

Three novel indirect indicators for the assessment of papers and authors based on generations of citations

Eleni Fragkiadaki · Georgios Evangelidis

Abstract A new indirect indicator is introduced for the assessment of scientific publications. The proposed indicator (fp^k – index) takes into account both the direct and indirect impact of scientific publications and their age. The indicator builds on the concept of generations of citations and acts as a measure of the accumulated impact of each scientific publication. A number of cases are examined that demonstrate the way the indicator behaves under well defined conditions in a *Paper-Citation graph*, like when a paper is cited by a highly cited paper, when cycles exist and when self-citations and chords are examined. Two new indicators for the assessment of authors are also proposed (fa – index and fas – index) that utilize the fp^k – index values of the scientific publications included in the Publication Record of an author. Finally, a comparative study of the fp^k and fa^k indices and a list of well known direct (Number of Citations, Mean number of citations, Contemporary h-index) and indirect (PageRank, SCEAS) indicators is presented.

1 Introduction

Scientific publications are responsible for disseminating the research results and achievements of scientists and scientific groups. The term describes any scientific document that has been peer reviewed and published in a way that can assist other researchers and be referenced in their work. Different types of scientific documents can be considered, like master and doctoral theses, review articles, conference papers and journal articles, technical reports and documents, books and book chapters, short communications and commentaries. In the rest of the paper, the term *paper* will be used to describe any of the above items and the term *author* for scientists and researchers that publish papers

Published papers do carry knowledge and their content has passed through a review process prior to their publication. Therefore, there is value attached to every

Department of Applied Informatics, University of Macedonia, 156 Egnatia St, GR-54636, Thessaloniki, Greece

E-mail: eleni.fra@gmail.com · gevan@uom.edu.gr

published paper, though not all published papers have the same impact on their respective field. Several bibliometric indicators have been proposed to evaluate the importance of a paper and/or its acceptance by the scientific community.

The most fundamental indicator for assessing the scientific impact of a paper is the total number of citations received. A number of researchers have argued that the importance of a paper should be considered by examining not only its direct impact but also the impact of the papers that have cited it (Rousseau 1987; Dervos and Kalkanis 2005; Sidiropoulos and Manolopoulos 2005; Walker, Xie, Yan, and Maslov 2007; Ma, Guan, and Zhao 2008; Maslov and Redner 2008; Yan, Ding, and Sugimoto 2011; Hu, Rousseau, and Chen 2011; Egghe 2011b; Su, Pan, Zhen, Ma, Yuan, Guo, Yu, Ma, and Wu 2011). By doing so, one considers not only the visibility of the paper but also its prestige.

Consequently, a number of indirect indicators have been proposed, some of which are alterations or adaptations of the PageRank algorithm that was originally defined for ranking pages on the web (Page, Brin, Motwani, and Winograd 1999). More specifically, Ma, Guan, and Zhao (2008) propose the application of PageRank to citation analysis and they have adapted the damping factor to better represent the walk of a random “researcher” rather than a random “surfer” (Chen, Xie, Maslov, and Redner 2007). CiteRank (Walker, Xie, Yan, and Maslov 2007; Maslov and Redner 2008) is another example of a PageRank based algorithm for assessing a paper that takes into account the age of the paper in order to increase its probability of being the starting point of a random walk. Prestige-Rank (Su, Pan, Zhen, Ma, Yuan, Guo, Yu, Ma, and Wu 2011) was proposed in order to account for the incompleteness of the *Paper-Citation graph*, which originates from the fact that no bibliometric database does actually include all the citations given to a particular paper. P-Rank (Yan, Ding, and Sugimoto 2011) is another PageRank based indicator that utilizes the *Paper-Citation graph* and information about the co-authors of the papers and the journals in which the papers have been published in.

SCEAS Rank (Sidiropoulos and Manolopoulos 2005) takes a similar approach to PageRank but introduces an indicator that defines the contribution of direct citations to be greater than the contribution of indirect citations. It also specifies that indirect citations should have a greater impact on papers in their neighborhood rather than to distant papers. We examine both of these principles in this paper. Another example is the Cumulative patent citations and the Weighted cumulative patent indicators (Atallah and Rodríguez 2006) that do not originate from PageRank but follow a different approach in evaluating indirect citations. These indicators were originally defined for a *Patent-Citation graph*, a network identical to the *Paper-Citation graph* if patents are replaced by papers. Their aim was to measure the impact of a patent by considering the direct and indirect citations received and the closeness of citations to the patent under scrutiny. Finally, another approach was followed in Fragkiadaki, Evangelidis, Samaras, and Dervos (2011) where the f -value indicator accounts for all indirect citations and includes a reducing factor that can be used to simulate the different citation patterns between different scientific fields.

Apart from the indirect indicators for the assessment of papers, a number of indirect indicators have also been proposed for the assessment of authors. SARA (Radicchi, Fortunato, Markines, and Vespignani 2009) is an indicator that follows a PageR-

ank approach applied to the a *Weighted Author-Citation graph* but with slight differences, mainly around the distribution of impact from dangling nodes (authors that do not appear to cite any other author in the graph). Another indicator that constructs and uses the *Author-Citation graph* has been proposed by Fiala, Rousselot, and Ježek (2008); Fiala (2012). The authors introduce a modification of PageRank where citations between authors are examined individually based on a number of factors, like the total number of publications of each author, the number of common publications between two authors, the number of distinct co-authors, the number of citations from one author to the other, as well as the year of each author to author citation . Another approach was followed by Kosmulski (2010) and Egghe (2011a,b). Both authors propose an indirect indicator based not only on the direct citations of a paper but also on the direct citations received by the citing papers (second generation citations). They choose to apply these indicators over a different set of papers included in the Publication Record of an author, thus, producing different results meant to be used either as standalone (*hfg-index*) or as complementary (Indirect *h-index*). Finally, Hu, Rousseau, and Chen (2011) propose the use of Generational indices as indirect indicators calculated per generation of citations with regards to a target paper and the use of Cross-generational indices as cumulative measurements of impact.

To summarize, there are a number of indirect indicators that one can use in order to assess the impact of a paper or author depending on the criteria at hand.

The first indicator proposed in this paper, *fp^k-index*, considers several aspects of the *Paper-Citation graph* like the existence of cycles, the existence of more than one citation paths of the same or different length from a source paper to a target paper as well as the scientific age of the paper in order to produce the individual paper scores. The next two indicators proposed, *fa-index* and *fas-index*, are based on the individual *fp^k-index* values of the papers included in the Publication Record of an author. These indicators provide the means for assessing an author and we demonstrate that they are time aware and, in most cases, size independent. In addition, *fas-index* also accounts for the existence of self-citations for the individual authors of a paper.

In Section 2, the *Paper-Citation graph* is presented in detail along with the different types of citation generations and some of the properties of the graph are discussed in more detail, like self-citations, chords and cycles. Section 3 further discusses citation generations and presents an example of the application of citation generations and citation generation counts in order to justify the reasons behind the type selected for the indicators introduced in this paper. In Section 4, the *fp^k-index* indicator is defined and two examples of its application are presented in Section 5. In that section, we compare *fp^k-index* to two well known indicators for the assessment of papers, namely, the Citation count and PageRank. The *fa-* and *fas-index* are defined in Section 6 and an application of both indicators is given in Section 7. Section 8 presents a comparative study of the proposed indicators to other well known indicators of direct and indirect impact found in the literature, along with experimental results for the rankings produced by each indicator based on the data provided by DBLP . Finally, the paper concludes in Section 9.

2 Theoretical background

We present an overview of the Citation graph along with a definition of the available meta-data information for each paper participating in a closed paper collection. In addition, the generations of citations are examined in detail and a thorough example of the four types of forward generations is discussed. Generations of self-citations and the concept of chords are also considered.

2.1 Citation graph

Citation graphs are constructed from the meta-data available for the papers included in a closed set of papers. The base form of a citation graph is the *Paper-Citation graph*, but there are other types of derived graphs like the *Author-Citation graph* and the *Journal-Citation graph*. Derived graphs are constructed from the *Paper-Citation graph* by applying appropriate transformations as presented in [Fragkiadaki and Evangelidis \(2014\)](#). Here, we only present the *Paper-Citation graph* along with the notations used throughout this paper to describe the different properties of this graph.

The *Paper-Citation graph* is a directed graph whose nodes are the papers included in the collection and edges are defined based on the citations present in the Reference lists of these papers. A directed edge from a source paper (S) to a target paper (T) exists if the source paper (S) includes the target paper (T) in its list of references. We denote this relationship between papers S and T as “S references T” or “T is cited by S”, and the corresponding notation for this edge is $S \rightarrow T$.

Apart from the papers and the citation data, the *Paper-Citation graph* includes additional information originating from the meta-data available for each paper. These information include the author list of each paper, the publication year and the publication journal. The different entities participating in this *Paper-Citation graph* along with the different properties of the graph are described by the following notations, as they were first presented in [Fragkiadaki and Evangelidis \(2014\)](#):

- $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_{\mathbf{NP}}\}$ denotes the closed set of papers participating in a *Paper-Citation graph* and \mathbf{NP} is the total number of papers included in the collection.
- $\mathbf{A} = \{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_{\mathbf{NA}}\}$ denotes the set of authors that have participated in any of the papers included in the *Paper-Citation graph*. \mathbf{NA} denotes the total number of authors participating in the *Paper-Citation graph*.
- $\mathbf{J} = \{\mathbf{J}_1, \mathbf{J}_2, \dots, \mathbf{J}_{\mathbf{NJ}}\}$ denotes the set of journals in which the papers of the *Paper-Citation graph* where published. \mathbf{NJ} denotes the total number of journals participating in the *Paper-Citation graph*.

An example of a *Paper-Citation graph* can be found in [Figure 1](#). Using the notations presented earlier the following for this graph:

- $P = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7\}$ is the set of papers in our collection and $NP = 7$
- $A = \{A_1, A_2, A_3, A_4, A_5\}$ is the set of authors and $NA = 5$
- $J = \{J_1, J_2, J_3\}$ is the set of journals and $NJ = 3$

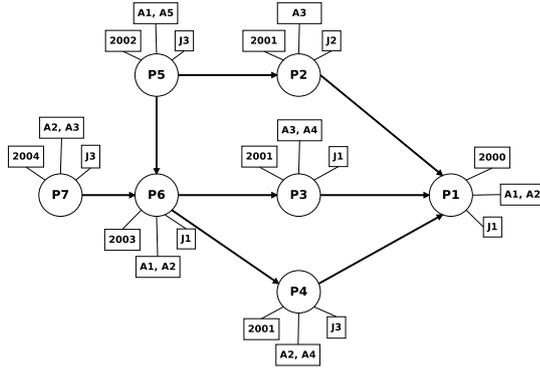


Fig. 1: Example Paper-Citation graph

Paper	Publication Year	Journal	Co-authors	References	Is cited by
P_1	2000	J_1	A_1, A_2	-	P_2, P_3, P_4
P_2	2001	J_2	A_3	P_1	P_5
P_3	2001	J_1	A_3, A_4	P_1	P_6
P_4	2001	J_3	A_2, A_4	P_1	P_6
P_5	2002	J_3	A_1, A_5	P_2, P_6	-
P_6	2003	J_1	A_1, A_2	P_3, P_4	P_5, P_7
P_7	2004	J_3	A_2, A_3	P_6	-

Table 1: *Paper-Citation table* for the *Paper-Citation graph* of Figure 1.

The *Paper-Citation graph* of Figure 1 may also be presented in the form of a table, which we call the *Paper-Citation table* and for our sample graph is shown in Table 1. Each row of the table describes a particular paper and includes the list of co-authors, the publication year and publication journal, the list of papers referenced by the paper and the list of papers that directly cite the paper.

2.2 Citation generations

We refer to citations received by a paper as direct citations and to the citations received via its citing papers as indirect citations. The term *citation path* is used to denote that a path exists in the *Paper-Citation graph* between a source and target paper. *Citation paths* can be categorized based on their length, which is the number of papers participating in the path excluding the target paper. Therefore, all direct citations are of length 1 since the path includes only one paper apart from the target paper, and, all indirect citations are of length greater than one. The *citation paths* for paper P_1 of Figure 1 are listed in Table 2a. We observe that paper P_1 has 3 citation paths of length 1 (or 3 1-gen citations), 3 citation paths of length 2 (or 3 2-gen citations) and 4 citation paths of length 4 (or 4 4-gen citations).

Citation path	Source paper	via	Target paper	
length 1	P_2		P_1	
	P_3		P_1	
	P_4		P_1	
length 2	P_5	P_2	P_1	
	P_6	P_3	P_1	
	P_6	P_4	P_1	
length 3	P_5	P_6	P_3	P_1
	P_5	P_6	P_4	P_1
	P_7	P_6	P_3	P_1
	P_7	P_6	P_4	P_1

(a)

	m (non-unique)	s (unique)
Restricted (G)	$G_0^m = \{P_1\}$	$G_0^s = \{P_1\}$
	$G_1^m = \{P_2, P_3, P_4\}$	$G_1^s = \{P_2, P_3, P_4\}$
	$G_2^m = \{P_5, P_6, P_6\}$	$G_2^s = \{P_5, P_6\}$
	$G_3^m = \{P_5, P_5, P_7, P_7\}$	$G_3^s = \{P_5, P_7\}$
Independent (H)	$H_0^m = \{P_1\}$	$H_0^s = \{P_1\}$
	$H_1^m = \{P_2, P_3, P_4\}$	$H_1^s = \{P_2, P_3, P_4\}$
	$H_2^m = \{P_5, P_6, P_6\}$	$H_2^s = \{P_5, P_6\}$
	$H_3^m = \{P_7, P_7\}$	$H_3^s = \{P_7\}$

(b)

Table 2: (a) Direct and indirect citation paths for paper P_1 of Figure 1; (b) Forward citation generations for paper P_1 of Figure 1.

The indirect citations are used to define the generations of citations originally proposed by Rousseau (1987). In that paper, generations are discussed from the references point of view and their influence over the current paper is examined. These generations are called backwards while generations created based on the citations received by a paper are called forward. Forward generations have also been discussed in the literature by Dervos and Kalkanis (2005); Dervos, Samaras, Evangelidis, and Foliás (2006); Atallah and Rodríguez (2006) and by Hu, Rousseau, and Chen (2011) where four different definitions of generations were proposed. The definitions take into account the existence or not of duplicate papers per generation and whether a paper already included in a generation participates or not in a higher rank generation. The following notations defined in Hu, Rousseau, and Chen (2011) are used throughout the rest of paper:

- Subscript $n = 0, \dots, M$ defines the individual generations for a particular paper, with M being the youngest generation or in other terms the longest path in the Citation graph leading to the current paper. Forward generations are denoted with a positive natural number whereas Backward generations are denoted with a negative whole number.

- G denotes that a citing paper can appear in many generations and H denotes that generations can only include papers not already included in a previous generation.
- Superscript s denotes that a paper can only be included once in a generation and superscript m denotes that a paper can be included more than once in a generation (definitions of sets and multi-sets from [Hu, Rousseau, and Chen \(2011\)](#)).

In the original paper of [Hu, Rousseau, and Chen \(2011\)](#), the definition of 0-gen set encapsulates the possibility of including more than one papers, like for example all papers co-authored by a single author, but we are going to consider Generation 0 to only include a single target paper.

The different sets of forward citation generations for target paper P_1 based on the four types of definitions one can get for the possible combinations of values $\{G, H\}$ and $\{m, s\}$ are listed in Table 2b. The table reveals that all definitions yield identical results for 0-gen and 1-gen sets. 0-gen set includes only the paper under scrutiny and 1-gen set includes papers directly citing the target paper. Since a paper cannot cite itself and can cite another paper only once, there are no duplicates in 1-gen set.

The four definitions produce different results starting from the 2-gen set and moving forward. In particular, the 2-gen set demonstrates the different results obtained based on whether a paper is allowed to be included more than once per generation or not (definitions of superscripts m and s respectively). In the former case (m), paper P_6 is included twice in the 2-gen set of citations, whereas in the latter (s) it is listed once. So, the s/m aspect of the definitions determines whether duplicates can be found within a generation. In other words, it determines if a generation is to be considered as the unique list of source papers that provide the target paper with at least one citation path of a particular length (s) or as a listing of the source papers of all citation paths of a particular length (m). Tables 2a and 2b better demonstrate the above statement. Paper P_6 is the source paper of two 2-gen citations for target paper P_1 , one via paper P_3 and one via paper P_4 . So, in the m definitions paper P_6 is included twice whereas in the s definitions it is included once.

The G/H aspect of the definitions is better illustrated by 3-gen citations and particularly by the citations originating from paper P_5 . When the generations are defined as G , paper P_5 is a 3-gen citation for paper P_1 , whereas if the generations are defined as H , it is not. In the second case, paper P_5 is not a 3-gen citation because it has already been counted as a 2-gen citation for paper P_1 . In other words, the G/H aspect of the definitions determines whether a source paper that provides more than one citation paths of different length for the target paper should be included in all generations based on its citation paths or if it should only be included in the generation closest to the target paper.

2.3 Generations of self-citations

When a *Paper-Citation graph* is examined from the paper point of view, the authors of the papers do not really participate in the process. But if we choose to examine the papers with regards to their contribution to the Publication Record of a particular author, one might wish to include extra information that relates to the author in ques-

tion. In that sense, we say that there exists a direct self-citation between papers P_1 and P_2 for author A_1 , if paper P_2 cites paper P_1 and A_1 has co-authored both papers.

Citation path	Source Paper	Co-authors	Via	Target		Self citation	
				Paper	Author		
Length 1	P_2	A_3		P_1	A_1 A_2		
	P_3	A_3, A_4		P_1	A_1 A_2		
	P_4	A_2, A_4		P_1	A_1 A_2	x	
Length 2	P_5	A_1, A_5	P_2	P_1	A_1 A_2	x	
	P_6	A_1, A_2	P_3	P_1	A_1 A_2	x	
	P_6	A_1, A_2	P_4	P_1	A_1 A_2	x	
Length 3	P_5	A_1, A_5	P_6	P_3	P_1	A_1 A_2	x
	P_5	A_1, A_5	P_6	P_4	P_1	A_1 A_2	x
	P_7	A_2, A_3	P_6	P_3	P_1	A_1 A_2	x
	P_7	A_2, A_3	P_6	P_4	P_1	A_1 A_2	x

Table 3: Direct and indirect citation paths for paper P_1 of Figure 1. Self-citations are considered at the (paper, author) level for the list of co-authors of paper P_1 .

When one wishes to account for the existence of self-citations, it is a common practice to examine a paper at the author level by either simply counting the number of self-citations and supplying this number alongside the full citation count or by completely removing the self-citations from the list of citations for the paper and author in question. So, in the same sense that self-citations are defined for a particular (paper, author) pair in the case of direct citations, we define the generations of self-citations for a (paper, author) pair for all indirect citations. This concept has been originally discussed in the Cascading-Citations Indexing Framework (cc-IF) defined in [Dervos, Samaras, Evangelidis, and Folias \(2006\)](#), where the generations of self-citations were defined as forward G^m .

In general, a n -gen self-citation for a (paper, author) pair (P, A) is defined by a citation path of length n originating from a source paper and ending at paper P , with author A being present in the author list of both papers. Therefore, in the definition of a self-citation the only points of interest are the source and target papers and the corresponding authors. For example, the citation path $P_6 \rightarrow P_3 \rightarrow P_1$ is considered a 2-gen self-citation for author A_1 , but the citation path $P_7 \rightarrow P_6 \rightarrow P_3 \rightarrow P_1$ is simply considered a 3-gen citation even though it passes through a paper co-authored by A_1 .

Thus, we may amend Table 2a to also include the authors of the papers, along with a characterization of which citation paths are considered self-citations for each of the authors in the author list of paper P_1 . The results are presented in Table 3.

	m (non-unique)	s (unique)
Restricted (G)	$G_0^m = \{P_1\}$	$G_0^s = \{P_1\}$
	$G_1^m = \{P_2, P_3, P_4\}$	$G_1^s = \{P_2, P_3, P_4\}$
	$G_2^m = \{\}$	$G_2^s = \{\}$
	$G_3^m = \{P_7, P_7\}$	$G_3^s = \{P_7\}$
Independent (H)	$H_0^m = \{P_1\}$	$H_0^s = \{P_1\}$
	$H_1^m = \{P_2, P_3, P_4\}$	$H_1^s = \{P_2, P_3, P_4\}$
	$H_2^m = \{\}$	$H_2^s = \{\}$
	$H_3^m = \{P_7, P_7\}$	$H_3^s = \{P_7\}$

(a)

	m(non-unique)	s (unique)
Restricted (G)	$G_0^m = \{P_1\}$	$G_0^s = \{P_1\}$
	$G_1^m = \{P_2, P_3\}$	$G_1^s = \{P_2, P_3\}$
	$G_2^m = \{P_5\}$	$G_2^s = \{P_5\}$
	$G_3^m = \{P_5, P_5\}$	$G_3^s = \{P_5\}$
Independent (H)	$H_0^m = \{P_1\}$	$H_0^s = \{P_1\}$
	$H_1^m = \{P_2, P_3\}$	$H_1^s = \{P_2, P_3\}$
	$H_2^m = \{P_5\}$	$H_2^s = \{P_5\}$
	$H_3^m = \{\}$	$H_3^s = \{\}$

(b)

Table 4: (a) Forward citation generations for paper P_1 and author A_1 of Figure 1 and, (b) Forward citation generations for paper P_1 and author A_2 of Figure 1.

We propose that when a paper is examined as part of the Publication Record of an author it should be determined whether self-citations should be included or not in the generations of citations. If self-citations are included, then the results for the four definitions of citations are the same as the ones shown in Table 2b. If self-citations are to be excluded from the citation generations for a particular author, then the results are shown in Tables 4a and 4b for authors A_1 and A_2 of paper P_1 .

It is interesting to examine 2-gen and 3-gen citations for author A_1 in Table 4a. After removing all self-citation paths for author A_1 , there is no citation path of length 2 left, which means that all 2-gen citations originate from papers co-authored by A_1 . This has as a consequence that generation 2 of citations for A_1 is empty. This does not necessarily imply that A_1 will not have any 3-gen citations since, as we have already mentioned, self-citations are only defined using the starting and ending points of the citation paths without examining the intermediate papers. Thus, even though A_1 has no 2-gen citations (by any definition), he still has some 3-gen citations.

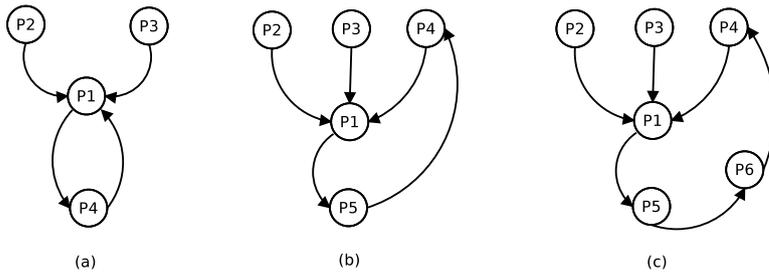


Fig. 2: Examples of different levels of citation cycles encountered in *Paper-Citation graphs* (a) Level 1 cycle (b) Level 2 cycle and, (c) Level 3 cycle.

2.4 Chords

Another aspect of the *Paper-Citation graph* that is related to the generations of citations is the existence of chords within the graph. Chords (Dervos and Kalkanis 2005) are defined as citations in the *Paper-Citation graph* of rank greater than one that co-exist with a 1-gen citation. So, a chord of rank 2, or 2-chord, exists between papers A and B when there is a 2-gen citation from paper A to paper B while at the same time there is also a 1-gen citation from A to B . This models the situation where a paper cites both directly and indirectly another paper in the citation graph.

2.5 Cycles

The *Paper-Citation graph* is a directed graph due to the nature of the connections between papers. While one might expect that the *Paper-Citation graph* is also acyclic, this is not always true. It is not uncommon for a paper to cite a version of another paper appearing in draft mode on the personal web page of one of the authors or to cite an online first edition of a paper (a paper made available online prior to its original publication). This may create cycles in the *Paper-Citation graph* Sidiropoulos and Manolopoulos (2005) and these cycles may be of different levels.

We define a *Level 1 cycle* to be any path of the form $S \rightarrow T \rightarrow S$ and a *Level n cycle* any path of the form $S \rightarrow \dots \rightarrow S$ where $n + 1$ papers participate in the formation of the path with $n \geq 1$. Figure 2 presents three different levels of cycles with regards to paper P_1 .

In Figure 2 we observe that in (a), P_1 participates in a *Level 1 cycle* via the path $P_1 \rightarrow P_4 \rightarrow P_1$, in (b), P_1 participates in a *Level 2 cycle* via the path $P_1 \rightarrow P_5 \rightarrow P_4 \rightarrow P_1$ and, finally, in (c), P_1 participates in a *Level 3 cycle* via the path $P_1 \rightarrow P_5 \rightarrow P_6 \rightarrow P_4 \rightarrow P_1$.

3 The meaning of generations of citations

So far, we have examined the different types of generations that can be defined based on the data included in a *Paper-Citation graph*, but we have not explored the meaning of indirect citations. We believe that a direct citation clearly indicates that a paper has been influenced in some way by the papers that it cites. The way that the referenced papers have affected the research of an author might not always be the preferred one, for example one might mention negative results based on another author's work but nevertheless the citation does mean that the cited paper has had an impact on the citing paper.

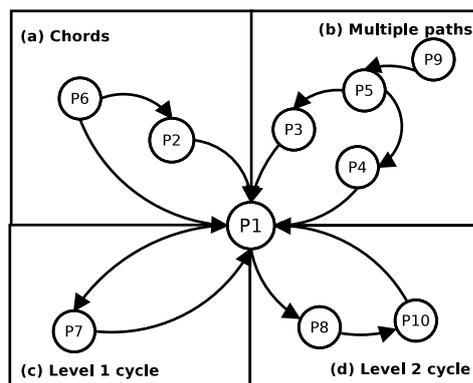


Fig. 3: A *Paper-Citation graph* that demonstrates four different types of citations paths (a) chords, (b) multiple citation paths of length n , $n > 1$ from a source paper to the target paper P_1 , (c) a *Level 1 cycle* and, (d) a *Level 2 cycle*.

But what do indirect citations mean and how should they be counted for? From the point of view that direct citations express a connection (or some form of influence) between two papers, we believe that indirect citations should carry the same meaning. In particular, an indirect citation should represent an imaginary connection between a source and a target paper with citations closer to the target paper (of lower rank) representing a stronger relationship between the papers. Based on the above and building on the concept of the Medal Standings Output table (MSO table) presented in [Dervos and Kalkanis \(2005\)](#), it is possible to create a table of the papers included in a *Paper-Citation graph* along with counts of the first n -gen citations of the papers based on the desired definition of generations.

The only question remaining now is which definition should one use for the generations of citations and how does that affect the output of the MSO table. Let us consider the *Paper-Citation graph* of Figure 3 that consists of ten papers, $P = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}\}$ and 13 edges that represent the 13 direct cita-

Citation path	Source paper	Via	Target paper	
Length 1	P_2		P_1	
	P_3		P_1	
	P_4		P_1	
	P_6		P_1	
	P_7		P_1	
	P_{10}		P_1	
Length 2	P_1	P_7	P_1	
	P_5	P_3	P_1	
	P_5	P_4	P_1	
	P_6	P_2	P_1	
	P_8	P_{10}	P_1	
Length 3	P_1	P_8	P_{10}	P_1
	P_2	P_1	P_7	P_1
	P_3	P_1	P_7	P_1
	P_4	P_1	P_7	P_1
	P_6	P_1	P_7	P_1
	P_7	P_1	P_7	P_1
	P_9	P_5	P_3	P_1
	P_9	P_5	P_4	P_1
	P_9	P_5	P_4	P_1
	P_{10}	P_1	P_7	P_1

Table 5: *Paper-Citation table* for paper P_1 presented in Figure 3.

tions that exist between the papers. The *Paper-Citation table* for paper P_1 is shown in Table 5.

Figure 3 demonstrates four different citation paths that a paper may participate in. In the lower left corner, paper P_1 is part of a *Level 1 cycle* via the path $P_1 \rightarrow P_7 \rightarrow P_1$, whereas in the lower right corner, P_1 is part of a *Level 2 cycle* via the path $P_1 \rightarrow P_8 \rightarrow P_{10} \rightarrow P_1$. In the top left corner, P_1 is the target of a 2-gen citation originating from P_6 , which also provides a 1-gen citation to P_1 . Thus, the 2-gen citation from P_6 to P_1 is also a 2-chord. Finally, in the top right corner, paper P_5 provides two 2-gen citations to P_1 via papers P_3 and P_4 respectively, whereas, P_9 provides two 3-gen citations to P_1 via paths $P_9 \rightarrow P_5 \rightarrow P_3 \rightarrow P_1$ and $P_9 \rightarrow P_5 \rightarrow P_4 \rightarrow P_1$.

In order to compare the four types of generation definitions we produce the MSO table for paper P_1 for each type of definition. The results are shown in Table 6, which shows the four different types of definitions in the vertical columns along with the citation counts of the first three generations of citations. The rows of the table represent the four sections of the *Paper-Citation graph* of Figure 3. The last line of the table contains the total number of citations for each generation for each type of definition. For example for the G^m definition, section (b) of the *Paper-Citation graph* provides two 1-gen citations from papers P_3 and P_4 , two 2-gen citations from paper P_5 , and four 3-gen citations, two from paper P_9 (paths $P_9 \rightarrow P_5 \rightarrow P_3 \rightarrow P_1$ and $P_9 \rightarrow P_5 \rightarrow P_4 \rightarrow P_1$) and two from papers P_3 and P_4 via paths $P_3 \rightarrow P_1 \rightarrow P_7 \rightarrow P_1$ and $P_4 \rightarrow P_1 \rightarrow P_7 \rightarrow P_1$ respectively.

The four definitions produce the same counts only for the 1-gen citations (direct citations). The largest citation counts are produced by the G^m definition and the numbers presented in the table equal the total number of the respective citation paths shown in Table 5, with 5 2-gen citations and 9 3-gen citations. Next comes the G^s definition, which eliminates duplicate papers from within each generation, thus producing a total of 4 2-gen citations and 8 3-gen citations by only counting P_5 once as a

P_1	G^m			G^s		
	1-gen	2-gen	3-gen	1-gen	2-gen	3-gen
a	2	1	2	2	1	2
b	2	2	4	2	1	3
c	1	1	1	1	1	1
d	1	1	2	1	1	2
Total	6	5	9	6	4	8

(a)

P_1	H^m			H^s		
	1-gen	2-gen	3-gen	1-gen	2-gen	3-gen
a	2	0	0	2	0	0
b	2	2	2	2	1	1
c	1	0	0	1	0	0
d	1	1	0	1	1	0
Total	6	3	2	6	2	1

(b)

Table 6: MSO table for the G (a) and H (b) definitions of citation generations for paper P_1 of Figure 3.

2-gen citation and paper P_9 once as a 3-gen citation. The H^m definition follows, which allows a paper to appear exactly once in the generation with the lowest possible rank. The counts produced from this definition are 3 2-gen citations (after removing paper P_6 as a 1-gen and paper P_1 as a 0-gen) and 2 3-gen citations (after removing paper P_1 as a 0-gen and P_2, P_3, P_4, P_6, P_7 and P_{10} as a 1-gen). Finally, the H^s definition produces 2 2-gen citations and 1 3-gen citation after removing all papers appearing in lower rank generations (same as H^m) plus all duplicate papers from within each generation (P_5 is only counted for once as a 2-gen and P_9 is only counted for once as a 3-gen).

To summarize, we observe that the G^m definition produces the largest counts of citations, by counting all the individual citation paths. As a result, it does not capture the nature of the individual citations. For example, in cases where a source paper provides citation paths of different lengths (like paper P_6), that paper, which is a single publication, also provides more than one indirect citations of different ranks. The same is true, when a paper provides more than one citation paths of the same length like papers P_5 and P_9 , which also provide more than one indirect citation but of the same rank. In addition this definition does not cope well with citation path cycles since indirect citations are always counted for no matter which paper provides them.

The G^s definition copes better with cases where a paper provides more than one indirect citation paths of the same length, since now a paper can only be included once per generation. Examples of this case are papers P_5 and P_9 each providing two citation paths of length 2 and 3 respectively, but now they are counted for only once per generation. Still, this definition does not distinguish between citation paths of different lengths originating from the same paper, like paper P_6 , nor it corrects for the cycles present in a *Paper-Citation graph*.

On the other hand, the H^m definition can handle cycles, since if a paper has been included in a generation of lower rank it is not included again in a higher rank generation. For example, paper P_1 is included in the 0-gen set, thus it does not provide a 2-gen citation to itself via P_7 . The same is true for P_1 and a 3-gen citation that it could provide to itself if papers were not restricted between generations. Finally, this definition also copes with citation paths of different length originating from a single paper like paper P_6 . Again, paper P_6 is included in the 1-gen set, thus, it does not also provide a 2-gen citation via P_2 . The only case that H^m does not handle is the existence of multiple citation paths of the same length originating from a single paper, like papers P_5 and P_9 .

All cases mentioned so far, are handled by the H^s definition, which is the one we propose for counting indirect citations. With this definition an indirect citation indicates a connection between two papers and not merely the existence of at least one citation path between the papers in a *Paper-Citation graph*.

4 Definition of the fp^k – index

We propose a new indicator for the assessment of a paper that accounts for both the direct and indirect impact of the paper as well as for the scientific age of the paper. The indicator can be described as a cross-generational index [Hu, Rousseau, and Chen \(2011\)](#), in the sense that it uses individual values generated for each generation of citations and then uses these values in order to calculate the cross-generational index that attempts to quantify the scientific value of a paper. Part of the definition of the indicator is the type of generation of citations used to produce the values to describe the generation of citations. The fp^k – index is calculated as

$$fp^k = \frac{1 + \sum_i^k \left(\frac{1}{i} * gen_i\right)}{n_p} \quad (1)$$

In general, indirect citations should indicate that there is a connection between the paper under scrutiny and the papers included in each generation. This connection should be stronger the closer it is to the target paper [Sidiropoulos and Manolopoulos \(2005\)](#). A connection between two papers is indicated by a single indirect citation rather than a count of all the indirect citation paths targeting the examined paper. In the proposed indicator, citations are weighted depending on the generation they belong to (gen_i), with citations of lower rank being more important and indicating that the target paper had a higher impact on the source paper. The indicator assigns a value 1 to each published paper and it uses the scientific age of the paper (n_p) to produce scores that can be used to compare papers of different scientific age. Once published, a paper is considered to have a scientific age of 1.

The proposed indicator considers the first k generations of citations of the H^s definition but the number of generations that one should consider is a subject that requires further investigation. If we assume that individual citation graphs are generated for publications belonging to different scientific fields then there are a number of characteristics that could affect the number of generations of citations that one

should examine. The following list provides just an overview of some of them and the authors consider it to be neither complete nor exhaustive.

- *Number of publications per year:*

Small number of papers published in a particular scientific field could mean that the density of the citation graph examined is high with a relatively small number of participating papers and many citations among them. On the other hand, large number of papers published each year could mean that the length of the citation paths is small therefore not providing many generations to base our calculations on.

- *Average number of citations received or references provided:*

A large average number of citations could indicate a citation pattern where authors reference not only new papers but also papers published several years ago, thus possibly producing large number of chords in the citation graph.

- *Average elapsed time from the date of publication until a paper receives its first citation:*

If the observed times are high it could be that several years may pass before published papers receive citations in which case the time is the limiting factor in our calculations.

- *Average age of citations:*

The average age of the citations received could also affect the number of generations considered since a large average citation age could mean that it could be several years before long citation paths could be generated within the graph.

For the calculations included later in this paper we have chosen $k = 3$, thus considering the first three generations of citations of the H^s definition. This number has been chosen based on the authors sentiment that three generations (similar to friends of friends of friends in social networks) are enough to illustrate the usability and validity of the indicator under different circumstances.

5 Application and comparison of fp^k – index with Number of citations (NC) and PageRank

In this section, we examine two applications of the fp^k – index. The first one is to the *Paper-Citation graph* of Figure 3. In this graph, we consider all papers to be of equal scientific age (age 1). The second one is on the *Paper-Citation graph* of Figure 4, where we provide the scientific age of the papers included in the graph.

The purpose of these examples is to demonstrate how the fp^k – index reacts to the different citation patterns present in the graphs, especially when compared to the two other indicators, namely the Number of citations (NC) and PageRank (Page, Brin, Motwani, and Winograd 1999; Ma, Guan, and Zhao 2008). The Number of citations (NC) is the most commonly used indicator and measures the impact of a paper by counting the number of direct citations received. This indicator produces values that are identical to the first generation citation counts we have discussed so far.

On the other hand, PageRank is an indicator originally used to rank pages on the web and was initially inspired by citation analysis. The indicator has found its way

back to citation analysis with multiple applications, modifications and adaptations that aim at providing a more accurate representation of scientific impact whether it is for a paper, author or journal. PageRank imitates the “random surfer” model, where a person navigates through the web by a number of random hops. The surfer, after randomly selecting one of the available pages, randomly chooses to follow one of the outgoing links of the page and continues to do so until he gets “bored”, at which point he completely stops his current navigation path and moves to a newly selected random page from where he starts a new navigation path. The number of hops performed is determined by a damping factor. PageRank is calculated as follows

$$PR(A) = (1 - d) + d * \sum \frac{PR(i)}{N(i)} \quad (2)$$

where d is the damping factor, which in the original implementation of PageRank was set to be 0.85, $PR(i)$ is the PageRank score of the i_{th} page that links to page A , and $N(i)$ is the number of outgoing links of page i . For the calculations included in this section of the paper we use $d = 0.5$ as defined in [Ma, Guan, and Zhao \(2008\)](#). We refer to this version of PageRank as *Base*.

A normalized version of PageRank also exists, where the first component is divided by the total number of nodes present in the network, or papers in the *Paper-Citation* graph.

$$PR(A) = \frac{(1 - d)}{N} + d * \sum \frac{PR(i)}{N(i)} \quad (3)$$

By implementing PageRank as shown in [3](#), the sum of the PageRank values of all nodes included in a particular graph should be 1.0. As discussed in the literature though, this is not the case in graphs that include nodes that do not provide any reference to any of the nodes included in the graph. These nodes are named dangling nodes [Yan and Ding \(2011\)](#) and their behaviour would cause the sum of the PageRank values to decline after a number of iterations. In the second version of PageRank, we accommodate these dangling nodes by equally re-distributing their value to all the nodes in the graph and we refer to this version of PageRank as *Normalized*.

5.1 First example

The purpose of this example is to demonstrate the usage of the $fp^k - index$, $k = 3$ and the way it reacts in a graph that includes the four distinct cases of citation patterns discussed earlier. [Table 7a](#) presents the citation generation counts for the ten papers included in the graph along with the calculated values of the three indicators (Number of citations, PageRank and $fp^3 - index$). As already mentioned, a damping factor $d = 0.50$ has been used for the PageRank calculations. *Base* PageRank required 26 iterations to converge and the *Normalized* PageRank required 14 (with a convergence criterion set to 0.00001). [Table 7b](#) presents the different categories created by the calculated values of each indicator and the papers that fit each category. It is interesting

to note that both versions of PageRank produce the same categories for the papers included in the graph, even though their calculated values are different.

	gen1	gen2	gen3	PageRank			
				NC	Base	Normalized	$fp^3 - index$
P_1	6	2	1	6	2.769	0.277	8.333
P_2	1	0	0	1	0.625	0.063	2.000
P_3	1	1	0	1	0.688	0.069	2.500
P_4	1	1	0	1	0.688	0.069	2.500
P_5	1	0	0	1	0.750	0.075	2.000
P_6	0	0	0	0	0.500	0.050	1.000
P_7	1	5	2	1	1.192	0.119	5.167
P_8	1	6	1	1	1.192	0.119	5.333
P_9	0	0	0	0	0.500	0.050	1.000
P_{10}	1	1	5	1	1.096	0.110	4.167

(a)

Number of citations		PageRank			$fp^3 - index$	
Score	Papers	Score (B)	Score (N)	Papers	Score	Papers
6	P_1	2.769	0.277	P_1	8.333	P_1
1	$P_2 - P_5$ $P_7 - P_{10}$	1.192	0.119	P_7 P_8	5.333	P_8
0	P_6 P_9	1.096	0.110	P_{10}	5.167	P_7
		0.750	0.075	P_3	4.167	P_{10}
		0.688	0.069	P_3 P_4	2.500	P_3 P_4
		0.625	0.063	P_2	2.000	P_2 P_5
		0.500	0.050	P_6 P_9	1.000	P_6 P_9

(b)

Table 7: (a) On the left, we list the citation generation counts of the papers included in the *Paper-Citation graph* of Figure 3, and on the right we list the values of the three indicators (Number of citations (*NC*), PageRank (Base and Normalized) and $fp^3 - index$), (b) the categories defined by each indicator based on the available values are presented along with the papers that fit each category.

It turns out that all three indicators agree that the most important paper in the graph is P_1 and the less important ones are P_6 and P_9 that have not received any direct (and therefore indirect) citations. The less sensitive indicator is the Number of citations since it only considers the direct impact of the papers and thus produces the less distinctive categories for the papers in the graph, placing all papers that have received one citation in the same category with the same score. PageRank and $fp^3 - index$ seem to be able to better distinguish the remaining papers in the graph.

In particular, PageRank considers papers P_7 and P_8 to be the second most important papers in the graph whereas P_{10} occupies the third most important position. $fp^3 - index$ also considers paper P_8 as the second more important paper in the graph but it distinguishes it from P_7 which occupies the third most important position, with P_{10} moving one position down in the list, ranked fourth. According to $fp^3 - index$, P_8

is ranked higher even though it has one 3-gen citation less than P_7 because at the same time it has one 2-gen citation more than P_7 , and as we have seen so far gen2-citations have a greater impact on the calculated score when compared to 3-gen citations under the same conditions.

Moving further down the list, according to PageRank the next more important paper is P_5 (ranked fourth) since even though it only receives a single 1-gen citation from paper P_9 , paper P_9 does not provide any other citation to any of the other papers included in the graph.

According to fp^3 -index paper P_5 is ranked sixth, below papers P_3 and P_4 and it is considered of equal importance to P_2 . If we look at the number of citations received by these papers we can state that P_5 receives only one 1-gen citation (from paper P_9) and P_2 also receives one 1-gen citation (from paper P_6), whereas papers P_3 and P_4 receive one 1-gen citation each from P_5 and one 2-gen citation each from paper P_9 , thus ranking higher than P_5 .

5.2 Second example

The second application is to the *Paper-Citation graph* of Figure 4, that contains a graph with 22 papers. The graph is constructed using paper P_1 as the target paper. All citation paths of length lower than or equal to four have been included. For simplicity, we consider all papers within the same citation path length area to have the same scientific age. The oldest papers are P_1, P_2, P_3 and P_4 with scientific age 4.

Table 8 presents the gen1, gen2 and gen3 citation counts for the 22 papers of the

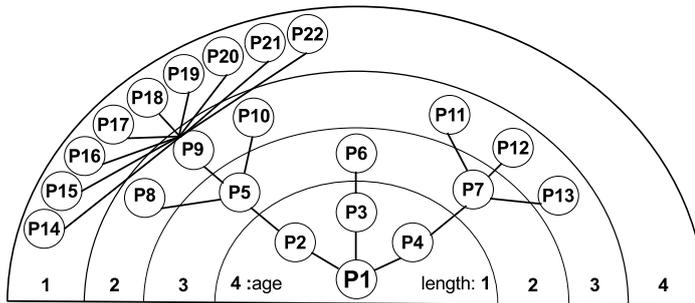


Fig. 4: Example of a *Paper-Citation graph*. All citation paths of length lower than or equal to four are included in the graph. For simplicity we consider all papers within the same citation path length to have the same scientific age.

graph along with the scientific age of each paper and the calculated values for the three indicators under examination. For PageRank, we are displaying the scores for both the *Base* and *Normalized* version. The *Base* version required 7 iterations to converge whereas the *Normalized* one required 17 (the convergence criterion has again been set to 0.000001). The papers are ordered in increasing order based on

their name and no other sorting has been applied. The PageRank and $fp^3 - index$ values have been rounded to three decimal places whereas the Number of citations are always integer values.

There are nine papers (P_{14} to P_{22}) that have an $fp^3 - index$ of 1.000 since they have not received any direct or indirect citations and their scientific age is 1. We can compare the $fp^3 - index$ values of these papers to the $fp^3 - index$ values of papers P_8 , P_{10} , P_{11} , P_{12} and P_{13} that also have not received any direct or indirect citations but whose scientific age is 2, and thus their $fp^3 - index$ value is 0.500. We consider this to be a valid result since if a paper has not received any direct or indirect citations its value should decline as it is getting older since (with the exception of sleeping beauties) it becomes more and more unlikely that it receives many citations in the future. The same logic applies to paper P_6 as well, whose value is 0.333, since it has not received any direct citations and its scientific age is 3.

	age	gen1	gen2	gen3	PageRank			
					$fp^3 - index$	NC	Base	Normalized
P_1	4	3	3	6	1.875	3	2.281	0.116
P_2	4	1	3	9	1.675	1	1.688	0.086
P_3	4	1	0	0	0.500	1	0.750	0.038
P_4	4	1	3	0	0.875	1	1.125	0.057
P_5	3	3	9	0	2.833	3	2.375	0.120
P_6	3	0	0	0	0.333	0	0.500	0.025
P_7	3	3	0	0	1.333	3	1.250	0.063
P_8	2	0	0	0	0.500	0	0.500	0.025
P_9	2	9	0	0	5.000	9	2.750	0.140
$P_{10} - P_{13}$	2	0	0	0	0.500	0	0.500	0.025
$P_{14} - P_{22}$	1	0	0	0	1.000	0	0.500	0.025

Table 8: On the left the 22 papers of the *Paper-Citation graph* of Figure 4 are listed along with their scientific age and citation generation counts. On the right the calculated values based on the Number of Citations (NC), PageRank (Base and Normalized) and $fp^3 - index$ indicators are presented.

Another interesting comparison is between papers P_3 , P_4 and P_2 of scientific age 4. P_3 has only received a single 1-gen citation, P_4 has received a single 1-gen citation along with 3 2-gen citations and, P_2 has received a single 1-gen citation along with 3 2-gen citations and 9 3-gen citations. Since all these papers have the same scientific age, the factor that determines the acquired score is the number of 1-gen, 2-gen and 3-gen citations. In addition, the 1-gen citation count is the same for all papers. Therefore, the one that should gather the lower score is the one that has no 2-gen and 3-gen citations, which is paper P_3 . From the remaining papers the one that should follow is the one that has 2-gen citations but no 3-gen citations. And, finally, the paper that should gather the greatest score is P_2 since it has more 3-gen citations than P_4 .

In order to make the comparison easier, the scores and the corresponding papers per indicator are presented in Table 9. The Number of citations (NC) indicator is the

less sensitive one since it only creates 4 different score based categories for score values 9, 3, 1 and 0.

PageRank also categorizes all papers that have no impact in the same category with a score of 0.500 for the Base version and 0.025 for the Normalized one. PageRank is clearly better than NC distinguishing between papers that have had some impact, indicated by the fact that these papers have received at least one citation. The remaining 7 papers received distinct scores, with P_9 being the most important paper in this graph.

Method	Score	Papers	
Number of citations	9	P_9	
	3	P_1 P_5 P_7	
	1	P_2 P_3 P_4	
	0	P_6 P_8 P_{10} P_{11} P_{12} P_{13} P_{14} P_{15}	
		P_{16} P_{17} P_{18} P_{19} P_{20} P_{21} P_{22}	
PageRank	Base	Norm	
	2.750	0.140	P_9
	2.375	0.120	P_5
	2.281	0.116	P_1
	1.688	0.086	P_2
	1.250	0.063	P_7
	1.125	0.057	P_4
	0.750	0.038	P_3
	0.500	0.025	P_6 P_8 P_{10} P_{11} P_{12} P_{13} P_{14} P_{15}
			P_{16} P_{17} P_{18} P_{19} P_{20} P_{21} P_{22}
$fp^3 - index$	5.000	P_9	
	2.833	P_5	
	1.875	P_1	
	1.675	P_2	
	1.333	P_7	
	1.000	P_{14} P_{15} P_{16} P_{17} P_{18} P_{19} P_{20} P_{21}	
	0.875	P_4	
	0.500	P_3 P_8 P_{10} P_{11} P_{12} P_{13}	
	0.333	P_6	

Table 9: Scores and Papers distribution per indicator. The three indicators included are the Number of citations, PageRank and the $fp^3 - index$.

$fp^3 - index$ generates 9 different categories. Papers P_9 , P_5 , P_1 , P_2 and P_7 are ranked similarly by both PageRank and $fp^3 - index$. $fp^3 - index$ takes into consideration the scientific age of a paper and young papers rank higher than older papers with identical properties.

6 Definition of the fa^k and fas^k indices

We have defined an indirect indicator, the $fp^k - index$, that can be used to calculate the current cumulative value of a paper based on the first three generations of citations

as defined by the H^s definition. Based on these values a new indicator is proposed for the scientific assessment of an author called $fa^k - index$.

$fa^k - index$ is defined as the sum of all $fp^k - index$ values of all papers co-authored by an author divided by the total number of papers (N) in the Publication Record of the author and is equal to

$$fa^k = \frac{\sum_i^N fp^k - index(i)}{N} \quad (4)$$

where $fp^k - index(i)$ is the $fp^k - index$ of the i_{th} paper of the author. Since the $fp^k - index$ of a paper represents the current value of a paper the $fa^k - index$ represents the average $fp^k - index$ value of the author's papers at the time when the evaluation occurs.

We might say that this indicator is independent of the scientific age of the author since the value of each paper is normalized based on its age. We believe that only the paper's age should be used to distinguish between younger and older papers that share the same properties and that younger papers that have attracted a considerable number of citations quickly should be rewarded. In addition the proposed indicator is size-independent since the cumulative value of the $fp^k - index$ scores of the papers is divided by the number of papers included in the Publication Record of an author. By doing so, authors with different productivity levels could more easily be compared based on the scientific impact of their papers.

Summarizing, the $fa^k - index$ is an indirect indicator that takes into account the first k generations of citations, the scientific age of each individual paper as well as the productivity of the author in order to produce the author's score and it is independent of the scientific age of the author.

An additional aspect that we could consider for an indicator used to assess authors is the number of self-citations. Another indicator is therefore proposed that considers the citations in the *Paper-Citation graph* at the (author, paper) level named $fas^k - index$. $fas^k - index$ is calculated using the same formula as the $fa^k - index$ with the only difference being the way the citation generations are produced for the calculations of the $fp^k - index$ values for the papers in the Publication Record of the author. For the $fa^k - index$ all citations based on the H^s definition are counted for, but for the $fas^k - index$ the citation generations should be constructed in the way described in Section 2.3.

The $fas^k - index$ is always smaller than or equal to the $fa^k - index$ of an author. The two indices are equal only when the author has zero self-citations in his first three generations of citations.

7 Application of the fa^k and fas^k indices

We present an example of the application of the fa^k and fas^k indices on the *Paper-Citation graph* of Figure 1 in order to demonstrate the differences in the calculated scores for the authors included in the graph. The graph consists of seven papers that

have been co-authored by five distinct authors. The graph also includes the publication year of each paper from which we calculate its scientific age with regards to 2014. Table 10a presents the papers listed in alphabetical order based on their name, the scientific age of each paper, the gen1, gen2 and gen3 citation counts and the $fp^k - index$ for each individual paper. Table 10b presents the papers each author has participated in along with the $fa^k - index$ value for the author calculated by equation 4.

In Table 11, we can see the citation generations for each (author, paper) pair. The

	age	gen1	gen2	gen3	$fp^3 - index$
P_1	14	3	2	1	0.356
P_2	13	1	0	0	0.143
P_3	13	1	2	0	0.214
P_4	13	1	2	0	0.214
P_5	12	0	0	0	0.077
P_6	11	2	0	0	0.250
P_7	10	0	0	0	0.091

(a)

	Papers				Age range		$fa^3 - index$
A_1	P_1	P_5	P_6		11	14	0.227
A_2	P_1	P_4	P_6	P_7	10	14	0.228
A_3	P_2	P_3	P_7		10	13	0.149
A_4	P_3	P_4			13	13	0.214
A_5	P_5				12	12	0.077

(b)

Table 10: (a) The papers included in Figure 1 along with their publication dates, citation generation counts and $fp^3 - index$ values and (b) The authors of the papers along with the papers each author has co-authored, the age range of the papers along with the $fa^3 - index$ values for the authors for year 2014.

citation generation counts are presented with all self-citations excluded, which is the reason why for the same paper the counts vary from author to author. With these new, refined citation counts the $fp^3 - index$ of the papers is calculated again and the results are presented in Table 11.

Table 12 presents the authors with the papers in their Publication Record along with the age range of the papers and the $fas^3 - index$ for each author. For the calculation of the $fas^3 - index$ equation 4 was used with the $fp^3 - index$ values presented in Table 11, where self-citations have been removed from the citation generation counts.

Comparing the calculated values for fa^3 and the fas^3 indices of the authors, we observe that the author scores become lower when removing self-citations. The calculated value for author A_5 remains the same since he has already received the maximum value for the single paper that he co-authored 12 years ago and which has

Author	Paper	age	gen1	gen2	gen3	$fp^3 - index$
A_1	P_1	14	3	0	1	0.310
	P_5	12	0	0	0	0.083
	P_6	11	1	0	0	0.182
A_2	P_1	14	2	1	0	0.250
	P_4	12	0	0	0	0.083
	P_6	11	1	0	0	0.182
	P_7	10	0	0	0	0.100
A_3	P_2	13	1	0	0	0.154
	P_3	13	1	1	0	0.192
	P_7	10	0	0	0	0.100
A_4	P_3	13	1	2	0	0.231
	P_4	13	1	2	0	0.231
A_5	P_5	12	0	0	0	0.083

Table 11: The (author, paper) pairs included in Figure 1, along with the age of the papers, the gen1, ge2 and gen3 citation generation counts (self-citations are excluded) and the $fp^3 - index$ value of each paper per author.

	Papers				Age range		$fas^3 - index$
A_1	P_1	P_5	P_6		11	14	0.192
A_2	P_1	P_4	P_6	P_7	10	14	0.154
A_3	P_2	P_3	P_7		10	13	0.149
A_4	P_3	P_4			13	13	0.231
A_5	P_5				12	12	0.083

Table 12: The authors of the papers along with the papers each author has co-authored, the age range of the papers along with the $fas^3 - index$ values for the authors for year 2014.

attracted no citations. In addition, the value of author A_4 also remains constant since none of the citations received belongs to papers co-authored by A_4 . The values for authors A_1 , A_2 and A_3 are lower and the calculated value for A_1 has the greatest drop since she has received many self-citations. The exclusion of self-citations from the citation generation counts can severely affect an author's score.

8 Comparative study

In order to compare the indicators discussed in this paper, we performed a comparative study utilizing the citation data provided by DBLP, a Computer Science Bibliography database that provides an online index of scientific publications. The underlying data is formatted in XML and is released under the ODC-BY 1.0 license. The XML formatted file can be downloaded from the DBLP website. PHP (DOM extension) was used in order to parse the XML file and store the data in a relational DBMS (MySQL) for easier retrieval and access.

8.1 DBLP Data

The different types of publications included in the DBLP dataset are presented in (DBLP) and mainly include articles (published in a journal or magazine), papers from conferences or workshops and Proceeding volumes. Other publication types, like authored monographs, parts or chapters in a monograph, PhD and master theses, are also included but in smaller numbers.

Like in previous studies (Sidiropoulos and Manolopoulos 2005; Fiala, Rousselot, and Ježek 2008), we chose to only consider articles and papers in our study. During parsing, we considered records to be *complete* if apart from the DBLP Key (uniquely identifies a publication within the DBLP dataset), they also provided a Title, Year of Publication and a list of Authors.

It is worth noting that DBLP uses the WWW record type to provide details about a particular author, such as the list of synonyms of an author’s name. DBLP’s methodology of identifying and mapping authors to their respective publications is described in dbi (2009). For the purposes of our study we have not made any attempt to identify any author type synonyms or distinguish between authors with the same name. This means that metrics presented for some authors may be misleading since publications of two authors with the same name are attributed to a single author.

Publication type	# records	No authors		No references	
		# records	% total	# records	% total
Article	1308552	6565	0.50%	1306765	99.86%
In Proceedings	1641467	2419	0.15%	1640414	99.94%

Table 13: Imported DBLP records per publication type along with the percentage compared with the original set of publication records.

Finally, wherever available we also considered the List of References for each publication, which essentially is a list of publication keys. Each key uniquely identifies a publication in the DBLP database and is a reference to the actual publication record. Table 13 presents the data imported from the XML file along with some statistics about the corresponding numbers of authors and references. With regards to the number of references, we observe that most publications do not provide references to other publications. This means that if we were to represent the dataset as a citation graph we would indeed have most of the publications appear as isolated nodes with no incoming or outgoing edges. Thus, we decided the citation graph to include all journal articles and conference papers that provide at least one reference to any other publication or receive at least one citation from any of the publications in the original dataset. This data was then extracted to a different database and Table 14 displays the summary statistics.

Publication type	Count	# Publications with		Authors	
		References	Citations	Count	Distinct
Article	8087	1767	7406	17646	8304
In Proceedings	12786	6177	10912	30473	11487

Table 14: Records included in the Citation Graph along with the number of references provided and citations received. The table also presents the total number of co-authors and the distinct count of authors per publication type.

We observe that the number of publications that provide references to other publications included in the data-set is smaller than the number of publications that receive citations. This means that the publications that include references, reference more than one publication each (not necessarily of the same type).

For the remaining of this paper, we will not distinguish between the two publication types, i.e., Article and InProceedings, and we will refer to all publications included in the *Paper-Citation graph* as papers.

8.2 Paper indicators

From the definition of $fp^k - index$ it follows that the indicator values can vary depending on the number of citation generations considered in the calculations. As previously mentioned, we argue that three generations of citations are adequate in producing an $fp^k - index$ value that is representative of the accumulated impact of a particular paper, but as part of our analysis we recursively calculated all generations of citations included in the graph according to the definition of generations we defined earlier. These values were stored in a separate Medal Standings Output (MSO) table in the relational DBMS and are presented in Figure 5.

The generations present in the citation graph are displayed on the x-axis of Figure 5. On the primary y-axis we plot the number of papers that have received at least one citation of the specified generation, and, on the secondary y-axis, we plot the total number of citations per generation.

We notice that the Publications series starts high with many papers receiving a gen-1 citation. The values gradually reduce to eventually reach 0 for generations 29 and 30, since no paper in our citation graph is part of a citation path of that length. With regards to the total number of citations for each generation, we notice that the number increases substantially from generation 1 to generation 5 and then it decreases down to 0 for generations 29 and 30.

Following the analysis of the citation graph, we selected a list of indicators to be implemented and compiled against the citation database. A description of each of the indicators considered in this study can be found in the following paragraphs.

Number of Citations (NC)

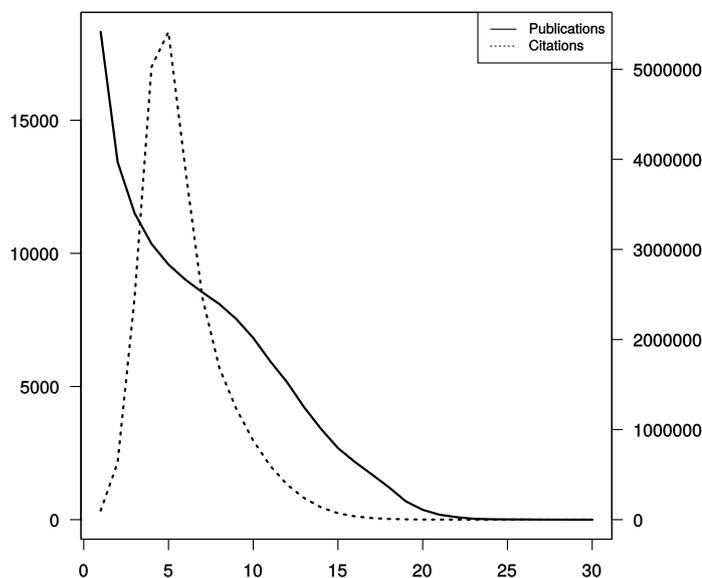


Fig. 5: Summary statistics of the publications included in the Paper-Citation graph and the citations received for each generation of citations identified.

The Number of Citations is perhaps the most widely used indicator for the assessment of papers. It has been used in many studies and its main benefit is that is easily calculated for each publication. It is generally defined as the number of citations received by a given paper.

Contemporary h-index score (h^c – index)

The contemporary h-index (h^c – index) is an author based indicator proposed by [Sidiropoulos, Katsaros, and Manolopoulos \(2007\)](#) and it is a variation of the well known h-index indicator. h-index uses the number of citations received by the publications a particular author has (co-) authored and is defined as follows:

“An author has index h , if h of his/her N_p papers have at least h citations each and the other $(N_p - h)$ papers have no more than h citations each.”

Contemporary h-index builds on this concept but instead of using the number of citations received by a publication it calculates a score for the publication that also considers its scientific age. All papers in the publication record of the researcher are listed in descending order based on the scoring function

$$S_i = \gamma \cdot (n_i^p + 1)^{-\delta} \cdot x_i \quad (5)$$

In the scoring function, γ is an arbitrarily chosen coefficient so that the resulting h^c - index is not too small. In Sidiropoulos, Katsaros, and Manolopoulos (2007), γ was selected to be 4. In addition, δ defines the strength of the time penalty. The greater the value of δ the more the age of a paper reduces its score. The h^c - index is then defined as the largest number h^c such that the value of the scoring function for that paper is greater than or equal to h^c and the remaining $N - h^c$ papers have a score of no more than h^c each.

SCEAS rank

The SCEAS indicators Sidiropoulos and Manolopoulos (2005) consider both the direct and indirect impact of citations by following an approach similar to PageRank whilst trying to minimize some of its side effects. According to the authors, the proposed score meets the following two conditions: (a) the factor that should have the greatest influence over the score of a particular paper should be the number of direct citations and, (b) the addition of new citations in the *Paper-Citation graph* should have a greater effect in the scores of nearby rather than distant papers. The SCEAS 1 scoring for papers is given by the following formula:

$$S_a = \sum_i \frac{S_i + b}{N_i} a^{-1} \quad (a \geq 1, b > 0) \quad (6)$$

where, S_a is the score of the current paper (paper a), S_i is the score of the individual papers directly citing paper a, N_i is the total number of papers cited by each paper i , b denotes the *direct citation enforcement factor* (which controls the effect that direct citations have to the calculated score) and a denotes the speed with which an indirect citation enforcement converges to zero.

The authors also propose a generalization of the above formula (SCEAS 1) and the original PageRank algorithm that introduces a dumping factor in the SCEAS rank (SCEAS 2):

$$S_a = (1 - d) + d \cdot \sum_i \frac{S_i + b}{N_i} a^{-1} \quad (a \geq 1) \quad (7)$$

PageRank

The PageRank score has also been calculated for the citation graph. As previously mentioned, PageRank in its *Base* form uses a damping factor of 0.85 as defined by the original authors. In bibliographic networks a damping factor of 0.50 has also been used.

In the calculations presented in the rest of the paper, we will be showing four different rankings for the PageRank indicator, two for the *Base* version and two for the *Normalized* one (with damping factors of $d = 0.50$ and $d = 0.85$).

8.3 Author indicators

In the Citation graph database, we also hold information about the list of co-authors for each paper. Using the list of co-authors it is possible to generate the Publication

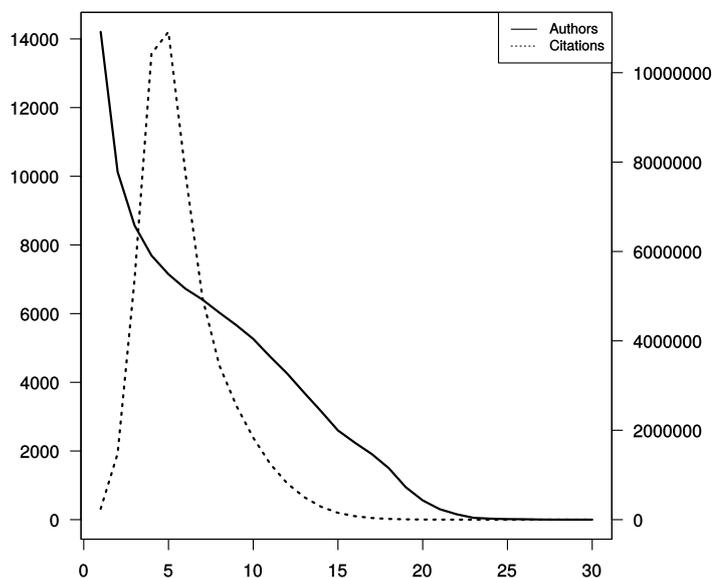


Fig. 6: Summary statistics of the authors and the citations received for each generation of citations identified.

Record of each author, and, then, using the values generated from the paper indicators for each individual paper, we can calculate the corresponding values for the author indicators.

We should mention, though, that the Publication Record for each author is far from complete since the DBLP database does not contain the complete list of papers for the examined authors. In addition, we do not distinguish between authors with the same name, so, it is possible that papers from two or more authors have been attributed to the same person. For these reasons, we do not consider the rankings presented later in this section as the absolute rankings of the authors but as indicators of the relative position that authors with the given publication records would achieve using each of the author indices under scrutiny.

Figure 6 presents some summary statistics about the authors that have (co-) authored the papers of the citation graph. The generations are displayed on the x-axis. On the primary y-axis we plot the number of authors with at least one publication that has received at least one citation of the specified generation and on the secondary y-axis we plot the total number of citations per generation received by all the papers the authors have co-authored. The numbers of citations appear to be higher than the

ones presented in Figure 5, but this is to be expected since a publication with several co-authors will have its citations accounted for more than once.

We selected a number of author specific indicators to implement and compile against the citation database, a description of which can be found in the following paragraphs.

Number of Citations (NC)

In Section 8.2, we presented the Number of Citations (*NC*) as an indicator for a single publication. The Number of Citations (*NC*) has been defined as the total number of citations received by all the papers a researcher has (co-)authored during his whole scientific career. The total Number of Citations (*NC*) has also been referred to as the *s-index* van Eck and Waltman (2008) and the *c-method* Wu (2010).

Using the values calculated by the Number of Citations indicator we can also produce a ranking for an author as follows: for a particular author, retrieve his/her publication record along with the number of direct citations received by each paper, which is now the score received by the author. All authors are then listed in descending order based on their cumulative citation count for all of their papers and this ordered list is then used to produce the ranking for the authors in the citation graph.

Mean number of citations (MNC)

The mean number of citations received by the papers the author has (co-) authored during his whole scientific career (Hirsch 2005, 2007; Costas and Bordons 2008) is expressed as

$$MNC = \frac{\sum_{i=1}^N x_i}{N}, N \geq 1 \quad (8)$$

where x_i is the number of citations for paper i , and it is defined only when the researcher has (co-)authored at least one paper. It has also been referred to as the *m-method* (Wu 2010). Here, the cumulative count of citations received by the publications included in the Publication Record is divided by the number of publications to produce the mean number of citations for the papers an author has co-authored.

h-index

See Section 8.2 for a definition of the h-index.

g-index

For the calculation of *g-index*, the papers in the publication record are listed in descending order based on their citation count. Then, the *g-index* is defined as the largest number g of papers that have together received at least g^2 citations Egghe (2006). The *g-index* uses the cumulative sum of the citations received by the papers of the researcher.

Contemporary h-index (h^c - index)

See section 8.2 for a definition of the Contemporary h-index (h^c - index).

SCEAS Rank

See section 8.2 for a definition of the SCEAS rank. In the original paper (Sidiropoulos and Manolopoulos 2005), the author ranking is produced as the average SCEAS score of an author’s papers. It is worth noting though that the average is not calculated across the full Publication Record for an author but using the top 25 publications from the author’s publication record. When an author has less than 25 papers in the *Paper-Citation graph*, we consider all of them in the calculations of the SCEAS rank.

PageRank

See Section 5 for a definition of the PageRank. As with SCEAS rank, we calculated the PageRank of an author based on the average PageRank of a set of publications from the author’s publication record. The rankings produced for PageRank use either the *Base* or *Normalized* version of PageRank, with a damping factor of either 0.50 or 0.85, and the final ranking is based either on the full publication record of an author or his/her top 25 papers.

8.4 Experimental results

8.4.1 Paper indicators

For each indicator discussed we have calculated the raw value for the indicator as well as the ordinal ranking of all papers included in the citation graph. Since the values produced by each indicator do not always provide enough granularity for each paper to receive a distinct ranking, we assign a ranking based on the following rules. For all papers with the same value, we sum the ranks they would have been assigned if their values were distinct and divide by the number of papers with the identical score. All papers examined are then assigned the same score.

Table 15 shows the number of distinct values produced by each indicator for the 20873 papers included in the citation graph. We observe that the indicators that only consider the direct impact of a publication in their calculations have low granularity, with the Number of Citations (*NC*) producing 144 distinct values and the Contemporary h-index score (*h^c score*) 929.

The PageRank variations provide more granularity with distinct values ranging from 9150 (for the *Normalized* version with $d = 0.50$ - PageRank N50), to 11365 (for the base version with $d = 0.50$ - PageRank B50). The convergence criterion was set to 0.000001 for all four versions of PageRank and for the *Base* version the algorithm required 15 iterations for $d = 0.50$ and 19 iterations for $d = 0.85$. For the *Normalized* versions, 9 and 10 iterations were performed for the damping factors $d = 0.50$ and $d = 0.85$, respectively. SCEAS1 and SCEAS2 produce 11687 and 10293 distinct values, respectively. Finally, the *fp³ - index* produces 6776 distinct values.

In Table 16, we present the top 10 papers based on the ranking produced by the *fp³ - index* indicator, along with the rankings these papers hold in the ranks of all the paper indicators described in the previous section. Each paper is usually referred to by the last part of its DBLP key (i.e. Chen76) or if that does not provide sufficient information to uniquely identify the paper within the citation graph, we have also

		# Distinct values
Direct impact	Number of Citations (<i>NC</i>)	144
	Contemporary h-index score (<i>h^c</i> score)	929
Indirect impact	PageRank	
	Base, d = 0.50 (B50)	11365
	Base, d = 0.85 (B85)	11251
	Normalized, d = 0.50 (N50)	9150
	Normalized, d = 0.85 (N85)	11344
	SCEAS	
SCEAS 1	11687	
SCEAS 2	10293	
<i>fp³ - index</i>		6776

Table 15: Number of distinct values generated by the paper indicators

included the second part of the key (i.e. tods/SmithS77). In the same table, we also present the citation counts for the first three generations, calculated using the H^s definition, along with a column that reports the longest citation path for each paper.

Paper	Year	fp^3	NC	h^c	SCEAS		PR		Citation counts			Max
					1	2	B & N		g1	g2	g3	
							50	85				
Codd70	1970	1	2	2	1	1	1	1	580	3150	2580	7
Astrahan BCEGGKLM MPTWW76	1976	2	9	13	5	7	4	4	239	2653	2991	7
Stonebraker WKH76	1976	3	11	15	8	8	7	6	228	2490	2924	7
Chen76	1976	4	1	1	2	2	2	2	604	1583	2471	8
SelingerACLP79	1979	5	3	3	4	4	5	7	370	1671	2541	9
Stonebraker75	1975	6	25.5	49	17	17	17	15	140	1815	3394	8
tods/SmithS77	1977	7	5	6	6	5	6	8	313	1672	2690	9
tods/Codd79	1979	8	7	8	9	9	10	12	280	1623	2491	8
EswarranGLT76	1976	9	4	5	3	3	3	5	326	1180	3304	8
Cod72	1972	10	17.5	40	11	11	11	9	170	1620	3662	8
Best rank		1	1	1	1	1	1	1				
Worst rank		10	25.5	49	17	17	17	15				
Median		5.5	6	7	5.5	6	5.5	6.5				
Standard Deviation		2.9	7.4	15.9	4.6	4.6	4.6	4.1				

Table 16: Top 10 papers based on the $fp^3 - index$ indicator. The table includes the positions these papers have received in the rankings of all other indicators described in section 8.4.1 along with the citation counts for their first three generations and the length of their longest citation path. In this table the PageRank Base and Normalized produce the same rankings

The top 10 papers according to fp^3 – *index* populate high positions on all indicator rankings. In particular, there seems to be an agreement across all indicators that Codd70 is the most influential publication and it populates either the 1st or 2nd position on all rankings. All the indirect indicators seem to agree that it should be the top paper, whereas the direct impact indicators (*NC* and h^c – *index* score) seem to place the publication at the second position, since it has received less direct citations than the Chen76 publication (580 versus 604).

In general, the paper from the top 10 listing that populates the lower position in the other ranks is Stonebraker75 that holds the 6th position in fp^3 – *index* but populates positions 15 to 49 on the other ranks (still very high positions in the overall ranking but not part of the top 10 publications). The lowest positions are assigned by the Number of Citations (*NC*) and the Contemporary h-index score (25.5 and 49 respectively), which is to be expected since there are papers with more direct citations included in the graph. This again highlights the effect that indirect citation counting can have on the rankings produced by the indicators.

With regards to the four versions of PageRank and the two different damping factors, it seems that the damping factor has had a stronger influence for these top 10 publications than whether we considered the total number of publications or the dangling nodes in the graph, since if we look at the ranking positions they follow the same pattern for the same values of the damping factor. In some cases the four rankings are in agreement (Codd70, Chen76 and AstrahanBCEGGKLMPTWW76), whereas in others the base version ranks the papers higher (SelingerACLP79) or lower (tods/SmithS77).

	fp^3	NC	PageRank						
			h^c	SCEAS		Base		Normalized	
				1	2	50	85	50	85
fp^3	1.000	0.847	0.792	0.749	0.743	0.763	0.797	0.756	0.795
NC	0.847	1.000	0.961	0.893	0.892	0.893	0.891	0.892	0.892
h^c	0.792	0.961	1.000	0.847	0.848	0.845	0.834	0.847	0.836
SCEAS1	0.749	0.893	0.847	1.000	0.999	0.999	0.993	0.991	0.993
SCEAS2	0.743	0.892	0.848	0.999	1.000	0.999	0.991	0.991	0.992
PRB50	0.763	0.893	0.845	0.999	0.999	1.000	0.996	0.990	0.997
PRB85	0.797	0.891	0.834	0.993	0.991	0.996	1.000	0.984	0.999
PRN50	0.756	0.892	0.847	0.991	0.991	0.990	0.984	1.000	0.985
PRN85	0.795	0.892	0.836	0.993	0.992	0.997	0.999	0.985	1.000
Top Cor.	NC	h^c	NC	SCEAS2	SCEAS1	SCEAS1	PR N85	SCEAS1	PR B85
Low Cor.	SCEAS2	fp^3							

Table 17: Spearman rank correlation matrix for the paper indicators

In Table 17, we present the Spearman rank correlation matrix for all the combinations of paper indicator ranks. For each indicator, the bottom two rows of the table report the indicators that have the highest and lowest correlation with the indicator under scrutiny. $fp^3 - index$ has the highest correlation (0.8468) with the Number of Citations and the lowest (0.7433) with SCEAS2. All other indicators appear to be less correlated with $fp^3 - index$ whereas the strongest correlation appears to be shared between the SCEAS1 and SCEAS2 scores with both of them reporting values of 0.9999. It is also worth noting that both the *Base* and the *Normalized* version of PageRank with a damping factor of 0.50 appear to have the strongest correlation with SCEAS1, in contrast to the *Base* and *Normalized* versions of PageRank with a damping factor of 0.85 that report a high correlation amongst themselves.

8.4.2 Author indicators

Table 18 shows the number of distinct values produced by each indicator for the 15862 (co-) authors of the papers. We observe that, in general, the direct indicators have low granularity, with $h - index$ generating only 24 distinct values and the Mean number of Citations (*MNC*), the most granular in this category, 939 distinct values. The indirect indicators, in general, produce many more distinct values ranging from 7239 for SCEAS2 to 8413 for SCEAS1. All other indirect indicators produce distinct values that fall in between the previous two counts.

In Table 19, we present the top 10 authors based on the $fa^3 - index$ along with some summary information about their Publication Record. For each author, we note the year of their first and last publication included in the set along with their total number of publications.

In Table 20 we present the rankings of the top 10 authors according to the $fa^3 - index$ along with their corresponding ranks for the list of direct and indirect impact author indicators.

The Mean number of citations (*MNC*) also places these authors in high positions that range from 1 to 24.5. The $h - index$, $g - index$ and $h^c - index$ indicators place the authors further down the ranking list with the worst ranks being close to the bottom of the list (14733 out of 15862 authors for $h^c - index$). These differences are to be expected since most of these authors have just one publication and based on the definition of these indicators their corresponding values and, therefore, rankings can not be high.

		# Distinct values
Direct impact	Number of Citations (<i>NC</i>)	368
	Mean number of Citations (<i>MNC</i>)	939
	<i>h</i> – index	24
	<i>g</i> – index	39
	<i>h^c</i> – index	466
Indirect impact	Base, d = 0.50, All (B50A)	8125
	Base, d = 0.50, Top (B50T)	8127
	PageRank Base, d = 0.85, All (B85A)	8003
	Base, d = 0.85, Top (B85T)	8007
	Normalized, d = 0.50, All (N50A)	7271
	Normalized, d = 0.50, Top (N50T)	7268
	Normalized, d = 0.85, All (N85A)	8218
	Normalized, d = 0.85, Top (N85T)	8220
	SCEAS SCEAS1	8413
	SCEAS2	7239
	fa³ <i>fa³</i> – index All	7532
	<i>fa³</i> – index Top	7531
	fas³ <i>fas³</i> – index All	7515
	<i>fas³</i> – index Top	7515

Table 18: Number of distinct values generated by the author indicators

Author	Publication Year		Publication count
	First	Last	
Vera Watson	1976	1976	1
Daniel Frank	1986	1986	1
C. G. Hoch	1987	1987	1
E. C. Chow	1987	1987	1
H. P. Cate	1987	1987	1
J. W. Davis	1987	1987	1
T. A. Ryan	1987	1987	1
Christopher L. Reeve	1980	1981	2
Paul R. McJones	1976	1981	3
Patricia P. Griffiths	1976	1976	4

Table 19: Top 10 authors according to the *fa³* – index along with the year of first and last publication included in the dataset and the total number of publications

Author	fa ³	NC	MNC	h – index	g – index	h ^c – index	PageRank										
							Base				Normalized				SCEAS		
							fa ³	50A	50T	85A	85T	50A	50T	85A	85T	1	2
Vera Watson	1	187.5	1	8762	9133.5	14733	1	1	1	1	1	1	1	1	1	1	1
Daniel Frank	2	740.5	24.5	8762	9133.5	14733	2	69	73	74	78	70	74	73	77	81	84
C. G. Hoch	5	383	5	8762	9133.5	14733	5	11	11	19	19	11	11	19	19	9	9
E. C. Chow	5	383	5	8762	9133.5	14733	5	11	11	19	19	11	11	19	19	9	9
H. P. Cate	5	383	5	8762	9133.5	14733	5	11	11	19	19	11	11	19	19	9	9
J. W. Davis	5	383	5	8762	9133.5	14733	5	11	11	19	19	11	11	19	19	9	9
T. A. Ryan	5	383	5	8762	9133.5	14733	5	11	11	19	19	11	11	19	19	9	9
Christopher L. Reeve	8	325	18	2401	3157	14733	8	29	32	34	36	31	34	34	36	29	27
Paul R. McJones	9	143	11	2401	1859.5	13152	9	3	3	3	3	3	3	3	3	4	4
Patricia P. Griffiths	10	149	15.5	1148	1260.5	315.5	10	5	5	5	5	5	5	5	5	5	5
Best rank	1	1	1	1148	1260.5	315.5	1	1	1	1	1	1	1	1	1	1	1
Worst rank	10	740.5	24.5	8762	9133.5	14733	10	69	73	74	78	70	74	73	77	81	84
Median	5	383	5	8762	9133.5	14733	5	11	11	19	19	11	11	19	19	9	9
Standard Deviation	2.7	163.72	7.13	3123.17	3255.66	4298.47	2.7	19.0	20.3	19.9	21.1	19.4	20.8	19.7	20.9	22.4	23.4

Table 20: Top 10 authors according to the fa^3 – index along with the direct and indirect impact indicator rankings

We observe that when looking at the rankings produced by the indirect impact indicators, the rankings of the authors have improved considerably, now ranging from positions 1 to 84. In particular, there are two authors that the indicators place in lower ranks, Daniel Frank (rankings range from 2 for $fa^3 - index$ to 84 in SCEAS2) and Christopher L. Reeve (rankings range from 8 in $fa^3 - index$ to 36 for the *Base* and *Normalized* versions of PageRank with a damping factor of 0.85 and whilst using the top 25 publications per author in order to produce the ranking).

The indicators appear to be in agreement for the remaining 8 authors that are placed in positions 1 to 19, whereas, all indirect impact indicators seem to agree that the most influential author in the citation graph is Vera Watson, even though she has co-authored only one paper titled “System R: Relational Approach to Database Management” and published in 1976. The particular paper has been co-authored by Vera Watson and 13 other authors all of which have more than one papers included in the *Paper-Citation graph* (publication record counts range from 3 to 46). It is very interesting to note that all indicators place these authors further down the ranking list with maximum three authors appearing at the different top 10 rankings across all examined indicators. This leads us to assume that all the indicators examined are indeed sensitive to the number of publications included in the publication record of an author. It is also worth noting that the fa^3 and fas^3 rankings of the authors are identical. This is to be expected for all the authors with only one publication, since they cannot receive a self-citation.

In order to present some comparative results with the ones found in the literature when the DBLP data-set is being used, we present the SIGMOD Edgar F. Codd Innovations Award winners (1992 - 2004) rankings in Table 21. Almost all of these authors do have a publication record that includes more than 25 publications, thus, looking at the rankings produced by $fa^3 - index$, we observe that using the top 25 publications improves the rankings of almost all the authors in Table 21.

As a whole, the authors included in Table 21 rank higher in the direct impact indicators with positions that range from 1 to 122. The Mean number of citations is the only direct impact indicator that places the authors further down the ranking list with assigned positions ranging from 91 to 1062.

Regarding the indirect impact indicators, the authors rank higher when we only consider their top 25 publications. In particular the authors hold higher positions in the SCEAS 1 & 2 ranks, followed by the rankings produced by PageRank ($d = 0.85$, base and normalized for the top 25 publications), followed by the $fa^3 - index$ (again when using the top 25 publications). The indirect indicators that use the full publication record for these authors place them in lower positions in their ranks.

Finally, the Spearman rank correlation matrix of the author indicators is shown in Table 22, where, we can see that there is a positive correlation among all indicators. The direct impact indicators present their highest correlation with other direct impact indicators and the lowest correlations are split between the h^c and $fas^3 - index$ indicators. Similarly, all indirect impact indicators are highly correlated with their variation (A vs T). All indirect impact indicators report their lowest correlation with the Contemporary $h - index$ indicator ($h^c - index$).

Author	Publications										PageRank							
	Year		Count	fa ³		fas ³		NC	MNC	h	g	h ^c	Base		Norm		SCEAS	
	First	Last		All	Top	All	Top						85T	85A	85T	85A	1	2
C. Mohan	1982	1999	58	891	265	926	279	33	564	21.5	26.5	29.5	327	829	325	825	270	271
David J. DeWitt	1978	2000	95	438	47	458	57	2	207	2.5	2	1	61	290	61	288	41	40
David Maier	1978	2000	76	741	141	745	143	12	392	21.5	8.5	6	116	515	118	520	101	106
Donald D. Chamberlin	1974	2000	25	106	113	113	120	21	91	70.5	32	51	33	31	33	31	49	51
Hector Garcia-Molina	1978	2000	155	2349	525	2424	561	10	1062	14	16.5	17.5	188	1291	187	1287	140	137
Jim Gray	1975	2000	46	282	85	284	85	4	98	14	4	19.5	24	62	23	62	19	18
Michael Stonebraker	1972	1999	131	327	19	336	20	1	282	1	1	2	29	247	28	252	31	30
Patricia G. Selinger	1979	1998	17	223	240	223	240	53	106	165.5	122	112	129	120	129	120	145	146
Philip A. Bernstein	1975	1999	69	328	49	329	50	5	232	4	5	9.5	56	205	56	205	51	50
Rakesh Agrawal	1983	2000	85	1399	486	1442	511	13	487	5.5	12	7	130	591	130	596	108	107
Ronald Fagin	1976	1998	41	450	180	450	180	22	218	14	16.5	17.5	108	218	109	218	125	125
Rudolf Bayer	1970	1999	30	621	493	622	493	90	550	165.5	75.5	117.5	222	261	225	268	293	298
Serge Abiteboul	1983	1999	97	972	170	1012	189	8	479	9	8.5	12	220	878	218	884	162	157
			Best rank	106	19	113	20	1	91	1	1	1	24	31	23	31	19	18
			Worst rank	2349	525	2424	561	53	1062	165.5	122	112	327	1291	325	1287	270	271
			Median	450	170	458	180	22	218	14	16.5	17.5	108	218	109	218	125	125
			Standard Deviation	659.6	171	608.3	177.9	74.1	345.6	134.5	60.9	93.4	139.8	367.7	142.6	365.4	210.2	214.9

Table 21: SIGMOD Edgar F. Codd Innovations Award winners (1992 - 2004) rankings

		PageRank																	
		Base				Normalized				SCEAS		fa ³		fas ³					
		50T	50A	85T	85A	50T	50A	85T	85A	1	2	T	A	T	A				
NC	1.000	0.894	0.809	0.843	0.746	0.770	0.768	0.772	0.770	0.773	0.772	0.771	0.769	0.769	0.768	0.775	0.773	0.754	0.753
MNC	0.894	1.000	0.697	0.719	0.678	0.849	0.848	0.851	0.851	0.852	0.852	0.852	0.852	0.847	0.845	0.833	0.832	0.832	0.831
h	0.809	0.697	1.000	0.943	0.743	0.625	0.622	0.626	0.623	0.629	0.626	0.626	0.623	0.624	0.624	0.630	0.628	0.613	0.611
g	0.843	0.719	0.943	1.000	0.757	0.636	0.633	0.637	0.635	0.642	0.639	0.636	0.633	0.635	0.635	0.643	0.641	0.626	0.623
h ^c	0.746	0.678	0.743	0.757	1.000	0.558	0.556	0.554	0.551	0.561	0.559	0.554	0.552	0.560	0.560	0.572	0.570	0.551	0.550
PR B50T	0.770	0.849	0.625	0.636	0.558	1.000	0.999	0.996	0.996	0.989	0.989	0.996	0.996	0.999	0.999	0.705	0.704	0.699	0.698
PR B50A	0.770	0.848	0.622	0.633	0.556	0.999	1.000	0.996	0.996	0.989	0.989	0.996	0.996	0.999	0.998	0.704	0.703	0.698	0.698
PR N50T	0.772	0.851	0.626	0.637	0.554	0.996	0.996	1.000	0.999	0.983	0.983	0.999	0.999	0.992	0.990	0.743	0.742	0.738	0.737
PR N50A	0.770	0.851	0.623	0.635	0.551	0.996	0.996	0.999	1.000	0.983	0.983	0.999	0.999	0.992	0.990	0.742	0.742	0.737	0.736
PR B85T	0.773	0.852	0.629	0.642	0.561	0.989	0.989	0.983	0.983	1.000	0.999	0.984	0.984	0.989	0.989	0.701	0.699	0.696	0.695
PR B85A	0.772	0.852	0.626	0.639	0.559	0.989	0.989	0.983	0.983	0.999	1.000	0.984	0.984	0.9892	0.989	0.699	0.699	0.695	0.694
PR N85T	0.771	0.852	0.626	0.636	0.554	0.996	0.996	0.999	0.999	0.984	0.984	1.000	0.999	0.993	0.991	0.740	0.739	0.735	0.734
PR N85A	0.769	0.852	0.623	0.633	0.552	0.996	0.996	0.999	0.999	0.984	0.984	0.999	1.000	0.992	0.990	0.739	0.739	0.734	0.734
SCEAS 1	0.769	0.847	0.624	0.635	0.560	0.999	0.999	0.992	0.992	0.989	0.989	0.993	0.992	1.000	0.999	0.689	0.688	0.6827	0.682
SCEAS 2	0.768	0.845	0.624	0.635	0.560	0.999	0.998	0.989	0.990	0.989	0.989	0.991	0.990	0.999	1.000	0.682	0.681	0.673	0.675
fa ³ T	0.775	0.833	0.630	0.643	0.572	0.707	0.704	0.743	0.742	0.701	0.6997	0.740	0.739	0.689	0.682	1.000	0.999	0.995	0.995
fa ³ A	0.773	0.832	0.628	0.641	0.570	0.704	0.703	0.742	0.742	0.700	0.6992	0.739	0.739	0.688	0.681	0.9998	1.000	0.995	0.995
fas ³ T	0.754	0.832	0.613	0.626	0.551	0.699	0.698	0.738	0.737	0.696	0.6950	0.735	0.734	0.683	0.676	0.995	0.995	1.000	0.999
fas ³ A	0.753	0.831	0.611	0.623	0.550	0.698	0.698	0.737	0.736	0.695	0.6944	0.734	0.734	0.682	0.675	0.995	0.995	0.999	1.000
Top Cor.	MNC	NC	g	h	g	PR B50A	PR B50T	PR B85A	PR B85T	PR N50A	PR N50T	PR N85A	PR N85T	SCEAS2	SCEAS1	fa ³ A	fa ³ T	fas ³ A	fas ³ T
Low Cor.	h ^c	h ^c	fas ³	fas ³	fas ³	h ^c	h ^c	h ^c	h ^c	h ^c	h ^c	h ^c	h ^c						

Table 22: Spearman rank correlation matrix for the author based indicators. An A appended to the name of an indicator denotes *All* and a T denotes *Top 25*.

9 Conclusions

In this paper, we presented three new indirect indicators that can be used for scientific evaluation. The first one applies to papers ($fp^k - index$) and the remaining two can be used for the evaluation of an author ($fa^k - index$ when ignoring self-citations and $fas^k - index$ when excluding self-citations). The indicators are based on the paper, the most fundamental entity of citation analysis. Papers are connected with other papers either directly (via the references list) or indirectly via one or more citation paths of varying lengths. An indirect citation between a source paper and a target paper exists if there is a citation path of length greater than one that connects the two papers. Citations provided by citation paths of the same length are considered to belong to the same generation.

The generations of citations are defined in such a way that citations closer to the target paper are considered more important. Papers provide indirect citations of greater generations only if they have not been included in a generation of lower rank (thus representing a stronger relation with the target paper) and if they have not yet been considered in the current generation. This follows the H^s definition for citation generations. The $fp^k - index$ value of a paper is then calculated by the weighted sum of the first three citation generation counts normalized by the scientific age of the paper. The $fp^k - index$ score represents the direct and indirect impact of the paper and reflects the value of a paper. If the paper ceases to receive citations its value eventually declines over time.

Both the new indirect indicators for evaluating authors, i.e., the fa and fas indices, are calculated as the average $fp^k - index$ value of the Publication Record of an author. The difference between the two is that the $fas^k - index$ also accounts for self-citations, which are excluded for each individual (paper, author) pair when constructing the citation generations for the calculations of the $fp^k - index$ scores.

As demonstrated by the comparative study and experimental results, the indicators depend on the number of publications included in the Publication Record of an author when one considers cases where a very high impact paper is the only publication an author has (co-) authored. We have also demonstrated that the indicators can be used to distinguish between authors with similar publication records but different scientific age spans.

We believe that all three indicators take advantage of the indirect citations in order to better distinguish authors with different Publication Records in a way that can be focused at a specific section of the *Paper-Citation graph* and to a specific author. The calculations require partial knowledge of the graph, which may even be acquired manually, although we do consider this task to be labor intensive for authors with large Publication Records and a vast number of citations. More investigation into the applications of the indicators to real citation data, while considering different citation depths and varying number of papers included in the Publication Record of an author, should better reveal the strengths and possible weaknesses of the proposed indicators.

References

- ODC-BY 1.0 license, a. URL <http://opendatacommons.org/licenses/by/summary>.
- DBLP Website, b. URL <http://dblp.uni-trier.de/db>.
- MySQL Website. URL <https://www.mysql.com>.
- PHP Website. URL <https://secure.php.net>.
- DBLP - Some lessons learned, volume 2, 2009.
- Gamal Atallah and Gabriel Rodríguez. Indirect patent citations. *Scientometrics*, 67:437–465, 2006. ISSN 0138-9130. doi: 10.1007/s11192-006-0063-7. 10.1007/s11192-006-0063-7.
- P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google’s PageRank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007. ISSN 1751-1577. doi: 10.1016/j.joi.2006.06.001.
- Rodrigo Costas and María Bordons. Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, 77(2):267–288, November 2008. ISSN 0138-9130. doi: 10.1007/s11192-007-1997-0. 10.1007/s11192-007-1997-0.
- DBLP. DBLP - Data description. URL <http://dblp.uni-trier.de/faq/How+to+parse+dblp+xml.html>.
- D. Dervos, N. Samaras, G. Evangelidis, and T. Foliás. A New Framework for the Citation Indexing Paradigm. In *Proc. of the ASSIST 2006 Annual Meeting, Austin, Texas, USA*, November 2006.
- Dimitris A. Dervos and T. Kalkanis. cc-IFF: A Cascading Citations Impact Factor Framework for the Automatic Ranking of Research Publications. In *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2005. IDAACS 2005. IEEE*, pages 668–673, sept. 2005. doi: 10.1109/IDAACS.2005.283070.
- L. Egghe. The single publication H-index and the indirect H-index of a researcher. *Scientometrics*, 88:1003–1004, sep 2011a. ISSN 0138-9130.
- L. Egghe. The single publication H-index of papers in the Hirsch-core of a researcher and the indirect H-index. *Scientometrics*, 89:727–739, December 2011b. ISSN 0138-9130.
- Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006. ISSN 0138-9130. doi: 10.1007/s11192-006-0144-7. 10.1007/s11192-006-0144-7.
- Dalibor Fiala. Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, 6(3):370–388, 2012.
- Dalibor Fiala, François Rousselot, and Karel Ježek. PageRank for bibliographic networks. *Scientometrics*, 76(1):135–158, 2008. ISSN 0138-9130. doi: 10.1007/s11192-007-1908-4.
- Eleni Fragkiadaki and Georgios Evangelidis. Review of the indirect citations paradigm: theory and practice of the assessment of papers, authors and journals. *Scientometrics*, 99(2):261–288, May 2014. ISSN 0138-9130. doi: 10.1007/s11192-013-1175-5.
- Eleni Fragkiadaki, Georgios Evangelidis, Nikolaos Samaras, and Dimitris A. Dervos. f-Value: measuring an article’s scientific impact. *Scientometrics*, 86(3):671–686, March 2011. ISSN 0138-9130. doi: 10.1007/s11192-010-0302-9.
- J. E. Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, November 2005. ISSN 1091-6490. doi: 10.1073/pnas.0507655102.
- J. E. Hirsch. Does the h-index have predictive power?, August 2007.
- Xiaojun Hu, Ronald Rousseau, and Jin Chen. On the definition of forward and backward citation generations. *Journal of Informetrics*, 5(1):27–36, 2011. ISSN 1751-1577. doi: 10.1016/j.joi.2010.07.004.
- Marek Kosmulski. Hirsch-type approach to the 2nd generation citations. *Journal of Informetrics*, 4(3):257–264, 2010. ISSN 1751-1577. doi: 10.1016/j.joi.2010.01.003.
- Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing PageRank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008. ISSN 0306-4573. doi: 10.1016/j.ipm.2007.06.006.
- Sergei Maslov and Sidney Redner. Promise and Pitfalls of Extending Google’s PageRank Algorithm to Citation Networks. *The Journal of Neuroscience*, 28(44):11103–11105, oct 2008.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
- Filippo Radicchi, Santo Fortunato, Benjamin Markines, and Alessandro Vespignani. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E*, 80:056103, Nov 2009.
- R. Rousseau. The Gozinto theorem: Using citations to determine influences on a scientific publication. *Scientometrics*, 11:217–229, 1987. ISSN 0138-9130. doi: 10.1007/BF02016593.

- Antonis Sidiropoulos and Yannis Manolopoulos. A citation-based system to assist prize awarding. *SIGMOD Rec.*, 34(4):54–60, December 2005. ISSN 0163-5808. doi: 10.1145/1107499.1107506.
- Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72:253–280, 2007. ISSN 0138-9130. doi: 10.1007/s11192-007-1722-z. 10.1007/s11192-007-1722-z.
- Cheng Su, YunTao Pan, YanNing Zhen, Zheng Ma, JunPeng Yuan, Hong Guo, ZhengLu Yu, CaiFeng Ma, and YiShan Wu. PrestigeRank: A new evaluation method for papers and journals. *Journal of Informetrics*, 5(1):1–13, 2011. ISSN 1751-1577. doi: 10.1016/j.joi.2010.03.011.
- Nees Jan van Eck and Ludo Waltman. Generalizing the h- and g-indices. *Journal of Informetrics*, 2(4): 263–271, 2008. ISSN 1751-1577. doi: 10.1016/j.joi.2008.09.004.
- Dylan Walker, Huafeng Xie, Koon-Kiu Yan, and Sergei Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.
- Qiang 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*, 61(3):609–614, mar 2010. ISSN 1532-2890. doi: 10.1002/asi.21276.
- Erjia Yan and Ying Ding. The effects of dangling nodes on citation networks. In *Proceedings of the 13th International Conference on Scientometrics and Informetrics*, pages 4–8, 2011.
- Erjia Yan, Ying Ding, and Cassidy R. Sugimoto. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 62(3):467–477, 2011. ISSN 1532-2890. doi: 10.1002/asi.21461.