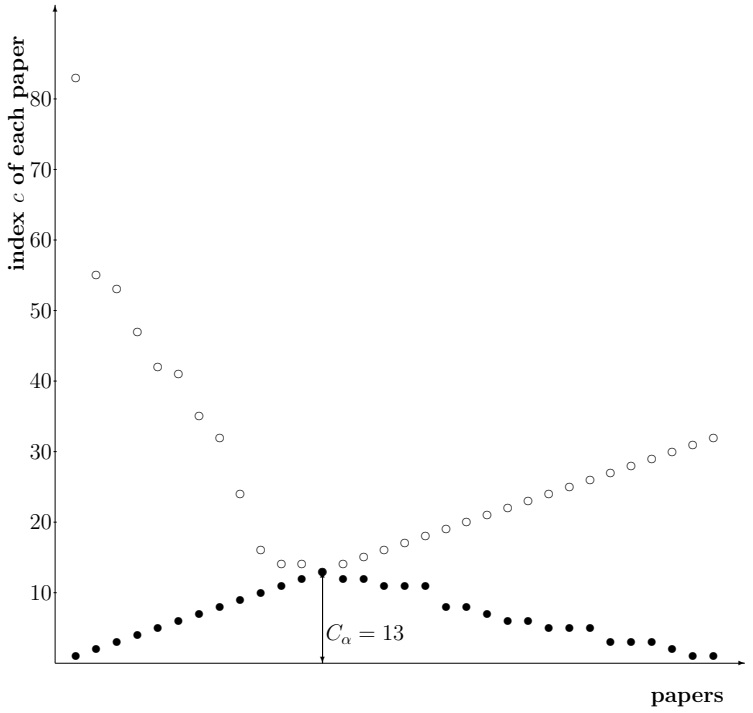

Fig. 7. C-Index with $\alpha = 1$ for EUROCRYPT 2004

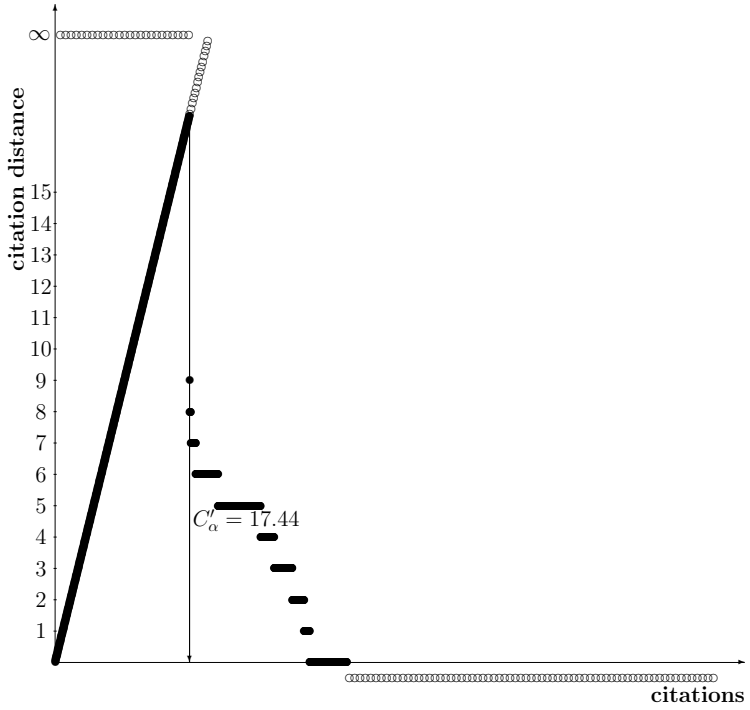

Fig. 8. C'-Index with $\alpha = 1$ for EUROCRYPT 2004

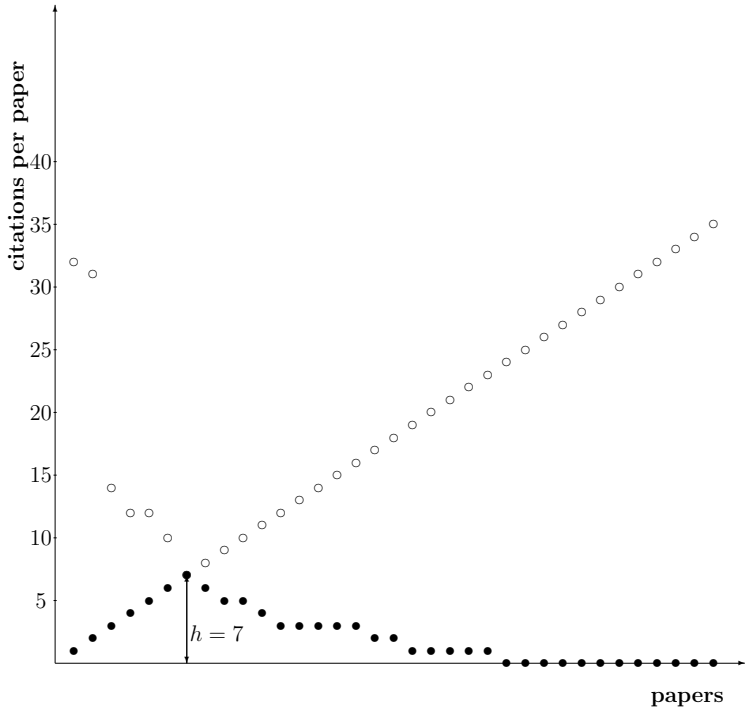

Fig. 9. Braun-Glänzel-Schubert's h index for SEC 2004

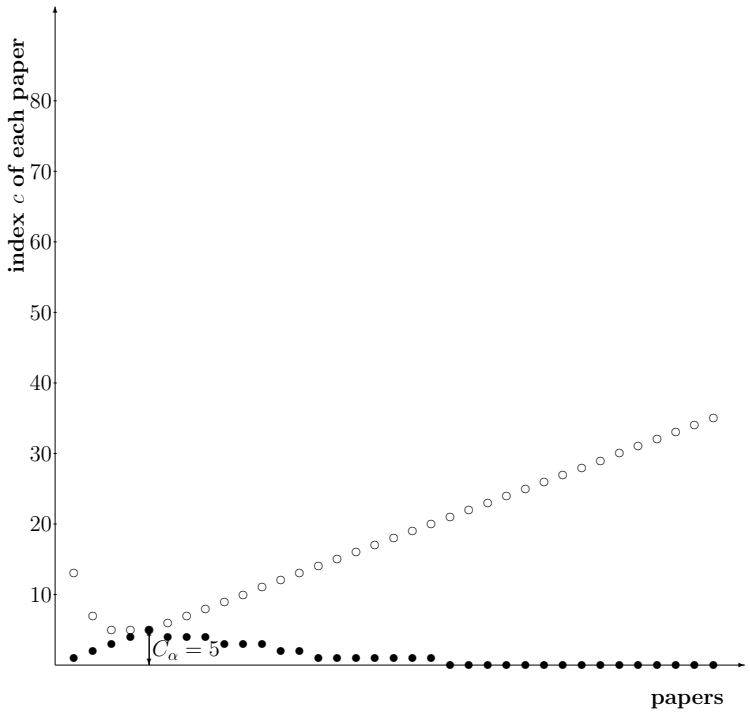

Fig. 10. Index C_α for SEC 2004

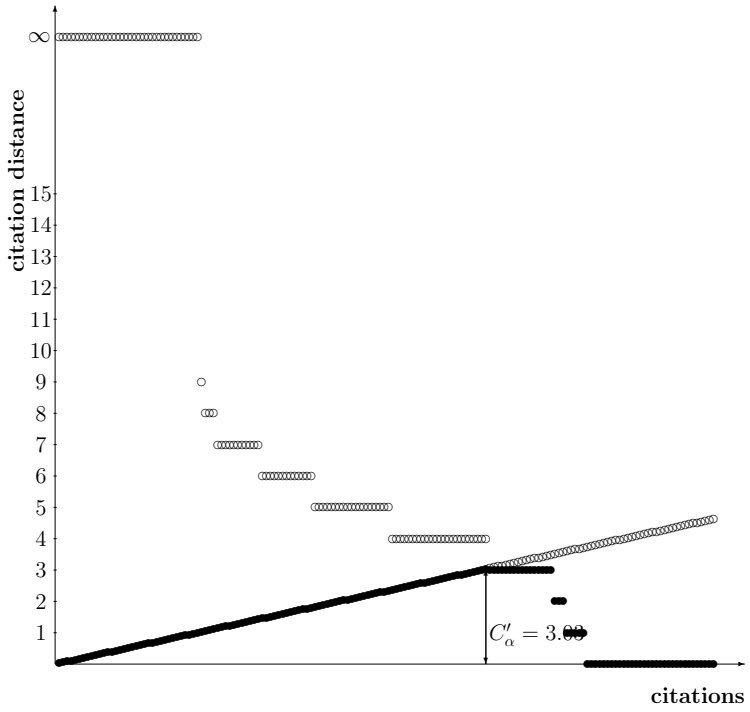

Fig. 11. Index C'_α for SEC 2004

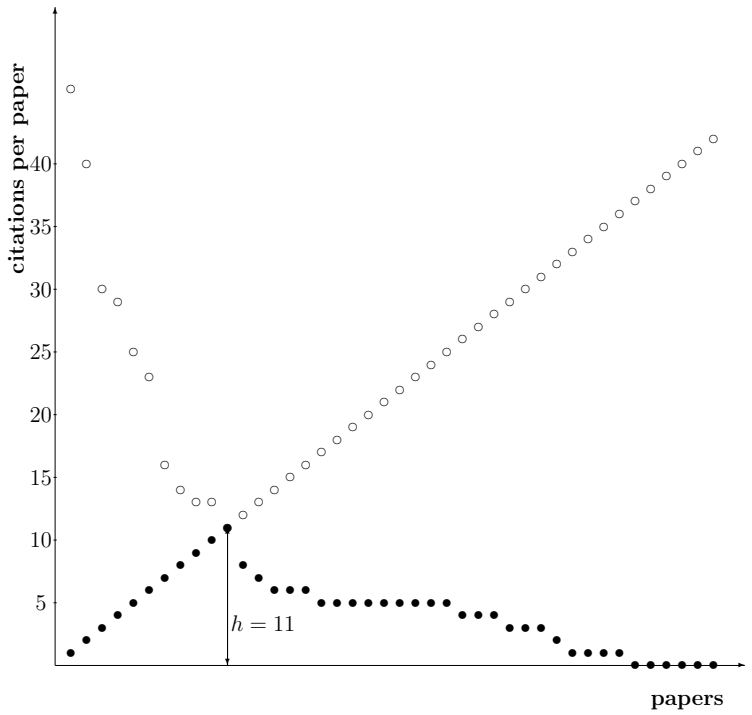

Fig. 12. Braun-Glänzel-Schubert's h index for ICICS 2004

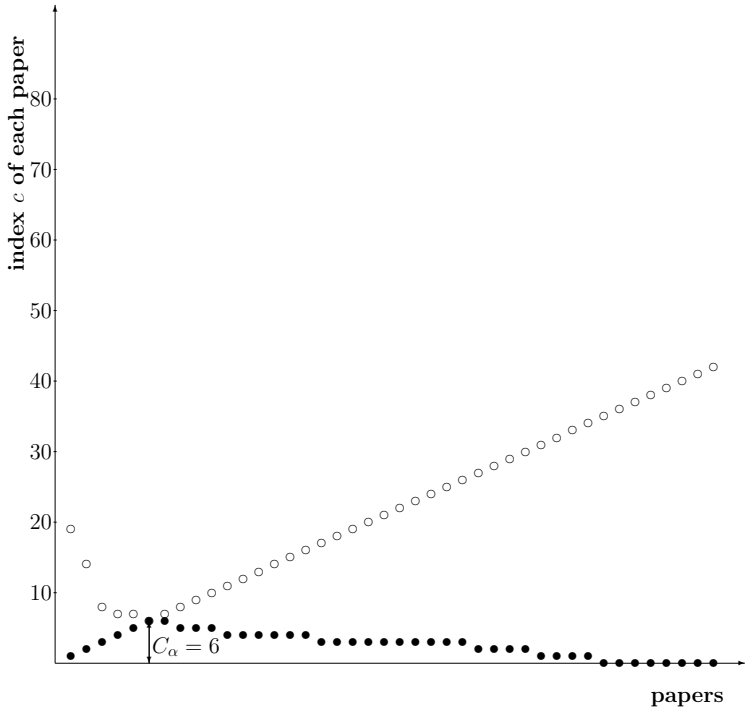

Fig. 13. Index C_α for ICICS 2004

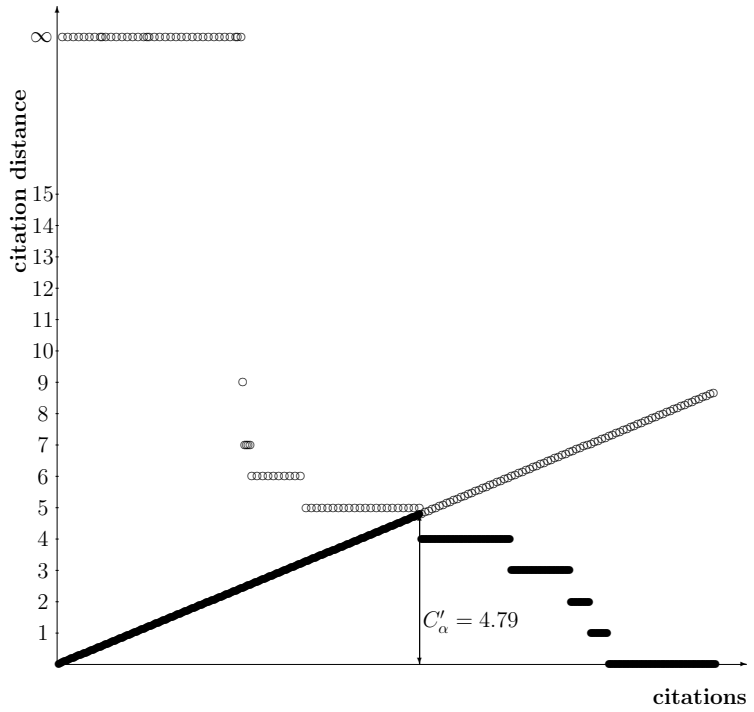

Fig. 14. Index C'_α for ICICS 2004

Acknowledgments

The authors would like to thank the anonymous referees for many interesting comments that led to substantial improvement of the paper.

References

[1] American Mathematical Society. MathSciNet: Mathematical Reviews on the Web. `http://www.ams.org/mathscinet/`

[2] P. Ball. Index aims for fair ranking of scientists. Nature, vol. 436, p. 900, 2005.

[3] L. Bornmann and H.-D. Daniel. What do we know about the h index? Journal of the American Society for Information Science and Technology, vol. 58, no. 9, pp. 1381-1385, 2007.

[4] T. Braun, W. Glänzel and A. Schubert. A Hirsch-type index for journals. Scientometrics, vol. 69, no. 1, pp. 169-173, 2006.

[5] CiteSeer.IST: Scientific Literature Digital Library. `citeseer.ist.psu.edu`

[6] The CORE Conference Rankings, `http://core.edu.au/index.php/categories/conference%20rankings`

[7] B. Cronin and L. I. Meho. Using the h-index to rank influential information scientists. Journal of the American Society for Information Science and Technology, vol. 57, no. 9, pp. 1275-1278, 2006.

[8] The DBLP Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/db/

[9] B. Derby. H-factors research metrics and self-citation. Nature Blogs, April 25, 2008. http://network.nature.com/people/U24D269FC/blog/2008/04/25/h-factors-research-metrics-and-self-citation

[10] N. J. Van Eck and L. Waltman. Generalizing the h- and g-indices. Journal of Informetrics, vol. 2, no. 4, pp. 263-271, 2008.

[11] L. Egghe. An improvement of the h-index: the g-index. ISSI Newsletter, vol. 2, no. 1, pp. 8-9, 2006.

[12] L. Egghe and R. Rousseau. An h-index weighted by citation impact. Information Processing and Management, vol. 44, no. 2, pp. 770-780, 2008.

[13] L. Egghe. The Hirsch-index and related impact measures. Annual Review of Information Science and Technology, vol. 44, pp. 65-144, 2010.

[14] R. W. Floyd. Algorithm 97: shortest path. Communications of the ACM, vol. 5, no. 6, p. 345, 1962.

[15] W. Glänzel. On the opportunities and limitations of the h-index. Science Focus, vol. 1, no. 1, pp. 10-11, 2006.

[16] J. E. Hirsch. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the USA, vol. 102, no. 46, pp. 16569-16572, 2005.

[17] B. H. Jin, L. Liang, R. Rousseau and L. Egghe. The R- and AR-indices: complementing the h-index. Chinese Science Bulletin, vol. 52, no. 6, pp. 855-863, 2007.

[18] D. Katsaros, L. Akritidis and P. Bozanis. The f-index: quantifying the impact of coterminal citations on scientists' ranking. Journal of the American Society for Information Science and Technology, vol. 60, no. 5, pp. 1051-1056, 2009.

[19] C. D. Kelly and M. D. Jennions. The h index and career assessment by numbers. Trends in Ecology and Evolution, vol. 21, no. 4, pp. 167-170, 2006

[20] S. Milgram. The small world problem. Psychology Today, vol. 2, pp. 60-70, 1967.

[21] A. Purvis. The h index: playing the numbers game. Trends in Ecology, vol. 21, no. 8, p. 422, 2006.

[22] R. Rousseau. A case study: evolution of JASIS' h-index. Science Focus, vol. 1, no. 1, pp. 16-17, 2006.

[23] M. Schreiber. Self-citation corrections for the Hirsch index. Europhysics Letters (EPL), vol. 78, pp. 30002p1-30002p6, 2007. doi:10.1209/0295-5075/78/30002

[24] A. Serenko. The development of an AI journal ranking based on the revealed preference approach. Journal of Informetrics (in press). doi:10.1016/j.joi.2010.04.001

[25] M. Sugeno. Theory of fuzzy integrals and its applications. Ph. D. Dissertation, Tokyo Institute of Technology, Tokyo, Japan, 1974.

[26] V. Torra, Y. Narukawa. The h-index and the number of citations: two fuzzy integrals. IEEE Trans. on Fuzzy Systems, vol. 16, no. 3, pp. 795-797, 2008.

[27] Thomson's Scientific. Web of Knowledge. http://isiknowledge.com

[28] Q. Wu. The w-index: a measure to assess scientific impact by focusing on widely cited papers. Journal of the American Society for Information Science and Technology, vol. 61, no. 3, pp. 609-614, 2009.

[29] L. A. Zhivotovsky and K. V. Krutovsky. Self-citation can inflate h-index. Scientometrics, vol. 77, no. 2, pp. 373-375, 2008.