Journal of Informetrics 10 (2016) 392–407

Contents lists available at ScienceDirect

Journal of Informetrics

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

Evaluating paper and author ranking algorithms using impact and contribution awards

Marcel Dunaiski[a,*], Willem Visser[b], Jaco Geldenhuys[b]

[a] *Media Lab, Stellenbosch University, 7602 Matieland, South Africa*
[b] *Department of Computer Science, Stellenbosch University, 7602 Matieland, South Africa*

ARTICLE INFO

Article history:
Received 21 September 2015
Accepted 22 January 2016
Available online 16 March 2016

Keywords:
Bibliometrics
Ranking
Citation analysis

ABSTRACT

In the work presented in this paper, we analyse ranking algorithms that can be applied to bibliographic citation networks and rank academic entities such as papers and authors. We evaluate how well these algorithms identify important and high-impact entities.

The ranking algorithms are computed on the Microsoft Academic Search (MAS) and the ACM digital library citation databases. The MAS database contains 40 million papers and over 260 million citations that span across multiple academic disciplines, while the ACM database contains 1.8 million papers from the computing literature and over 7 million citations.

We evaluate the ranking algorithms by using a test data set of papers and authors that won renowned prizes at numerous computer science conferences. The results show that using citation counts is, in general, the best ranking metric to measure high-impact. However, for certain tasks, such as ranking important papers or identifying high-impact authors, algorithms based on PageRank perform better.

1. Introduction

Citation analysis is an important tool in the academic community. It can aid universities, funding bodies, and individual researchers to evaluate scientific work and direct resources appropriately. With the rapid growth of the scientific enterprise and the increase of online libraries that include citation analysis tools, the need for a systematic evaluation of these tools becomes more important.

In bibliometrics, citation counts or metrics that are based directly on citation counts are still the de facto measurements used to evaluate an entity's quality, impact, influence and importance. However, algorithms that only use citation counts or are based only on the structure of citation networks can only measure quality and importance to a small degree. What they are in fact measuring is their impact or popularity which are not necessarily related to their intrinsic quality and the importance of their contribution to the scientific enterprise. The difficulty is to obtain objective test data that can be used with appropriate evaluation metrics to evaluate ranking algorithms in terms of how well they measure a scientific entity's impact, quality or importance.

[*] Corresponding author.
E-mail address: marcel@ml.sun.ac.za (M. Dunaiski).

http://dx.doi.org/10.1016/j.joi.2016.01.010
1751-1577/© 2016 Elsevier Ltd. All rights reserved.

In Section 2 background information about the used ranking algorithms is given and related work, in which appropriate test data sets are used, is outlined. It shows that in previous research only small test data sets have been used to validate proposed ranking methods that only apply to one or two fields within computer science.

In this paper we use four different test data sets that are based on expert opinions each of which is substantially larger than those in previous research and apply them in different scenarios:

- 207 papers that won high-impact awards (usually 10–15 years after publication) from 14 difference computer science conferences are used to evaluate the algorithms on how well they identify high-impact papers.
- 464 papers from 32 venues that won best-paper awards at the time of publication are used to see how well venues predict future high-impact papers.
- From a list of 19 different awards, 268 authors that won one or more prizes for their innovative, significant and enduring contributions to science were collected. This data set is used to evaluate author-ranking algorithms.
- A list of 129 important papers, sourced from Wikipedia, is used to evaluate how well the algorithms identify important scientific work.

Therefore, this paper focuses on algorithms that are designed to measure a paper's or an author's impact and are described in Section 3. In Section 4 the MAS (Microsoft, 2013) and ACM (Association for Computing Machinery, 2014) citation data sets are described which are used for the experiments in this article. Section 5 shows the results of evaluating the various ranking algorithms with the above mentioned test data sets followed by a discussion of the results in Section 6.

2. Background information

The idea of using algorithms based on the PageRank algorithm has been applied to academic citation networks frequently. For example, Chen, Xie, Maslov, and Redner (2007) apply the algorithm to all American Physical Society publications between 1893 and 2003. They show that there exists a close correlation between a paper's number of citations and its PageRank score but that important papers, based purely on the authors' opinions, are found by the PageRank algorithm that would not have easily been identified by looking at citation counts only.

Hwang, Chae, Kim, and Woo (2010) modify the PageRank algorithm by incorporating two additional factors when calculating a paper's score. Firstly, the age of a paper is taken into consideration and secondly, the impact factor of the publication venue associated with a paper is also included in the computation. The algorithm was proposed in an article called "Yet Another Paper Ranking Algorithm Advocating Recent Publications". For brevity this algorithm is referred to as YetRank and is described in Section 3.4.

Dunaiski and Visser (2012) propose an algorithm, NewRank, that also incorporates the publication dates of papers similar to YetRank. They compare the NewRank algorithm to PageRank and YetRank and find that it focuses more on recently published papers. In addition, they evaluate the algorithms using papers that won the "Most Influential Paper" award at ICSE conferences and find that PageRank identifies the most influential papers the best.

Sidiropoulos and Manolopoulos (2005) propose an algorithm that is loosely based on PageRank. The authors call their algorithm SceasRank (Scientific Collection Evaluator with Advanced Scoring). SceasRank places greater emphasis on citations than the underlying network structure compared to PageRank. Sidiropoulos and Manolopoulos use a data set of computer science papers from the DBLP library (The DBLP Team, 2014) and compare different versions of the SceasRank algorithm with PageRank and rankings according to citation counts. They evaluate the algorithms using papers that won impact awards at one of the two venues. Firstly, papers that won the 10 Year Award (Very Large Data Base Endowment Inc., 2014) at VLDB conferences, and secondly, the papers that won SIGMOD's Test of Time Award (ACM Special Interest Group on Management of Data, 2014) are used as evaluation data to judge the ranking methods in ranking important papers. Their results show that SceasRank and PageRank perform the best in identifying these high-impact papers but that using citation counts directly performs very close to those methods. They also rank authors by using the best 25 papers of each author and use the "SIGMOD Edgar F. Codd Innovations Award" (ACM Special Interest Group on Management of Data, 2014) as evaluation data. Their results show that SceasRank performs equally well compared to PageRank and improves over the method of simply counting citations to find important authors.

The above mentioned algorithms are designed to rank individual papers and authors or venues. The ranking scores produced by these algorithms can be aggregated to author or venue entities but this entails considerable biases towards certain entities. For example, taking the average score of authors' publications favours authors unfairly that have only published a few highly cited papers which does not reflect their overall contribution or significance.

Therefore, metrics specifically designed for ranking authors are discussed in Sections 3.5 and 3.6. The metrics that are considered and evaluated are the *h*-index (Hirsch, 2005), the *g*-index (Egghe, 2006), the $i10-$ index (Connor, 2011) and the Author-Level Eigenfactor metric (West, Jensen, Dandrea, Gordon, & Bergstrom, 2013).

A lot of research has been conducted on variations of PageRank to rank author entities. Fiala, Rousselot, and Ježek (2008), for example, also use the Edgar F. Codd award to evaluate their version of PageRank that includes co-authorship graph information. They find that simply using citation counts performs best at ranking authors.

Similar research has been conducted by Yan and Ding (2011), using the Derek de Solla Price award (International Society for Scientometrics & Informetrics, 2014), showing that PageRank with co-authorship graph information included performs better than the basic PageRank algorithm.

By using researchers that won ACM's A. M. Turing (Association for Computing Machinery, 2012) and Edgar F. Codd awards, Fiala (2012) shows that incorporating publication years into the PageRank computation yields better results over the basic PageRank algorithm.

Similarly, Nykl, Ježek, Fiala, and Dostal (2014) use the ACM's A. M. Turing and Edgar F. Codd awards, ISI Highly Cited authors, and ACM Fellows as test data to evaluate PageRank variants to rank researchers. They find that the best ranking is achieved when author self-citations are ignored and all authors of a paper are treated equally. In (Nykl, Campr, & Ježek, 2015) their research is continued using ACM's Fellowships for researchers in the categories of Artificial Intelligence and Hardware and show that the best result is obtained by including the journals' impact factors in the PageRank computations.

Fiala, Šubelj, Žitnik, and Bajec (2015) use three computer science categories of the Web Of Science database to evaluate 12 different author ranking methods of which 9 are PageRank variants. As test data they use a list of editorial board members of the top 10 journals in the fields of artificial intelligence, software engineering, and theory and methods based on the journals' impact factors reported by the 2012 edition of the Journal Citation Report. They find that no PageRank variant outperforms the baseline citation counts of authors. When comparing the PageRank variants against 28 ACM Turing Award winners and settings PageRank's damping factor to 0.5 instead of 0.9, they find that PageRank performs slightly better but is still far from outperforming citation counts.

3. Ranking algorithms

In this paper CountRank (CR) refers to the method of simply ranking papers according to their citation counts. Let $G=(V, E)$ be a directed citation graph containing n papers in the vertex set V and m citations in the edge set E. A CountRank score $CR(i)$ for each paper $i \in V$ can then be calculated using the equation

$$CR(i) = \frac{\mathrm{id}(i)}{m} \tag{1}$$

where $\mathrm{id}(i)$ is the in-degree of vertex i which corresponds to the number of citation that the paper associated with vertex i has received. The citation counts of papers are normalised by the total number of citations in the network in order for the CountRank scores to be comparable to the other ranking algorithms discussed in this section. This results in scores between 0 and 1 for each paper, with the norm[1] of the result vector equal to 1.

This is also true for all algorithms described in this section that rank individual papers. They can be described by using an analogy of a random researcher and are based on the same idea of calculating the predicted traffic to the articles in citation networks. The intuition behind these algorithms is that random researchers start a search at some vertices in the network and follow references until they eventually stop their search, controlled by a damping factor α, and restart their search on a new vertex. The result vectors of the paper ranking algorithms described in this section converge after a sufficient number of iterations, which is controlled by a predefined precision threshold δ.

Therefore, the ranking algorithms differ in only two aspects:

- How are the random researchers positioned on the citation network when they start or restart their searches? Should a random researcher be randomly placed on any vertex in the network or does the random researcher choose a vertex corresponding to a recent paper with a higher probability?
- Which edge (citation) should the random researcher follow to the next vertex (paper)? Should the decision depend on the age of the citation? Should the impact factor of the venue at which the citing or cited paper was published contribute to the decision?

3.1. PageRank

In the case of the standard PageRank algorithm the random researchers are uniformly distributed on the citation network and select the edge to follow at random. In other words, all articles and references are treated equally and a random researcher does not have any preference in selecting a certain paper or following a reference to another paper.

Let $\mathrm{od}(i)$ be the out-degree of the vertex associated with paper i. Then A is defined as the matrix of a citation graph G, where $a_{ij} = 1/\mathrm{od}(i)$ if $(i, j) \in E$ and zero otherwise. Furthermore, let $\boldsymbol{d}$ be a vector with values $d_i = 1$ if the vertex corresponding to paper i is a dangling vertex (no outgoing edges) and zero otherwise.

[1] Throughout this paper the norm refers to the L^1-norm and is explicitly indicated with a subscripted 1. It is defined as $\|\boldsymbol{x}\|_1 = |x_1| + |x_2| + \ldots + |x_n|$.

The PageRank algorithm is initialised with $\boldsymbol{x_0 = 1/n}$ and every subsequent iteration is described by the following equation:

$$\boldsymbol{x}_t = \underbrace{\frac{(1-\alpha)}{n}\cdot\mathbf{1}}_{\text{RandomRestarts}} + \alpha\cdot(A^T + \underbrace{\frac{1}{n}\cdot\mathbf{1}\cdot\boldsymbol{d}^T}_{\text{DanglingVertices}})\cdot\ \boldsymbol{x}_{t-1} \tag{2}$$

It should be noted that the PageRank algorithm defined here adds n edges from each dangling vertex to all other vertices in the graph and evenly distributes the weight between the added edges. This is modelled by the "Dangling Vertices" term in Eq. (2), while the first part of the equation, $(1-\alpha)/n\cdot\boldsymbol{1}$, models the evenly distributed placement of random researchers when they restart a search which is controlled by α whose default value is 0.85.

The computation stops when the predefined precision threshold δ is reached, i.e.:

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_1 < \delta \tag{3}$$

The time complexity to compute one iteration of PageRank is $O(n)$. Furthermore, two values have to be stored in memory for each vertex in the graph, the current PageRank score of a vertex and that of the previous iteration. Therefore, the space requirement for PageRank is also $O(n)$.

3.2. SceasRank

The *Scientific Collection Evaluator with Advanced Scoring* (SCEAS) ranking method introduced by Sidiropoulos and Manolopoulos (2005) and used in (Sidiropoulos & Manolopoulos, 2006) is the PageRank algorithm as described above with alterations by introducing two parameters a and b. According to the authors, b is called the *direct citation enforcement factor* and a is a parameter controlling the speed at which an indirect citation enforcement converges to zero.

In addition to the previously defined parameters, let K be a matrix that contains $k_{ij} = 1$ if $(i, j) \in E(G)$ and zero otherwise. Then SceasRank is defined as follows:

$$\boldsymbol{x}_t = \frac{(1-\alpha)}{N}\cdot\mathbf{1} + \frac{\alpha}{a}\cdot\left(A^T + \frac{1}{N}\cdot\mathbf{1}\cdot\boldsymbol{d}^T\right)\cdot(\boldsymbol{x}_{t-1} + b\cdot K^T\cdot\mathbf{1}) \tag{4}$$

For $b = 0$ and $a = 1$ the above equation is equivalent to PageRank's formula given in (2).

According to the authors, b is used because citations from papers with scores of zero should also contribute to the score of the cited paper. Furthermore, the indirect citation factor a is used to control the weight that a paper x citations away from the current paper has on the score and is a contribution that is proportional to a^{-x}. SceasRank's time and space complexity is also $O(n)$ for each iteration of the algorithm. However, its main advantage is that it converges faster than algorithms that are more similar to PageRank (Dunaiski, 2014, p. 69).

3.3. NewRank

The NewRank algorithm (Dunaiski & Visser, 2012) is based on the PageRank algorithm but incorporates the age of publications into the computation. This is based on the intuition that researchers usually start investigating a new research topic by reading recently published papers in journals or conference proceedings and then follow references to older publications. Therefore, when the random researchers are initially distributed on the citation network their chances are higher to select a paper that was published recently. Moreover, when choosing an edge to follow, the probability of choosing a citation to a recently published paper is higher than a citation to an old paper.

Therefore, let $\boldsymbol{\rho}$ be the vector containing the probabilities of selecting a paper, where $\rho_i = e^{-age(i)/\tau}$ which takes the age of a paper, $age(i)$, into consideration and defines τ to be the characteristic decay time of a citation network with a default value of 4.0.

Furthermore, let $D(i)$ be the probability of following a reference from paper i which is defined as

$$D(i) = \frac{\rho_i}{\sum_{j\in N^+(i)}\rho_j} \tag{5}$$

where $N^+(i)$ is the out-neighbourhood of vertex i which is the set of papers that are cited by the paper i. The above equation simply normalizes the initial value of paper i by the initial values of all papers in its reference list. It follows from this equation that the likelihood of the random researcher following a young citation is greater than following a citation to a paper that is older.

The matrix A of Eq. (2) is updated such that it contains the elements $a_{ij} = (D(i))/(\mathrm{od}(i))$. In addition, let the initial probability distribution be given by $\boldsymbol{x_0} = \mathbf{r}$ where $r_i = \rho_i/||\boldsymbol{\rho}||_1$.

For each iteration $i = 1, 2, \ldots$ the NewRank values are computed, similar to the PageRank algorithm, using the following equation

$$\boldsymbol{x}_t = (1-\alpha)\cdot\boldsymbol{r} + \alpha\cdot\left(A^T + \boldsymbol{r}\cdot\boldsymbol{d}^T\right)\cdot\boldsymbol{x}_{t-1} \tag{6}$$

with the same stopping criteria as given in Eq. (3). NewRank converges at the same rate at PageRank and also has time and space complexities of $O(n)$.

3.4. YetRank

YetRank is an algorithm that incorporates the impact factors of venues into its computation and was proposed by Hwang et al. (2010). The idea of including the impact factor of venues is based on the assumption that a citation from a paper that is published at a prestigious venue should be weighted more than a citation from a less renowned venue.

The definition of the Journal Impact Factor that is currently used by Thomson Reuters is the following (Garfield, 1994):

> In a given year, the Impact Factor of a journal is the average number of citations received per paper published in that journal during the two preceding years.

In order to generalise the formulation of the Journal Impact Factor, two time frames have to be defined. Firstly, the *census window* (*CW*) is a time frame that is defined to include all the papers whose outgoing citation should be considered. Secondly, the *target window* (*TW*) is a year range directly before the census window. All papers published in journals during the target window are potential citable items and references to these papers are used for measuring the importance of journals. In other words, all references originating from papers in the census window and citing papers in the target window are considered when computing impact factor scores for journals.

Let $\mathcal{P}(v, (t_1, t_2))$ be the set of papers that are published by venue v during the time frame $[t_1 ; t_2]$. Furthermore, let $G(V, E)$ be the underlying citation network and $\mathcal{V}$ the set of venues associated with the papers in G. In a weighted graph $w(i, j)$ denotes the weight associated with the edge from vertex i to j which in this case are single citations and therefore all weights are equal to 1.

The following equation denotes the number of citations from any paper in $\mathcal{V}$ during the CW to papers that fall within the TW and are published at venue v:

$$\text{Cited}(v, CW, TW) = \sum_{\{(i,j)\in E | i\in \mathcal{P}(\mathcal{V},CW)\wedge j\in \mathcal{P}(v,TW)\}} w(i, j) \tag{7}$$

If the impact factors for journals were measured by using the above equation, then venues that publish a larger set of papers would be unfairly advantaged since they would have more citable items which is the set $\mathcal{P}(v, TW)$ in Eq. (7). Therefore, the value is normalised by the number of articles associated with a venue during the target window as described by the following equation:

$$IF(v, CW, TW) = \frac{\text{Cited}(v, CW, TW)}{|\mathcal{P}(v, TW)|} \tag{8}$$

Now that the Impact Factor metric is formally defined, the YetRank algorithm is described below. Similarly to NewRank, let $\rho_i = (1/\tau) \cdot e^{-age(i)/\tau}$, where τ is the characteristic decay time and $age(i)$ is the age of the paper i. The impact factor of a venue v for a certain year y is calculated by the Impact Factor method as described by Eq. (8) with parameters: $IF(v, [y, y], [y - 5, y - 1])$. It should be noted that the target window size is 5 and not the default value of 2 years as used by Thomson Reuters.

Then the initial score for paper i published in the year y_i and at venue v_i is $s_i = IF(v_i, [y_i, y_i], [y_i - 5, y_i - 1]) \cdot \rho_i$. Furthermore, let $\boldsymbol{r}$ be the normalised vector such that $r_i = s_i/||\boldsymbol{s}||_1$.

As in the PageRank algorithm let A be the adjacency matrix where $a_{ij} = 1/\text{od}(i)$ if paper i cites paper j and zero otherwise. The YetRank algorithm is initialised with $\boldsymbol{x_0} = \mathbf{r}$ and uses Eq. (6) to compute each following iteration until the stopping criteria, given in Eq. (3), is reached.

By taking the impact factor of publishing venues into consideration the random researchers are more likely to start and restart their searches with papers that were published recently and in more renowned venues.

YetRank's time and space complexity is also $O(n)$ for each iteration but requires an expensive once-off computation to compute the impact factors for each venue for each year.

3.5. The i10-, h-, and g-indices

The i10-index is a simple author impact measure developed by Google and introduced in 2011 on the Google Scholar website. An author has an i10-index value of i if the author has published i papers that have received at least 10 citations each (Connor, 2011). Intrinsically, the i10-index only measures the impact of an author and is highly dependent on publication counts of authors.

The h-index is a relatively new method developed by Hirsch (2005) and was first published in 2005. It was developed for measuring the quality of theoretical physicists' research output but has since gained a lot of popularity in the academic community for computing the impact of researchers in general.

The h-index is based on citation counts solely and considers the distribution of citations of a researcher's publications. The h-index is defined as follows:

> An author has an index h if their h most-cited publications have h or more citations each.

More formally, let $\{p_1, p_2, p_3, \ldots | \mathrm{id}(p_i) \geq \mathrm{id}(p_{i+1})\}$ be an author's set of papers that is sorted in descending order of citations counts. The h-index is then computed by stepping through this set and finding the largest value for h such that:

$$h \leq \mathrm{id}(p_h) \tag{9}$$

The h-index tries to improve on simply counting the total number of papers and the total number of citations that an author has received since the total number of papers does not measure the impact of the work and the total citation count of an author can easily be skewed by co-authoring a small number of highly cited papers which does not accurately reflect the authors overall contribution to science.

Therefore, it was devised to capture both the quality (number of citations of most cited papers) and quantity (the number of papers published over the years) of an authors work.

The g-index was developed in 2006 by Egghe (2006) and tries to overcome some of the drawbacks of the h-index. It is one of the more popular variations of the h-index.

> An author has a g-index value of g if their top g articles in sum have received at least g^2 citations.

As with the h-index, the g-index is computed by stepping through an author's sorted set of papers and finding the largest value for g such that:

$$g \leq \frac{1}{g} \cdot \sum_{i \leq g} \mathrm{id}(p_i) \tag{10}$$

Similarly to the h-index, the g-index measures two quantities. Firstly, it indicates the amount of research output an author has produced and secondly, it also gives an indication of the quality of the author's work. The g-index allows citations from highly cited papers to push up the g-index while not affecting the h-index therefore lowering the quality threshold. Therefore, g is at least the value of h but usually greater than the h-index value.

3.6. The author-level eigenfactor metric

The Eigenfactor project, created by Bergstrom, West, and Wiseman (2008), ranks academic journals using a PageRank-like algorithm on a journal cross-citation graph. It has recently gained a lot of attention and has been included in the Thomson Reuters "Journal Citation Report" (Thomson Reuters, 2014) since 2007.

West et al. (2013) demonstrate how to apply the Eigenfactor metric to author citation graphs. The Eigenfactor metric is simply the PageRank algorithm applied to a normalised author citation graph that is constructed from a data set that contains information about authors in addition to articles and references.

Let G_C be a paper citation graph and $\mathcal{A}$ be the set of authors, where $\mathcal{A}(p_i)$ is the set of authors that authored paper p_i. Similarly, let $\mathcal{P}(a_i)$ be the set of papers written by author a_i. The author citation graph G_A, used as input for the Author-Level Eigenfactor method, is then constructed as follows:

Step 1 - Normalising the citation network G_C:

$$w_{G_C}(p_i, p_j) = \frac{1}{|\mathcal{A}(p_i)| \cdot |\mathcal{A}(p_j)| \cdot \mathrm{od}_{G_C}(p_i)} \tag{11}$$

The equation above normalises the weight of an edge (p_i, p_j) by the product of the number of authors in the citing paper p_i, the number of authors in the cited paper p_j, and the number of references in the bibliography of paper p_i.

Eq. (11) divides the credit of an incoming citation equally between the co-authors of a paper because the average sizes of collaboration groups differ between various academic disciplines. Otherwise, authors that commonly work in larger groups of collaboration would be unfairly advantaged because they would receive full accreditation of a citation.

Step 2 - Constructing the author citation graph G_A:

$$w_{G_A}(a_i, a_j) = \sum_{\left\{(p_i, p_j) \in E(G_C) | p_i \in \mathcal{P}(a_i) \wedge p_j \in \mathcal{P}(a_j)\right\}} w_{G_C}(p_i, p_j) \tag{12}$$

The author citation graph is constructed by inserting edges $w_{ij} = (a_i, a_j)$ whose weights correspond to the sum of the edges from the citation network G_C of papers p_i associated with author a_i that cite papers p_j written or co-authored by author a_j.

Step 3 - Normalizing the co-author adjacency matrix $A(G_A)$:

$$\begin{aligned} A_{ij} &= \frac{w_{G_A}(i, j)}{\sum_{k \in N^+_{G_A}(i)} w_{G_A}(i, k)} && \forall i \neq j \\ A_{ij} &= 0 && \forall i = j \end{aligned} \tag{13}$$

The diagonal values of the matrix A are set to zero so that author self-citations are omitted. For multi-authored papers, this step only removes the citation credit for the authors who are self-citing. The citation is still counted for authors that only co-authored either the cited article or the citing article.

Let the vector $\boldsymbol{r}$ contain the number of articles written by each author normalised by the total number of articles n in the graph. Formally, let $r_i = |\mathcal{P}(i)|/n$ for each author i. Initially, the random researchers are distributed over the author citation graph depending on the number of articles published by authors (ie. $\boldsymbol{x_0} = \boldsymbol{P(i)/|P|}$). Each subsequent iteration is computed with Eq. (6) until it converges and reaches the stopping criteria as given by Eq. (3).

It should be noted that the probabilities related to the restarts of the random researchers are weighted by $\boldsymbol{r}$, which contains values proportional to the number of articles written by an author. This is required to ensure that the random restarts do not favour authors with only a few articles published.

To compensate for the bias that is introduced by the restarts of the random researchers that favour authors that are rarely cited, the result scores of authors are normalised by the incoming citations for each author. The final Author-Level Eigenfactor (AF) ranking scores are therefore computed as follows:

$$AF = 100 \cdot \frac{A^T \cdot \boldsymbol{x}_t}{\|A^T \cdot \boldsymbol{x}_t\|_1} \tag{14}$$

The above equation computes scores for authors between 0 and 100 and can be interpreted as the overall impact or importance of an author. The Author-Level Eigenfactor method has a time and space complexity of $O(n)$ where n is the number of authors in the citation network.

4. The data sets

Microsoft Academic Search (MAS) (Microsoft Research, 2013) is an academic search engine developed by Microsoft Research. The source data set is an integration of various publishing sources such as Springer and ACM.

The entities that are extracted from the data set and processed for the experiments and analyses in the following sections are papers, authors, publication venues and references. The raw count of these entities are as follows; 39,846,004 papers, 19,825,806 authors and 262,555,262 references. Furthermore, it includes information about 21,994 journals and 5190 conferences.

Publication venues and each paper published there are assigned to exactly one domain. For example, all papers published at the *International Conference on Software Engineering* (ICSE) are associated with the computer science (CS) domain.

In order to use this data set for citation analysis it has to be preprocessed and cleaned up. Firstly, 20.58% of papers do not have a publication venue and therefore are not associated with a domain. Secondly, papers that do not contain a year value have to be excluded from the experiments as well since some ranking algorithms make use of these values. Furthermore, papers that contain erroneous year values such as −1 and 2050 were excluded as well.

For all the experiments described in this article only the domain of computer science is considered. However, when constructing the CS citation network, all non-CS papers citing CS papers have to be included.

The final MAS CS citation network consists of 2,394,976 papers (of which 1,573,679 are CS papers), 12,907,440 references, 3152 conferences and 1351 journals. When constructing the associated author citation graph 823,858 distinct authors are found.

The second data set used is a copy of the ACM's digital library data set (Association for Computing Machinery, 2014) that includes papers up to March 2015. All papers published in periodicals and proceedings are included, while PhD dissertations and books are not part this data set.

Similar preprocessing was performed on this data set. The final ACM citation graph consists of 1,159,137 articles and 6,703,224 references with 927,677 unique authors.

5. Evaluation

For the experiments in this paper four different types of test data sets are used that are based on expert opinions and collected by hand from Internet sources. Firstly, papers that won high-impact awards at conferences are used to train and evaluate the paper ranking algorithms on how well they identify and rank high-impact papers. The results are shown in Section 5.1. Secondly, a list of papers that won best paper awards at conferences was compiled and used to evaluate how well these conferences predict future high-impact papers (see Section 5.2). Thirdly, in Section 5.3, authors that won contribution awards in their fields were used to evaluate the author ranking algorithms. And lastly in Section 5.4, a set of important papers listed on Wikipedia (Wikipedia, 2014) is used to evaluate how well the paper ranking algorithms rank these papers that are said to have had a large influence in their fields.

Table 1
Results of evaluating the ranking algorithms using the MAS CS and ACM citation networks as input against the set of high-impact award papers from 14 CS conferences. The optimal parameters are found for each algorithm by training them on 70% of the award papers. The evaluation set (15%) is used to calculate a MAP@10 value for each venue and their average is shown in columns "AMAP".

Algorithm	Parameters	AMAP (MAS)	Parameters	AMAP (ACM)
CountRank	–	0.647	–	0.658
PageRank	$\alpha = 0.55$	0.632	$\alpha = 0.25$	0.624
NewRank	$\alpha = 0.35, \tau = 32$	0.605	$\alpha = 0.25, \tau = 32$	0.628
YetRank	$\alpha = 0.45, \tau = 32$	0.607	$\alpha = 0.15, \tau = 32$	0.603
SceasRank	$\alpha = 0.95, a = 2.5, b = 0$	0.635	$\alpha = 0.85, a = e, b = 0$	0.622

5.1. Evaluating the paper ranking algorithms

A list of 207 academic papers that received accolades as important and high-impact papers was compiled for 14 different computer science (CS) conferences. These prizes are awarded to papers post-publication, usually 10–15 years after their initial publication. The complete list can be found in Table A.9.

The prizes signify that a paper has had the most impact over the intervening years in terms of research, methodology or application. Conferences that hand out these types of awards are predominantly in the CS domain with varying guidelines on the selection processes, but the prizes represent the same meaning of influence and impact. The prizes are selected by reviewing panels of the various venues and therefore can be assumed to be picked by experts in their fields.

Usually a single paper is awarded this prize at a conference in a given year but it does occur that two or more papers tie in the selection process. Therefore, for some conferences more than one paper that won a high-impact prize can be found in the data set for a certain year.

In the following discussions these papers are referred to as **award papers** and are used to measure the performance of the algorithms in identifying and ranking high-impact papers.

Since the award papers all belong to the CS domain, only the subset of CS papers and their citing non-CS papers from the MAS data set are used as input for the ranking algorithms. Therefore, the citation network used consists of 2,394,976 papers, 12,907,440 citations and 4503 venues. The complete ACM citation network is used since it contains predominantly computing literature.

Except CountRank all algorithms have parameters that have to be fitted to the MAS citation network. Therefore, the set of award papers is split into a training set (70%), a validation set (15%), and a test set (15%). In addition, the papers are stratified across these three sets such that the publication years and venues of the award papers are evenly distributed between them to improve their representativeness. Furthermore, it should be noted that the precision threshold was set to $\delta = 1.0 \times 10^{-6}$ for all algorithms.

The parameter ranges over which the algorithms are optimised depend on the algorithms and the experiment which is conducted. In general, the damping factor α ranges from 0.05 to 0.95 with intervals of 0.1. The range for the time decay parameter τ was chosen to start with 2 and grow exponentially according to $\tau = 2^x$, where $x = 1, 2, \ldots$.

For evaluation purposes the **mean average precision** (MAP) is used as the performance measure and is described below. The **average precision** is a single value that encompasses both the precision and recall accuracy of m ranked elements in a query that returns a result set of size n. It is often used in the field of information retrieval and is defined as follows:

$$\mathrm{AP@}n = \frac{1}{\min(m, n)} \cdot \sum_{k=1}^{n} \frac{\mathrm{P}(k) \cdot \mathrm{rel}(k)}{k} \tag{15}$$

where $\mathrm{P}(k)$ is the precision at cut-off k in the result set (described below) and $\mathrm{rel}(k)$ is a function that returns 1 if the element with rank k is relevant and 0 otherwise. $\mathrm{P}(k)$ is the number of relevant elements found in the first k ranked elements. For example, consider three ICSE award papers that were published in 1990 and ranked in positions 1, 5 and 11 in a list of all publications published at ICSE in 1990. The average precision (AP@10) for ICSE for 1990 would be (1/1 +2/5)/3 =0.56.

The **mean average precision** is the mean of a set of N queries, therefore

$$\mathrm{MAP@}n = \frac{1}{N} \cdot \sum_{i=1}^{N} \mathrm{AP@}n(i) \tag{16}$$

In the context of this experiment the MAP@10 is used to calculate the precision of the algorithms in ranking the award papers per venue. More precisely, the mean average precision (MAP@10) is computed for each venue where the average precision (AP@10) of each publication year of the award papers for that venue is averaged. In the following sections **AMAP** refers to the average MAP@10 scores over all venues. Using 10 as the cut-off value for the MAP is somewhat arbitrary, however, most search engines return 10 results per page and therefore empirically 10 seems like the most appropriate value.

Table 1 shows the results of training the algorithms on the training set and evaluating them using the evaluation set for both the MAS and ACM data sets. When considering the results for the MAS data set in this table one can see that CountRank

Table 2
Results of evaluating the ranking algorithms using the adjusted MAS CS and ACM citation networks as input against the set of high-impact award papers. The optimal parameters are found for each algorithm by training them on 70% of the award papers. The evaluation set (15%) is used to calculate a MAP@10 value for each venue and their average is shown in columns "AMAP".

Algorithm	Parameters	AMAP (MAS)	Parameters	AMAP (ACM)
CountRank	–	0.573	–	0.657
PageRank	$\alpha = 0.55$	0.574	$\alpha = 0.45$	0.629
NewRank	$\alpha = 0.35, \tau = 8$	0.588	$\alpha = 0.35, \tau = 32$	0.627
YetRank	$\alpha = 0.45, \tau = 16$	0.586	$\alpha = 0.35, \tau = 16$	0.575
SceasRank	$\alpha = 0.35, a = 3.5, b = 0$	0.564	$\alpha = 0.95, a = e, b = 0$	0.626

performs the best with an AMAP value of 0.647 followed by SceasRank (0.635) and PageRank (0.632). The two algorithms that incorporate the publication years of papers into their computations, namely YetRank (0.607) and NewRank (0.605), perform the worst. The result of testing CountRank on the test set is 0.494.

Similar results are obtained when the ACM data set is used. Again, CountRank performs the best with an AMAP value of 0.658 and YetRank (0.603) performs the worst. However, this time NewRank (0.628) performs better than PageRank (0.624) and SceasRank (0.622). The result of testing CountRank on the test set is 0.591.

It should be noted that the results are computed on the entire data sets of papers with publication dates ranging until 2013 (MAS) and 2015 (ACM). It seems reasonable to assume that after articles win a high-impact award their visibility increases making them more likely to be cited in the years following the prizewinning. In order to avoid this bias, the citation graphs are truncated to only include papers up to the years of award consideration for each award paper. For example, given that an award paper wins a high-impact award in 2008 and was published in 1998, the input citation graph only contains references from papers published in or before 2008 and therefore excludes all references produced after 2008.

Using this normalisation strategy, Table 2 shows the results that the algorithms obtain for both the adjusted MAS and ACM data sets.

On the MAS adjusted data set NewRank and YetRank perform the best with AMAP values of 0.588 and 0.586 respectively. However, using the ACM data set CountRank remains the best performing algorithm with an AMAP value of 0.657 followed by PageRank and NewRank.

Evaluating NewRank on the MAS citation graph with the trained parameters $\alpha = 0.35$ and $\tau = 8$ using the test set, it achieves an AMAP value of 0.532. Similarly, computing CountRank on the ACM citation graph and evaluating it using the test set it obtained an AMAP value of 0.613. These values should be interpreted as conservative upper bounds of the predictive capabilities of the algorithms when applied to unknown citation graphs.

Lastly, it should be noted that using different values for SceasRank's parameter b does not have an effect on the ranking results of the award papers and therefore did not influence the results of this evaluation. Moreover, when comparing the damping factor value of SceasRank to the other algorithms dividing α by a gives an approximation of the "real" damping factor. This can be done when the value of b is close to zero. For example in Table 1 dividing $\alpha = 0.95$ by $a = 2.5$ yields 0.38 which is close to the damping factor values obtained by the other algorithms. The interpretation of the damping factor is further discussed in Section 6. Although, NewRank and YetRank perform better than CountRank on the adjusted MAS citation graph, using citation counts appears to be the best approach in general when trying to identify high-impact papers.

5.2. How well do venues predict high-impact papers?

The second type of data that was collected consists of articles that were awarded the prize of best paper at a conference in the year that they were published. At conferences this prize is usually awarded to one or more articles that are considered to be of the highest quality in the given year by a review panel. Usually all papers presented in a year are considered for this award. Either a review panel of experts choose the best paper or the reviewers of the peer review processes give their recommendations on the quality of the papers to the conference panel from which the best papers are then chosen.

There are varying guidelines on how many best-paper awards are awarded. For example, at ICSE not more than 10% of papers are allowed to receive the prize. Alternatively, some conferences award a best-paper prize per track.

In the following discussions these papers are referred to as **best papers**. In total 464 papers from 32 different venues were collected and matched to the corresponding entries in the MAS data set. The list of venues is given in Table A.10. The best papers are used to evaluate these venues on how well they predict future high-impact papers.

The CountRank algorithm is chosen for this experiment since it performed the best in identifying high-impact papers (see Section 5.1). For each year that a conference awards best-paper prizes, the AP@10 of the best papers is calculated from the ranks of all papers published at the conference in that year. The MAP@10 is then calculated over all years in which best-paper awards were handed out at that conference.

The results are shown in Table 3. The number of best papers in the test data for each conference is given in column "Count" and the average citation count that the best papers received is given in column "In-Deg.".

It should be noted that the year ranges for which the best-paper awards were handed out are not identical for each conference. For example, AAAI lists best papers since 1996 while for SIGMOBILE the data set only has best papers since 2008. In order to ensure that the varying publication dates of the best papers do not have an impact on the analysis, the MAS CS

Table 3
The precision of the award committees in identifying high-impact papers based on the papers that won best-paper awards at the associated conferences. The input network is truncated to 5 years after the papers won the best-paper awards. Only the top 5 venues are listed in this table. The entire listing of all 32 venues can be found in (Dunaiski, 2014).

Conference	Count	In-Deg.	MAP
SOSP	19	66.89	0.577
OSDI	12	78.42	0.544
SIGMETRICS	7	54.71	0.525
FOCS	10	57.90	0.495
ACL	11	68.00	0.457

network is truncated to 5 years after the publication of the papers that won the best-paper awards. For example, given a paper that won the best-paper prize at AAAI in 1996, the network is truncated to only include papers up to 2001 and used by the algorithms to compute ranking scores. The rank that this paper achieves in the list of all papers published at AAAI in 1996 is used to evaluate the venue for that year. In addition, for all conferences, the best papers published after 2008 are ignored since the MAS data only contains papers up to 2013.

From the table one can see that the venues SOSP (ACM Symposium on Operating Systems Principles), OSDI (Operating Systems Design and Implementation) and SIGMETRICS (Special Interest Group on Measurement and Evaluation) predict high-impact papers the most accurately with a MAP@10 of over 0.5.

On the one hand, it could be argued that the more papers that are awarded the best-paper prize, the higher the chances of choosing papers that will not receive high citation counts. This can result in a lower precision. In order to account for this bias, only one best paper per year can be chosen for each conference. On the other hand, a venue that awards more best-paper prizes in a year has a higher chance to choose the paper which receives the most citations in the following years.

One possible way of choosing a single best paper per year for each venue is by only considering the paper with the highest citation count and comparing it to all other papers published that year. The results of choosing only one best paper for each year per conference are given in Table 4 which shows the top 5 venues that predicted the high-impact papers the most accurately. The complete list of all 32 venues is given in (Dunaiski, 2014).

The column "Nr. Years" shows for how many years the venues awarded best paper prizes. These values therefore indicate how many best papers are considered when computing the precision of how well the venues predict high-impact papers. Similarly to Table 3, the column "In-Degree" shows the average number of citations of the best papers that are chosen as test data. In this case, it shows the average citation count of the best papers with the most citations for each year at a venue. Lastly, the column "MAP" shows the MAP@10 of the venues in the prediction of high-impact papers.

One can see that the top conferences stay roughly the same. Again "SOSP" achieves the highest precision with 0.640. However, it should be noted that the precision values in Table 4 are notably higher than the values obtained in Table 3 where all best papers are considered. This is expected since only the papers with the highest citation counts are chosen for each year at each conference which are ranked higher than the other best papers that are ignored.

5.3. *Evaluating author ranking algorithms*

In order to assess the performance of the venue ranking algorithms, test data that contains qualitative information about authors or journals is required. Since this type of data is not readily available for journals and conferences, only the ranking algorithms that rank authors are evaluated with appropriate test data. This is possible since the ranking algorithms for venues can also be adapted to rank authors since both entities publish one or more articles. The main difference is that authors can publish at different venues while journals and conferences intrinsically publish at a unique venue.

Therefore, in order to evaluate the author ranking algorithms, 19 lists (see Table A.11) of in total 268 researchers that won an award for their innovative, highly significant and enduring contributions to their fields were collected. Of the 268 prize recipients, 17 authors have won two different awards while "Karen Spärck Jones" won three awards, namely, the "ACM – AAAI Allen Newell Award", the "ACL Lifetime Achievement Award", and the "Gerard Salton Award" handed out by the "Special Interest Group on Information Retrieval" (SIGIR).

Table 4
The precision of the top 5 award committees in identifying high-impact papers based on the single papers that won a best-paper award with the highest citation counts for each year in which the best-paper prize was awarded.

Venue	Nr. Years	In-Deg.	MAP
SOSP	7	106.50	0.640
SIGMETRICS	6	51.50	0.483
ACL	8	78.50	0.458
FOCS	6	68.33	0.410
FSE	7	71.57	0.405

Table 5
The results of evaluating the author ranking algorithms against the list of 249 authors that won innovation and contribution awards. The median rank of the award authors is used to measure the algorithms' precision.

Algorithm	MAS	ACM
CountRank (w/o. self-citations)	907	1010
CountRank (w. self-citations)	925	1044
Author-Level Eigenfactor	728	722
h-index	1035	1282
g-index	940	1115
*i*10-index	1371	1448
Publication Count	3201	4017

Therefore, in total 249 distinct authors were matched to corresponding entries in the MAS data set. This set of authors is referred to as **award authors** in the following discussion. A detailed description of the awards handed out at various conferences can be found in (Dunaiski, 2014).

Since the authors that won the author awards are from various disciplines and the awards fall into different domains, all authors in the entire citation networks are considered when evaluating the author ranking algorithms. Therefore, median ranks of the award authors are computed. Authors that won multiple awards are only counted once.

Table 5 lists the median ranks as produced by the various author ranking algorithms. The Author-Level Eigenfactor method achieves the best results with a median rank of 728 and 722 for both the MAS and ACM citation graphs respectively.

Using citation counts with self-citations omitted performs second best (907 and 1010) followed by citation counts with author-self citations included. This indicates that self-citations do not necessarily increase an author's chance of receiving contribution awards. This corroborates the findings by Nykl et al. (2014) who show that, using different test data, PageRank performs the best when self-citations are ignored. Further investigation is required to measure the impact that author collaboration has on these results.

The *g*-index ranks the award authors higher than the *h*-index. The worst indicator is using the publication counts of authors which is expected since the number articles that authors have published rather reflects their life-time achievement and not innovativeness of their contributions or the impact that their articles have had on a field.

It was found that the Author-Level Eigenfactor method ranks the award authors the highest with a median rank of 720 and 704, respectively, when the damping factor is set to 0.84 and 0.92 for the MAS and ACM citation graph.

5.4. Identifying important papers

Lastly, a list of **important papers** in the CS domain was compiled. The source for this list is Wikipedia (Wikipedia, 2014) where papers that are regarded important to a research field were selected by Wikipedia editors. According to the guidelines on the Wikipedia webpages themselves, an important paper can be any type of academic publication given that it meets at least one of the following three conditions. Firstly, a publication led to a significant, new avenue of research in the domain in which it was published. Secondly, a paper is regarded as a breakthrough publication if it changed the scientific knowledge significantly and is therefore judged noteworthy enough to be granted a place on this list. Thirdly, influential papers that changed the world or had a substantial impact on the teaching of the domain, are also included in the list of important papers. This data set is used to evaluate how well the various ranking algorithms can identify these important papers.

From the papers listed on Wikipedia 129 were matched against paper entries in the MAS data set of which 115 contain venue and publication year information. For the ACM data set only 103 papers were matched.

Since the set of important papers span various fields in computer science and are published in different journals and conferences, the overall ranks of the papers are used as a metric to evaluate the ranking algorithms independent of the publication years of the papers. Therefore, the median rank of the important papers is computed on the whole citation graphs of the MAS and ACM data. It should be noted that the average publication year of the important papers is 1981 which is relatively old.

Using the default parameter values for the algorithms on the MAS citation graph, PageRank ranks the important papers the highest with a median rank of 990 as shown in Table 6, followed by YetRank (1078) and CountRank (1652). NewRank performs the worst (9566) which can be explained by the fact that the average publication year of the important papers is 1981 and NewRank gives higher priority to recently published papers. When evaluating the algorithms on the ACM data SceasRank performs the best with a median rank of 818 followed by PageRank (893). Again NewRank performs the worst (6755).

Table 7 shows the median ranks of the important papers when the algorithms are executed using the trained parameters from Section 5.1. For the MAS data set, only NewRank (3624) and SceasRank (1858) improve on the results over using the default parameters, while all but YetRank improve on the median rank when used with the ACM data set. In both cases PageRank performs the best with a median rank of 1708 and 805 for the MAS and ACM data sets respectively. The reason for displaying the results of Table 7 is to show that the trained parameters using the award papers should not be used in a different application, in this case, ranking the overall important papers in computer science.

Table 6
Results of evaluating the ranking algorithms against a set of important papers in Computer Science. The median rank of the important papers is used to measure the algorithms' precision with their default parameters.

Algorithm	Default parameters	Median (MAS)	Median (ACM)
CountRank	–	1652	1257
PageRank	$\alpha = 0.85$	990	893
NewRank	$\alpha = 0.85$, $\tau = 4$	9566	6755
YetRank	$\alpha = 0.85$, $\tau = 4$	1078	1165
SceasRank	$\alpha = 0.85$, $a = e$, $b = 1$	2153	818

Table 7
Results of evaluating the ranking algorithms in identifying the set of important papers. The median rank of the important papers, given in columns MAS and ACM, is used as evaluation indicator when the trained parameters are used for the algorithms.

Algorithm	Trained parameters	MAS	Trained parameters	ACM
CountRank	–	1652	–	1257
PageRank	$\alpha = 0.55$	1708	$\alpha = 0.25$	805
NewRank	$\alpha = 0.35$, $\tau = 32$	3624	$\alpha = 0.25$, $\tau = 32$	2012
YetRank	$\alpha = 0.45$, $\tau = 32$	1285	$\alpha = 0.15$, $\tau = 32$	8011
SceasRank	$\alpha = 0.95$, $a = 2.5$, $b = 0$	1858	$\alpha = 0.85$, $a = e$, $b = 0$	808

Table 8
Results of evaluating the ranking algorithms in identifying the set of important papers. The median rank of the important papers is used as evaluation indicator when the optimal parameters are used for the algorithms.

Algorithm	Optimal parameters	MAS	Optimal parameters	ACM
CountRank	–	1652	–	1257
PageRank	$\alpha = 0.85$	990	$\alpha = 0.59$	702
NewRank	$\alpha = 0.85$, $\tau = 10,000$	990	$\alpha = 0.59$, $\tau = 10,000$	702
YetRank	$\alpha = 0.85$, $\tau = 40$	807	$\alpha = 0.63$, $\tau = 10,000$	801
SceasRank	$\alpha = 0.88$, $a = 1.05$, $b = 0$	990	$\alpha = 0.62$, $a = 1.05$, $b = 0$	702

Table 8 shows the results of the algorithms' optimal parameters for identifying the important papers. For all algorithms, the optimal α values are relatively large with α at around 0.85 (MAS) and 0.60 (ACM) compared to the trained values obtained from using the award papers. Furthermore, the influence that the age of papers have on the ranks, which is controlled by the parameter τ, can be set very high for NewRank and YetRank so that they do not play a role. For example, NewRank becomes identical to PageRank with a large enough τ value and therefore performs exactly as well as PageRank.

Since the important papers are relatively old, these values are expected since they shift the focus towards older publications in the citation network as shown in (Dunaiski, 2014, pp. 97–99).

Using the MAS data set, YetRank manages to outperform PageRank with a median rank of 807. However, with the ACM data set, YetRank does not achieve the same accuracy as the other algorithms.

It should be noted that when keeping α the same value, the median rank decreases by choosing larger τ values. By increasing the τ values, the effect that the age of a publication has on the resulting scores of papers is decreased. This indicates that for this set of important papers, the age of publications is not as important as the citations they receive. SceasRank performs the best when α/a is close to the damping factor α of PageRank. As seen in previous experiments the value of b has no effect on the results. All algorithms perform better than CountRank after finding optimal parameters for each.

6. Discussion

The results shown in the following discussion are the ones obtained from the experiments using the MAS data set. However, the conclusions drawn from this discussion hold true for the results using the ACM data set as well.

The damping factor of PageRank has multiple uses and implications. The same properties hold true for algorithms that are based on PageRank such as NewRank, YetRank and the Author-Level Eigenfactor metric.

Firstly, when $\alpha \to 1$ more focus is placed on the characteristics of the underlying network structure. Using the analogy of the random researcher, the closer α is to 0, the more random restarts occur and the more likely the random researcher stops following citations and chooses a new random paper. Conversely, if $\alpha = 1$, then the random researcher does not stop a search until reaching a dangling vertex.

Secondly, it should be noted that the nature and structure of the hyperlink graph of the Internet (webgraph) and academic citation networks differ in important ways. Webgraphs are dynamic since hyperlinks can be added or removed by updating webpages at any point in time. Outgoing edges of vertices in a citation network are fixed since references cannot be added to a paper after it has been published. In addition, webpages can be deleted from the webgraph but papers, once integrated into the academic corpus, are permanent. Vertices in a citation network can only acquire new incoming edges over time by citations from papers that are published at a later point in time.

This introduces an inherent time variable in citation networks which has to be considered separately and influences the use of the damping factor. More precisely, α controls the distribution of the ranking scores over the publication years of papers in citation networks. The smaller the value of α, the more evenly the scores are distributed over the years (Dunaiski, 2014, p. 97). Alternatively, a larger value of α has the effect that older papers are prioritised and receive larger ranking scores on average compared to recently published papers.

In addition to the damping factor, the NewRank and YetRank algorithms have a second parameter (τ) controlling the characteristic decay of a citation network. Therefore, the parameters α and τ in conjunction control the score distribution over the publication years.

When constructing an author citation graph from citation data, this intrinsic time-arrow exhibited by paper citation networks falls away. With this in mind, it is not surprising that the optimal α value for the Author-Level Eigenfactor metric is 0.84 for the MAS data set which is very close to PageRank's default value of 0.85 initially used by Google for the Internet's hyperlink graph (Brin & Page, 1998).

When considering the results of PageRank in this paper, different optimal α values were found for different purposes. The optimal damping factor for finding papers that won high-impact prizes is 0.55 and for identifying overall important papers it is 0.85.

Empirically, these parameter values are consistent with the observation that α controls the score distribution over the years. The larger the value of α, the higher the scores of older papers. Papers receive high-impact prizes about 10–15 years after their publication and fall within the mid-range of all published papers in the data set. Accordingly, the optimal α value was found to be 0.55.

Furthermore, when the citation network was truncated to only include references produced up until the award considerations, the τ values for NewRank and YetRank decreased. This is expected since the citation network becomes "younger" and more emphasis has to be placed on recently published papers.

Lastly, the set of important papers are relatively old, with an average publication year of 1981, and hence the optimal damping factor value of 0.85 is comparatively large with corresponding large τ values.

Chen et al. (2007) who were the first to use the PageRank algorithm on citation networks used a damping factor of $\alpha = 0.5$ instead of 0.85. They argued that entries in the bibliographies of papers are compiled by authors by searching citation paths of length two on average. Choosing a damping factor of 0.5 leads to an average citation path length of 2 in the PageRank model which seems more appropriate for citation networks. They base this choice on the observation that about 42% of the papers that are referenced by a paper *A* have at least one reference directly to another paper that is also in the reference list of *A*. Their choice of a damping factor value is appropriate for finding high-impact papers as shown with the findings in this paper. It should be noted however, that the choice of α is highly dependent on the underlying network structure. More importantly, Chen et al computed the above mentioned values from a data set containing only physics publications and may be different to data sets containing other academic domains.

7. Threats to validity

For all the experiments in Section 5, the CS subset of the MAS data set was used. Therefore, only citations are used that originate from CS papers or are citations that directly cite CS papers. This means that all citations that originate from outside the CS domain are weighted the same, which does not reflect the true weight if the entire citation network would have been considered. Therefore, using the CS citation network has to be seen as an approximation of the entire academic citation network structure. Because of the time and space complexity of the algorithms it was not feasible to compute the various ranking algorithms on the entire citation network. Furthermore, the validity of the results discussed in this paper is dependent on the data quality of the citation database used.

The MAS and the ACM data sets cannot be seen as two distinct data sets since the ACM database is one of the many sources from which the MAS database is constructed. Therefore, the ACM citation graph should be interpreted as a subgraph of the MAS citation graph which can be problematic when it is used for checking the reproducibility of the results. However, the citation structures of the two data sets vary significantly because the ACM data set is restricted to internal citations and is therefore less comprehensive.

The use of award papers that won prizes retrospectively for their high impact is not perfect test data. For most venues the selection process requires someone to submit potential papers manually to the review panel. The selection of the final award papers is therefore subject to the submission process. High-impact papers might not be considered since they were not submitted for evaluation in the first place.

The set of author awards used as test data are awarded to authors for their long-lasting, significant and innovative contributions to their field of study. This is also not perfect evaluation data. The selection of award authors is very subjective and takes other aspects of impact into consideration, in addition to the objective measures such as publication counts or the intrinsic quality of an author's work. For example, teaching duties and administrative work are also considered as contributions of a researcher and cannot be measured based on his or her publication record. Furthermore, all author awards are treated equally but some prizes might be more prestigious than others.

8. Conclusion

Simply counting citations is the best metric for ranking high-impact papers in general. This suggests that citation counts, although surrounded by controversy on their fairness and interpretation (Garfield, 1955), are a good measurement of a paper's impact.

However, when the goal is to find important papers and influential authors, metrics based on PageRank outperform the use of citation counts. This was shown by evaluating the author ranking algorithms using a set of authors that won contribution awards and identifying the Author-Level Eigenfactor metric as the most accurate method for ranking influential authors.

Using the MAS citation graph it was found that YetRank, the method that includes the impact factor of venues in its computation, ranks the overall important papers the highest outperforming all other PageRank-like algorithms and the use of citation counts.

The interpretation of this result is tricky since the causation is unclear. On the one hand, the choice of where to publish matters and publishing at prestigious venues does have an advantageous impact on future success of the paper. On the other hand, it could be argued that since the contents of the articles are important, they were accepted at renowned venues in the first place.

It should be noted that the score of a paper according to YetRank is dependent on the prestige of the venues of articles citing the current article. Therefore, it could be argued that important papers are cited more likely by prestigious venues. Consequently, an article could be considered important if highly cited by papers published at prestigious venues.

Moreover, it was shown that the impact of self-citations does not contribute an advantage to the overall success of authors. Further analysis is required to evaluate whether self-citations have an impact on the rankings of authors within small speciality fields where an author or a group of authors focus on narrow specialities and therefore produce high self-citation rates.

Author contributions

Conceived and designed the analysis: MD.
Collected the data: MD.
Contributed data or analysis tools: MD.
Performed the analysis: MD.
Wrote the paper: MD.
Supervisor: WV, JG.
Proof-read: WV.

Appendix A. Evaluation data information

Table A.9
List of conferences and Special Interest Groups for which award papers (high-impact papers) were selected.

Venue	Award name	Nr. high-impact papers
AAAI	Classic Paper Award	21
ASE	Most Influential Paper Award	5
ICFP	Most Influential ICFP Paper Award	8
ICSE	Most Influential Paper Award	25
ISCA	Influential ISCA Paper Award	11
OOPSLA	Most Influential OOPSLA Paper Award	8
PLDI	Most Influential PLDI Paper Award	14
POPL	Most Influential POPL Paper Award	11
SIGEVO	SIGEVO Impact Award	3
SIGCOMM	Test of Time Paper Award	29
SIGMETRICS	Test of Time Award	7
SIGMOD	Test of Time Award	19
SIGSOFT	Impact Paper Award	29
VLDB	VLDB 10 Years Award	17

Table A.10
Conferences, learned societies or Special Interest Groups for which best paper awards were collected.

Venue	Nr. of best papers	Venue	Nr. of best papers
AAAI	21	NSDI	6
ACL	14	OSDI	12
ASE	76	PLDI	10
CHI	38	PODS	16
CIKM	6	S&P	3
CVPR	11	SIGCOMM	3
FOCS	10	SIGIR	15
FSE	19	SIGMETRICS	8
ICCV	12	SIGMOBILE	3
ICDM	9	SIGMOD	13
ICML	7	SODA	3
ICSE	31	SOSP	22
IJCAI	16	STOC	14
INFOCOM	16	UIST	12
KDD	12	VLDB	6
LISA	7	WWW	13

Table A.11
Number of authors who received lifetime achievement or contribution award per venue.

Venue	Award	Nr. of Authors
AAAI	ACM – AAAI Allen Newell Award	20
ACL	ACL Lifetime Achievement Award	11
CHI	SIGCHI Lifetime Research Award	15
ICCV	PAMI Azriel Rosenfeld Lifetime Achievement Award	4
ICDM	Research Contributions Award	10
IJCAI	Award for Research Excellence	14
ISCA	ACM SIGARCH Maurice Wilkes Award	14
KDD	SIGKDD Innovations Award	13
PLDI	Programming Languages Achievement Award	24
SIGACT	Knuth Prize	12
SIGCOMM	Lifetime Contribution Award	21
SIGIR	Gerard Salton Award	10
SIGMETRICS	Achievement Award	11
SIGMOBILE	Outstanding Contributions Award	14
SIGMOD	SIGMOD Edgar F. Codd Innovations Award	22
SIGOPS	Mark Weiser Award	14
SIGSIM	ACM SIGSIM Distinguished Contributions Award	6
SIGSOFT	ACM SIGSOFT Outstanding Research Award	23
USENIX	USENIX Lifetime Achievement Award	10

References

ACM Special Interest Group on Management of Data. (2014). *SIGMOD awards*. http://www.sigmod.org/sigmod-awards/,. Online; Accessed 19.01.16

Association for Computing Machinery. (2012). *A. M. Turing Award*. http://amturing.acm.org/,. Online; Accessed 12.10.15

Association for Computing Machinery. (2014). *ACM Digital Library*. http://dl.acm.org/,. Online; Accessed 08.12.15

Bergstrom, C. T., West, J. D., & Wiseman, M. (2008). The eigenfactor metrics? *The Journal of Neuroscience*, *28*(45), 11433–11434.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web, WWW '07* (pp. 107–117). Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm? *Journal of Informetrics*, *1*(1), 8–15.

Connor, J. (2011). *Google Scholar Citations Open To All*. http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html. Online; Accessed 19.01.16

Dunaiski, M. (2014). *Analysing ranking algorithms and publication trends on scholarly citation networks*. Stellenbosch University. Master's thesis.

Dunaiski, M., & Visser, W. (2012). Comparing paper ranking algorithms. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, SAICSIT '12* (pp. 21–30). New York, USA: ACM.

Egghe, L. (2006). Theory and practise of the g-index? *Scientometrics*, *69*(1), 131–152.

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, *6*(3).

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks? *Scientometrics*, *76*(1), 135–158.

Fiala, D., Šubelj, L., Žitnik, S., & Bajec, M. (2015). Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*, *9*(2).

Garfield, E. (1955). Citation indexes for science? A new dimension in documentation through association of ideas. *Nature*, *122*(3159), 108–111.

Garfield, E. (1994). *The Thomson Reuters Impact Factor*. http://wokinfo.com/essays/impact-factor/. Online; Accessed 19.01.16

Hirsch, J. (2005). 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.

Hwang, W., Chae, S., Kim, S., & Woo, G. (2010). Yet another paper ranking algorithm advocating recent publications. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10* (pp. 1117–1118). New York, USA: ACM.

International Society for Scientometrics and Informetrics. (2014). *Derek John de Solla Price award of the journal Scientometrics*. http://www.issi-society.org/price.html,. Online; Accessed 12.10.15

Microsoft. (2013). *Microsoft Academic Data*. https://datamarket.azure.com/dataset/mrc/microsoftacademic. Online; Accessed 19.01.16

Microsoft Research. (2013). *Microsoft Academic Search*. http://academic.research.microsoft.com,. Online; Accessed 19.01.16

Nykl, M., Campr, M., & Ježek, K. (2015). Author ranking based on personalized PageRank? *Journal of Informetrics*, *9*(4), 777–799.

Nykl, M., Ježek, K., Fiala, D., & Dostal, M. (2014). PageRank variants in the evaluation of citation networks? *Journal of Informetrics*, *8*(3), 683–692.

Sidiropoulos, A., & Manolopoulos, Y. (2005). A citation-based system to assist prize awarding? *SIGMOD Records*, *34*(4), 54–60.

Sidiropoulos, A., & Manolopoulos, Y. (2006). Generalized comparison of graph-based ranking algorithms for publications and authors? *Journal of Systems and Software*, *79*(12), 1679–1700.

The DBLP Team. (2014). *The DBLP Computer Science Bibliography*. http://dblp.uni-trier.de/. Online; Accessed 19.01.16

Thomson Reuters. (2014). *Journal citation reports*. http://thomsonreuters.com/journal-citation-reports. Online; Accessed 19.01.16

Very Large Data Base Endowment Inc. (2014). *VLDB 10 years awards*. http://vldb.org/archives/10year.html,. Online; Accessed 19.01.16

West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., & Bergstrom, C. T. (2013). Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, *64*(4), 787–801.

Wikipedia. (2014). *Lists of important publications in science*. http://en.wikipedia.org/wiki/Lists_of_important_publications_in_science. Online; Accessed 19.01.16

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective? *Information Processing & Management*, *47*(1), 125–134.